future internet

MDPI

Article

Data Is the New Oil–Sort of: A View on Why This Comparison Is Misleading and Its Implications for Modern Data Administration

Christoph Stach

Institute for Parallel and Distributed Systems, University of Stuttgart, Universitätsstraße 38, 70569 Stuttgart, Germany; christoph.stach@ipvs.uni-stuttgart.de

Abstract: Currently, data are often referred to as the oil of the 21st century. This comparison is not only used to express that the resource data are just as important for the fourth industrial revolution as oil was for the technological revolution in the late 19th century. There are also further similarities between these two valuable resources in terms of their handling. Both must first be discovered and extracted from their sources. Then, the raw materials must be cleaned, preprocessed, and stored before they can finally be delivered to consumers. Despite these undeniable similarities, however, there are significant differences between oil and data in all of these processing steps, making data a resource that is considerably more challenging to handle. For instance, data sources, as well as the data themselves, are heterogeneous, which means there is no one-size-fits-all data acquisition solution. Furthermore, data can be distorted by the source or by third parties without being noticed, which affects both quality and usability. Unlike oil, there is also no uniform refinement process for data, as data preparation should be tailored to the subsequent consumers and their intended use cases. With regard to storage, it has to be taken into account that data are not consumed when they are processed or delivered to consumers, which means that the data volume that has to be managed is constantly growing. Finally, data may be subject to special constraints in terms of distribution, which may entail individual delivery plans depending on the customer and their intended purposes. Overall, it can be concluded that innovative approaches are needed for handling the resource data that address these inherent challenges. In this paper, we therefore study and discuss the relevant characteristics of data making them such a challenging resource to handle. In order to enable appropriate data provisioning, we introduce a holistic research concept from data source to data sink that respects the processing requirements of data producers as well as the quality requirements of data consumers and, moreover, ensures a trustworthy data administration.

Keywords: data characteristics; data administration; data refinement; reliability; security; privacy

check for updates

Citation: Stach, C. Data Is the New Oil–Sort of: A View on Why This Comparison Is Misleading and Its Implications for Modern Data Administration. *Future Internet* **2023**, *15*, 71. https://doi.org/10.3390/fi15020071

Academic Editor: Cheng-Chi Lee

Received: 3 January 2023
Revised: 6 February 2023
Accepted: 10 February 2023
Published: 12 February 2023

1. Introduction

In 2011, the World Economic Forum stated that “data will be the new oil” [1]. This metaphor is intended to express not only that the commodity ‘data’ has a steadily growing economic value [2] but also that this commodity is becoming one of the most important drivers in the industrial world [3]. This change is facilitated in particular by the Internet of Things (IoT). As a result, machines are equipped with sensors to collect data on all kinds of manufacturing aspects [4]. Furthermore, all machines are connected to data processing infrastructures, which continuously analyze the measured data to monitor the manufacturing process and, if necessary, intervene to optimize it [5]. Thus, the comparison of data with oil seems quite apt, as oil was a significant driver for the Technological Revolution in the 19th century and data are the catalyst of the Fourth Industrial Revolution in the modern age [6]. The World Economic Forum also recognized that their 2011 vision of the future has meanwhile become reality when they concluded in 2017 that “the world’s most valuable resource is no longer oil, but data” [7].

While the allegory of data and oil primarily invokes associations with industrial application domains—often coined as Industry 4.0 [8]—IoT-based data capturing also enables a general improvement of everyone's daily life. Such data-driven services are often referred to as Smart Living [9]. This umbrella summarizes, e.g., services in the area of Smart Mobility [10], Smart Health [11], and Smart Homes [12]. This also shares some similarities with oil, as the extensive use of oil initially transformed the industry but gradually found its way into the private sector in the form of new types of products and finally changed society as a whole [13]. Data also possess this potential—it is not a coincidence that IoT technologies are seen as a booster for the digital age [14].

Besides their strategic and economic importance, data and oil share another common feature: like oil, data initially have to be refined before they can be used profitably [15]. Peter Sondergaard, Senior Vice President of Gartner Research, addresses this fact with his statement that "information is the oil of the 21st century, and analytics is the combustion engine". So, in order to extract interpretable information from raw data, they must first be structured and put into context [16]. This requires processes to systematically transform the data [17] as well as tools and techniques to support such information management [18]. Only when the data have been properly refined can they reach their full potential [19].

While it therefore seems that the metaphorical comparison is sound and that data are just another commodity to be managed and processed like any other asset, a closer look reveals significant differences between the digital commodity 'data' and physical commodities such as oil [20]. These differences are so fundamental that it is necessary to rethink the way data are handled in order to be effective and efficient in the process [21].

In this paper, we therefore study and discuss which characteristics are unique to the intangible commodity 'data' and the resulting implications for modern data administration. To this end, we make the following three contributions:

(a) We discuss the key differences between data and oil. For that purpose, we identify ten specific characteristics of data that need to be taken into account in data administration. In the context of this work, we focus on inherent challenges that arise due to technical characteristics of big data, often referred to as Big Vs [22]. Ethical social aspects, e.g., data liberation and fair distribution [23], or green processing, e.g., energy-efficient data acquisition and storage [24], are out of scope since such factors must also be considered for any other commodity.

(b) For each identified special characteristic, we derive the resulting implications for data administration. In the context of this work, we take an end-to-end look at the data management process, i.e., we deal with data acquisition, data refinement, data storage, and data provision.

(c) We present a concept for a novel reliable information retrieval and delivery platform, called REFINERY Platform, which addresses the data administration challenges, we have identified. In this context, 'reliable' refers to both the data producers—in those terms, it is ensured that sensitive data are handled in a trustworthy manner—and the data consumers—in those terms, it is ensured that data have the promised quality.

The remainder of this paper is structured as follows: In Section 2, we elaborate on the characteristics of data that inherently distinguish them from tangible commodities and address specific and novel challenges that arise when handling such intangible assets. We provide an overview of the state of the art in modern data administration in Section 3 and discuss how it responds to these challenges. Based on these findings, we introduce our concept of the REFINERY Platform in Section 4. This holistic end-to-end concept not only takes into account the unique characteristics of data but also addresses the weaknesses of current approaches. We then assess our approach in Section 5. To this end, we present a security and privacy assessment, a feature discussion, a case study, and a performance evaluation. Finally, the lessons learned are summarized in Section 6.

2. Characteristics of Data and Resulting Consequences for Data Administration

As outlined in the introduction, besides the metaphorical level, there are also many similarities between oil and data when thinking about the handling of these two commodities. For instance, both have first to be discovered and then extracted. The extracted product, i.e., the crude oil or the raw data, must then be refined to transform it into a usable resource (i.e., value-added products such as kerosene or information). For this purpose, the commodity has to be cleansed and preprocessed. The resources refined in this way must then be stored and delivered to customers [25].

While the handling of both commodities involves the same steps, there are considerable differences in the actual implementation. These are due to special characteristics of data that distinguish them significantly from oil. From our point of view, ten characteristics must be taken into account to this end. In the following, we present these characteristics and discuss their consequences for data administration.

I. *Data are nonconsumable:* When tangible commodities are transformed into value-added products, they are consumed in the process. This is completely different for the intangible commodity 'data'. Even after data have been fully analyzed and a value-added data product has been generated (e.g., information or knowledge), the raw data are still available. They do not lose their meaning in the process, since they can be processed again, in another way, in order to derive new information or knowledge from it. Therefore, the volume of the data to be administered increases constantly, since processed data are neither consumed nor become worthless. For this reason, data management systems are needed that store these volumes of raw data in a resource-efficient manner as well as concepts that enable efficient access to the data. Without such concepts, data can no longer be retrieved as needed, resulting in an economic loss of value.

II. *Data can be duplicated losslessly:* The supply of tangible commodities is finite. For instance, every single drop of oil is unique and can only be consumed once. Tangible commodities can be thinned down to a certain degree, but this reduces their quality and thus their value. Data, on the other hand, can be duplicated indefinitely and even without any loss. This initially sounds promising to data producers and data providers since it means that their product can never run out. However, this fact also means that the value of data is measured differently than that of oil. While the price of oil is determined primarily by supply and demand, the value of data is determined by how original and unique their content is. If previously unknown correlations can be determined with them, they represent a clear competitive advantage and thus a particularly high value. Whereas the more the data are reproduced—i.e., their content becomes common knowledge—the lower their value becomes. Concepts are therefore needed to ensure that the contents of the data remain as confidential as possible and that only certain data consumers gain insight.

III. *Data are generated at high velocity:* Tangible commodities are available wherever they arise naturally. For instance, crude oil remains in the Earth's crust until it is extracted. This can be done based on demand and free capacities in the refineries. Data, on the other hand, are generated at any point in time. A processable data object is obtained only if they are captured at exactly this point in time. If they are not captured, they are lost. However, since many IoT devices have limited memory resources, they cannot store the captured data indefinitely but rely on a stream-based processing concept, i.e., they process the data on the fly. Therefore, while oil requires a pull model (i.e., it is acquired from the source when needed), data require a push model (i.e., the source transmits the data when captured). Since data currently accumulate at a high velocity, data storage systems must either have large input buffers to temporarily store new data until screening and further processing or have the necessary capacities to handle voluminous data streams.

IV. *Data are volatile:* Oil has no expiration date, which is why processing is not time critical. Yet, data are volatile. Although a data object can be stored indefinitely, its content is sometimes only relevant for a very short time. For instance, if a sensor in a driverless car detects that there is an obstacle in the lane, this information must be processed immediately, since it is only relevant until the collision occurs. In other cases, data also become invalid. For instance, if the driverless car detects that a traffic light is green, this information is rendered invalid as soon as the light changes to red. Data storage systems must therefore be able to cope with this limited lifespan and have the capability to process data in (near) real time. While some tangible commodities also have a limited shelf life, the volatility of data dynamically differs from data object to data object, and it is often not possible to specify its expiration date in advance, i.e., how quickly the data must be processed.

V. *Data are heterogeneous:* Tangible commodities are usually homogeneous. Although each drop of oil is unique (i.e., it exists only once), all drops from one source are identical in terms of properties such as purity and quality. Therefore, all extracted oil can be stored in a common container and refined similarly. Data, meanwhile, are heterogeneous. For instance, they can have different data formats or schemata and have different levels of completeness and accuracy. Furthermore, the contents of the data differ. Therefore, data cannot be stored in a common storage. Either all data must initially be transformed into a common structure or data stores that support heterogeneous structures are required. Metadata management is also required to track the properties of the data so that they can be handled appropriately.

VI. *Data refinement has to be in accordance with the data source and intended use:* There are established refinement processes for tangible commodities in order to convert them into certain value-added products. Even though these processes may be adapted over time due to new findings or technical innovations, they can be seen as static processes. With data, this is completely different. On the one hand, new and improved cleansing and preparation techniques are constantly developed, which require adjustments to the data refinement process. On the other hand, due to the heterogeneity of the data, a variety of processing steps geared to the raw data are required. This is aggravated by the fact that there is no one-size-fits-all data refinement process. Rather, adjustments must be made to the steps of the data refinement process depending on the intended use of the data. Only if the process is tailored to both the raw data and the intended use can an optimal result can be achieved. Therefore, data refinement requires flexible adjustments to dynamically respond to changes in sources (i.e., the raw data) and sinks (i.e., the intended use).

VII. *The economic value of data is uncertain:* The value of tangible commodities is generally known. There are some fluctuations due to supply and demand, and over time, commodities can gain or lose value. However, these fluctuations tend to be rather small, while substantial changes are extremely rare. With data, this is completely different. Here, the economic value is initially completely unknown. Data that appear to be worthless today may prove to be needle-movers tomorrow. The reason for this is on the one hand that the derivable knowledge cannot be identified in advance but only when the data have been processed. On the other hand, in such a highly dynamic environment, new use cases for data are constantly emerging, which subsequently define the need and thus the value of the data. Since it is almost impossible to anticipate this need in advance, data administration must be able to manage and process data as cost-effectively as possible, since it is not feasible to distinguish between worthless and valuable data.

VIII. *Data can be manipulated indiscernibly:* Tangible commodities are usually relatively resilient to manipulation. For instance, crude oil could be deliberately contaminated, but this can be detected and subsequently purified. In the worst case, sources can be corrupted to such an extent that they become unusable. However, this problem is far worse in the case of intangible commodities and, in particular, data. Data can be manipulated indiscernibly and, above all, in a targeted manner. Malicious parties can falsify data either in their favor or to harm the data consumers, blend fake data with real data, or withhold data. This can happen both when transferring data from the sources and while storing the data. Since the manipulation generally goes unnoticed, it is also almost impossible to undo the contamination. To make matters worse, besides third parties, data producers themselves may have an interest in falsifying the data they provide. Measures must therefore be taken to verify the authenticity and genuineness of data and to prevent subsequent manipulation.

IX. *Data may be subject to special restrictions:* Tangible commodities such as oil are primarily subject to rights related to ownership. Whoever owns the oil well may extract, refine, and sell the oil. With regard to the last two issues, there may be further restrictions, e.g., regarding the environmental friendliness of the refining process or regarding sanctions that affect exports to certain customers. However, these restrictions always relate to the product as a whole. With data, the situation is much more complex. In particular, when it comes to personal data, the data subject (which is not necessarily the data producer) has far-reaching rights when it comes to data processing. For instance, the consent of the data subject is required for the processing of such data. This consent can be withdrawn at any time. However, even with the consent of the data subject, there are further restrictions to be observed when processing personal data, such as a purpose limitation or data minimization. Furthermore, the data subject has the right to request that all data about him or her be erased. Yet, this applies not only to the raw data themselves, but also to all data products in which the raw data in question have been incorporated. Data administration must therefore take measures to implement such privacy rights. These include, e.g., the use of privacy filters that either anonymize data or reduce the amount of contained information to a required minimum, or provenance mechanisms that make it possible to trace which raw data has been incorporated into which data products.

X. *Data require new trading concepts and infrastructures:* When trading tangible commodities, the main problem is to build distribution infrastructures that bring the goods to international trading partners in time. This is not the case with data. Thanks to the Internet, data can be made available in an instant anywhere in the world. Explicit distribution channels therefore do not need to be established. With data, however, three novel trade problems arise: First, due to the large amount of constantly emerging data, there is an elevated risk of losing track of the available data. However, potential customers must be able to find data that are relevant to them. Second, customers must be able to rely on the provided data. This means that they must be able to use the data for their purposes and that there are no conflicting confidentiality or privacy restrictions. For instance, if privacy filters have to be applied to the data in advance, this contaminates the data and reduces the quality of the data. Data administration must therefore ensure that a customer can rely on the authenticity and quality of the data despite the use of such privacy techniques. Third, the privacy requirements of data subjects as well as the quality requirements of data consumers change dynamically. Therefore, it is not possible to offer static data products, but the data refinement must be constantly adapted to offer tailored data products. A trading platform for data must therefore establish concepts to cope with these three problems.

In summary, these ten inherent differences we identified between oil and data also lead to a significant difference regarding the handling of these commodities. The differences are related to three pillars in particular: Novel concepts and techniques must be developed regarding the administration of data so that this can be done efficiently. That is, the large volumes of data that are generated at high velocity must be handled, tailored cleansing and transformation measures must be applied, and access structures must be established for facilitating the retrieval of the data products. Due to their economic value today, data must be protected against illegal access, manipulation, and unauthorized erasures. This requires end-to-end measures, from the authentication of data sources and the secure storage of data to an appropriate access control system for ready-to-use data products. Finally, data protection laws now make it essential to implement privacy by design and by default concepts whenever personal data are processed. In the following section, we therefore look at how related work addresses these challenges.

3. Related Work

In this section, we review the state of the art in data handling. In the context of our work, three research directions are of particular interest, namely data administration [26], data security [27], and data privacy [28]. We discuss these three research areas in Section 3.1 to Section 3.3 and identify research focuses within these areas. The resulting hierarchical classification of related work is shown in Figure 1. We summarize in Section 3.4 the findings regarding the state of research and discuss what open questions remain to be addressed.

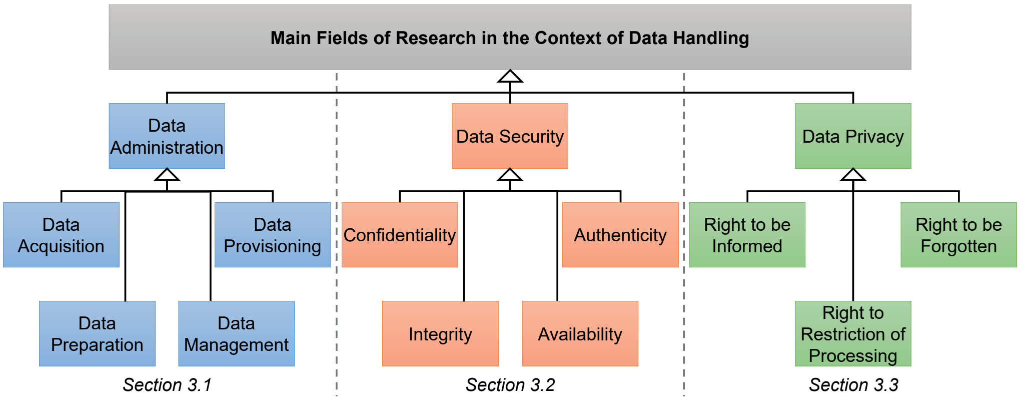


Figure 1. Hierarchical Classification of Research in the Context of Data Handling.

3.1. Data Administration

Data administration comprises all data science tasks in the context of data refinement, i.e., all steps necessary to gain knowledge from raw data [29]. We divide them into the core tasks of a data refinement process, namely selection and extraction of raw data (i.e., data acquisition), data cleansing and data transformation (i.e., data preparation), data storage and data curation (i.e., data management), and distribute refined data (i.e., data provisioning) [30]. Research approaches to assist with these tasks are discussed next.

Data Acquisition. In the context of this work, data acquisition refers to the process of selecting relevant data from a wide variety of sources and then gathering them in a central data management architecture [31]. This represents the first step in the big data value chain and thus enables data to become an asset in the first place [32]. Due to the prevailing heterogeneity among data sources and schemas in which the raw data are available, a systematic approach is required for data acquisition. The so-called ETL process (ETL stands for extraction, transformation, loading) represents a well-established three-step process in which adapter technologies are used to first gather the selected data from the sources, then sanitize them, and finally store them in a structured form [33]. However, this process assumes that there is a central data management architecture with a uniform data schema and that the data in the sources are at rest, i.e., can be retrieved at any point in time [34]. However, due to the IoT, such a conception is outdated. Here, data have a high variety of features even within a single source. Therefore, a rigid target schema is not practicable, since information would be lost in the merger [35]. Moreover, the data are in motion, i.e., they are sent out as a data stream immediately after they are captured by the source, and they accrue in a large volume and at a high velocity, which requires adjustments to the ETL process [36]. Thus, modern-day data acquisition approaches must offer a combination of near real-time processing for streaming data (e.g., via Kafka [37]) and traditional batch-based processing for data at rest [38]. To store the collected raw data, modern data management systems such as Apache Hive (see https://hive.apache.org/; accessed on 6 February 2023) are suitable, as they are not only able to store large volumes of data efficiently, but also cope with heterogeneity within the data [39]. This way, an acquisition infrastructure for big data can be implemented based on the lambda architecture [40]. In the lambda architecture principle, a batch processing layer and a stream processing layer acquire and preprocess data independently of each other and then make them available via a merging layer [41]. A fundamental problem with this architecture is that two separate implementations of the same preprocessing logic need to be maintained for the batch processing layer and the stream processing layer, respectively. In the kappa architecture, all data are therefore gathered and preprocessed as micro-batches by a single stream processing system and made available in a schemaless mass storage system [42]. However, the composition of the micro-batches results in either latency (if the batches are too big which results in long waiting times until sufficient data are available) or a high overhead (if the batches are too small and therefore a lot of batches have to be processed). So, there are sacrifices to be made with both approaches. Therefore, approaches such as the delta architecture aim to combine these two architectures to achieve real-time processing with the ability to handle bulk data efficiently [43]. Nevertheless, more comprehensive preprocessing operations should be performed detached from data acquisition as part of subsequent data preparation [44].

Data Preparation. Once data acquisition has been completed, the collected raw data must be converted into a machine-processable form via data cleansing, transforming, adding metadata, and harmonizing the schemas. These activities are collectively referred to as data preparation [45]. To carry out data preparation effectively, both data knowledge and domain knowledge are urgently needed [46]. The term 'human in the loop' encompasses approaches that empower experts without IT knowledge to actively participate in the data preparation process and contribute their expertise [47]. Research approaches therefore aim to cluster thematically related data sources and thereby represent the sources as a knowledge network so that users can easily identify further relevant sources [48]. Alternative approaches aim to group sources based on the features of their data, as they may need similar data preparation steps [49] or to suggest which data transforming operations are appropriate for such data [50]. Furthermore, the knowledge of the experts can be persisted in the form of a knowledge base that users can leverage in data preparation [51]. Sampling approaches aim directly at facilitating the work of experts by presenting them with only a representative sample of the complete base data. On this manageable sample, the expert defines cleansing and transforming steps, which are subsequently applied to the entire dataset [52]. Here, efficiency and effectiveness can be significantly increased if the data are initially divided into semantically related blocks, which are then taken into account for sampling [53]. Such defined cleansing and transforming steps can be converted into data preparation rules, which can be applied semi-automatically also to new datasets in the future [54]. These rules describe how to obtain processable data from raw data. For reasons of transparency, however, the backward direction must also be provided in order to be able to disclose later on which base data a result was obtained [55]. Why- and how-provenance can be used for this purpose [56]. In addition to human-in-the-loop approaches, however, there is also the countertrend, namely AI-assisted fully automated data preparation [57]. However, this results in a chicken-and-egg problem, since good and reliable training data are required to train the AI—this training data, however, also requires sound data preparation [58].

Data Management. For the management of the processed data, data warehouses were state-of-the-art technology for a long time. Here, the data from multiple data sources are organized in a unified structure that is optimized for tailorable but predefined analysis purposes [59]. However, due to the IoT, the heterogeneity of data sources as well as the amount of semistructured or outright unstructured data increased drastically. Moreover, as data became an essential asset, there is a need for comprehensive and flexible data analysis. The rigid structure of data warehouses is not designed for either [60]. Although there are approaches to describe the semantics of inherently unstructured data to make them processable in a data warehouse [61], they are always limited to specific types of data. Since a rigid data structure is an inherent property of a data warehouse, such approaches do not solve the fundamental issues when dealing with IoT data. Data lakes are intended to overcome these challenges. The basic idea is that all raw data are stored (almost) untouched and data preparation takes place dynamically depending on the respective use case [62]. Thus, a data lake pursues a schema-on-read philosophy, i.e., only when data are processed, a schema that is appropriate for the data and the intended usage is defined and applied [63]. To reduce the resulting overhead that occurs with every data access and to facilitate data governance in general, a zone architecture for the data lake is highly recommended. Each zone provides data at a certain processing stage [64]. However, data lakes are rather concepts than clearly specified architectures [65]. Research approaches therefore attempt to create a reference model for a data lake in which, besides the raw data, harmonized data (i.e., data with a consolidated schema) and distilled data (i.e., aggregated data) are also kept in dedicated zones. In a sandbox zone, data scientists can play around with the data at will, in order to enable exploratory data analytics and provide the full flexibility of a data lake [66]. While this unlimited freedom initially sounds appealing, this might flood the data lake with too much irrelevant data—the data lake increasingly degenerates into a data swamp in which no useful data can be found. This can be prevented on the one hand by a systematic metadata management to keep an overview of all collected data [67] and on the other hand by sanitizing the raw data in terms of detecting integrity violations in the data and dealing with them to maintain the quality [68]. In practice, however, such monolithic data stores are prone to be exceedingly difficult to manage in terms of governance and operation. Research therefore aims to develop a data mesh in which the central data lake is split into distributed, independently managed data silos [69]. Other approaches focus on achieving an optimal trade-off between the flexibility of data lakes (i.e., support for all current and future use cases) and the structured organization of a data warehouses (i.e., efficient data processing and effective information retrieval) [70]. To this end, the data lakehouse architecture supports both classic BI and exploratory data analytics by means of a transaction layer that provides a logical ETL process on top of a data lake [71].

Data Provisioning. In recent years, self-service BI has become increasingly relevant, i.e., users should be able to conduct customized analytics autonomously for their individual use cases [72]. However, a basic requirement to this end is that there is simple access to the relevant data of the required quality [73]. Due to the large amount of data required for today's analyses, the data available internally in a corporation is often not sufficient. Therefore, data from external providers are required as well. As a consequence, data is not only a commodity but has also become a tradable good. To address this new strategic role of data, infrastructures for data marketplaces are being developed to allow customers to find and obtain data products [74]. However, a data marketplace is not a traditional warehouse, but because of the intangible nature of data, rather, a storefront for the available data products. Customers can select the data they want from a data catalog and the marketplace then acts as an interface to the respective data store [75]. From a data provider perspective, one of the most important functionalities that a data marketplace has to offer for this purpose is a comprehensive metadata management system that allows them to describe their data. This includes descriptive information about the data themselves (e.g., their data model or content descriptions) as well as information about the conditions under which they are permitted to be accessed (e.g., their price or their permitted usage) [76]. Since a marketplace usually represents the storefront for multiple third-party data providers, the metadata of all these providers must be merged to assemble a holistic data catalog [77]. From a customer perspective, the data marketplace must facilitate data retrieval. To this end, two main functionalities must be supported: On the one hand, it must be possible to find relevant data from all available sources (e.g., content- or quality-wise), and on the other hand, data acquisition has to be simple [78]. Comprehensive metadata management is required to this end as well [79]. One of the most important aspects of a data marketplace for both sides, however, is trust. Only if data providers can assume that confidentiality and privacy are guaranteed with regard to their data and customers can rely on the authenticity and quality of the offered data, they will use a data marketplace [80]. Therefore, data security and data privacy are central issues in the context of data provisioning.

3.2. Data Security

In modern data administration, four protection goals in particular have to be addressed, namely confidentiality, integrity, availability, and authenticity [81]. Next, we discuss research approaches designed to fulfill these protection goals. We look at the protection goals separately. In the practical application, however, there are correlations with other protection goals. For instance, effective authenticity is a cornerstone of access control and thus a prerequisite for all the other protection goals. The protection goals are also contradictory to some extent, e.g., the highest level of confidentiality can be achieved if no one has access to the data. Yet, this conflicts with availability [82]. Such mutual effects have to be taken into account when adopting solutions to ensure the protection goals.

In our discussion, we refer to the definitions of these four protection goals given in the ISO/IEC 27000-series [83], which is the internationally recognized standard for information security management systems. Confidentiality means that information is not disclosed to unauthorized third parties. Integrity ensures that the information made available is complete and uncorrupted. Availability means that authorized parties have access to the information at all times. Finally, authenticity ensures that both the origin of the information and the identity of the parties interacting with the information are verified.

Confidentiality. To protect against the disclosure of sensitive information, cryptography approaches are typically used. That is, data are available in encrypted form and can only be decrypted (and thus read) with the appropriate key. Both symmetric encryption—in which the same key is used for encryption and decryption—and asymmetric encryption—in which a key pair with different keys for encryption and decryption is used—can be applied to this end. While symmetric encryption approaches generally require less encryption time, asymmetric encryption approaches facilitate key distribution, since the private key always remains with the key owner, while the corresponding public key is shared with anybody without compromising confidentiality [84]. Combinations of these two techniques are also found, particularly in IoT environments, in order to reconcile simple key management with reduced hardware requirements [85]. To reduce the overall decryption effort required for each data access, homomorphic encryption can be used. Here, encrypted data are also unreadable without a corresponding key, but certain predefined operators can still be applied to them, such as aggregation functions, for statistical surveys [86] or search queries [87]. That is, the data can be preprocessed and even analyzed without being fully exposed [88]. An access control policy can be used to specify who has access to which data and for what purpose [89]. However, since the IoT is a dynamic environment, this must also be reflected in an access control system [90]. Thus, policy rules also have to consider the current context in which data are accessed (e.g., a spatiotemporal context in which the access takes place or a role-based context of the accessor) [91]. As a result, such a policy becomes highly complex, which is why access control systems in practice have to solve problems regarding conflicting policy rules [92] and scalable access management [93].

Integrity. Currently, data integrity is often ensured by the use of blockchain technologies. A blockchain can be regarded as an immutable and tamper-resistant data store. By organizing the data in blocks that are inseparably linked to each other via cryptographic hashing, it can be ensured that neither individual data items within a block nor entire blocks can be manipulated. As this chain of blocks is managed in a distributed manner, i.e., multiple parties manage an equivalent copy of the chain, manipulations can be easily detected and reversed [94]. In addition to providing a secure storage for IoT data [95], however, blockchain technologies also facilitate the trustworthy sharing of sensitive data in inherently semi-trusted or unreliable environments [96]. Yet, inherent problems of blockchain-based data stores are their low throughput due to their serial operating principle in terms of query processing [97] and their limited data access support which results in minimalistic query capabilities [98]. Therefore, there are a variety of research approaches to improve the query performance as well as the query capabilities of blockchain-based data stores. For instance, SQL-based query languages are being developed for blockchain systems to improve usability [99]. In addition, there are query extensions for specific application domains that exceed the SQL standard, such as spatiotemporal queries [100] or top-k queries [101]. Other approaches aim to create schemata for the data in the blocks [102] or cross-block index structures, in order to improve the performance of query processing [103]. However, as blockchain systems not only have low throughput but also high storage costs [104], it is necessary to keep the volume of stored data as low as possible. Therefore, an off-chain strategy is often applied. In this case, the actual payload data are stored in an external data store. The blockchain itself only stores references to the payload data and digital fingerprints in the form of hash codes that can be used to verify the integrity of the data [105]. This way, even traditional relational databases can be extended by the integrity properties of blockchain storages by means of a lightweight blockchain-based verification layer on top of the database [106]. While such an approach can ensure the integrity of the payload data, the same cannot be said for the queries and the query results [107]. For this reason, there are also approaches aimed at deeper integration of blockchain technologies in a relational database system to enable more holistic integrity assurances [108].

Availability. IoT devices generally do not have the capability to permanently store the vast amounts of data they collect, let alone the computing power to adequately process them. As a result, IoT applications rely on cloud providers to store, manage, and analyze their data [109]. Despite the undeniable advantages that cloud services offer in this context, the nontransparent nature of cloud computing requires blind trust on the part of the data owner in the cloud provider [110]. A key concern for data owners is that they have to hand over control of their data to the provider. This also includes where the provider stores the data and whether there are enough replicas of the data to ensure permanent availability [111]. In general, a semihonest provider is assumed in the cloud environment, i.e., a basic level of trust is appropriate, but a provider will always act to maximize its own benefit [112]. For instance, a provider could keep significantly fewer replicas of the data than promised in order to cut storage costs. Initially, there is no noticeable disadvantage for the data owner, but protection against data loss deteriorates considerably as a result [113]. Data owners therefore need tools to enable them to verify whether a cloud provider is storing their data reliably. So-called Proofs of Ownership and Retrievability (PoOR) are one option for this purpose [114]. Here, digital fingerprints of the data managed in the cloud are stored in the form of homomorphic verifiable tags [115] in a Merkle tree [116]. A data owner can pose challenges to the cloud provider, which the provider can only solve if it is in possession of the data. The user can verify the provider's answers using the homomorphic verifiable tags. If this is successful, proof is provided that the data are available without having to download the full data. However, this does not ensure that there is also the promised number of replicas available. Proof of Retrievability and Reliability (PoRR) approaches can be applied to verify this as well [117]. Here, a verifiable delay function (VDF) is applied to the data, which is slow to compute but easy to verify [118]. Therefore, if a data owner poses challenges to instances of the cloud provider that are supposed to hold the data in question that relate to this function, the provider can only answer them if the data are actually at rest here. If the response takes too long, this is proof that there is no replica on the instance and the cloud provider needs to compute the VFD on the fly. In addition to unreliable cloud providers, a central server always represents a bottleneck and thus an inherent weak point with regard to the availability of data in traditional client-server structures. If such a server is flooded with requests, e.g., due to a distributed denial of service attack (DDoS), and thus becomes unavailable, all data managed by it are also no longer available to the clients. IoT environments in particular are severely vulnerable to such attacks [119]. In order to ensure data availability, it is therefore necessary to replace such centralized structures with highly distributed models that store the data in multiple replicas on different nodes to be resilient in the event of a single node failure [120]. Consequently, the use of blockchain technologies is also suitable for ensuring data availability as the blockchain is based on the distributed ledger technology [121]. This refers to technologies that enable data to be stored and shared over distributed computer networks. In simplified terms, a distributed ledger is a data storage system that manages data on multiple computer nodes with equal rights [122]. Due to the decentralized nature, no central authority has control and interpretational sovereignty over the data. Moreover, the collective of nodes can decide which data should be available and thus keep false or harmful data out of the data store [123]. As blockchain-based database systems typically require each node to manage the entire blockchain, this incurs excessive storage costs. Therefore, there are approaches in which only a few server nodes need to store the entire blockchain, while clients can still verify the authenticity of the data by means of authenticated data structures (ADS) [124]. Since this constrains the data distribution, which can endanger availability if the number of expected malicious or compromised nodes is remarkably high, other approaches rely on data partitioning. In sharding, the complete data stock of a blockchain is divided into several parts and distributed to the available nodes according to well-defined rules [125].

Authenticity. In order to identify users, i.e., to verify their authenticity, passwords are commonly used. These can either be real words or PIN codes or lock patterns that have to be entered for authentication [126]. The IoT also offers other authentication options based on biometrics features such as voice, fingerprints, or facial expressions [127]. All of these methods have in common, however, that they can be easily exploited by shoulder surfing during input [128] or replay attacks [129]. To reduce the number of authentications required and thus mitigate some of these threats, there are OAuth-based approaches for the IoT. Here, an authentication service issues a token that authorizes the use of devices or services for a certain period of time [130]. However, this only shifts the problem of illegally acquired authentication data to the OAuth service. To address this, the ownership model relies on using a technical device for authentication that has a unique hardwired fingerprint [131], e.g., by means of physical unclonable functions (PUF) [132]. Yet, the loss of such a device inevitably enables another person to gain possession of the authentication data. In the IoT, this issue is further exacerbated as devices are not linked to specific users but are used by several people. Moreover, IoT devices have limited input capabilities, which means that users cannot enter their credentials like on a traditional computer [133]. For these reasons, there are trends away from 'what you know' (e.g., password-based approaches) or 'what you have' (token-based approaches) authentication toward 'what you are' authentication [134]. In attribute-based approaches, an entity is authenticated based on certain properties it has in the current situation. Especially in dynamic and rapidly changing environments, such a context-based description is advantageous [135]. Due to the high flexibility of attribute-based approaches, they are particularly suitable for IoT applications [136] or cloud-based applications [137]. In addition to users, however, it must also be ensured that data are authentic, in terms of, they come from the specified sources and have not been falsified [138]. For digital media, such as image, sound, or video data, digital watermarking can be used for this purpose. That is, an identification tag is inseparably burned into the carrier medium. If it can be ensured that no unauthorized third parties have access to the identification tag, authenticity can be verified by the presence of the digital watermark [139]. Similar techniques can also be applied to data in relational databases. Here, individual marker bits are inserted into the payload data [140]. While the origin of the data can be proven in this way, watermarking approaches inevitably contaminate the actual data with the inserted identification tags. Digital signatures represent a noise-free approach to ensure the authenticity of data [141]. Here, methods of asymmetric cryptography are used. While asymmetric cryptography uses the public key of a recipient for encryption—thereby ensuring that the message can only be decrypted with the corresponding private key, i.e., only by the recipient—the sender uses his or her own private key for signing. Since the sender's public key is freely available, anyone can decrypt the message. However, this verifies that the sender has used the corresponding private key, which proves the origin of the data beyond doubt [142]. In the IoT, attribute-based signatures are suitable. Here, the attributes of a data source are stored in the signature (e.g., technical specifications of a sensor) and the receiver can verify whether these attributes are sufficient for the data to be authentic (e.g., does the sender have the capabilities to capture the data in the required quality) [143]. Yet, attribute-based signatures are computationally expensive, which is a particular problem in the context of lightweight IoT devices. Thus, research approaches aim to outsource part of the heavy workload to the cloud [144]. Other approaches deal with the problem that the attributes in the signature might contain sensitive information. To this end, a trusted intermediary is installed that preverifies the signatures and removes the sensitive attributes from it [145]. For areas like social media, where such an authentication of sources is not possible, the authenticity of data can be verified based on their content [146]. Then, fake news can either be blocked [147] or overridden by authentic data [148].

3.3. Data Privacy

Due to the Quantified Self movement, i.e., the continuous self-tracking with IoT technology, the amount of personal data is growing exponentially [149]. The analysis of such data is also of high economic value for the industry, as it reveals a lot about potential customers [150]. Therefore, one aspect of data security is becoming increasingly relevant in the context of data processing and management, namely data privacy. The idea of data privacy originates from philosophical theories that predate the information age. These theories reflect the basic need of humans to keep certain information about themselves secret. Historical examples include the Hippocratic Oath [151], the seal of confession [152], and the secrecy of correspondence [153]. In these examples, the owner of the secret trusts that the person keeping the secret will adhere to his duty of professional secrecy [154]. In a broader sense, however, data privacy is motivated by the fundamental human desire not to be at the mercy of external control and to have autonomy over one's personal data [155]. A pioneering work that attempts to define this philosophical idea describes privacy as follows: "Privacy is the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" [156].

To this end, two aspects are particularly noteworthy: On the one hand, it is evident that privacy, unlike the protection goals discussed in Section 3.2, is an individual right—that is, each data subject must be able to decide individually what information s/he wants to shares with society. On the other hand, this definition implies a necessity to provide controlling and regulating measures for the sharing of personal information [157]. Therefore, in our digitalized world, it is particularly important that laws comprehensively preserve the privacy of all individuals. This includes a plethora of organizational, technical, and legal measures necessary to enable data subjects to enforce these two aspects in the information society of today [158].

Data protection laws such as the General Data Protection Regulation (GDPR) [159] therefore demand technical measures that give data subjects full control over their data. However, the so-called privacy paradox arises here—although data subjects have fundamental privacy concerns, they do not want to give up the comfort and respective benefit they experience from the analysis of their data. In the context of the IoT and its smart services, this problem is even more aggravated [160]. For effective data protection, it is therefore not only important to protect private data, but also to maintain the perceived quality of service. Otherwise, the privacy measures would be rejected by users. Technical privacy measures can be divided into three categories: measures that inform data subjects about potential privacy risks, measures that restrict the sharing and processing of personal data, and measures that erase data permanently [161]. Only with such technical measures it is possible to comply with the requirement for data protection by design and default [162]. Besides these technical measures, data protection laws also deal with organizational measures (e.g., the appointment of a dedicated contact person for data subjects or the obligation that data must not leave a predefined territorial scope) [163]. In the context of this work, however, we are focusing on technical challenges, only.

Right to be Informed. Data protection regulations give data subjects the right to be informed about the processing of their data. Three main research directions can be identified in this context. First, techniques are being developed to explain what knowledge can be derived from certain types of data. For instance, IoT data sources are described with the help of ontologies. Thus, it can be disclosed which data is collected by a sensor (including, e.g., the accuracy and frequency) and which information is derived from the respective data [164]. Based on such ontologies, reasoning can be done about what knowledge can be derived from this information [165]. These ontologies can be regarded as persisted knowledge of technical experts and domain experts. By involving privacy experts, the ontology can be extended in order to identify potential privacy threats and thereby make data subjects aware of critical issues. From this information, abnormal and questionable data usage patterns can be detected [166]. Second, in addition to informing about correlations between data and knowledge, research approaches also aim to describe the actual data processing. For this purpose, a holistic view of data-driven applications and processes is provided in order to identify inherent privacy threats. In the field of system safety, the STAMP framework (STAMPS stands for System-Theoretic Accident Model and Processes) provides such a top-down approach. Here, causal relationships between hazards and actions are initially modeled. Based on this model, the System-Theoretic Process Analysis (STPA) can be performed to identify problems in the design of the system that lead to or facilitate these hazardous actions [167]. This approach can be adapted to the field of privacy threats. Here, potential privacy threats posed by a data-processing application have to be identified initially, e.g., using the ontologies described earlier. The components of the application are then analyzed to determine whether they are sufficiently protected against these privacy threats [168]. Such a top-down analysis is ideal for modeling and analyzing complex data-driven systems and informing data subjects about inherent privacy-related vulnerabilities with respect to the data sources involved [169]. Third, data subjects also have a right to be informed about the data products for which their data was used as input. Machine learning models represent a prime example. Although these models can be processed by machines easily, they are usually a black box for humans. When applying such a model, it is generally nontransparent why it came to a certain result (e.g., a prediction or a suggestion). Since the models may be flawed or unfair due to an insufficient amount of appropriate training data, it is crucial that the results (and thus the models themselves) are comprehensible to data subjects [170]. To this end, there are three research areas: In the field of interpretable models, methods are developed which generate explainable models. In the field of model induction, models, which represent a black box, are transformed into an explainable model. Yet, both approaches have limitations when it comes to deep learning, as models in this context are far more complex. In deep explanation, deep learning algorithms are therefore adapted in such a way that not the model but at least the relevance of individual input factors are identified across the layers of the model, in order to determine what eventually led to a decision [171]. Yet, there are still many open questions to be solved, especially in the area of deep learning, before full transparency is achieved [172].

Right to Restriction of Processing. However, all this information is of little use to a data subject in exercising his or her digital self-determination if s/he cannot also object to the data processing and enforce this technically. To this end, however, a binary consent system is far too restrictive. Here, a rejection leads to a massive reduction in service quality, which tempts data subjects to agree to all requests. Therefore, this is not an actual informed consent [173]. Instead, privacy-enhancing technologies (PET) should be used to minimize the data—or rather the amount of information they contain—in a target-oriented manner in accordance with the privacy requirements of the data subjects. Three fundamentally different types of PET can be identified: obfuscation techniques for users, statistical disclosure control for mass data providers, and distribution of data across multiple trusted third parties [174]. The obfuscation techniques for users apply privacy filters to the data. Very simple filters allow horizontal filtering [175]—which corresponds to the selection operator in relational algebra, i.e., filtering out specific data items from a dataset—or vertical filtering [176]—which corresponds to the projection operator in relational algebra, i.e., filtering out specific features of a data item. Also, an aggregation, i.e., condensing many data items to a single representative data item (e.g., the mean), can be used as a privacy filter [177]. However, such generic privacy filters are rather coarse-grained and therefore severely impair the data quality. Privacy filters that are tailored to a specific type of data are therefore more appropriate, as they are fine-grained and thus less invasive [178]. For instance, there are privacy filters that are tailored to location data and can optionally obfuscate individual locations or entire trajectories [179], privacy filters for health data that enable certain types of examinations only, while rendering the data unusable for all other analyses [180], or privacy filters that mask the voice or remove revealing background noise in speech data [181]. Other approaches focus on complex events that represent a specific sequence of data items instead of individual data items. Privacy-critical events are concealed, e.g., by reordering the data items or dropping or inserting data items. This preserves privacy without reducing the data quality of the data themselves [182]. Statistical disclosure controls for mass data providers include methods in which individual data subjects are hidden in the faceless masses formed by all users [183]. For instance, k-anonymity approaches ensure that a data item can only be mapped to a group of k users. With a sufficiently large k, no private information can be derived from the data about each of the k individuals [184]. Differential privacy is intended to protect an individual even more extensively. It ensures, e.g., by adding noise, that statistical analyses cannot determine whether the data of an individual data subject contributed (significantly) to the outcome [185]. While this sounds tempting in theory, differential privacy often turns out to be complex to implement and very costly to compute in practice [186]. If parameterized inappropriately, it even offers little protection and therefore leads to a false sense of security [187]. Federated learning can be regarded as an approach in the field of distribution of data across multiple trusted third parties. Here, the data with which a machine learning model is to be trained is distributed among several parties. Each party calculates its own local model, which is then incorporated into a global model. By partitioning the data appropriately, it can be ensured that no party gains a comprehensive insight into the data [188]. By using privacy filters when creating the local models, it can also be ensured that no unintended conclusions can be drawn from the global model [189].

Right to be Forgotten. The right to be informed and the right to restriction of processing are, however, not sufficient to provide all-round protection. As data can become outdated, the privacy requirements of data subjects can change, or data may have been unlawfully transferred to data providers in the first place, there is also a need for a right to have all personal data erased. This also includes all related metadata as well as all data products in which these data have been integrated. In addition, personal data also have an inherent lifespan, after which they must also be completely deleted, as they may not be stored longer than required for the intended purpose. In terms of today's data protection regulations, data erasure has to ensure that the data cannot be restored in any way or form [190]. From a technical point of view, therefore, measures are needed to enable secure deletion of all the data concerned. In the context of providers of big data, the large volume of data from a vast number of users means that, from an economic point of view, it is not feasible to apply deletion methods that destroy the data carriers. To this end, there are nondestructive approaches for both hard disk drives and solid-state drives. While overwriting-based approaches are effective for hard disk drives, where the sectors containing the data in question are overwritten several times with other data, erasure-based approaches for solid-state drives ensure that the flash block that provides access to the flash page containing the data is erased. As a result, the data can no longer be accessed and is finally dumped by the garbage collection [191]. So, while in theory, there are approaches to enable secure deletion for those physical data carriers, these approaches have limitations in terms of logical data storage. For instance, in the case of databases, such an approach is not possible as abstraction layers prevent a direct mapping of data to sectors or flash pages [192]. Blockchain-based data stores represent another example where this type of secure erasure is unsuccessful, as here the data are immutable, i.e., a deletion renders the blockchain invalid [193]. Cloud-based data stores also require other approaches, since only the cloud provider has control over the physical storages and the knowledge of how many replicas are held on which nodes [194]. In these cases, encryption-based approaches can be used. Here, the data are stored fully encrypted. This way, it is sufficient to delete the decryption key, which renders the encrypted data unreadable in the process. Hierarchical key management enables fine-grained deletion of individual data items, clusters of related data, or all data of a data subject at once [195]. Provenance analyses can be used to identify all data products that have been created based on a specific data item [196]. These data products must also be erased if a data subject requests to have their base data (or any part of it) deleted. It is evident that such a provenance analysis also generates a lot of additional data that potentially disclose private information [197]. To this end, when answering this kind of provenance queries, it is therefore necessary to rely on special privacy-aware approaches, such as intensional provenance answers [198].

3.4. Lessons Learned

As illustrated in Section 2, there are ten key differences between data and tangible commodities such as oil that make them special assets. These special characteristics must also be taken into account when handling and managing data in order to use this commodity effectively and efficiently. Our review of the state of the art shows that there are many research approaches to this end. They can be divided into three categories: In the area of data administration, methods and techniques are being explored that enable or facilitate data acquisition, data preparation, data management, and data provisioning. Thereby, the special challenges in today's big data context (namely, volume, variety, and velocity) are addressed. In the area of data security, approaches are explored that ensure the confidentiality, integrity, and authenticity of the data in order to verify their veracity and thus preserve their value. Furthermore, methods and techniques are developed that guarantee high availability of the data and ensure that only authorized entities have access to them. Data privacy is a branch of data security that plays an important role currently due to the increasing importance of personal data. In the context of this work, we focus on technical issues, whereas legal, ethical, and organizational research is out of scope. In this regard, there are approaches that provide data subjects with more information regarding the processing of their data as well as identifying potential privacy threats. Other approaches address how data processing can be restricted in a fine-grained manner. PET ensure that no sensitive knowledge can be gained from the data without compromising the overall data quality. Finally, methods are being developed to erase data permanently to ensure that they can no longer be processed. In addition to the actual data, this also includes all replicas and metadata as well as all data products that have been created based on those data.

However, all of these research approaches are island solutions to individual data administration, security, and privacy problems. Yet, these three aspects are highly interrelated. For instance, a data marketplace cannot provide accountability for data if their origin cannot be traced through provenance and authentication of the sources, while privacy can only be guaranteed if confidentiality is ensured as well. Moreover, data administration can only be effective if availability is ensured. Thus, these aspects must not be considered independently of each other. Even within the individual research areas, the isolated approach represents a disadvantage. For instance, obfuscation techniques for users are immensely powerful because they give data subjects full control over their data. With statistical disclosure control for mass data providers, they lose this control. However, due to the holistic view on all available data, data providers are able to apply privacy techniques in a much more target-oriented manner. Therefore, the optimal solution would be a mutually coordinated mix that first provides data subjects with a prefiltering option and then allows data providers to make comprehensive readjustments from their side. Island solutions cannot achieve synergy effects and, even worse, some of them are mutually exclusive. For instance, privacy can easily be achieved by completely randomizing all data. This, however, minimizes data quality and thus renders the data worthless. Another example is the use of blockchain-based data stores to ensure integrity and availability. Yet, immutable data storages inevitably prevent data subjects from exercising their right to be forgotten.

Although these individual solutions are efficient for the respective application context, a holistic end-to-end view from data acquisition to data delivery is required to enable trustworthy and demand-driven data provisioning. 'Trustworthy' refers to enabling data producers to make their data available without losing control over them and without having to fully disclose their assets. 'Demand-driven' refers to ensuring that data consumers are provided with authentic data of the required quality. To this end, it is required to implement and apply appropriate security and privacy mechanisms in all data administration processing steps. Furthermore, such an approach must be both generic and flexible to cope with the great heterogeneity in today's big data landscape.

For this purpose, we have developed a set of solutions for all data processing steps. The novelty of our research is its holistic approach. That is, all our solutions can be integrated seamlessly into an end-to-end platform. Therefore, by combining these concepts appropriately, the result is significantly more than the sum of the individual parts. In the following section, we present this integrated concept called REFINERY Platform as a whole for the first time and describe how its individual components interact with each other.

4. The REFINERY Platform

In our REFINERY Platform, we aim to provide comprehensive support for data administration, starting from any data source and ending with tailored data products. Despite this holistic view of the data administration process, the REFINERY Platform consists of a wide range of individual solutions for each process step, e.g., data acquisition from smart devices and databases, customizable data preparation rules, secure data management, and making the data products visible to data consumers. The main focus here is on ensuring that all concepts are geared to the specific characteristics of data (see Section 2). This includes in particular that the commodity 'data' is handled reliably. On the one hand, this means that it must be feasible for data producers or data subjects to regulate data processing in accordance with their data privacy and confidentiality requirements. On the other hand, the customers of the data products, i.e., the data consumers, must be able to fully trust the authenticity of the data and be reassured that the data are of the promised quality. A high-level overview of the general concept of the REFINERY Platform is shown in Figure 2.

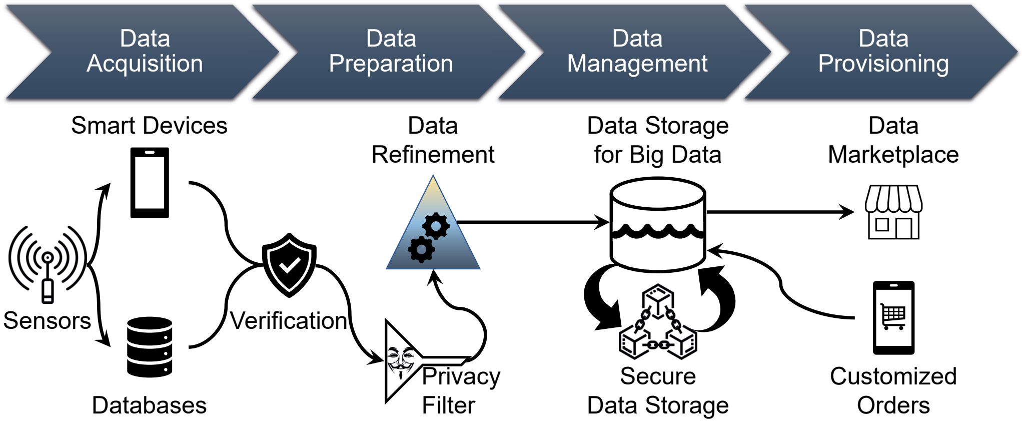


Figure 2. High-Level Overview of the General Concept of the REFINERY Platform.

Due to the IoT, two different types of data sources need to be supported by a data administration platform currently. On the one hand, there is a multitude of smart devices that are able to collect a plethora of different data via connected sensors and share them due to their connectivity. On the other hand, there are mass storage devices in the form of databases, both relational databases and NoSQL data stores, which provide data at rest. In the REFINERY Platform, both types of data sources can be integrated. During acquisition, the data are verified by the REFINERY Platform in terms of whether it possesses the specified properties, e.g., regarding data quality. In addition, the data sources can assert privacy requirements, which are then enforced in the REFINERY Platform in the form of PET, e.g., privacy filters. All three types of PET are supported, i.e., obfuscation techniques for users, statistical disclosure control for mass data providers, and distribution of data across multiple trusted third parties.

The received data prepared in this way are subsequently processed semiautomatically in the REFINERY Platform and data products are generated. For this purpose, domain experts specify rules for the data cleansing and data transformation steps to be applied to the data. These rules can then be automatically applied to the acquired raw data. The semi-automatic approach is necessary to generate customized and high-quality data products, which would not be possible without the involvement of human experts. However, the resulting rule base increasingly reduces the experts' workload in this context, as they can draw on already specified preparation rules (either in parts or as a whole).

Then, both the raw data and the data products have to be managed. The REFINERY Platform uses a big data storage system that enables efficient management of the raw data and the processed data, as well as demand-oriented access to the actual data products. Due to the high economic value of these intangible commodities and digital products, we apply blockchain technologies to secure them against malicious tampering and deletion.

Finally, it must be possible for customers to find the products they need from the vast amount of available data products. For this purpose, the REFINERY Platform has data management structures that support an electronic storefront, which can be used to search the product range efficiently. The metadata gathered and generated by the REFINERY Platform enables straightforward information retrieval. In addition, customers can use the metadata to inform themselves about existing raw data and place orders for customized data products. That is, they describe their own data refinement steps, which are then carried out by the REFINERY Platform to provide data products tailored to their needs.

The concepts used in the REFINERY Platform to this end are described in more detail in the following. The structure of the section reflects the data administration tasks as defined in Section 3.1 since these tasks represent the value generation in terms of data refinement. First, the concepts of data acquisition are outlined in Section 4.1. Subsequently, Section 4.2 describes the data preparation in the REFINERY Platform. The management concepts are covered in Section 4.3, while data provisioning is addressed in Section 4.4. The protection goals regarding data security and privacy constitute value-added services, only. Yet, these matters are taken into account in every component of the REFINERY Platform.

4.1. Data Acquisition

In the area of data acquisition, it is important to give data producers extensive options to control which of their data they want to share with third parties, i.e., which information they want to disclose and which knowledge they want to reveal. Without such options, they would be faced with the binary choice of releasing all or none of their data. However, this would inevitably mean that the REFINERY Platform would have to exclude many potential data sources upfront.

To this end, we have developed a fine-grained permission system that allows data producers to specify which information content they want to share with the REFINERY Platform at all, and which knowledge third parties are allowed to derive from it for a specific purpose. As observed in Section 3.3, there are two fundamentally different approaches to regulating the information content of data, namely the application of PET either on the user side (e.g., obfuscation techniques for users or a distribution of data across multiple trusted users, respectively) or on the side of the mass data providers (e.g., statistical disclosure control). Both approaches have their intrinsic strengths and weaknesses. For instance, if data regulation takes place in the user-controlled area, the data producer has full control over his or her data. In contrast, privacy measures applied by mass data providers are often much more effective as here all available data sources are known. Moreover, a mass data provider can select PET which are adjusted to the subsequent data processing steps.

As a result, PET can be applied to the data in a target-oriented manner, which reduces the impact on data quality.

In the REFINERY Platform, we therefore combine these two approaches to obtain the best of both worlds. Our solution is shown in Figure 3. First, the data producer specifies his or her privacy requirements (more on this in Section 4.4). These privacy requirements are then split using a tailorable metric based on a data quality and privacy budget into policies that have to be deployed in the user-controlled area and policies that have to be deployed at the mass data provider. That is, it is determined how much data can be disclosed unmodified to the mass data provider in order to achieve a higher data quality and which data are so confidential that they have to be distorted already by the data producers themselves. As a mass data provider usually receives data from more than one data producer, this also achieves data distribution, since some raw data never leave the sphere of influence of the respective data producer in unprocessed form [199]. How the privacy policies are applied in the REFINERY Platform at the mass data provider is described in Sections 4.2 and 4.3. In this subsection, we discuss how the policies are applied to the data during acquisition, i.e., in the user-controlled area.

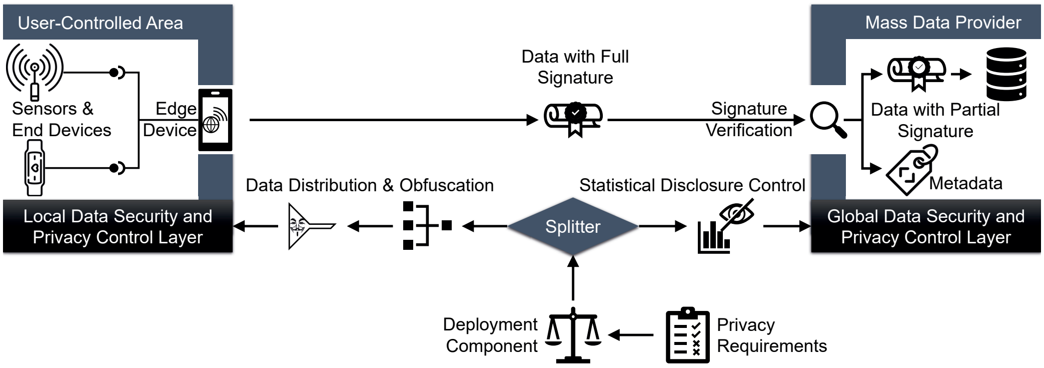

User-Controlled Area
Sensors & End Devices
Edge Device
Data with Full Signature
Signature Verification
Mass Data Provider
Data with Partial Signature
Metadata
Local Data Security and Privacy Control Layer
Data Distribution & Obfuscation
Splitter
Statistical Disclosure Control
Global Data Security and Privacy Control Layer
Deployment Component
Privacy Requirements

Figure 3. Deployment Concept of PET as Part of the Data Acquisition in the REFINERY Platform.

For this purpose, we have developed the Privacy Management Platform (PMP) for Android-based IoT edge devices (see https://www.android.com/; accessed on 6 February 2023). In this context, an IoT edge device is a device that is capable of running third-party applications (i.e., not just a fixed set of applications hardcoded into it by the manufacturer) and has sufficient computing power to process a reasonable amount of data. In addition, an edge device has the ability to connect to the Internet or cloud services. Thus, in our use case, it represents the interface to the mass data provider for all connected sensors and IoT end devices of a single user.

Our PMP is able to enforce privacy policies that describe which data consumer (or which application, respectively) is allowed to have access to which data. In the most basic case, this is implemented by means of horizontal or vertical filtering [200]. However, since such an approach is restrictive and severely compromises either the data volume or the data quality, more fine-grained privacy filters can also be applied. For this purpose, we have studied privacy filters that can remove different information contents from time-series data (i.e., the predominant type of data in the IoT domain). For instance, individual data items can be concealed while preserving the overall temporal progression, or resilient noise can be added to all data to reduce the accuracy of all values (and thus the information content). These filters are rather generic, so they can be applied to data from any sensor [201]. In addition to these generic privacy filters, specialized filters can also be applied, which are adjusted to specific types of data. For instance, in the case of location data, the accuracy can be decreased so that only the approximate location can be determined. Our approach is also extensible in the sense that additional specialized filters for other types of data can be added to the PMP retroactively [202].

From a logical perspective, the PMP therefore represents an intermediate layer that isolates the data-consuming applications from the data-producing operating system (OS). In other words, any data flow has to be handled by the PMP, which means that the PMP has full data sovereignty. In order to assure this, two properties must apply to the PMP: First, the PMP has to be able to provide all kinds of data that can be requested by applications. Second, the PMP has to be able to prevent applications from bypassing it, i.e., there must be no data flow that is not controlled by the PMP.

The former is programmatically realized by means of an adapter concept. There is an adapter in the PMP for each data source that is accessible via the underlying OS. In this context, it is irrelevant whether streaming data (e.g., data originating from a sensor) or data at rest (e.g., data stored in a database or a file system) are involved. Each adapter essentially replicates the interfaces of its underlying data source. This way, any data from the data source can be simply passed through. In addition, however, privacy filters are implemented in the adapters, which are applicable to the respective type of data. This allows the adapter also to forward modified versions of the data instead of the original data to meet the privacy requirements of the data subject. The interactions between these components are shown in Figure 4 as a sequence diagram.

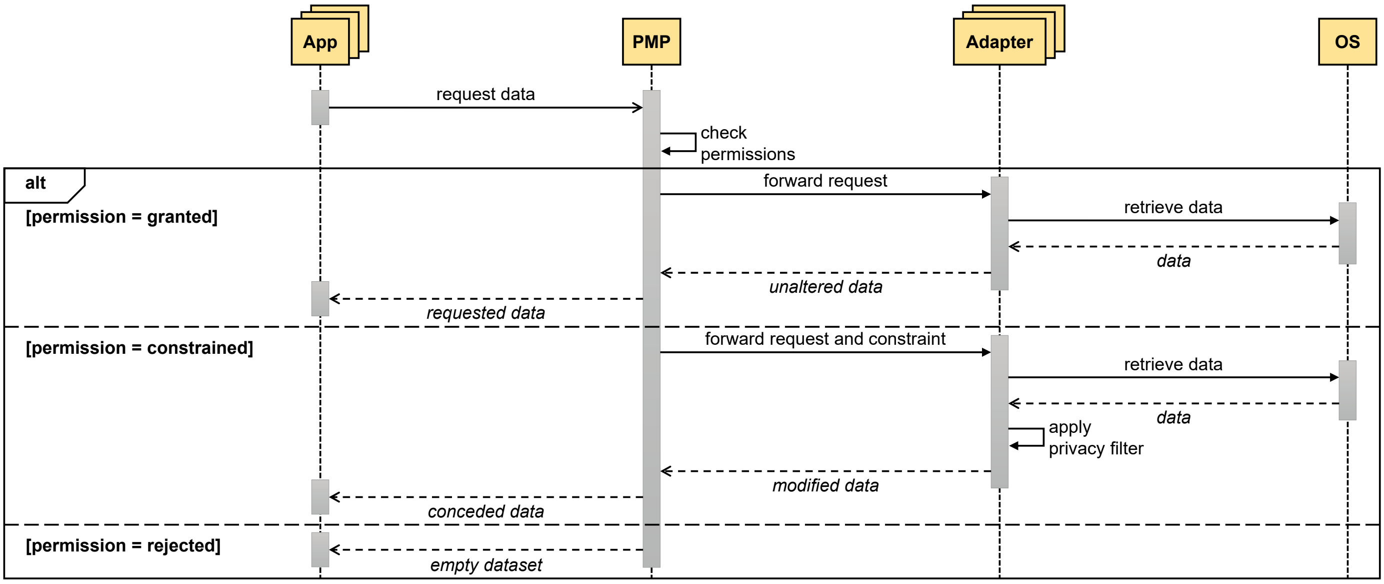


Figure 4. Sequential Processing of a Data Request Including All Interactions Between Applications, the PMP, Adapters, and the OS.

When an application (referred to as 'app' in the figure) requests data from the PMP, the PMP checks whether the data subject has permitted this access. Data subjects can assign one of three permissions: a request can be granted, constrained, or rejected. The latter is the default if no permission is specified for the requesting application. If the request is granted, the PMP forwards it to the corresponding adapter, which fetches the data from the OS, and the unaltered data are provided to the requesting application. If constraints are specified for the data access, these constraints are forwarded to the adapter along with the request. In this case, the adapter also retrieves the corresponding data from the OS but applies the appropriate privacy filters to the data. The modified data are then provided to the application. Whereas if the request is rejected, the PMP returns an empty dataset to the application. This way, the application cannot tell whether the data access has been denied or not. Otherwise, the application could penalize the data subject (e.g., by means of a reduced functionality) to force him or her to disclose the data.

In order to ensure that an application cannot bypass this data request process, the PMP is also deeply integrated into the Android installation process. During installation, an application in Android is assigned all the permissions required to interact with the OS, e.g., access a data source. The PMP revokes all requested Android permissions from an application. As a result, even if an application tries to bypass the PMP, it cannot request data directly from the OS due to the lack of permissions [203].

In addition to applying privacy filters to data, the extensible adapter concept of the PMP also facilitates the integration of complementary external data sources. That is, adapters can be used to connect external IoT components to an edge device in a hub-and-spoke architecture. These external components include, among others, sensors or IoT end devices such as fitness wristbands, which themselves do not have a direct connection to the Internet. Instead, they transfer their data to a more powerful IoT device, e.g., a smartphone, via technologies such as Bluetooth. Our adapters ensure that such external data providers are fully integrated into the edge device on a logical level. This eliminates the need to distinguish between data sources that are embedded in the edge device and those that originate from external IoT components. In both cases, the edge device acts as a gateway for the data to the outside world. It also allows the privacy filters to operate on the data of the external IoT components directly. Otherwise, this would not be possible since such components neither have the necessary computing power nor offer the possibility of running third-party software, such as the PMP with its privacy filters [204].

While stream-based data processing is commonly used in the IoT as the lightweight IoT devices do not have sufficient memory to persistently store the captured data, in the context of a data delivery platform such as the REFINERY Platform, a different approach is required. To counteract the volatility of the IoT data, they are buffered on the edge device and made available via the PMP. To this end, we have developed a secure data store for IoT edge devices. This data store is secured against illegal access and manipulation from the outside by the fact that the data it contains are fully encrypted. Only the PMP has the key to decrypt the data. This ensures that data can only be accessed via the PMP, i.e., in a fully controlled and regulated manner. Internally, our data store manages the data in a relational database [205] or a NoSQL data store (e.g., a key-value store or a document store) [206], depending on the type of data. This way, it is able to handle both structured and unstructured data efficiently. However, the PMP completely abstracts from this internal data management, as the secure data store on a logical level is just another of its data sources. Furthermore, via the PMP, we enable a systematic synchronization mechanism for our data store. This allows data from multiple IoT edge devices to be synchronized with a mass data provider, i.e., as soon as an edge device has connectivity, it propagates all changes in the secure data store to its mass data provider [207].

For reliable information delivery, as we are targeting in the REFINERY Platform, two factors are crucial in data acquisition: On the one hand, the data quality must be transparent to data consumers. That is, if the quality has deteriorated, e.g., due to the application of privacy filters, this must be communicated in a transparent manner. On the other hand, confidentiality must also be maintained with respect to the data producers. In particular, this entails that data consumers must not know, e.g., what data have been withheld, as this could disclose what information has been concealed. This also means that data consumers do not gain complete insight into how the data have been tampered with. In addition, it must be ensured that the mass data provider meets its obligations with regard to privacy policies. To master this balancing act, the REFINERY Platform applies a two-stage attribute-based signature procedure.

To this effect, data are digitally signed on the edge devices. The full signature used here contains both the information on how the data were captured and which privacy filters were applied to them, as well as which privacy requirements still have to be addressed by the mass data provider. This means that the information about the data (e.g., data quality or accuracy) and the privacy policies that still need to be applied are inseparably linked to the payload data. The mass data provider checks and verifies the signature of the incoming data and uses the information contained in the signature for metadata management, as information on data quality is mandatory for subsequent processing and provision to data consumers. However, not all of this information is intended for the data consumer, as it can be used to draw conclusions about the data. If, e.g., it is evident that a privacy filter was applied to spoof certain location data, then assumptions can be made as to why these locations are considered to be compromising by the data subject. After the signature has been verified, all of this information is therefore removed from the signature. On a technical level, this is realized by means of a second key pair, the delegated keys. The resulting partial signature only contains the privacy policies that have to be considered during data preparation (see Section 4.2) and applied during data management (see Section 4.3). Yet, the remaining information is not lost as it is still available as part of the metadata in the REFINERY Platform. It was only detached from the payload data [208].

4.2. Data Preparation

Data preparation encompasses all activities required to cleanse raw data, transform them into a processable form, and finally turn them into a marketable data product. In the REFINERY Platform, we split data preparation into two separate steps. In the first step, we transform data into information. In contrast to raw data, information is organized and has a structure. General data impurities (e.g., missing values or outliers) are also addressed in this first step. In a second step, we transform information into knowledge. Unlike the rather generic information, knowledge is geared to a specific use case. This means that data products on the knowledge level can be applied by data consumers according to their purposes, e.g., as base data for their analyzes or as training data for their machine learning models. For both data preparation steps, we have developed techniques for the REFINERY Platform that are tailored to the specific characteristics of the commodity 'data'.

In order to bring raw data to the information level, the first step is to improve the data quality. For this purpose, missing data or attributes must be filled in, outliers or integrity violations must be identified as well as treated, and the data must be harmonized (e.g., by the unification of the value units). Although some of this can be done automatically (e.g., identification of null values), human participation is essential for successful data preparation, so that they can contribute their data knowledge to resolve the data impurities. For this purpose, human experts, often referred to as data stewards, need extensive insight into the base data in order to be able to identify the problems and apply appropriate countermeasures. However, it is necessary to comply with the privacy requirements of the data producers in the process.

Therefore, we have developed a sample-based approach for the REFINERY Platform, in which the data steward only has access to a representative sample of the data and works on this sample [209]. In synthesizing this sample, statistical disclosure techniques are applied. In a first step, a data sample is automatically generated from the stock of raw data. According to the privacy requirements that are still available in the partial signature of the data (i.e., the privacy requirements that have not yet been applied by the PMP in the user-controlled area), it is assessed whether the sample meets these requirements. Thresholds for various metrics (e.g., regarding the uncertainty, information gain, or accuracy of the sample) can be used to this end. Only if the sample meets these requirements, it is forwarded to the data steward. If it does not, a new sample must be selected. In addition to privacy metrics, fairness metrics can also be specified to ensure that the sample is not biased, i.e., that it is truly representative for the base data.

The data steward then cleanses the sample. Based on the applied techniques, data cleansing rules are derived, which are then applied to the complete base data. The data steward has access to metrics about the base data and can request and cleanse additional samples until s/he is fully satisfied with the overall data quality. User studies show that this dynamic sample-based approach even helps to fix more data issues compared to a data steward working on the entire base data. Since the volume of base data is large, the data steward cannot inspect all of the data. In contrast, the representative (and significantly smaller) sample can be analyzed in much greater detail, enabling the steward to identify and correct any errors it contains. That is, our approach is not only privacy-friendly but also leads to better overall data quality. With this approach, the data steward can also transform raw data into a uniform (and thus processable) structure.

In order to bring the data from the information level to the knowledge level, individual processing tailored to the respective intended use cases is required. This necessitates extensive domain knowledge. However, domain experts generally do not have the necessary IT knowledge to implement the processing steps themselves. For the REFINERY Platform, we have therefore developed an ontology approach that allows domain experts to specify in a simple way how the data needs to be processed for a given use case. In doing so, we use RDF/XML (see https://www.w3.org/TR/rdf-syntax-grammar/; accessed on 6 February 2023) for the internal representation of the ontology.

We have implemented a processing engine for our ontology that automatically applies the specified data processing rules to the data in question [210]. Figure 5 shows a simplified excerpt from such an ontology.

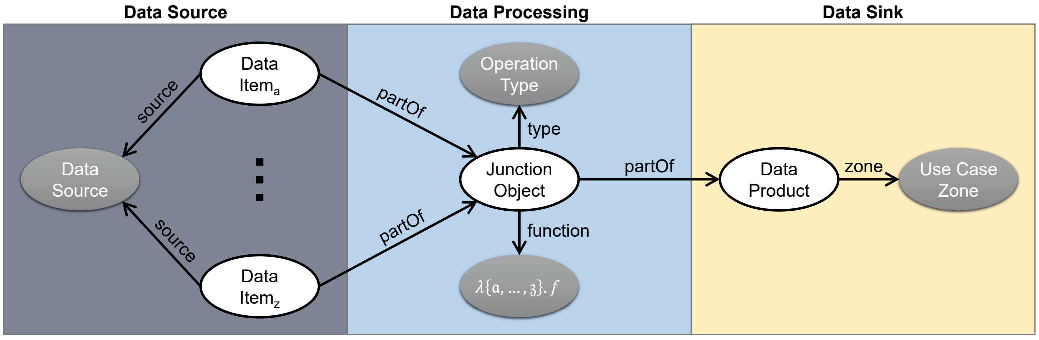


Figure 5. A Data Processing Rule Specified in the Ontology Provided by the REFINERY Platform.

A data processing rule of the ontology always consists of three segments. These three segments correspond to the three questions 'What?', 'How?', and 'To what end?'. First, data items must be selected to which the processing rules should be applied. Basically, all cleansed data on the information level can be selected. For instance, all temperature data should be converted into a data product. With this selection, however, the domain expert gains no insight into actual data items but only into which types of data are available and which attributes these items have. Therefore, this does not pose a threat to the privacy requirements of individual data producers.

After the source data have been selected, the core part of the data processing rule must describe which kind of processing has to be performed. To this end, we support the three key operators from functional programming, namely the map operator, the filter operator, and the reduce operator. These operators are each applied to a data sequence (e.g., to all data items from temperature sensors). In the following, D and E denote two arbitrary data types and $\mathbb{B}$ stands for Boolean data whereas a_i and b_i are instances of these data types.

$$\text{map operator} : (D \to E) \times (a_0, \ldots, a_n) \to (b_0, \ldots, b_n)$$

The map operator applies a unary function to all n elements of the sequence. This results in a new sequence consisting also of n elements—namely the n elements from the original sequence after they have been processed. The data type of the elements may change during processing. In our example, the domain expert could use a map operator to change the unit of temperature data from Celsius to Fahrenheit.

$$\text{filter operator} : (D \to \mathbb{B}) \times \underbrace{(a_0, \ldots, a_n)}_{n} \to \underbrace{(a_0, \ldots, a_n)}_{m} \mid m \leq n$$

The filter operator validates all n elements of a sequence using a unary predicate logical expression. The result is a sequence with the m elements from the original sequence for which the expression is evaluated to true ($0 \leq m \leq n$). However, the elements themselves are not changed in the process. In our example, the domain expert could use a filter operator to filter out the temperature data that are below a certain threshold.

$$\text{reduce operator}: (E \times D \to E) \times (a_0, \ldots, a_n) \times E \to E$$

The reduce operator aggregates the n elements of a sequence to a single result value. Unlike the map and filter, reduce applies a binary function. With this function, the reduce operator initially combines the first element of the sequence with an initial value. The result is then combined with the second element of the sequence using the same function. This is repeated until all elements of the sequence have been condensed to a single value. In our example, a reduce operator could calculate the average temperature of the n data items.

We apply the functional programming paradigm since functional programming is very well suited for composing such functions at runtime and applying them to arbitrary data streams. The actual program logic is defined via lambda expressions. As lambda expressions are essentially a concise mathematical description of functions and require a simplified syntax only, even non-IT experts can handle them.

In addition, our ontology approach also supports user-defined procedures, if the expression power of these three operators is not sufficient. To this end, arbitrary program code can be specified in the ontology, which is then applied to the data.

Finally, a data sink has to be selected, i.e., it has to be specified which data product was generated with this processing rule and for which use case it is intended. Intermediate data products can also be specified in our ontology, i.e., a data sink can be the data source for another data processing rule.

These data products can be further refined, e.g., they can be used to train machine learning and AI models. Such models also represent data products in their own right. In contrast to the data products addressed by the REFINERY Platform, however, such models have a more complex life cycle [211]. Maintaining these complex data products therefore requires additional measures, e.g., monitoring the validity of the models or providing the models in different data formats [212]. Yet, this is not in the scope of our work. Rather, the REFINERY Platform is the precursor to such a model management platform, providing it with the data needed to train these models. Furthermore, with our privacy filters, it is also possible to provide the data foundation for several variants of a model in which certain privacy-relevant aspects are not included [213]. For more information on such a model management platform, please refer to the work of Weber and Reimann [214].

However, even without considering such complex data products, a large number of (intermediate) data products need to be managed. Section 4.3 describes how this is solved in the REFINERY Platform. In addition, data consumers can retrieve available data products and also tailor them to their needs, which is outlined in Section 4.4.

4.3. Data Management

Data lake architectures are well suited for managing big data. They can hold not only any heterogeneous raw data but also processed variants of these data that have been refined for specific purposes. However, it is extremely important that the stored data are organized appropriately, because otherwise, the lake can easily degenerate into a data swamp, i.e., although the data are basically available, users are unable to retrieve them.

For the REFINERY Platform, we have therefore used the concepts of a data lake architecture and extended its basic zone concept. We apply pass-through zones, in which data are only temporarily buffered until they are processed, and persistent zones, in which data are permanently stored. Figure 6 illustrates the zone architecture we designed for the REFINERY Platform. Pass-through zones are shown in light blue and persistent zones are shown in white. In addition to these zones, there is a Data Security and Privacy Control Layer that manages the privacy requirements and access policy for the data lake as a whole. More details on the access control are given in Section 4.4.

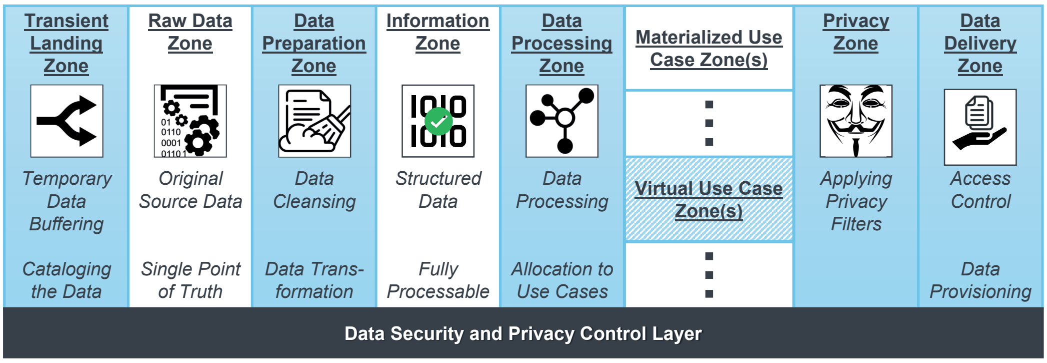


Figure 6. Data Lake Zone Architecture Applied in the REFINERY Platform.

The Transient Landing Zone represents the entry point of the data lake. Any incoming data is initially buffered here. This is also where the verification of the full signature takes place. If successful, the metadata contained in the signature is used to catalog the data so that they can be retrieved later. Then, the partial signature is added, to retain the privacy requirements to be observed in further processing. Via the event streaming platform Kafka (see https://kafka.apache.org/; accessed on 6 February 2023), the Transient Landing Zone then forwards the raw data to the Raw Data Zone for persistent storage.

In the first persistent zone, the incoming data are stored in an appropriate storage system. Since these raw data are heterogeneous and partly unstructured, a distributed file system like HDFS (see https://hadoop.apache.org/; accessed on 6 February 2023) is suitable for this purpose. To live up to the promise of being a reliable information retrieval and delivery platform, the original raw data must be protected as they represent the single point of truth. On the one hand, it must be ensured that they cannot be deleted, and on the other hand, it must be possible to prove to data consumers that the data products are based on authentic facts which have not been tampered with. To ensure both, the Raw Data Zone is made immutable and tamper-resistant by means of blockchain technologies. Whether the data are stored completely on-chain or only a digital fingerprint is stored in the blockchain depends on the respective data (e.g., their volume or their required level of protection) [215]. In case the data are stored on-chain, we have developed privacy-by-design concepts for blockchain systems [216], as well as concepts to improve the query capabilities of blockchain systems [217].

These base data are subsequently considered ground truth for the entire REFINERY Platform. The Raw Data Zone is also the basis for the Data Preparation Zone, in which our sample-based concepts are applied to cleanse and transform the data (see Section 4.2). The result of this preparation (i.e., the data at the information level) is then persisted in the Information Zone. As information is structured (i.e., a predefined data schema exists), relational databases such as PostgreSQL (see https://www.postgresql.org/; accessed on 6 February 2023) are suitable for storage. The schema provided by these databases facilitates the handling of the data in the following zones. Special protection measures such as blockchain technologies are not required here—the contents of this zone can be restored at any time based on the raw data and the defined cleansing and transformation rules.

In the Data Processing Zone, the structured data are converted into a data product using our ontology with the processing rules (see Section 4.2). Since some of the data products might be tailored to rather uncommon use cases, our concept allows them to be stored in Virtual Use Case Zones in addition to Materialized Use Case Zones. Materialized Use Case Zones store the data products in a fully persistent manner. The choice of storage technology depends on the data product, e.g., document stores such as MongoDB (see https://www.mongodb.com/; accessed on 6 February 2023). In contrast, data products in Virtual Use Case Zones are maintained only temporarily, e.g., in a pure in-memory database such as Redis (see https://redis.io/; accessed on 6 February 2023), or via Spark Structured Streaming (see https://spark.apache.org/streaming/; accessed on 6 February 2023) as a data stream that is generated live on demand. To enable such a mix of batch processing and stream processing, we have developed a hybrid processing model [218].

As described in Section 4.2, statistical disclosure techniques are applied during data preparation in accordance with the privacy requirements. However, due to data refinement activities (e.g., by combining data from various sources), data products can still violate privacy requirements. Therefore, there is a dedicated Privacy Zone in our REFINERY Platform before any data product is made available. In this Privacy Zone, privacy filters are applied to the data products if necessary. Since mass data providers have extensive processing capabilities, computationally intensive data obfuscation techniques can also be applied here to conceal highly specific information content [219]. This zone represents the final audit, to ensure that all privacy requirements of the data producers have been met. If this is the case, the partial signature is removed, and the data products are made available to data consumers via the Data Delivery Zone. We address data provisioning in Section 4.4.

4.4. Data Provisioning

There are two main tasks to be fulfilled in data provisioning: There are two main tasks to be fulfilled in data provision: On the one hand, the available data products must be retrievable for data consumers, and on the other hand, a control mechanism must ensure that only authorized entities (in accordance with the privacy requirements of the data producers) are granted access to the products. Both tasks are facilitated by our ontology of processing rules.

The ontology maps the complete lineage of a data product. This lineage includes where the raw data came from, what was done with them (i.e., what processing steps were taken during data preparation, but also what PET were applied to them), and where the data products are stored (i.e., in which Use Case Zone they can be found). With the help of this data lineage, data consumers can retrieve the data products they are looking for. As the ontology is machine-processable and can be analyzed, e.g., it can be used to feed a recommender system that can suggest related data products to data consumers that are also suitable for their purposes. In addition, data consumers can also design their own tailored data products. Since the ontology is extensible, data consumers can define their own data preparation processes if the existing data products do not meet their requirements. Either existing processing rules can be customized, or entirely new ones can be specified. Existing partial solutions can be used for this purpose, i.e., any subset of the processing rule in the ontology as well as any data product can be reused as a starting point for the new rules.

An access control system regulates which data products a data consumer has access to. We have developed a purposed-based access policy for the REFINERY Platform. The primary focus here is that it is comprehensible for data producers. That is, it must be transparent what knowledge a certain data product might expose about them. The model of our access control policy is shown in Figure 7 in a notation based on UML (see https://www.omg.org/spec/UML/; accessed on 6 February 2023).

From a technical perspective, access control is about defining who—in terms of which smart service—is allowed to access which data sources. Yet, humans cannot grasp what these data expose about them, as they are often abstract and only reveal meaningful knowledge after processing and analysis. That is why our access policy focuses on precisely that wisdom that can be disclosed due to the processing of data. To quantify this, we have studied a privacy risk elicitation procedure based on STPA, which allows privacy experts to systematically audit data processes. In this way, they can identify potential exposure threats in terms of disclosed wisdom [220]. The data processes are defined by our ontology as the sum of all processing rules. Therefore, the threats identified by the privacy experts can be mapped to one or more knowledge patterns. For this purpose, domain experts have to specify for which purposes the data products can be used, e.g., which analyses can be performed with them. These knowledge patterns are composed of the information prepared by the data stewards who process the raw data provided by the data producers. In the IoT, these data generally originate from sensors. Therefore, sensor experts are needed to describe from which sensors—or more generally, from which data sources—this kind of data can originate. This hierarchical top-down approach enables a comprehensible mapping of disclosed wisdom to data sources. That is, data subjects define their privacy requirements at a level of abstraction they can understand, whereas the rules are mapped to a technical level and applied to the respective components [221].

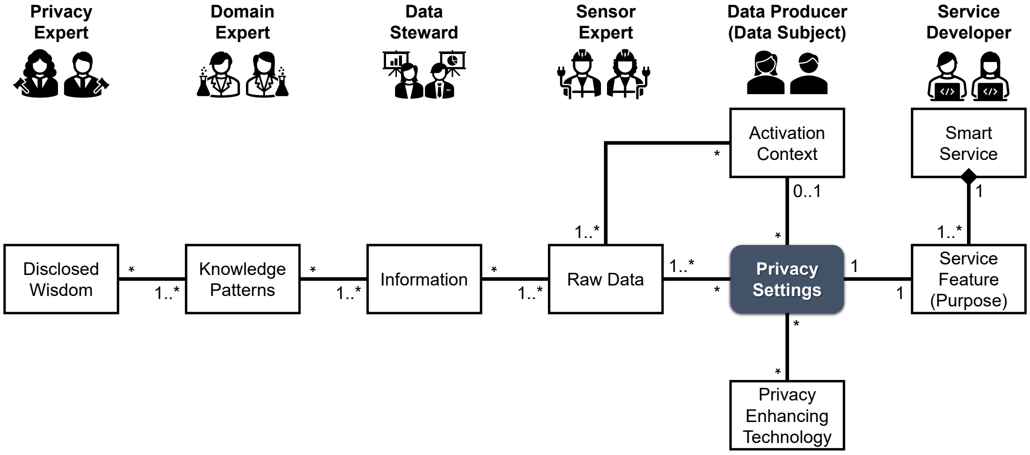


Figure 7. Model of the Purposed-Based Access Control Policy Designed for the REFINERY Platform.

On the opposite side, service developers in their role as data consumers must define which data their services need to access. Our policy allows to break down a smart service into individual service features. These service features each correspond to a purpose as the GDPR stipulates that data access is restricted to a specific purpose. With the help of our policy model, a data subject can thus clearly identify the purpose for which s/he exposes what kind of knowledge. Finally, PET can be attached to each policy rule. To this end, we evaluated a variety of privacy filters for specific types of data (e.g., location data, health data, or audio data) [222]. These reflect the privacy requirements expressed by data producers. The PET are applied in the Privacy Zone before a data product is made available to the smart service in question.

As static access control rules are too restrictive in a dynamic environment like the IoT, each rule in our model can be subject to an activation context. This could be, e.g., a temporal context (e.g., data collected during free time is subject to stricter access rules than data collected during working hours) or a spatial context (e.g., data collected at home is subject to stricter access rules than data collected at the workplace). Any type of data that is available and evaluable can be used to define this activation context [223]. This way, demand-driven data provisioning is made possible in the REFINERY Platform.

5. Assessment

After introducing the REFINERY Platform as our end-to-end approach toward reliable information retrieval and delivery and outlining how it carries out data administration tasks while complying with data security and data privacy, we now critically review our work. For this purpose, we will first perform a security and privacy assessment for the individual stages of the REFINERY Platform in Section 5.1. Subsequently, we provide a feature discussion in Section 5.2, in which we assess whether the REFINERY Platform has the required functionality to address the special data characteristic, i.e., whether it is able to effectively handle the commodity 'data'. Then, in Section 5.3, we discuss the practicality of the two key components of the REFINERY Platform with which a user primarily interacts, namely its privacy control measures and the selection and specification of data products, based on two case studies. Since the best functionality is of little use if it cannot be provided efficiently, we also perform a performance evaluation for the REFINERY Platform in Section 5.4. These four assessments reflect whether the REFINERY Platform is able to manage data in a secure, effective, practical, and efficient manner, i.e., whether it is a useful tool for modern data administration.

5.1. Security and Privacy Assessment

As described in Section 4, the REFINERY Platform fulfills all functional data administration tasks that arise in the context of data refinement, namely data acquisition, data preparation, data management, and data provisioning. In addition, data security and data privacy precautions are intrinsically integrated into each of these process steps, as required to ensure a reliable handling of the valuable commodity 'data'. Since these precautions are coordinated and seamlessly intertwine due to the holistic approach, end-to-end data protection is guaranteed. Since the application of these precautions is coordinated and all individual measures are closely intertwined, end-to-end data protection is achieved. In the following, we assess whether the protection goals (see Section 3.2) and privacy requirements (see Section 3.3) in this context are addressed in the REFINERY Platform.

The confidentiality of the data is ensured on the side of the data producer since they are stored solely in encrypted form in our secure data store on the edge devices. That is, they are protected against illegal access right after they have been captured. The data producer defines via the privacy requirements which of these data are forwarded by the PMP to the REFINERY Platform. This further promotes confidentiality, since only a portion of the data leaves the data producer's sphere of influence and can thus be leaked in the first place. For transmission and during data refinement, the data are signed. In general, a digital signature does not protect against unwanted access, since the data are encrypted with the private key of the data owner and thus anyone who has the public key can verify the signature, i.e., decrypt the data. However, in our case, only the REFINERY Platform has the necessary public keys. That is, the digital signature automatically guarantees confidentiality during transmission and storage in the REFINERY Platform as well. Since for the data preparation only samples that meet the privacy requirements can be accessed by the data stewards, confidentiality is also maintained in this process step. Even the data processing rules are specified without granting third parties deep insights into the data. External third parties (e.g., data consumers) can only access the data products via restricted interfaces, namely the Data Delivery Zone. Here, an access policy set by the data producers regulates who may access which data products and for what purpose.

Two main techniques are used to verify the integrity of the data: On the one hand, the REFINERY Platform can use the digital attribute-based signature to check how the data were captured, as it contains, e.g., information about the sensor used for this purpose and its accuracy. Furthermore, the privacy requirements are included in the signature as well. That is, the REFINERY Platform knows which distortions have been made to the data on the part of the data producer and which constraints have to be respected during data refinement. This transparency makes it possible to determine whether the quality of the raw data meets the quality requirements of a data consumer. Furthermore, manipulations by any third parties (e.g., during transmission) can be detected due to the signatures. While this does not prevent such manipulations, it can be ensured that no corrupted data can make its way into the data store of the REFINERY Platform. On the other hand, the integrity of the raw data after they have been transmitted is ensured by blockchain technologies. By means of the information stored in the blockchain (either the data items themselves or their digital fingerprints), it is possible to verify that the data in the Raw Data Zone have not been tampered with. Since the ontology with the processing rules provides a complete lineage of each data product, this also ensures the integrity of these products. If there is any doubt about the integrity of a data product, the raw data can first be verified and then the data product can be reproduced using the ontology.

No custom-made approaches for availability and authenticity are introduced in the REFINERY Platform. Rather, established techniques are also used for this purpose. For instance, the raw data are managed using HDFS. This distributed file system ensures high availability, as the data are redundantly distributed on different nodes. The relational databases in the Information Zone also have a high fault tolerance and allow recovering the stored data in the unlikely event of a failure. Thus, this also applies to the data products in the Use Case Zones, as they can be rebuilt based on the data from the Raw Data Zone and the Information Zone via the ontology. For IoT devices (i.e., the data sources), permanent availability cannot be achieved. However, our synchronization mechanism for the secured data store ensures that in case of connection failure, all new data are transmitted to the REFINERY Platform as soon as the connection is re-established. The digital signatures ensure the authenticity of the data as they certify the origin of the data. We have not addressed issues related to the authenticity of data consumers. For us, it is only important that the REFINERY platform is able to identify them. There are numerous mature approaches such as attribute-based credentials that can be used for this purpose. Authentication of data consumers is not within the scope of our work, as it has no implications for the REFINERY Platform.

Our access policy approach describes the insights that can be gained from certain data by mapping the rather abstract raw data to human-comprehensible knowledge patterns. This illustrates to data producers what can be extracted from their data. Furthermore, the ontology reflects the full data processing workflow. In this way, we enable data subjects to obtain detailed information about the processing of their data in a way that even non-IT experts can understand. Thus, our REFINERY Platform inherently implements the right to be informed. In addition, due to the PET which are geared to the privacy requirements, e.g., privacy filters or statistical disclosure control, we also offer technical support to implement the right to restriction of processing. This is particularly effective in our approach for two reasons: On the one hand, we use a variety of dedicated privacy filters and techniques. Thus, certain information contents in the data can be specifically concealed without degrading the overall data quality. On the other hand, we deploy the privacy filters in both the user-controlled area (i.e., on the edge devices in the form of the PMP) and in the REFINERY Platform mass data storage (i.e., in the Data Preparation Zone and in the Privacy Zone). This distribution allows us to apply the privacy filter in a much more targeted manner. It also leverages the strengths of both approaches. As a result, the usability of the data can be maintained without having to make any sacrifices in terms of privacy. Our access policy model also contributes to demand-driven data provisioning, as data accesses are mapped to actual purposes instead of a smart service as a whole. The introduction of activation contexts enables an even more fine-grained permission management. This reduces the privacy paradox since users are no longer faced with a more or less binary choice between privacy and service quality. We have not taken any explicit measures to enforce the right to be forgotten. Our concept is based on the assumption that the REFINERY Platform is operated by a trusted party. If a data subject therefore makes use of the right to be forgotten, we provide technical support for this, e.g., our ontology to easily identify and subsequently delete all data products related to the raw data in question. Explicit means for data subjects to verify that the REFINERY Platform actually deletes all data in question are not necessary given our basic assumption.

Table 1 summarizes the key data security and data privacy concepts integrated into our REFINERY Platform for the four data administration tasks.

5.2. Feature Discussion

As this initial investigation demonstrated that the REFINERY Platform provides comprehensive solutions for all protection goals regarding data security and privacy, we now assess to what extent the ten data characteristics, which we have identified and discussed in Section 2, are addressed in our approach.

The fact that data are not consumed during processing, i.e., the data volume grows continuously (Characteristic I), is addressed by the deployment of a distributed file system, namely HDFS, for the management of raw data. Since HDFS distributes the data across a multitude of nodes, it is suitable for the efficient handling of big data. In addition to the raw data, the number of data products to be managed in the REFINERY Platform is also growing steadily. With the introduction of Virtual Use Case Zones, i.e., a way to store data products only temporarily in volatile storages, data products that are tailored to rather uncommon use cases and therefore rarely in demand are automatically removed after usage. However, the knowledge gained during production is not lost, since the processing steps required to manufacture these data products are still available in the ontology. That is, if required, the data products can be restored from the raw data at any time.

Table 1. Summary of the Key Contributions of the REFINERY Platform with Regard to Data Security and Data Privacy.

Data Administration Task	Key Contributions to Data Security	Key Contributions to Data Privacy
Data Acquisition	On edge devices, data are fully encrypted. The data producers stipulate which data are transmitted. Digital signatures secure the transmission.	Data subjects specify privacy requirements that are applied by means of PET (e.g., privacy filters) on edge devices and in the REFINERY Platform.
Data Preparation	The applied sample-based data preparation ensures that no unauthorized insights into the data can be gained.	Data subjects specify a privacy threshold that is respected when selecting the data samples for the data stewards.
Data Management	Blockchain technologies ensure the integrity of the transmitted data via digital fingerprints.	Additional PET can be applied to a data product before it is distributed.
Data Provisioning	Our access control model enables to define who is allowed to access which data. Data access may be subject to additional privacy restrictions.	Since our model maps retrieved data to revealed knowledge patterns, this approach allows for truly informed consent.

Since data can be losslessly duplicated, which inevitably leads to a loss of value (Characteristic II), data consumers are not given direct access to the raw data or the processable information retrieved from them. Only the data products can be accessed by data consumers via restricted interfaces. This preserves the value of the raw data. The REFINERY Platform does not provide any special protection against the duplication of data products after they have been delivered to data consumers. However, this is not necessary from our point of view, as it is in the interest of data consumers that the value of the data products acquired by them is not diminished, e.g., by means of unregulated reproduction. The assets of the REFINERY Platform (i.e., the raw data, the information, and the production rules) remain unaffected by such a reproduction in any case.

In the REFINERY Platform, we first address the fact that data are generated at high velocity and are partially volatile (Characteristics III & IV) by developing a secure data store for edge devices that serves as a buffer until the data can be transmitted to the mass data provider. Our synchronization mechanism for this data store ensures that changes are always transmitted in a timely manner, as soon as connectivity is available. The Virtual Use Case Zones take into account the fact that raw data—and thus the data products derived from them—are volatile. The lifespan of these zones is limited by design, which acts as a kind of garbage collection in our data storage.

The fact that data are heterogeneous (Characteristic V) is addressed in the REFINERY Platform on the one hand by using HDFS in the Raw Data Zone. That is, even unstructured data can be managed without having to transform them first. In addition, metadata about the acquired data are provided by the digital signatures, which facilitate further processing of the data. For the data preparation, we rely on a human-in-the-loop approach. A data steward defines the necessary preparation steps. In contrast to a fully automated approach, our approach therefore does not require a predefined data schema.

Our approach to data preparation also addresses the fact that data refinement has to be in accordance with the data source and intended use (Characteristic VI). While the data preparation by the data steward is rather generic (whereby s/he can also take special characteristics of data sources into account), the subsequent data processing is fully geared to the intended use. The processing rules in our ontology describe how the processable information is turned into a data product. Data consumers can add further processing rules or adapt existing ones if no available data product meets their needs. Our PET include specialized privacy filters that are tailored to certain types of data. These filters can be used to conceal specific information contents without rendering the data unusable for an intended purpose.

HDFS is used for the implementation of the Raw Data Zone. With this file system, it is possible to add further nodes at any time in order to increase capacity. Due to decreasing prices for hard disk space, it is therefore possible to store all available raw data, even if their economic value is initially uncertain (Characteristic VII).

We address the fact that data can be manipulated indiscriminately (Characteristic VIII) by means of our data integrity measures. Blockchain technologies are used to verify that the stored raw data, which represent the ground truth for all data products, have not been manipulated. Digital signatures ensure that they have not been secretly falsified by third parties during transmission. These signatures also describe which PET have already been applied to the data in the user-controlled area. Therefore, it is not only possible to prevent all illegitimate data manipulations but also to communicate these justified distortions by the data producers (in accordance with their privacy requirements) to the data consumers in a transparent manner. In addition, the selection of PET is matched to the intended use case, i.e., their impact on the data attributes relevant for the data consumer is as low as possible.

In addition to the PET, our purposed-based access policy model ensures that special restrictions regarding the handling of certain data can be complied with (Characteristic IX). This access policy is used in the Data Delivery Zone and maps for which purpose which data consumer has access to which data products. Access can be further restricted by means of activation contexts that can be attached to each access policy rule.

All data products are described by means of our machine-processable ontology. Recommender systems can therefore use the ontology to point data consumers to similar data products that might also be relevant to them. Data consumers can also extend the ontology to design customized data products for their particular use cases. This forms the foundation for a data marketplace. That is, our REFINERY Platform provides the required concepts and infrastructures to trade the commodity 'data' (Characteristic X). To this end, only an implementation of an electronic storefront as an interface to the data marketplace is missing and has to be addressed in future work.

This feature discussion demonstrates that our REFINERY Platform is effective in handling the commodity 'data'. The concepts responsible for this are recapped in Table 2.

5.3. Case Study

After this feature discussion, showing that our REFINERY Platform provides all required functionalities to enable an appropriate modern data administration, we now focus on the practicality of our solution by means of two case studies.

This study is divided into two parts, as we aim to evaluate the two user interfaces. On the one hand, this involves the privacy control capabilities enabled by the PMP, which provide the foundation for demand-driven data acquisition (see Section 4.1). On the other hand, this also concerns the specification of tailored data products in the course of data preparation (see Section 4.2). These two case studies also reflect the two key user groups of the REFINERY Platform, namely data producers and data consumers.

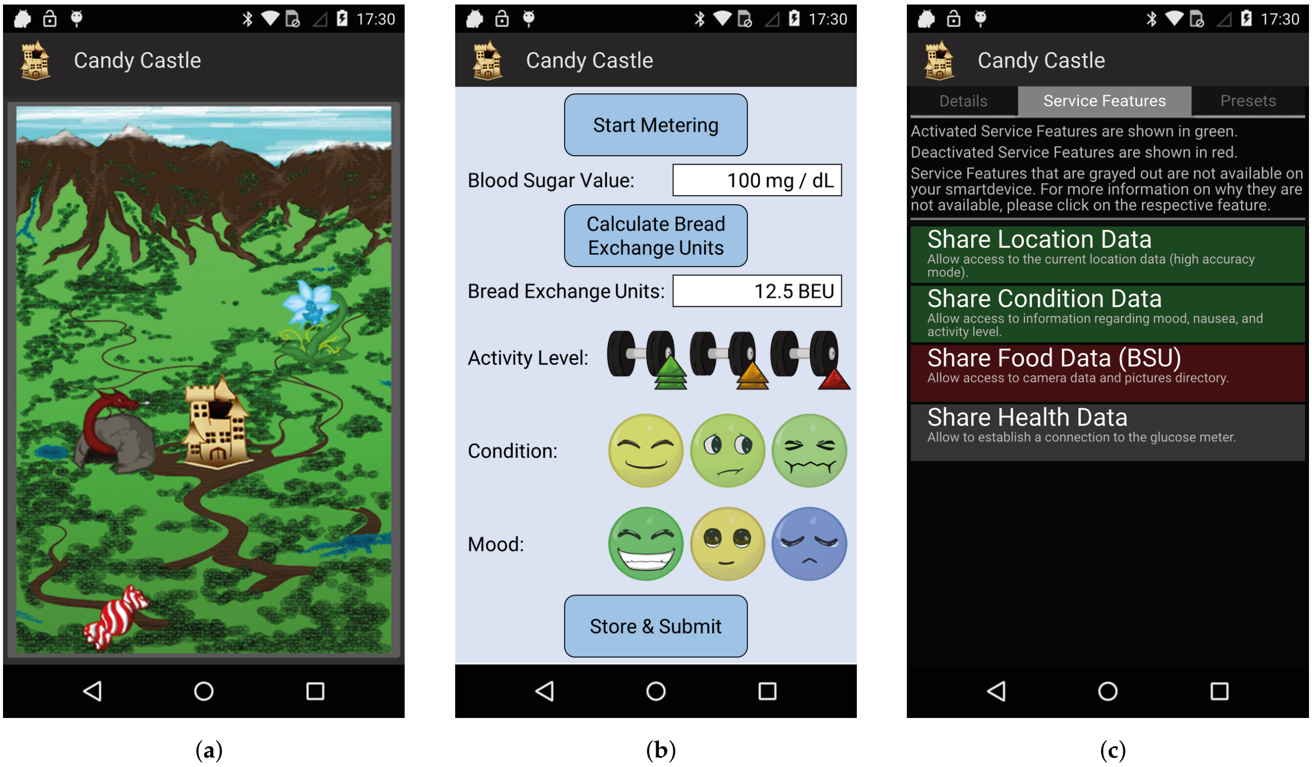
Privacy Control Capabilities. In our first case study, we focus on data acquisition of health applications for smartphones. Since it is evident that such applications are particularly beneficial, there is a mobile health application for literally all aspects of life [224]. One of the main drivers for these applications is the fact that the built-in sensors in standard smartphones can capture many health-related factors without a great deal of user input. For instance, the stress level [225] or the mood [226] can be determined passively (i.e., without explicit user interaction) by recording and analyzing the user's voice via the microphone. Or image analysis techniques can be used to analyze pictures of a food product on a smartphone in order to determine its ingredients, such as bread units [227]. Furthermore, there is a large variety of IoT-enabled medical metering devices that can be connected to a smartphone and thus provide these applications with more specific health data [228].

Table 2. Summary of the Key Concepts Applied in the REFINERY Platform to Address the Special Characteristics of Data.

Data Characteristic	**Concept in the REFINERY Platform to Handle this Characteristic**
I. Data are nonconsumable.	All incoming data are stored in the Raw Data Zone in an expandable big data storage. Data products that are not permanently demanded can be made available temporarily via a Virtual Use Case Zone and reproduced when needed.
II. Data can be duplicated losslessly.	Third parties such as data consumers only have access to data products, not the underlying raw data or the refined information.
III. Data are generated at high velocity.	Data are buffered on edge devices in our secure data store and updates are synchronized with the mass data storage automatically.
IV. Data are volatile.	Data products that are based on volatile data can be stored in Virtual Use Case Zones to ensure automatic sanitization of the provided products.
V. Data are heterogeneous.	The data storage in the Raw Data Zone is schemaless. Data preparation is based on a human-in-the-loop approach, i.e., no strict data schema is required here either.
VI. Data refinement has to be in accordance with the data source and intended use.	Data products are generated by means of an ontology, which can be extended if needed. Besides generic privacy filters, specialized filters tailored to specific types of data can be applied in order to preserve the data quality for the intended usage.
VII. The economic value of data is uncertain.	Virtually unlimited amounts of data can be stored in the Raw Data Zone at almost no cost. Therefore, their potential value does not need to be known in advance.
VIII. Data can be manipulated indiscernibly.	Digital signatures assure the integrity of the data in transit and the data in use while blockchain technologies enable to verify the integrity of the data at rest.
IX. Data may be subject to special restrictions.	The access policy model enables data subjects to define who can access which data and for what purpose. Access can be further constrained, e.g., via privacy filters.
X. Data require new trading concepts and infrastructures.	The information available in the Delivery Zone (e.g., metadata on raw data, specification of data products, and access policy) provide the foundation for a data marketplace.

While this kind of data is an obvious choice in the context of a health application, smartphones provide another very relevant piece of information that is often overlooked. It can be observed that the location of a user is also relevant to health applications. This spatiotemporal aspect is important for the interpretation of health data, because, e.g., a higher stress level in a noisy and hectic environment has to be evaluated differently than if it occurs in a peaceful place [229]. A smartphone can usually determine the current location accurately and therefore put every medical reading into its context.

By combining this data collection with gamification aspects, it is possible to address children in particular and encourage them to regularly capture and document their health values, which are otherwise often perceived as a chore [230]. With Candy Castle [231], we have therefore developed such a game aimed at young children suffering from diabetes. The main functionality of this application is shown in Figure 8.



Start Metering

Blood Sugar Value: 100 mg / dL

Calculate Bread Exchange Units

Bread Exchange Units: 12.5 BEU

Activity Level:

Condition:

Mood:

Store & Submit

Details | Service Features | Presets

Activated Service Features are shown in green.
Deactivated Service Features are shown in red.
Service Features that are grayed out are not available on your smartdevice. For more information on why they are not available, please click on the respective feature.

Share Location Data
Allow access to the current location data (high accuracy mode).

Share Condition Data
Allow access to information regarding mood, nausea, and activity level.

Share Food Data (BSU)
Allow access to camera data and pictures directory.

Share Health Data
Allow to establish a connection to the glucose meter.

Figure 8. Screenshots of the Health Game 'Candy Castle' Enhanced with PMP Features for Privacy Control. (**a**) Main Map of the Game. (**b**) Data Collection. (**c**) Privacy Control.

The main playing field shows the Candy Castle (see Figure 8a), which represents the condition of the child. Periodically, the castle is attacked by 'dark forces'. To ward off these forces, the child must take a blood glucose reading. Other symbols, such as a dragon or a flower, indicate particularly harmful or healthy locations in the surroundings (based on previous health readings in these areas). With each blood glucose measurement, additional factors relevant to people with diabetes are collected via the smartphone, such as the current activity level, condition, and mood (see Figure 8b). This way, a comprehensive and complete digital diabetes diary is kept. This is particularly useful for physicians, as such an electronic record is not only less prone to errors but also much easier to read than a manually kept diabetes diary—i.e., physicians can rely on the accuracy of the data [232].

However, a lot of data about the child are collected in the process, which represents a considerable invasion of privacy. Keep in mind that such mobile applications are usually developed and provided not by physicians or authorities but by unknown third parties. Therefore, in Candy Castle, we provide support for an integration into the PMP (see Figure 8c). The children (respectively their parents) can decide which data are collected by the application and for which purpose these data can be used. Since our adapter concept enables us to provide privacy filters tailored to each type of data source, any data restriction is done in an appropriate manner. This ensures that certain types of data are only shared correctly or not at all (e.g., blood glucose data), while for others the accuracy can be reduced (e.g., location data). This way, the data meet the privacy requirements of the data subjects as well as the quality requirements of the physicians. In a discussion with parents and physicians at a diabetes workshop, both sides were generally satisfied with this, as the PMP transparently communicates the data usage of an application, while its privacy control does not render the application unusable.

Specification of Tailored Data Products. Our second case study focuses on how effectively domain experts can specify tailored data products with our approach. For this purpose, we collaborated with food chemists. The number of people suffering from food allergies is constantly increasing. These allergies are in some cases life-threatening, which is why it is crucial that all ingredients are correctly indicated on a food product. Furthermore, control authorities are required to check the food products on a regular basis. Even the smallest particles of an allergen must be reliably detected even at the molecular level [233]. In our specific use case, we want to identify nut seeds in chocolate samples, since nut seeds are among the most prevalent food allergens which can trigger severe allergic shocks [234].

As part of the data preparation, the food chemists first determine the mass-to-charge ratio of the food samples with a mass spectrometer and use a chromatographic system to separate, identify, and quantify each component of the sample. The analysis data generated in this way are then checked against a protein sequence database to identify hazelnut or walnut peptides in the sample. Peptides are fragments of proteins, i.e., even the smallest traces of nut seeds can be detected this way. The samples for which the database indicates a match are marked. Only the marked samples need to be further analyzed, e.g., using peptide analysis software for manual analysis by a food chemist [235].

The data pipeline, which therefore has to be specified, has to identify and isolate all samples with marker peptides from the bulk of all samples and forward them to the peptide analysis software for in-depth analysis. Listing 1 shows the corresponding part of the ontology that is used to configure the data preparation in the REFINERY Platform as RDF/XML code.

First of all, the partition of the Raw Data Zone has to be selected in which the captured peptide data on the chocolate samples are available, namely the partition with the label 'chocolate' (line 4 resp. line 9). For the data contained in this partition, the attribute 'walnut' (line 2) and the attribute 'hazelnut' (line 7) have to be processed. The processing logic to obtain high-level information is defined in line 14 as a simple lambda expression. The expression evaluates to true if and only if a data object has one of the two markers 'walnut' or 'hazelnut' or both of them. This expression is applied as a filter (line 15). Therefore, only the samples that have at least one of the two markers are stored in the Use Case Zone 'allergens' (line 20). This zone then serves as input data for the peptide analysis software, which can access it via the Data Delivery Zone.

The feedback from the food chemists was positive, as this model-based description of complex data pipelines is feasible even without extensive programming knowledge. A significant advantage is that the domain experts do not have to deal with the different programming interfaces of the different data systems. Instead, they can describe the data preparation process at a higher level that abstracts from such programming details. Moreover, for the creation of RDF/XML ontologies, there are graphical editors such as VizBrick [236]. By means of such an editor, the specification of an ontology for the REFINERY Platform could be made even more user-friendly. More details on this matter can be found in Stach et al. [210].

5.4. Performance Evaluation

Since these two case studies indicate the practicality of the two main components with which users interact with the REFINERY Platform, we now evaluate whether its performance is also reasonable. To this end, we focus on the processing engine that applies the rules defined in the ontology to the data and creates the data products. All core functionalities of the REFINERY Platform depend on this processing engine, which is why the feasibility of the entire REFINERY Platform depends significantly on its performance in terms of data throughput.

For our performance evaluation, we therefore define different processing rules and apply them to artificially generated data. We focus on simple general-purpose data processing tasks, namely a selection task, a projection task, and an aggregation task. In the selection task, the data are processed by filtering out items based on their attributes (modeled as a filter operator). The projection task prepares the data by removing certain attributes (modeled as a map operator). Aggregation groups the data based on an attribute and condenses the data to the mean values of each group (modeled as a reduce operator).

Listing 1. Ontology Excerpt to Specify a Data Preparation Process in the Food Chemistry Domain.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 ↪  xmlns:dl="http://barents.dl/">
   <rdf:Description rdf:about="http://barents.dl/walnut">
      <dl:layer>Raw Data Zone</dl:layer>
      <dl:source>chocolate</dl:source>
      <dl:partOf rdf:resource="http://barents.dl/peptides"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://barents.dl/hazelnut">
      <dl:layer>Raw Data Zone</dl:layer>
      <dl:source>chocolate</dl:source>
      <dl:partOf rdf:resource="http://barents.dl/peptides"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://barents.dl/peptides">
      <dl:layer>Information Zone</dl:layer>
      <dl:function>lambda x : x.hazelnut or x.walnut</dl:function>
      <dl:type>filter</dl:type>
      <dl:partOf rdf:resource="http://barents.dl/results"/>
   </rdf:Description>
   <rdf:Description rdf:about="http://barents.dl/results">
      <dl:layer>Use Case Zone</dl:layer>
      <dl:zone>allergens</dl:zone>
   </rdf:Description>
</rdf:RDF>
```

Such tasks represent worst-case scenarios for our processing engine since the accesses to the Information Zone account for the majority of the processing costs compared to the actual data processing. In this process, the data must not only be read from the relational database used in the Information Zone but must also be converted into the data structure on which our processing engine operates, namely a data frame. As soon as the data are contained in this data structure, we can make use of indexes, which enables the execution environment to perform computations efficiently. This overhead caused by reading the data is inherent to any type of processing. Therefore, in general, it can be assumed that for more complex processing tasks, the overall overhead is lower, since in these cases, those access costs are negligible compared to the actual computation costs. The latter, however, also accrues without the use of the REFINERY Platform when manufacturing the respective data product.

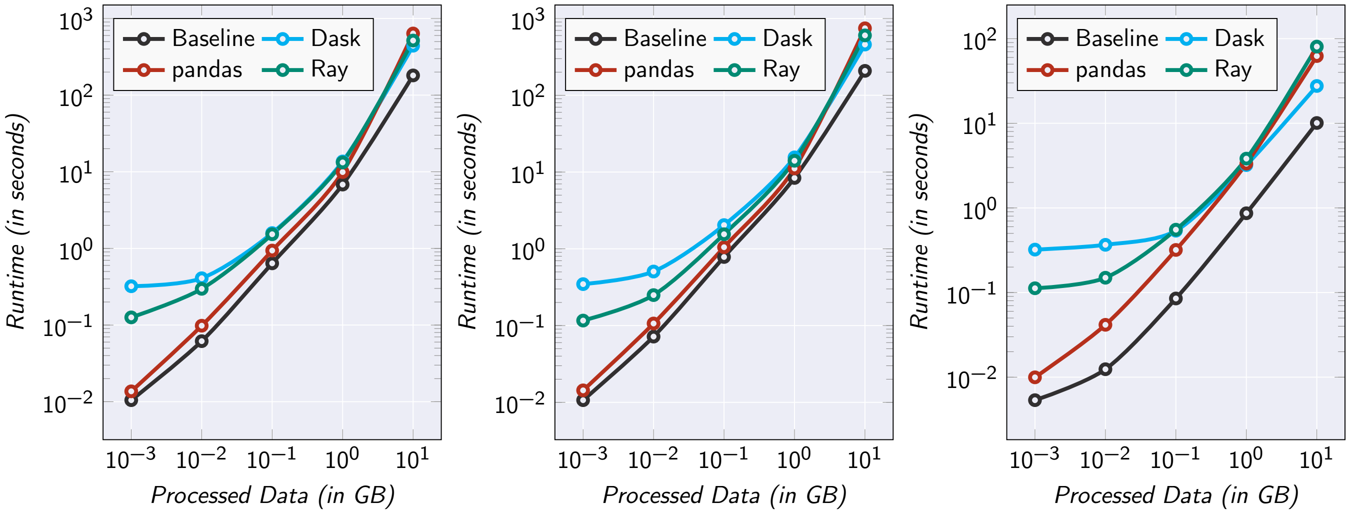
For our performance evaluation, the Information Zone, which contains the base data, is implemented using SQLite DB in version 3.39.4 (see https://www.sqlite.org/; accessed on 6 February 2023). The data products are stored in a Virtual Use Case Zone, which is implemented using TinyDB in version 4.7.0 (see https://tinydb.readthedocs.io/; accessed on 6 February 2023). However, the choice of these two databases does not affect the evaluation results. They can be replaced by any other relational database or NoSQL data store without loss of generality since the actual processing is fully decoupled from the input and output technologies. Our prototype of the processing engine is implemented in Python 3.10.9 (see https://www.python.org/; accessed on 6 February 2023) and uses pandas 1.5.2 (see https://pandas.pydata.org/; accessed on 6 February 2023) for data processing. To this end, the data from the Information Zone are initially loaded into a pandas DataFrame, a tablelike data structure. All operators specified in the ontology are then applied to this DataFrame and the result is forwarded to the Virtual Use Case Zone. Pandas is well suited for this purpose since in addition to import and export functions that support a variety of data sources and data sinks, it also provides index structures that allow efficient computations on a DataFrame. For this reason, pandas is a de facto standard in the field of data science for these kinds of tasks [237].

However, when it comes to processing large amounts of data, there are a few decisive limitations in pandas. On the one hand, pandas is by design only able to use a single threat on a single CPU for processing the data. Yet, especially with large amounts of data, significant performance improvements can be achieved by splitting the data into smaller chunks that are processed in parallel by multiple cores. On the other hand, pandas holds the entire DataFrame as well as all intermediate processing artifacts in main memory. Therefore, this represents another bottleneck in terms of an upper limit for the maximum amount of data that can be processed. Regardless of the available main memory, pandas is not designed to handle more than 100 GB of data [238]. Modin addresses these scalability issues by providing its own distributed DataFrame that can be processed by multiple cores. The Modin DataFrame is almost fully compatible with the pandas API, which makes it easy to parallelize pandas applications [239]. To this end, a Modin DataFrame can be partitioned either horizontally (i.e., by data item) or vertically (i.e., by attribute) [240].

Modin uses either Dask [241] or Ray [242] as its execution engine. Apart from some minor differences, the main distinctive feature is their respective scheduling strategies. While Dask uses a centralized scheduler that distributes the tasks to the workers and monitors the progress, Ray applies a distributed bottom-up scheduling strategy. Here, local schedulers distribute the tasks independently to the workers assigned to them. Workers can exchange data with each other via a shared-memory object store. Local schedulers can also forward tasks to a global scheduler, which can assign them to another local scheduler to achieve load balancing between the local schedulers. In effect, Dask is particularly suited for general-purpose data science tasks such as standard data analytics and data wrangling, while Ray shows its strengths in complex machine learning and AI-related tasks [243].

Therefore, we also implemented our prototype of the processing engine in a parallelized version using Modin 0.18.0 (see https://modin.readthedocs.io/; accessed on 6 February 2023) with Dask 2022.12.1 (see https://www.dask.org/; accessed on 6 February 2023) and Ray 2.2.0 (see https://www.ray.io/; accessed on 6 February 2023) as execution engines. As a baseline, we implemented the three data processing tasks as SQL commands that are executed directly by the SQLite DB, i.e., in a highly optimized manner that is beyond the implementation knowledge of the domain experts or the data consumers. This baseline represents the minimum processing cost. The more our processing machine approximates this baseline, the better its efficiency. Keep in mind that a certain overhead is inevitable. This is the price to pay for supporting tailorable processing rules and a wide range of data sources and sinks.

For the evaluation, we first adopted a deployment scenario in which a data producer runs an instance of the REFINERY Platform on his or her hardware for his or her own data only. That is, it is a limited amount of data and a significant limitation regarding computational power. To reflect this, we generated stepwise from 500 to 5000k data items, i.e., a data volume between 1 MB and 10 GB. In each step, we increased the amount of data by a factor of ten. We applied the three data processing tasks to these synthetic base data, using a desktop computer with an Intel Core i7-1165G7 with four cores and 16 GB DDR4-3200 of main memory. For each data volume and processing task, we measured the time it takes to process all data items. We carried out these measurements ten times each and after each run, we fully rolled back the SQLite DB and the TinyDB to exclude distortions due to warm caches. The medians of these runtime measurements are shown in Figure 9. Due to the use of medians, outliers (e.g., due to influences of concurrent background processes) do not skew the results.



(**a**) (**b**) (**c**)

Figure 9. Performance Evaluation Results Regarding the Runtime Overhead of the REFINERY Platform on a Desktop Computer. (**a**) Selection Task; (**b**) Projection Task; (**c**) Aggregation Task.

Two things can be observed across all three tasks: On the one hand, Modin causes a basic overhead with both execution engines due to the initial partitioning and distribution as well as the merging of the results. Especially for small data volumes, this overhead is excessive, since the data processing tasks are not time-consuming. On the other hand, the costs caused by pandas increase significantly for larger data volumes. Since the complete DataFrame must be kept permanently in main memory in this case, a lot of memory paging is required once the remaining free main memory runs out. Especially with the highly parallelizable aggregation task—here, the mean value for each group can be computed independently—one can see that for large data volumes, Modin has the advantage. Apparently, for such rather simple tasks, the centralized scheduling strategy of Dask is more advantageous. For more complex tasks, however, in particular in the area of machine learning, the Ray-based execution engine should clearly outperform the other implementation approaches. Apart from the small data volumes, both parallelized approaches are even for these simple processing tasks in $\mathcal{O}(\text{baseline})$, i.e., the asymptotic behavior of the runtime costs incurred by the REFINERY Platform are identical to the one of the highly optimized baseline.

In a second deployment scenario, we assumed a large mass data provider running the REFINERY Platform to refine data from many data producers on a high-performance server cluster. To this end, we incrementally generated from 1250 k to 20,000 k data items, i.e., a data volume between 5 GB and 80 GB. In each step, we increased the data volume by a factor of two. We applied the three data processing tasks to these synthetic base data, using a server cluster with 188 CPUs and 3 TB of main memory, organized as one master and ten worker nodes. Again, for each data volume and processing task, we measured the time it takes to process all data items. Figure 10 also presents the median of ten consecutive runs for each runtime measurement.

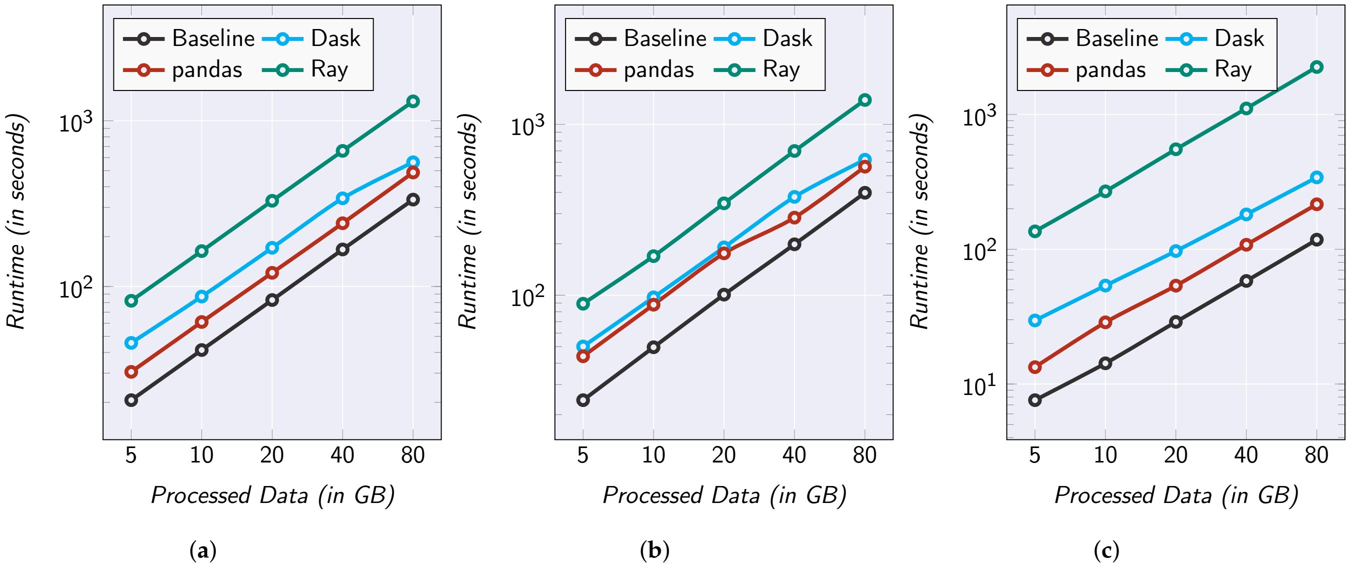


Figure 10. Performance Evaluation Results Regarding the Runtime Overhead of the REFINERY Platform on a Server Cluster. (**a**) Selection Task; (**b**) Projection Task; (**c**) Aggregation Task.

Here, the result is much more uniform for all three tasks. All three implementation variants as well as the baseline show a linear growth in processing costs across the board. Since all datasets easily fit into main memory, the runtime behavior of pandas does not deteriorate, even for the large data volumes. The processing tasks are so simple that the organizational overhead of splitting the dataset and distributing the chunks in parallel does not pay off. For more complex tasks, however, this would be the case. In addition, the pandas DataFrame already reaches nearly maximum capacity with the largest set of base data. Thus, a partitioning strategy is needed for larger volumes anyway. In any case, the runtime behavior of the REFINERY Platform is also in $\mathcal{O}(\text{baseline})$. As it therefore only causes a constant overhead in relation to the baseline in both deployment scenarios, this can be considered a great success. In return for this overhead, our approach offers the possibility to model data products in our ontology without requiring IT knowledge. Furthermore, our approach offers maximum flexibility in terms of the involved data sources and sinks.

If the latter is not required, the implementation of the Information Zone could be limited to relational databases. In this case, it is possible to use a tool such as Grizzly (see https://github.com/dbis-ilm/grizzly; accessed on 6 February 2023) for the implementation of our processing engine in order to further reduce the processing costs [244]. Grizzly also operates on pandas-like DataFrames. However, those DataFrames are not realized as actual data objects in memory that contain the data. Instead, all operations are translated into SQL commands that are executed directly by the data source. Lazy evaluation ensures that processing only occurs when a result has to be output, which can further reduce processing costs [245]. Since all operations used by our processing engine are fully supported by Grizzly, a migration to this execution engine is also possible if flexibility regarding the supported data sources is not required.

This performance evaluation demonstrates that our REFINERY Platform is efficient in handling the commodity 'data'. As it is therefore effective, practically applicable, and efficient in data administration and, due to its security and privacy features, respects both the interests of the data producers (i.e., protection of their sensitive data) and the interests of the data consumers (i.e., compliance with the promised data quality), it can be concluded that we have achieved our research goal to develop a reliable information retrieval and delivery platform.

6. Conclusions

Currently, data are boldly pithily referred to as the oil of the 21st century. On a metaphorical level, this statement is quite accurate, as the IoT and the resulting large-scale systematic collection of data not only enabled the fourth industrial revolution but also marked a major evolutionary step in the information age as it led to the digitization of society. Data-driven services significantly shape our everyday lives.

Even on a less figurative level, there are similarities between oil and data when it comes to their handling. Both commodities first have to be discovered and extracted, then refined, and finally delivered to consumers. A closer look, however, reveals inherent differences between the intangible commodity 'data' and a tangible commodity such as oil, which must be addressed when processing them.

Therefore, the goal of this work was to elaborate a modern data administration strategy that takes into account the strategic and economic importance of this special commodity. To this end, we made three contributions:

(a) Our investigation of the commodity 'data' revealed that ten unique characteristics have to be taken into account when handling this intangible resource. For instance, data are not only nonconsumable but can also be duplicated losslessly, which means that their volume is constantly growing. Furthermore, data accumulate at high velocity and have to be processed quickly, as they are partially volatile. Their heterogeneous nature and the need to apply individual refinement techniques to the data further complicate this endeavor. Since the economic value of data cannot be estimated in advance, and indiscernibly data manipulations can impair the quality of the data, it is essential to avoid unreasonably high handling costs. Finally, data products can be subject to special restrictions in terms of processing and provisioning. Therefore, there is a fundamental need for new trading concepts and infrastructures for the commodity 'data'.

(b) Based on this knowledge base, our review of state-of-the-art techniques related to data administration indicated that there are four aspects in particular where these characteristics need to be taken into account in order to enable effective and efficient data handling. First, data have to be acquired appropriately (in terms of, e.g., quality and quantity) from heterogeneous sources. These data must then be cleansed and made processable by means of data preparation and transformed into custom-made data products. The data products, together with all the high-volume data artifacts generated during manufacturing, must be managed and made retrievable. Only then can they be offered to data consumers in a digital storefront as part of data provisioning. In addition to these data administration tasks, security and privacy aspects also have to be taken into account in each of these work steps.

(c) Our review of related work revealed that there are many island solutions to individual aspects of these data administration problems. However, there is no holistic end-to-end solution addressing all data characteristics at once. This is necessary in order to achieve synergy effects and thus exploit the full potential of the commodity 'data'. To this end, we presented our own concept toward a reliable information retrieval and delivery platform called REFINERY Platform. Our REFINERY Platform not only addresses all the challenges we identified in the area of data administration but also provides both data producers and data consumers with assertions regarding data security and privacy on the one hand and data quality on the other hand. An in-depth assessment confirms that our approach is effective (in terms of provided functionality), practicable (in terms of operability), and efficient (in terms of data throughput) in this respect.

Despite its undeniable advantages in terms of modern data administration—namely its capability to deal with the challenges of big data while satisfying both the privacy requirements of data subjects and the data quality demands of data consumers—our presented REFINERY Platform also has some limitations. In this regard, it is important to keep in mind that we are presenting a concept. That is, although the various components of the REFINERY Platform have been implemented and their isolated application shows good results in terms of practicality and performance, it is an open task to implement and comprehensively evaluate a full-fledged prototype of the REFINERY Platform. However, since the more than promising results presented in this paper demonstrate the soundness of our approach and the effectiveness, practicality, and efficiency of its key components, we are confident about this future work.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this paper:

ADS	Authenticated Data Structures
AI	Artificial Intelligence
App	(Mobile) Application
BI	Business Intelligence
CPU	Central Processing Unit
DB	Database
DDoS	Distributed Denial of Service Attack
DDR	Double Data Rate
ETL	Extraction, Transformation, Loading
GB	Gigabyte
GDPR	General Data Protection Regulation
IoT	Internet of Things
IT	Information Technology
MB	Megabyte
NoSQL	Not only SQL
OAuth	Open Authorization
OS	Operating System
PET	Privacy-Enhancing Technologies
PIN	Personal Identification Number
PMP	Privacy Management Platform
PoOR	Proofs of Ownership and Retrievability
PoRR	Proofs of Retrievability and Reliability
PUF	Physical Unclonable Function
RDF	Resource Description Framework
REFINERY Platform	Reliable Information Retrieval and Delivery Platform
SQL	Structured Query Language
STAMP	System-Theoretic Accident Model and Processes
STPA	System-Theoretic Process Analysis

TB	Terabyte
VDF	Verifiable Delay Function
XML	Extensible Markup Language

References

1. Schwab, K.; Marcus, A.; Oyola, J.R.; Hoffman, W.; Luzi, M. Personal Data: The Emergence of a New Asset Class. An Initiative of the World Economic Forum. 2011. pp. 1–40. Available online: https://www.weforum.org/reports/personal-data-emergence-new-asset-class/ (accessed on 6 February 2023).
2. Javornik, M.; Nadoh, N.; Lange, D. Data Is the New Oil. In *Towards User-Centric Transport in Europe: Challenges, Solutions and Collaborations*; Müller, B., Meyer, G., Eds.; Springer: Cham, Switzerland, 2019; pp. 295–308.
3. Klingenberg, C.O.; Borges, M.A.V.; Antunes, J.A.V., Jr. Industry 4.0 as a data-driven paradigm: A systematic literature review on technologies. *J. Manuf. Technol. Manag.* **2021**, *32*, 570–592. [CrossRef]
4. Sisinni, E.; Saifullah, A.; Han, S.; Jennehag, U.; Gidlund, M. Industrial Internet of Things: Challenges, Opportunities, and Directions. *IEEE Trans. Ind. Inform.* **2018**, *14*, 4724–4734. [CrossRef]
5. Singh, M.; Fuenmayor, E.; Hinchy, E.P.; Qiao, Y.; Murray, N.; Devine, D. Digital Twin: Origin to Future. *Appl. Syst. Innov.* **2021**, *4*, 36. [CrossRef]
6. Philbeck, T.; Davis, N. The Fourth Industrial Revolution: Shaping a New Era. *J. Int. Aff.* **2018**, *72*, 17–22.
7. Schwab, K. (Ed.) *The Fourth Industrial Revolution*, illustrated ed.; Crown Business: New York, NY, USA, 2017.
8. Lasi, H.; Fettke, P.; Kemper, H.G.; Feld, T.; Hoffmann, M. Industry 4.0. *Bus. Inf. Syst. Eng.* **2014**, *6*, 239–242. [CrossRef]
9. Leelaarporn, P.; Wachiraphan, P.; Kaewlee, T.; Udsa, T.; Chaisaen, R.; Choksatchawathi, T.; Laosirirat, R.; Lakhan, P.; Natnithikarat, P.; Thanontip, K.; et al. Sensor-Driven Achieving of Smart Living: A Review. *IEEE Sens. J.* **2021**, *21*, 10369–10391. [CrossRef]
10. Paiva, S.; Ahad, M.A.; Tripathi, G.; Feroz, N.; Casalino, G. Enabling Technologies for Urban Smart Mobility: Recent Trends, Opportunities and Challenges. *Sensors* **2021**, *21*, 2143. [CrossRef]
11. Al-rawashdeh, M.; Keikhosrokiani, P.; Belaton, B.; Alawida, M.; Zwiri, A. IoT Adoption and Application for Smart Healthcare: A Systematic Review. *Sensors* **2022**, *22*, 5377. [CrossRef] [PubMed]
12. Yar, H.; Imran, A.S.; Khan, Z.A.; Sajjad, M.; Kastrati, Z. Towards Smart Home Automation Using IoT-Enabled Edge-Computing Paradigm. *Sensors* **2021**, *21*, 4932. [CrossRef]
13. Taffel, S. Data and oil: Metaphor, materiality and metabolic rifts. *New Media Soc.* **2021**. [CrossRef]
14. Urbach, N.; Ahlemann, F. *IT Management in the Digital Age: A Roadmap for the IT Department of the Future*; Springer: Cham, Switzerland, 2019.
15. Possler, D.; Bruns, S.; Niemann-Lenz, J. Data Is the New Oil–But How Do We Drill It? Pathways to Access and Acquire Large Data Sets in Communication Science. *Int. J. Commun.* **2019**, *13*, 3894–3911.
16. Liew, A. Understanding Data, Information, Knowledge And Their Inter-Relationships. *J. Knowl. Manag. Pract.* **2007**, *8*, 1–10.
17. Sarker, I.H. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. *SN Comput. Sci.* **2021**, *2*, 377. [CrossRef]
18. Arfat, Y.; Usman, S.; Mehmood, R.; Katib, I. Big Data Tools, Technologies, and Applications: A Survey. In *Smart Infrastructure and Applications: Foundations for Smarter Cities and Societies*; Mehmood, R., See, S., Katib, I., Chlamtac, I., Eds.; Springer: Cham, Switzerland, 2020; Chapter 19, pp. 453–490.
19. Rowley, J. The wisdom hierarchy: Representations of the DIKW hierarchy. *J. Inf. Sci.* **2007**, *33*, 163–180. [CrossRef]
20. Mandel, M. *The Economic Impact of Data: Why Data Is Not Like Oil*; ppi Radically Pragmatic: London, UK, 2017; pp. 1–20. Available online: https://www.progressivepolicy.org/publication/economic-impact-data-data-not-like-oil/ (accessed on 6 February 2023).
21. Nolin, J.M. Data as oil, infrastructure or asset? Three metaphors of data as economic value. *J. Inf. Commun. Ethics Soc.* **2020**, *18*, 28–43. [CrossRef]
22. Katal, A.; Wazid, M.; Goudar, R.H. Big data: Issues, challenges, tools and Good practices. In Proceedings of the 2013 Sixth International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 404–409.
23. Mladenović, M.N. Data is not the new oil, but could be water or sunlight? From ethical to moral pathways for urban data management. In Proceedings of the 17th International Conference on Computational Urban Planning and Urban Management (CUPUM), Espoo, Finland, 9–11 June 2021; pp. 9–11.
24. Hirsch, D.D. The Glass House Effect: Big Data, the New Oil, and the Power of Analogy. *Maine Law Rev.* **2014**, *66*, 373–395.
25. van der Aalst, W.M.P. Data Scientist: The Engineer of the Future. In Proceedings of the 7th International Conference on Interoperability for Enterprises Systems and Applications (I-ESA), Albi, France, 24–28 March 2014; Springer: Cham, Switzerland, 2014; pp. 13–26.
26. Siddiqa, A.; Hashem, I.A.T.; Yaqoob, I.; Marjani, M.; Shamshirband, S.; Gani, A.; Nasaruddin, F. A survey of big data management: Taxonomy and state-of-the-art. *J. Netw. Comput. Appl.* **2016**, *71*, 151–166. [CrossRef]
27. Moreno, J.; Serrano, M.A.; Fernández-Medina, E. Main Issues in Big Data Security. *Future Internet* **2016**, *8*, 44. [CrossRef]

28. Binjubeir, M.; Ahmed, A.A.; Ismail, M.A.B.; Sadiq, A.S.; Khurram Khan, M. Comprehensive Survey on Big Data Privacy Protection. *IEEE Access* **2020**, *8*, 20067–20079. [CrossRef]
29. Löcklin, A.; Vietz, H.; White, D.; Ruppert, T.; Jazdi, N.; Weyrich, M. Data administration shell for data-science-driven development. *Procedia CIRP* **2021**, *100*, 115–120. [CrossRef]
30. Jeyaprakash, T.; Padmaveni, K. Introduction to Data Science—An Overview. *Int. J. Sci. Manag. Stud.* **2021**, *4*, 407–410. [CrossRef]
31. Lyko, K.; Nitzschke, M.; Ngonga Ngomo, A.C. Big Data Acquisition. In *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*; Cavanillas, J.M., Curry, E., Wahlster, W., Eds.; Springer: Cham, Switzerland, 2016; Chapter 4, pp. 39–61.
32. Curry, E. The Big Data Value Chain: Definitions, Concepts, and Theoretical Approaches. In *New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe*; Cavanillas, J.M., Curry, E., Wahlster, W., Eds.; Springer: Cham, Switzerland, 2016; Chapter 3, pp. 29–37.
33. Vassiliadis, P.; Simitsis, A.; Skiadopoulos, S. Conceptual Modeling for ETL Processes. In Proceedings of the 5th ACM International Workshop on Data Warehousing and OLAP (DOLAP), McLean, VA, USA, 8 November 2002; ACM: New York, NY, USA, 2002; pp. 14–21.
34. Simitsis, A. Modeling and managing ETL processes. In Proceedings of the VLDB 2003 PhD Workshop co-located with the 29th International Conference on Very Large Databases (VLDB), Berlin, Germany, 9–12 September 2003; CEUR-WS.org: Aachen, Germany, 2003; pp. 1–5.
35. Lau, B.P.L.; Marakkalage, S.H.; Zhou, Y.; Hassan, N.U.; Yuen, C.; Zhang, M.; Tan, U.X. A survey of data fusion in smart city applications. *Inf. Fusion* **2019**, *52*, 357–374. [CrossRef]
36. Diouf, P.S.; Boly, A.; Ndiaye, S. Variety of data in the ETL processes in the cloud: State of the art. In Proceedings of the 2018 IEEE International Conference on Innovative Research and Development (ICIRD), Bangkok, Thailand, 11–12 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–5.
37. D'silva, G.M.; Khan, A.; Gaurav.; Bari, S. Real-time processing of IoT events with historic data using Apache Kafka and Apache Spark with dashing framework. In Proceedings of the 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), Bangalore, India, 19–20 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1804–1809.
38. Geng, D.; Zhang, C.; Xia, C.; Xia, X.; Liu, Q.; Fu, X. Big Data-Based Improved Data Acquisition and Storage System for Designing Industrial Data Platform. *IEEE Access* **2019**, *7*, 44574–44582. [CrossRef]
39. Huai, Y.; Chauhan, A.; Gates, A.; Hagleitner, G.; Hanson, E.N.; O'Malley, O.; Pandey, J.; Yuan, Y.; Lee, R.; Zhang, X. Major Technical Advancements in Apache Hive. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD), Snowbird, UT, USA, 22–27 June 2014; ACM: New York, NY, USA, 2014; pp. 1235–1246.
40. Lee, G.; Lin, J.; Liu, C.; Lorek, A.; Ryaboy, D. The Unified Logging Infrastructure for Data Analytics at Twitter. *Proc. VLDB Endow.* **2012**, *5*, 1771–1780. [CrossRef]
41. Marz, N. How to Beat the CAP Theorem. Thoughts from the Red Planet. 2011. Available online: http://nathanmarz.com/blog/how-to-beat-the-cap-theorem.html (accessed on 6 February 2023).
42. Kreps, J. Questioning the Lambda Architecture. *O'Reilly*, 2 July 2014. Available online: https://www.oreilly.com/radar/questioning-the-lambda-architecture/ (accessed on 6 February 2023).
43. Kraetz, D.; Morawski, M. Architecture Patterns—Batch and Real-Time Capabilities. In *The Digital Journey of Banking and Insurance, Volume III: Data Storage, Data Processing and Data Analysis*; Liermann, V., Stegmann, C., Eds.; Palgrave Macmillan: Cham, Switzerland, 2021; pp. 89–104.
44. Lin, J. The Lambda and the Kappa. *IEEE Internet Comput.* **2017**, *21*, 60–66. [CrossRef]
45. Terrizzano, I.; Schwarz, P.; Roth, M.; Colino, J.E. Data Wrangling: The Challenging Journey from the Wild to the Lake. In Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, USA, 4–7 January 2015; pp. 1–9.
46. Ding, X.; Wang, H.; Su, J.; Li, Z.; Li, J.; Gao, H. Cleanits: A Data Cleaning System for Industrial Time Series. *Proc. VLDB Endow.* **2019**, *12*, 1786–1789. [CrossRef]
47. Behringer, M.; Hirmer, P.; Mitschang, B. A Human-Centered Approach for Interactive Data Processing and Analytics. In Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS), Porto, Portugal, 26–29 April 2017; Springer: Cham, Switzerland, 2018; pp. 498–514.
48. Diamantini, C.; Lo Giudice, P.; Potena, D.; Storti, E.; Ursino, D. An Approach to Extracting Topic-guided Views from the Sources of a Data Lake. *Inf. Syst. Front.* **2021**, *23*, 243–262. [CrossRef]
49. Bogatu, A.; Fernandes, A.A.A.; Paton, N.W.; Konstantinou, N. Dataset Discovery in Data Lakes. In Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, 20–24 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 709–720.
50. Megdiche, I.; Ravat, F.; Zhao, Y. Metadata Management on Data Processing in Data Lakes. In Proceedings of the 47th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM), Bolzano-Bozen, Italy, 25–29 January 2021; Springer: Cham, Switzerland, 2021; pp. 553–562.

51. Castro Fernandez, R.; Abedjan, Z.; Koko, F.; Yuan, G.; Madden, S.; Stonebraker, M. Aurum: A Data Discovery System. In Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE), Paris, France, 16–19 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1001–1012.
52. Behringer, M.; Hirmer, P.; Fritz, M.; Mitschang, B. Empowering Domain Experts to Preprocess Massive Distributed Datasets. In Proceedings of the 23rd International Conference on Business Information Systems (BIS), Colorado Springs, CO, USA, 8–10 June 2020; Springer: Cham, Switzerland, 2020; pp. 61–75.
53. Behringer, M.; Fritz, M.; Schwarz, H.; Mitschang, B. DATA-IMP: An Interactive Approach to Specify Data Imputation Transformations on Large Datasets. In Proceedings of the 27th International Conference on Cooperative Information Systems (CoopIS), Bozen-Bolzano, Italy, 4–7 October 2022; Springer: Cham, Switzerland, 2022; pp. 55–74.
54. Mahdavi, M.; Abedjan, Z. Semi-Supervised Data Cleaning with Raha and Baran. In Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR), Chaminade, CA, USA, 11–15 January 2021; pp. 1–7.
55. Wulf, A.J.; Seizov, O. "Please understand we cannot provide further information": Evaluating content and transparency of GDPR-mandated AI disclosures. *AI & Soc.* **2022**, 1–22. [CrossRef]
56. Auge, T.; Heuer, A. ProSA—Using the CHASE for Provenance Management. In Proceedings of the 23rd European Conference on Advances in Databases and Information Systems (ADBIS), Bled, Slovenia, 8–11 September 2019; Springer: Cham, Switzerland, 2019; pp. 357–372.
57. Lam, H.T.; Buesser, B.; Min, H.; Minh, T.N.; Wistuba, M.; Khurana, U.; Bramble, G.; Salonidis, T.; Wang, D.; Samulowitz, H. Automated Data Science for Relational Data. In Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 19–22 April 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 2689–2692.
58. Ilyas, I.F.; Rekatsinas, T. Machine Learning and Data Cleaning: Which Serves the Other? *J. Data Inf. Qual.* **2022**, *14*, 1–11. [CrossRef]
59. Devlin, B.; Cote, L.D. *Data Warehouse: From Architecture to Implementation*; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1996.
60. Aftab, U.; Siddiqui, G.F. Big Data Augmentation with Data Warehouse: A Survey. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2785–2794.
61. Wongthongtham, P.; Abu-Salih, B. Ontology and trust based data warehouse in new generation of business intelligence: State-of-the-art, challenges, and opportunities. In Proceedings of the 2015 IEEE 13th International Conference on Industrial Informatics (INDIN), Cambridge, UK, 22–24 July 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 476–483.
62. Mathis, C. Data Lakes. *Datenbank-Spektrum* **2017**, *17*, 289–293. [CrossRef]
63. Taniar, D.; Rahayu, W. Data Lake Architecture. In Proceedings of the 9th International Conference on Emerging Internet, Data & Web Technologies (EIDWT), Chiang Mai, Thailand, 25–27 February 2021; Springer: Cham, Switzerland, 2021; pp. 344–357.
64. Ravat, F.; Zhao, Y. Data Lakes: Trends and Perspectives. In Proceedings of the 30th International Conference on Database and Expert Systems Applications (DEXA), Linz, Austria, 26–29 August 2019; Springer: Cham, Switzerland, 2019; pp. 304–313.
65. Giebler, C.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B. Leveraging the Data Lake: Current State and Challenges. In Proceedings of the 21st International Conference on Big Data Analytics and Knowledge Discovery (DaWaK), Linz, Austria, 26–29 August 2019; Springer: Cham, Switzerland, 2019; pp. 179–188.
66. Giebler, C.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B. A Zone Reference Model for Enterprise-Grade Data Lake Management. In Proceedings of the 2020 IEEE 24th International Enterprise Distributed Object Computing Conference (EDOC), Eindhoven, The Netherlands, 5–8 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 57–66.
67. Hai, R.; Geisler, S.; Quix, C. Constance: An Intelligent Data Lake System. In Proceedings of the 2016 International Conference on Management of Data (SIGMOD), San Francisco, CA, USA, 26 June–1 July 2016; ACM: New York, NY, USA, 2016; pp. 2097–2100.
68. Farid, M.; Roatis, A.; Ilyas, I.F.; Hoffmann, H.F.; Chu, X. CLAMS: Bringing Quality to Data Lakes. In Proceedings of the 2016 International Conference on Management of Data (SIGMOD), San Francisco, CA, USA, 26 June–1 July 2016; ACM: New York, NY, USA, 2016; pp. 2089–2092.
69. Machado, I.A.; Costa, C.; Santos, M.Y. Data Mesh: Concepts and Principles of a Paradigm Shift in Data Architectures. *Procedia Comput. Sci.* **2022**, *196*, 263–271. [CrossRef]
70. Oreščanin, D.; Hlupić, T. Data Lakehouse—A Novel Step in Analytics Architecture. In Proceedings of the 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 9–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1242–1246.
71. Armbrust, M.; Ghodsi1, A.; Xin, R.; Zaharia, M. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR), Chaminade, CA, USA, 11–15 January 2021; pp. 1–8.
72. Alpar, P.; Schulz, M. Self-Service Business Intelligence. *Bus. Inf. Syst. Eng.* **2016**, *58*, 151–155. [CrossRef]
73. Lennerholt, C.; van Laere, J. Data access and data quality challenges of self-service business intelligence. In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm and Uppsala, Sweden, 8–14 June 2019; AIS: Atlanta, GA, USA, 2019; pp. 1–13.
74. Huang, L.; Dou, Y.; Liu, Y.; Wang, J.; Chen, G.; Zhang, X.; Wang, R. Toward a research framework to conceptualize data as a factor of production: The data marketplace perspective. *Fundam. Res.* **2021**, *1*, 586–594. [CrossRef]
75. Gröger, C. There is No AI without Data. *Commun. ACM* **2021**, *64*, 98–108. [CrossRef]

76. Eichler, R.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B. From Data Asset to Data Product – The Role of the Data Provider in the Enterprise Data Marketplace. In Proceedings of the 17th Symposium and Summer School On Service-Oriented Computing (SummerSOC), Heraklion, Greece, 3–9 July 2022; Springer: Cham, Switzerland, 2022; pp. 119–138.
77. Eichler, R.; Giebler, C.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B. Enterprise-Wide Metadata Management: An Industry Case on the Current State and Challenges. In Proceedings of the 24th International Conference on Business Information Systems (BIS), Hannover, Germany, 14–17 June 2021; pp. 269–279.
78. Eichler, R.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B. Data Shopping—How an Enterprise Data Marketplace Supports Data Democratization in Companies. In Proceedings of the 34th International Conference on Advanced Information Systems Engineering (CAiSE), Leuven, Belgium, 6–10 June 2022; Springer: Cham, Switzerland, 2022; pp. 19–26.
79. Eichler, R.; Giebler, C.; Gröger, C.; Schwarz, H.; Mitschang, B. Modeling metadata in data lakes—A generic model. *Data Knowl. Eng.* **2021**, *136*, 101931. [CrossRef]
80. Driessen, S.W.; Monsieur, G.; Van Den Heuvel, W.J. Data Market Design: A Systematic Literature Review. *IEEE Access* **2022**, *10*, 33123–33153. [CrossRef]
81. Chanal, P.M.; Kakkasageri, M.S. Security and Privacy in IoT: A Survey. *Wirel. Pers. Commun.* **2020**, *115*, 1667–1693. [CrossRef]
82. Samonas, S.; Coss, D. The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security. *J. Inf. Syst. Secur.* **2014**, *10*, 21–45.
83. *ISO/IEC 27000:2018(en)*; Information Technology—Security Techniques—Information Security Management Systems—Overview and Vocabulary. International Organization for Standardization: Geneva, Switzerland, 2018.
84. Maqsood, F.; Ali, M.M.; Ahmed, M.; Shah, M.A. Cryptography: A Comparative Analysis for Modern Techniques. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*, 442–448. [CrossRef]
85. Henriques, M.S.; Vernekar, N.K. Using symmetric and asymmetric cryptography to secure communication between devices in IoT. In Proceedings of the 2017 International Conference on IoT and Application (ICIOT), Nagapattinam, India, 19–20 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
86. Shafagh, H.; Hithnawi, A.; Burkhalter, L.; Fischli, P.; Duquennoy, S. Secure Sharing of Partially Homomorphic Encrypted IoT Data. In Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems (SenSys), Delft, The Netherlands, 6–8 November 2017; ACM: New York, NY, USA, 2017; pp. 1–14.
87. Do, H.G.; Ng, W.K. Blockchain-Based System for Secure Data Storage with Private Keyword Search. In Proceedings of the 2017 IEEE World Congress on Services (SERVICES), Honolulu, HI, USA, 25–30 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 90–93.
88. Sun, X.; Zhang, P.; Liu, J.K.; Yu, J.; Xie, W. Private Machine Learning Classification Based on Fully Homomorphic Encryption. *IEEE Trans. Emerg. Top. Comput.* **2020**, *8*, 352–364. [CrossRef]
89. Ouaddah, A.; Mousannif, H.; Abou Elkalam, A.; Ait Ouahman, A. Access control in the Internet of Things: Big challenges and new opportunities. *Comput. Netw.* **2017**, *112*, 237–262. [CrossRef]
90. Qiu, J.; Tian, Z.; Du, C.; Zuo, Q.; Su, S.; Fang, B. A Survey on Access Control in the Age of Internet of Things. *IEEE Internet Things J.* **2020**, *7*, 4682–4696. [CrossRef]
91. Alagar, V.; Alsaig, A.; Ormandjiva, O.; Wan, K. Context-Based Security and Privacy for Healthcare IoT. In Proceedings of the 2018 IEEE International Conference on Smart Internet of Things (SmartIoT), Xi'an, China, 17–19 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 122–128.
92. Alkhresheh, A.; Elgazzar, K.; Hassanein, H.S. Context-aware Automatic Access Policy Specification for IoT Environments. In Proceedings of the 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), Limassol, Cyprus, 25–29 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 793–799.
93. Novo, O. Blockchain Meets IoT: An Architecture for Scalable Access Management in IoT. *IEEE Internet Things J.* **2018**, *5*, 1184–1195. [CrossRef]
94. Raikwar, M.; Gligoroski, D.; Velinov, G. Trends in Development of Databases and Blockchain. In Proceedings of the 2020 Seventh International Conference on Software Defined Systems (SDS), Paris, France, 20–23 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 177–182.
95. Li, R.; Song, T.; Mei, B.; Li, H.; Cheng, X.; Sun, L. Blockchain for Large-Scale Internet of Things Data Storage and Protection. *IEEE Trans. Serv. Comput.* **2019**, *12*, 762–771. [CrossRef]
96. Chowdhury, M.J.M.; Colman, A.; Kabir, M.A.; Han, J.; Sarda, P. Blockchain as a Notarization Service for Data Sharing with Personal Data Store. In Proceedings of the 2018 17th IEEE International Conference on Trust, Security And Privacy In Computing and Communications/12th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), New York, NY, USA, 1–3 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1330–1335.
97. Gupta, S.; Hellings, J.; Rahnama, S.; Sadoghi, M. Building High Throughput Permissioned Blockchain Fabrics: Challenges and Opportunities. *Proc. VLDB Endow.* **2020**, *13*, 3441–3444. [CrossRef]
98. Li, Y.; Zheng, K.; Yan, Y.; Liu, Q.; Zhou, X. EtherQL: A Query Layer for Blockchain System. In Proceedings of the 22nd International Conference on Database Systems for Advanced Applications (DASFAA), Suzhou, China, 27–30 March 2017; Springer: Cham, Switzerland, 2017; pp. 556–567.
99. Bragagnolo, S.; Rocha, H.; Denker, M.; Ducasse, S. Ethereum Query Language. In Proceedings of the 1st International Workshop on Emerging Trends in Software Engineering for Blockchain (WETSEB), Gothenburg, Sweden, 27 May 2018; ACM: New York, NY, USA, 2018; pp. 1–8.

100. Qu, Q.; Nurgaliev, I.; Muzammal, M.; Jensen, C.S.; Fan, J. On spatio-temporal blockchain query processing. *Future Gener. Comput. Syst.* **2019**, *98*, 208–218. [CrossRef]
101. Hao, K.; Xin, J.; Wang, Z.; Yao, Z.; Wang, G. On efficient top-k transaction path query processing in blockchain database. *Data Knowl. Eng.* **2022**, *141*, 102079. [CrossRef]
102. Han, J.; Kim, H.; Eom, H.; Coignard, J.; Wu, K.; Son, Y. Enabling SQL-Query Processing for Ethereum-Based Blockchain Systems. In Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics (WIMS), Seoul, Republic of Korea, 26–28 June 2019; ACM: New York, NY, USA, 2019; pp. 1–7.
103. Przytarski, D. Using Triples as the Data Model for Blockchain Systems. In Proceedings of the Blockchain enabled Semantic Web Workshop and Contextualized Knowledge Graphs Workshop Co-Located with the 18th International Semantic Web Conference (BlockSW/CKG@ISWC), Auckland, New Zealand, 26–30 October 2019; pp. 1–2.
104. Kurt Peker, Y.; Rodriguez, X.; Ericsson, J.; Lee, S.J.; Perez, A.J. A Cost Analysis of Internet of Things Sensor Data Storage on Blockchain via Smart Contracts. *Electronics* **2020**, *9*, 244. [CrossRef]
105. Hepp, T.; Sharinghousen, M.; Ehret, P.; Schoenhals, A.; Gipp, B. On-chain vs. off-chain storage for supply- and blockchain integration. *IT Inf. Technol.* **2018**, *60*, 283–291. [CrossRef]
106. Schuhknecht, F.; Sharma, A.; Dittrich, J.; Agrawal, D. chainifyDB: How to get rid of your Blockchain and use your DBMS instead. In Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR), Chaminade, CA, USA, 11–15 January 2021; pp. 1–10.
107. Wang, H.; Xu, C.; Zhang, C.; Xu, J. vChain: A Blockchain System Ensuring Query Integrity. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD), Portland, OR, USA, 14–19 June 2020; ACM: New York, NY, USA, 2020; pp. 2693–2696.
108. Nathan, S.; Govindarajan, C.; Saraf, A.; Sethi, M.; Jayachandran, P. Blockchain Meets Database: Design and Implementation of a Blockchain Relational Database. *Proc. VLDB Endow.* **2019**, *12*, 1539–1552. [CrossRef]
109. Cai, H.; Xu, B.; Jiang, L.; Vasilakos, A.V. IoT-Based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges. *IEEE Internet Things J.* **2017**, *4*, 75–87. [CrossRef]
110. Habib, S.M.; Hauke, S.; Ries, S.; Mühlhäuser, M. Trust as a facilitator in cloud computing: A survey. *J. Cloud Comput. Adv. Syst. Appl.* **2012**, *1*, 19. [CrossRef]
111. Khan, K.M.; Malluhi, Q. Establishing Trust in Cloud Computing. *IT Prof.* **2010**, *12*, 20–27. [CrossRef]
112. Gilad-Bachrach, R.; Laine, K.; Lauter, K.; Rindal, P.; Rosulek, M. Secure Data Exchange: A Marketplace in the Cloud. In Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop (CCSW), London, UK, 11 November 2019; ACM: New York, NY, USA, 2019; pp. 117–128.
113. Gritti, C.; Chen, R.; Susilo, W.; Plantard, T. Dynamic Provable Data Possession Protocols with Public Verifiability and Data Privacy. In Proceedings of the 13th International Conference on Information Security Practice and Experience (ISPEC), Melbourne, Australia, 13–15 December 2017; Springer: Cham, Switzerland, 2017; pp. 485–505.
114. Du, R.; Deng, L.; Chen, J.; He, K.; Zheng, M. Proofs of Ownership and Retrievability in Cloud Storage. In Proceedings of the 2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), Beijing, China, 24–26 September 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 328–335.
115. Erway, C.C.; Küpçü, A.; Papamanthou, C.; Tamassia, R. Dynamic Provable Data Possession. *ACM Trans. Inf. Syst. Secur.* **2015**, *17*, 1–29. [CrossRef]
116. Merkle, R.C. A Digital Signature Based on a Conventional Encryption Function. In Proceedings of the Conference on the Theory and Applications of Cryptographic Techniques (CRYPTO), Santa Barbara, CA, USA, 16–20 August 1987; Springer: Berlin/Heidelberg, Germany, 1988; pp. 369–378.
117. Gritti, C.; Li, H. Efficient Publicly Verifiable Proofs of Data Replication and Retrievability Applicable for Cloud Storage. *Adv. Sci. Technol. Eng. Syst. J.* **2022**, *7*, 107–124. [CrossRef]
118. Boneh, D.; Bonneau, J.; Bünz, B.; Fisch, B. Verifiable Delay Functions. In Proceedings of the 38th International Cryptology Conference (Crypto), Santa Barbara, CA, USA, 17–19 August 2018; Springer: Cham, Switzerland, 2018; pp. 757–788.
119. Salim, M.M.; Rathore, S.; Park, J.H. Distributed denial of service attacks and its defenses in IoT: A survey. *J. Supercomput.* **2020**, *76*, 5320–5363. [CrossRef]
120. Zhang, H.; Wen, Y.; Xie, H.; Yu, N. *Distributed Hash Table: Theory, Platforms and Applications*; Springer: New York, NY, USA, 2013.
121. Singh, R.; Tanwar, S.; Sharma, T.P. Utilization of blockchain for mitigating the distributed denial of service attacks. *Secur. Priv.* **2020**, *3*, e96. [CrossRef]
122. Liu, X.; Farahani, B.; Firouzi, F. Distributed Ledger Technology. In *Intelligent Internet of Things: From Device to Fog and Cloud*; Firouzi, F., Chakrabarty, K., Nassif, S., Eds.; Springer: Cham, Switzerland, 2020; Chapter 8, pp. 393–431.
123. Zhu, Q.; Loke, S.W.; Trujillo-Rasua, R.; Jiang, F.; Xiang, Y. Applications of Distributed Ledger Technologies to the Internet of Things: A Survey. *ACM Comput. Surv.* **2019**, *52*, 1–34. [CrossRef]
124. Peng, Y.; Du, M.; Li, F.; Cheng, R.; Song, D. FalconDB: Blockchain-Based Collaborative Database. In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (SIGMOD), Portland, OR, USA, 14–19 June 2020; ACM: New York, NY, USA, 2020; pp. 637–652.
125. El-Hindi, M.; Binnig, C.; Arasu, A.; Kossmann, D.; Ramamurthy, R. BlockchainDB: A Shared Database on Blockchains. *Proc. VLDB Endow.* **2019**, *12*, 1597–1609. [CrossRef]

126. Barkadehi, M.H.; Nilashi, M.; Ibrahim, O.; Zakeri Fardi, A.; Samad, S. Authentication systems: A literature review and classification. *Telemat. Inform.* **2018**, *35*, 1491–1511. [CrossRef]
127. Ferrag, M.A.; Maglaras, L.; Derhab, A. Authentication and Authorization for Mobile IoT Devices Using Biofeatures: Recent Advances and Future Trends. *Secur. Commun. Netw.* **2019**, *2019*, 5452870. [CrossRef]
128. Cheong, S.N.; Ling, H.C.; Teh, P.L. Secure Encrypted Steganography Graphical Password scheme for Near Field Communication smartphone access control system. *Expert Syst. Appl.* **2014**, *41*, 3561–3568. [CrossRef]
129. Baig, A.F.; Eskeland, S. Security, Privacy, and Usability in Continuous Authentication: A Survey. *Sensors* **2021**, *35*, 5967. [CrossRef]
130. Sciancalepore, S.; Piro, G.; Caldarola, D.; Boggia, G.; Bianchi, G. OAuth-IoT: An access control framework for the Internet of Things based on open standards. In Proceedings of the 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, Crete, Greece, 3–6 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 676–681.
131. Kulseng, L.; Yu, Z.; Wei, Y.; Guan, Y. Lightweight Mutual Authentication and Ownership Transfer for RFID Systems. In Proceedings of the 2010 30th IEEE International Conference on Computer Communications (INFOCOM), San Diego, CA, USA, 14–19 March 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1–5.
132. Maes, R. *Physically Unclonable Functions*; Springer: Berlin/Heidelberg, Germany, 2013.
133. He, W.; Golla, M.; Padhi, R.; Ofek, J.; Dürmuth, M.; Fernandes, E.; Ur, B. Rethinking Access Control and Authentication for the Home Internet of Things (IoT). In Proceedings of the 27th USENIX Security Symposium (USENIX Security), Baltimore, MD, USA, 15–17 August 2018; USENIX Association: Berkeley, CA, USA, 2018; pp. 255–272.
134. Almuairfi, S.; Veeraraghavan, P.; Chilamkurti, N. A novel image-based implicit password authentication system (IPAS) for mobile and non-mobile devices. *Math. Comput. Model.* **2013**, *58*, 108–116. [CrossRef]
135. Hu, V.C.; Kuhn, D.R.; Ferraiolo, D.F.; Voas, J. Attribute-Based Access Control. *Computer* **2015**, *48*, 85–88. [CrossRef]
136. Hemdi, M.; Deters, R. Using REST based protocol to enable ABAC within IoT systems. In Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 13–15 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–7.
137. Hüffmeyer, M.; Schreier, U. Formal Comparison of an Attribute Based Access Control Language for RESTful Services with XACML. In Proceedings of the 21st ACM on Symposium on Access Control Models and Technologies (SACMAT), Shanghai, China, 6–8 June 2016; ACM: New York, NY, USA, 2016; pp. 171–178.
138. Liu, S.; You, S.; Yin, H.; Lin, Z.; Liu, Y.; Yao, W.; Sundaresh, L. Model-Free Data Authentication for Cyber Security in Power Systems. *IEEE Trans. Smart Grid* **2020**, *11*, 4565–4568. [CrossRef]
139. Arnold, M.; Schmucker, M.; Wolthusen, S.D. *Techniques and Applications of Digital Watermarking and Content Protection*; Artech House, Inc.: Norwood, MA, USA, 2003.
140. Agrawal, R.; Haas, P.J.; Kiernan, J. Watermarking relational data: Framework, algorithms and analysis. *VLDB J.* **2003**, *12*, 157–169.
141. Subramanya, S.; Yi, B.K. Digital signatures. *IEEE Potentials* **2006**, *25*, 5–8. [CrossRef]
142. Kaur, R.; Kaur, A. Digital Signature. In Proceedings of the 2012 International Conference on Computing Sciences (ICCS), Phagwara, India, 14–15 September 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 295–301.
143. Gritti, C.; Molva, R.; Önen, M. Lightweight Secure Bootstrap and Message Attestation in the Internet of Things. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC), Pau, France, 9–13 April 2018; ACM: New York, NY, USA, 2018; pp. 775–782.
144. Gritti, C.; Önen, M.; Molva, R. CHARIOT: Cloud-Assisted Access Control for the Internet of Things. In Proceedings of the 2018 16th Annual Conference on Privacy, Security and Trust (PST), Belfast, Ireland, 28–30 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
145. Gritti, C.; Önen, M.; Molva, R. Privacy-Preserving Delegable Authentication in the Internet of Things. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC), Limassol, Cyprus, 8–12 April 2019; ACM: New York, NY, USA, 2019; pp. 861–869.
146. Antoniadis, S.; Litou, I.; Kalogeraki, V. A Model for Identifying Misinformation in Online Social Networks. In Proceedings of the 2015 Confederated International Conferences on the Move to Meaningful Internet Systems: CoopIS, ODBASE, and C&TC (OTM), Rhodes, Greece, 26–30 October 2015; Springer: Cham, Switzerland, 2015; pp. 473–482.
147. Litou, I.; Kalogeraki, V.; Katakis, I.; Gunopulos, D. Efficient and timely misinformation blocking under varying cost constraints. *Online Soc. Netw. Media* **2017**, *2*, 19–31. [CrossRef]
148. Litou, I.; Kalogeraki, V. Influence Maximization in Evolving Multi-Campaign Environments. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 448–457.
149. Lupton, D. *The Quantified Self*; Polity: Cambridge, UK; Malden, MA, USA, 2016.
150. Jiang, B.; Li, J.; Yue, G.; Song, H. Differential Privacy for Industrial Internet of Things: Opportunities, Applications, and Challenges. *IEEE Internet Things J.* **2021**, *8*, 10430–10451. [CrossRef]
151. Perr, I.N. Privilege, confidentiality, and patient privacy: Status 1980. *J. Forensic Sci.* **1981**, *26*, 109–115. [CrossRef]
152. Read, G. The Seal of Confession. *Law Justice Christ. Law Rev.* **2022**, *188*, 28–37.
153. Roba, R.M. The legal protection of the secrecy of correspondence. *Curentul Jurid. Jurid. Curr. Le Courant Juridique* **2009**, *36*, 135–154.
154. Bok, S. The Limits of Confidentiality. *Hastings Cent. Rep.* **1983**, *13*, 24–31. [CrossRef] [PubMed]
155. Hirshleifer, J. Privacy: Its Origin, Function, and Future. *J. Leg. Stud.* **1980**, *9*, 649–664. [CrossRef]

156. Westin, A.F. *Privacy and Freedom*; Atheneum Books: New York City, NY, USA, 1967.
157. Margulis, S.T. On the Status and Contribution of Westin's and Altman's Theories of Privacy. *J. Soc. Issues* **2003**, *59*, 411–429. [CrossRef]
158. Diamantopoulou, V.; Lambrinoudakis, C.; King, J.; Gritzalis, S. EU GDPR: Toward a Regulatory Initiative for Deploying a Private Digital Era. In *Modern Socio-Technical Perspectives on Privacy*; Knijnenburg, B.P., Page, X., Wisniewski, P., Lipford, H.R., Proferes, N., Romano, J., Eds.; Springer: Cham, Switzerland, 2022; Chapter 18, pp. 427–448.
159. European Parliament and Council of the European Union. Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive). Legislative Acts L119. *Off. J. Eur. Union* **2016**. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 6 February 2023).
160. Williams, M.; Nurse, J.R.C.; Creese, S. The Perfect Storm: The Privacy Paradox and the Internet-of-Things. In Proceedings of the 2016 11th International Conference on Availability, Reliability and Security (ARES), Salzburg, Austria, 31 August–2 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 644–652.
161. Jung, F.; von Holdt, K.; Krüger, R.; Meyer, J.; Heuten, W. I Do. Do I?—Understanding User Perspectives on the Privacy Paradox. In Proceedings of the 25th International Academic Mindtrek Conference (Academic Mindtrek), Tampere, Finland, 16–18 November 2022; ACM: New York, NY, USA, 2022; pp. 268–277.
162. Rubinstein, I.S.; Good, N. The trouble with Article 25 (and how to fix it): The future of data protection by design and default. *Int. Data Priv. Law* **2020**, *10*, 37–56. [CrossRef]
163. Georgiopoulou, Z.; Makri, E.L.; Lambrinoudakis, C. GDPR compliance: Proposed technical and organizational measures for cloud provider. *Inf. Comput. Secur.* **2020**, *28*, 665–680. [CrossRef]
164. Gyrard, A.; Zimmermann, A.; Sheth, A. Building IoT-Based Applications for Smart Cities: How Can Ontology Catalogs Help? *IEEE Internet Things J.* **2018**, *5*, 3978–3990. [CrossRef]
165. Chen, G.; Jiang, T.; Wang, M.; Tang, X.; Ji, W. Modeling and reasoning of IoT architecture in semantic ontology dimension. *Comput. Commun.* **2020**, *153*, 580–594. [CrossRef]
166. Gheisari, M.; Najafabadi, H.E.; Alzubi, J.A.; Gao, J.; Wang, G.; Abbasi, A.A.; Castiglione, A. OBPP: An ontology-based framework for privacy-preserving in IoT-based smart city. *Future Gener. Comput. Syst.* **2021**, *123*, 1–13. [CrossRef]
167. Leveson, N.G. *Engineering a Safer World: Systems Thinking Applied to Safety*; MIT Press: Cambridge, MA, USA, 2011.
168. Shapiro, S.S. Privacy Risk Analysis Based on System Control Structures: Adapting System-Theoretic Process Analysis for Privacy Engineering. In Proceedings of the 2016 IEEE Security and Privacy Workshops (SPW), San Jose, CA, USA, 22–26 May 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 17–24.
169. Renganathan, V.; Yurtsever, E.; Ahmed, Q.; Yener, A. Valet attack on privacy: A cybersecurity threat in automotive Bluetooth infotainment systems. *Cybersecurity* **2022**, *5*, 30. [CrossRef]
170. Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and Explanation in AI-Informed Decision Making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579. [CrossRef]
171. Hagras, H. Toward Human-Understandable, Explainable AI. *Computer* **2018**, *51*, 28–36. [CrossRef]
172. Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; Samek, W. Explainable AI Methods—A Brief Overview. In Proceedings of the International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers Held in Conjunction with ICML 2020 (xxAI), Vienna, Austria, 18 July 2020; Springer: Cham, Switzerland, 2022; pp. 13–38.
173. Utz, C.; Degeling, M.; Fahl, S.; Schaub, F.; Holz, T. (Un)Informed Consent: Studying GDPR Consent Notices in the Field. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS), London, UK, 11–15 November 2019; ACM: New York, NY, USA, 2019; pp. 973–990.
174. Kaaniche, N.; Laurent, M.; Belguith, S. Privacy enhancing technologies for solving the privacy-personalization paradox: Taxonomy and survey. *J. Netw. Comput. Appl.* **2020**, *171*, 102807. [CrossRef]
175. Li, N.; Qardaji, W.; Su, D. On Sampling, Anonymization, and Differential Privacy or, k-Anonymization Meets Differential Privacy. In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security (ASIACCS), Seoul, Republic of Korea, 2–4 May 2012; ACM: New York, NY, USA, 2012; pp. 32–33.
176. Pattuk, E.; Kantarcioglu, M.; Ulusoy, H.; Malin, B. Privacy-aware dynamic feature selection. In Proceedings of the 2015 IEEE 31st International Conference on Data Engineering (ICDE), Seoul, Republic of Korea, 13–17 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 78–88.
177. Dou, H.; Chen, Y.; Yang, Y.; Long, Y. A secure and efficient privacy-preserving data aggregation algorithm. *J. Ambient. Intell. Humaniz. Comput.* **2022**, *13*, 1495–1503. [CrossRef]
178. Alpers, S.; Oberweis, A.; Pieper, M.; Betz, S.; Fritsch, A.; Schiefer, G.; Wagner, M. PRIVACY-AVARE: An approach to manage and distribute privacy settings. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, 13–16 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1460–1468.
179. Jiang, H.; Li, J.; Zhao, P.; Zeng, F.; Xiao, Z.; Iyengar, A. Location Privacy-Preserving Mechanisms in Location-Based Services: A Comprehensive Survey. *ACM Comput. Surv.* **2021**, *54*, 4. [CrossRef]
180. Wang, Z.; Wang, B.; Srivastava, M. Protecting User Data Privacy with Adversarial Perturbations: Poster Abstract. In Proceedings of the 20th International Conference on Information Processing in Sensor Networks Co-Located with CPS-IoT Week 2021 (IPSN), Nashville, TN, USA, 18–21 May 2021; ACM: New York, NY, USA, 2021; pp. 386–387.

181. Hernández Acosta, L.; Reinhardt, D. A Survey on Privacy Issues and Solutions for Voice-Controlled Digital Assistants. *Pervasive Mob. Comput.* **2022**, *80*, 101523. [CrossRef]
182. Palanisamy, S.M.; Dürr, F.; Tariq, M.A.; Rothermel, K. Preserving Privacy and Quality of Service in Complex Event Processing through Event Reordering. In Proceedings of the 12th ACM International Conference on Distributed and Event-Based Systems (DEBS), Hamilton, New Zealand, 25–29 June 2018; ACM: New York, NY, USA, 2018; pp. 40–51.
183. Kwecka, Z.; Buchanan, W.; Schafer, B.; Rauhofer, J. "I am Spartacus": Privacy enhancing technologies, collaborative obfuscation and privacy as a public good. *Artif. Intell. Law* **2014**, *22*, 113–139. [CrossRef]
184. Slijepčević, D.; Henzl, M.; Klausner, L.D.; Dam, T.; Kieseberg, P.; Zeppelzauer, M. k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Comput. Secur.* **2021**, *111*, 102488. [CrossRef]
185. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. *J. Priv. Confidentiality* **2017**, *7*, 17–51. [CrossRef]
186. Machanavajjhala, A.; He, X.; Hay, M. Differential Privacy in the Wild: A Tutorial on Current Practices & Open Challenges. In Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD), Chicago, IL, USA, 14–19 May 2017; ACM: New York, NY, USA, 2017; pp. 1727–1730.
187. Dwork, C.; Kohli, N.; Mulligan, D. Differential Privacy in Practice: Expose your Epsilons! *J. Privacy Confid.* **2019**, *9*, 1–22. [CrossRef]
188. Yang, Q.; Liu, Y.; Cheng, Y.; Kang, Y.; Chen, T.; Yu, H. *Federated Learning*; Morgan & Claypool: San Rafael, CA, USA, 2019.
189. Wu, X.; Zhang, Y.; Shi, M.; Li, P.; Li, R.; Xiong, N.N. An adaptive federated learning scheme with differential privacy preserving. *Future Gener. Comput. Syst.* **2022**, *127*, 362–372. [CrossRef]
190. István, Z.; Ponnapalli, S.; Chidambaram, V. Software-Defined Data Protection: Low Overhead Policy Compliance at the Storage Layer is within Reach! *Proc. VLDB Endow.* **2021**, *14*, 1167–1174. [CrossRef]
191. Wang, W.C.; Ho, C.C.; Chang, Y.M.; Chang, Y.H. Challenges and Designs for Secure Deletion in Storage Systems. In Proceedings of the 2020 Indo—Taiwan 2nd International Conference on Computing, Analytics and Networks (Indo-Taiwan ICAN), Rajpura, India, 7–15 February 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 181–189.
192. Zhang, Q.; Jia, S.; Chang, B.; Chen, B. Ensuring data confidentiality via plausibly deniable encryption and secure deletion—A survey. *Cybersecurity* **2018**, *1*, 1. [CrossRef]
193. Politou, E.; Alepis, E.; Virvou, M.; Patsakis, C. Privacy in Blockchain. In *Privacy and Data Protection Challenges in the Distributed Era*; Springer: Cham, Switzerland, 2022; Chapter 7, pp. 133–149.
194. Meng, W.; Ge, J.; Jiang, T. Secure Data Deduplication with Reliable Data Deletion in Cloud. *Int. J. Found. Comput. Sci.* **2019**, *30*, 551–570. [CrossRef]
195. Waizenegger, T.; Wagner, F.; Mega, C. SDOS: Using Trusted Platform Modules for Secure Cryptographic Deletion in the Swift Object Store. In Proceedings of the 20th International Conference on Extending Database Technology (EDBT), Venice, Italy, 21–24 March 2017; OpenProceedings.org: Konstanz, Germany, 2017; pp. 550–553.
196. Auge, T. Extended Provenance Management for Data Science Applications. In Proceedings of the VLDB 2020 PhD Workshop co-located with the 46th International Conference on Very Large Databases (VLDB), Tokyo, Japan, 31 August–4 September 2020; CEUR-WS.org: Aachen, Germany, 2020; pp. 1–4.
197. Davidson, S.B.; Khanna, S.; Roy, S.; Stoyanovich, J.; Tannen, V.; Chen, Y. On Provenance and Privacy. In Proceedings of the 14th International Conference on Database Theory (ICDT), Uppsala, Sweden, 21–24 March 2011; ACM: New York, NY, USA, 2011; pp. 3–10.
198. Auge, T.; Scharlau, N.; Heuer, A. Privacy Aspects of Provenance Queries. In Proceedings of the 8th and 9th International Provenance and Annotation Workshop (IPAW), Virtual Event, 19–22 July 2021; Springer: Cham, Switzerland, 2021; pp. 218–221.
199. Stach, C.; Alpers, S.; Betz, S.; Dürr, F.; Fritsch, A.; Mindermann, K.; Palanisamy, S.M.; Schiefer, G.; Wagner, M.; Mitschang, B.; et al. The AVARE PATRON—A Holistic Privacy Approach for the Internet of Things. In Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (SECRYPT), Porto, Portugal, 26–28 July 2018; SciTePress: Setúbal, Portugal, 2018; pp. 372–379.
200. Stach, C. How to Deal with Third Party Apps in a Privacy System—The PMP Gatekeeper. In Proceedings of the 2015 16th IEEE International Conference on Mobile Data Management (MDM), Pittsburgh, PA, USA, 15–18 June 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 167–172.
201. Stach, C. VAULT: A Privacy Approach towards High-Utility Time Series Data. In Proceedings of the Thirteenth International Conference on Emerging Security Information, Systems and Technologies (SECURWARE), Nice, France, 27–31 October 2019; IARIA: Wilmington, DE, USA, 2019; pp. 41–46.
202. Stach, C.; Mitschang, B. Privacy Management for Mobile Platforms—A Review of Concepts and Approaches. In Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management (MDM), Milan, Italy, 3–6 June 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 305–313.
203. Stach, C.; Mitschang, B. Design and Implementation of the Privacy Management Platform. In Proceedings of the 2014 IEEE 15th International Conference on Mobile Data Management (MDM), Brisbane, Australia, 14–18 July 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 69–72.

204. Stach, C.; Steimle, F.; Franco da Silva, A.C. TIROL: The Extensible Interconnectivity Layer for mHealth Applications. In Proceedings of the 23rd International Conference on Information and Software Technologies (ICIST), Druskininkai, Lithuania, 12–14 October 2017; Springer: Cham, Switzerland, 2017; pp. 190–202.
205. Stach, C.; Mitschang, B. The Secure Data Container: An Approach to Harmonize Data Sharing with Information Security. In Proceedings of the 2016 17th IEEE International Conference on Mobile Data Management (MDM), Porto, Portugal, 13–16 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 292–297.
206. Stach, C.; Mitschang, B. Curator—A Secure Shared Object Store: Design, Implementation, and Evaluation of a Manageable, Secure, and Performant Data Exchange Mechanism for Smart Devices. In Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC), Pau, France, 9–13 April 2018; ACM: New York, NY, USA, 2018; pp. 533–540.
207. Stach, C.; Mitschang, B. ECHOES: A Fail-Safe, Conflict Handling, and Scalable Data Management Mechanism for the Internet of Things. In Proceedings of the 23rd European Conference on Advances in Databases and Information Systems (ADBIS), Bled, Slovenia, 8–11 September 2019; Springer: Cham, Switzerland, 2019; pp. 373–389.
208. Stach, C.; Gritti, C.; Mitschang, B. Bringing Privacy Control Back to Citizens: DISPEL—A Distributed Privacy Management Platform for the Internet of Things. In Proceedings of the 35th Annual ACM Symposium on Applied Computing (SAC), Brno, Czech Republic, 30 March–3 April 2020; ACM: New York, NY, USA, 2020; pp. 1272–1279.
209. Stach, C.; Behringer, M.; Bräcker, J.; Gritti, C.; Mitschang, B. SMARTEN—A Sample-Based Approach towards Privacy-Friendly Data Refinement. *J. Cybersecur. Priv.* **2022**, *2*, 606–628. [CrossRef]
210. Stach, C.; Bräcker, J.; Eichler, R.; Giebler, C.; Mitschang, B. Demand-Driven Data Provisioning in Data Lakes: BARENTS—A Tailorable Data Preparation Zone. In Proceedings of the 23rd International Conference on Information Integration and Web Intelligence (iiWAS), Linz, Austria, 29 November–1 December 2021; ACM: New York, NY, USA, 2021; pp. 187–198.
211. Weber, C.; Hirmer, P.; Reimann, P.; Schwarz, H. A New Process Model for the Comprehensive Management of Machine Learning Models. In Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS), Heraklion, Crete, Greece, 3–5 May 2019; SciTePress: Setúbal, Portugal, 2019; pp. 415–422.
212. Weber, C.; Hirmer, P.; Reimann, P. A Model Management Platform for Industry 4.0 – Enabling Management of Machine Learning Models in Manufacturing Environments. In Proceedings of the 23rd International Conference on Business Information Systems (BIS), Colorado Springs, CO, USA, 8–10 June 2020; Springer: Cham, Switzerland, 2020; pp. 403–417.
213. Stach, C.; Giebler, C.; Wagner, M.; Weber, C.; Mitschang, B. AMNESIA: A Technical Solution towards GDPR-compliant Machine Learning. In Proceedings of the 6th International Conference on Information Systems Security and Privacy (ICISSP), Valletta, Malta, 25–27 February 2020; SciTePress: Setúbal, Portugal, 2020; pp. 21–32.
214. Weber, C.; Reimann, P. MMP—A Platform to Manage Machine Learning Models in Industry 4.0 Environments. In Proceedings of the 2020 IEEE 24th International Enterprise Distributed Object Computing Workshop (EDOCW), Eindhoven, The Netherlands, 5 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 91–94.
215. Stach, C.; Gritti, C.; Przytarski, D.; Mitschang, B. Trustworthy, Secure, and Privacy-aware Food Monitoring Enabled by Blockchains and the IoT. In Proceedings of the 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Austin, TX, USA, 23–27 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–4.
216. Stach, C.; Gritti, C.; Przytarski, D.; Mitschang, B. Assessment and Treatment of Privacy Issues in Blockchain Systems. *ACM SIGAPP Appl. Comput. Rev.* **2022**, *22*, 5–24. [CrossRef]
217. Przytarski, D.; Stach, C.; Gritti, C.; Mitschang, B. Query Processing in Blockchain Systems: Current State and Future Challenges. *Future Internet* **2022**, *14*, 1. [CrossRef]
218. Giebler, C.; Stach, C.; Schwarz, H.; Mitschang, B. BRAID—A Hybrid Processing Architecture for Big Data. In Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA), Lisbon, Portugal, 26–28 July 2018; SciTePress: Setúbal, Portugal, 2018; pp. 294–301.
219. Stach, C.; Bräcker, J.; Eichler, R.; Giebler, C.; Gritti, C. How to Provide High-Utility Time Series Data in a Privacy-Aware Manner: A VAULT to Manage Time Series Data. *Int. J. Adv. Secur.* **2020**, *13*, 88–108.
220. Mindermann, K.; Riedel, F.; Abdulkhaleq, A.; Stach, C.; Wagner, S. Exploratory Study of the Privacy Extension for System Theoretic Process Analysis (STPA-Priv) to Elicit Privacy Risks in eHealth. In Proceedings of the 2017 IEEE 25th International Requirements Engineering Conference Workshops (REW), Lisbon, Portugal, 4–8 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 90–96.
221. Stach, C.; Steimle, F. Recommender-Based Privacy Requirements Elicitation—EPICUREAN: An Approach to Simplify Privacy Settings in IoT Applications with Respect to the GDPR. In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC), Limassol, Cyprus, 8–12 April 2019; ACM: New York, NY, USA, 2019; pp. 1500–1507.
222. Stach, C.; Gritti, C.; Bräcker, J.; Behringer, M.; Mitschang, B. Protecting Sensitive Data in the Information Age: State of the Art and Future Prospects. *Future Internet* **2022**, *14*, 302. [CrossRef]
223. Stach, C.; Mitschang, B. ACCESSORS—A Data-Centric Permission Model for the Internet of Things. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), Funchal, Madeira, Portugal, 22–24 January 2018; SciTePress: Setúbal, Portugal, 2018; pp. 30–40.
224. Siewiorek, D. Generation smartphone. *IEEE Spectrum* **2012**, *49*, 54–58. [CrossRef]

225. Lu, H.; Frauendorfer, D.; Rabbi, M.; Mast, M.S.; Chittaranjan, G.T.; Campbell, A.T.; Gatica-Perez, D.; Choudhury, T. StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp), Pittsburgh, PA, USA, 5–8 September 2012; ACM: New York, NY, USA, 2012; pp. 351–360.
226. Spathis, D.; Servia-Rodriguez, S.; Farrahi, K.; Mascolo, C.; Rentfrow, J. Passive Mobile Sensing and Psychological Traits for Large Scale Mood Prediction. In Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth), Trento, Italy, 20–23 May 2019; ACM: New York, NY, USA, 2019; pp. 272–281.
227. Christ, P.F.; Schlecht, S.; Ettlinger, F.; Grün, F.; Heinle, C.; Tatavatry, S.; Ahmadi, S.A.; Diepold, K.; Menze, B.H. Diabetes60—Inferring Bread Units From Food Images Using Fully Convolutional Neural Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1526–1535.
228. Madrid, R.E.; Ashur Ramallo, F.; Barraza, D.E.; Chaile, R.E. Smartphone-Based Biosensor Devices for Healthcare: Technologies, Trends, and Adoption by End-Users. *Bioengineering* **2022**, *9*, 101. [CrossRef] [PubMed]
229. Knöll, M.; Neuheuser, K.; Cleff, T.; Rudolph-Cleff, A. A tool to predict perceived urban stress in open public spaces. *Environ. Plan. B Urban Anal. City Sci.* **2018**, *45*, 797–813. [CrossRef]
230. Moosa, A.M.; Al-Maadeed, N.; Saleh, M.; Al-Maadeed, S.A.; Aljaam, J.M. Designing a Mobile Serious Game for Raising Awareness of Diabetic Children. *IEEE Access* **2020**, *8*, 222876–222889. [CrossRef]
231. Stach, C. Secure Candy Castle—A Prototype for Privacy-Aware mHealth Apps. In Proceedings of the 2016 17th IEEE International Conference on Mobile Data Management (MDM), Porto, Portugal, 13–16 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 361–364.
232. Shan, R.; Sarkar, S.; Martin, S.S. Digital health technology and mobile devices for the management of diabetes mellitus: State of the art. *Diabetologia* **2019**, *62*, 877–887. [CrossRef] [PubMed]
233. Wangorsch, A.; Kulkarni, A.; Jamin, A.; Spiric, J.; Bräcker, J.; Brockmeyer, J.; Mahler, V.; Blanca-López, N.; Ferrer, M.; Blanca, M.; et al. Identification and Characterization of IgE-Reactive Proteins and a New Allergen (Cic a 1.01) from Chickpea (Cicer arietinum). *Mol. Nutr. Food Res.* **2020**, *64*, 2000560. [CrossRef] [PubMed]
234. Bräcker, J.; Brockmeyer, J. Characterization and Detection of Food Allergens Using High-Resolution Mass Spectrometry: Current Status and Future Perspective. *J. Agric. Food Chem.* **2018**, *66*, 8935–8940. [CrossRef] [PubMed]
235. Korte, R.; Bräcker, J.; Brockmeyer, J. Gastrointestinal digestion of hazelnut allergens on molecular level: Elucidation of degradation kinetics and resistant immunoactive peptides using mass spectrometry. *Mol. Nutr. Food Res.* **2017**, *61*, 1700130. [CrossRef] [PubMed]
236. Lee, S.; Cui, B.; Bhandari, M.; Luo, N.; Im, P. VizBrick: A GUI-based Interactive Tool for Authoring Semantic Metadata for Building Datasets. In Proceedings of the 21st International Semantic Web Conference (ISWC), Hangzhou, China, 23–27 October 2022; CEUR-WS.org: Aachen, Germany, 2022; pp. 1–5.
237. Molin, S. *Hands-On Data Analysis with Pandas: A Python Data Science Handbook for Data Collection, Wrangling, Analysis, and Visualization*,2nd ed.; Packt Publishing: Birmingham, UK; Mumbai, India, 2021.
238. McKinney, W. Apache Arrow and the “10 Things I Hate about Pandas”. Archives for Wes McKinney. 2017. Available online: https://wesmckinney.com/blog/apache-arrow-pandas-internals/ (accessed on 6 February 2023).
239. Petersohn, D.; Macke, S.; Xin, D.; Ma, W.; Lee, D.; Mo, X.; Gonzalez, J.E.; Hellerstein, J.M.; Joseph, A.D.; Parameswaran, A. Towards Scalable Dataframe Systems. *Proc. VLDB Endow.* **2020**, *13*, 2033–2046. [CrossRef]
240. Petersohn, D.; Tang, D.; Durrani, R.; Melik-Adamyan, A.; Gonzalez, J.E.; Joseph, A.D.; Parameswaran, A.G. Flexible Rule-Based Decomposition and Metadata Independence in Modin: A Parallel Dataframe System. *Proc. VLDB Endow.* **2022**, *15*, 739–751. [CrossRef]
241. Rocklin, M. Dask: Parallel Computation with Blocked algorithms and Task Scheduling. In Proceedings of the 14th Python in Science Conference (SciPy), Austin, TX, USA, 6–12 July 2015; pp. 126–132.
242. Moritz, P.; Nishihara, R.; Wang, S.; Tumanov, A.; Liaw, R.; Liang, E.; Elibol, M.; Yang, Z.; Paul, W.; Jordan, M.I.; et al. Ray: A Distributed Framework for Emerging AI Applications. In Proceedings of the 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Carlsbad, CA, USA, 8–10 October 2018; USENIX Association: Berkeley, CA, USA, 2018; pp. 561–577.
243. Sarkar, T. Parallelized Data Science. In *Productive and Efficient Data Science with Python: With Modularizing, Memory Profiles, and Parallel/GPU Processing*; Apress: Berkeley, CA, USA, 2022; Chapter 10, pp. 257–298.
244. Kläbe, S.; Hagedorn, S. Applying Machine Learning Models to Scalable DataFrames with Grizzly. In Proceedings of the 19. Fachtagung für Datenbanksysteme für Business, Technologie und Web (BTW), Dresden, Germany, 19 April–21 June 2021; GI: Bonn, Germany, 2021; pp. 195–214.
245. Hagedorn, S.; Kläbe, S.; Sattler, K.U. Putting Pandas in a Box. In Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR), Chaminade, CA, USA, 11–15 January 2021; pp. 1–6.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.