big data and cognitive computing

MDPI

Review

An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management

Athira Nambiar * and Divyansh Mundra

Department of Computational Intelligence, School of Computing, SRM Institute of Science and Technology, Chennai 603203, India

* Correspondence: athiram@srmist.edu.in

Abstract: Data is the lifeblood of any organization. In today's world, organizations recognize the vital role of data in modern business intelligence systems for making meaningful decisions and staying competitive in the field. Efficient and optimal data analytics provides a competitive edge to its performance and services. Major organizations generate, collect and process vast amounts of data, falling under the category of big data. Managing and analyzing the sheer volume and variety of big data is a cumbersome process. At the same time, proper utilization of the vast collection of an organization's information can generate meaningful insights into business tactics. In this regard, two of the popular data management systems in the area of big data analytics (i.e., data warehouse and data lake) act as platforms to accumulate the big data generated and used by organizations. Although seemingly similar, both of them differ in terms of their characteristics and applications. This article presents a detailed overview of the roles of data warehouses and data lakes in modern enterprise data management. We detail the definitions, characteristics and related works for the respective data management frameworks. Furthermore, we explain the architecture and design considerations of the current state of the art. Finally, we provide a perspective on the challenges and promising research directions for the future.

Keywords: big data; data warehousing; data lake; enterprise data management; OLAP; ETL tools; metadata; cloud computing; Internet of Things

Citation: Nambiar, A.; Mundra, D. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data Cogn. Comput.* **2022**, *6*, 132. https://doi.org/10.3390/bdcc6040132

Academic Editors: Domenico Talia and Fabrizio Marozzo

Received: 28 September 2022
Accepted: 2 November 2022
Published: 7 November 2022

1. Introduction

Big data analytics is one of the buzzwords in today's digital world. It entails examining big data and uncovering the hidden patterns, correlations, etc. available in the data [1]. Big data analytics extracts and analyzes random data sets, forming them into meaningful information. According to statistics, the overall amount of data generated in the world in 2021 was approximately 79 zettabytes, and this is expected to double by 2025 [2]. This unprecedented amount of data was the result of a data explosion that occurred during the last decade, wherein data interactions increased by 5000% [3].

Big data deals with the volume, variety, and velocity of data to process and provides veracity (insightfulness) and value to data. These are known as the 5 Vs of big data [4]. An unprecedented amount of diverse data is acquired, stored, and processed with high data quality for various application domains. These include business transactions, real-time streaming, social media, video analytics, and text mining, creating a huge amount of semi- or unstructured data to be stored in different information silos [5]. The efficient integration and analysis of these multiple data across silos are required to divulge complete insight into the database. This is an open research topic of interest.

Big data and its related emerging technologies have been changing the way e-commerce and e-services operate and have been opening new frontiers in business analytics and related research [6]. Big data analytics systems play a big role in the modern enterprise management domain, from product distribution to sales and marketing, as well as analyzing hidden trends, similarities, and other insights and allowing companies to analyze and optimize their data

to find new opportunities [7]. Since organizations with better and more accurate data can make informed business decisions by looking at market trends and customer preferences, they can gain competitive advantages over others. Hence, organizations invest tremendously in artificial intelligence (AI) and big data technologies to strive toward digital transformation and data-driven decision making, which ultimately leads to advanced business intelligence [6]. As per reports, the worldwide big data analytics and business intelligence software applications markets seem as though they will increase by USD 68 billion and 17.6 billion by 2024–2025, respectively [8].

Big data repositories exist in many forms, as per the requirements of corporations [9]. An effective data repository needs to unify, regulate, evaluate, and deploy a huge amount of data resources to enhance the analytics and query performance. Based on the nature and the application scenario, there are many different types of data repositories other than traditional relational databases. Two of the popular data repositories among them are enterprise data warehouses and data lakes [10–12].

A data warehouse (DW) is a data repository which stores structured, filtered, and processed data that has been treated for a specific purpose, whereas a data lake (DL) is a vast pool of data for which the purpose is not defined [9]. In detail, data warehouses store large amounts of data collected by different sources, typically using predefined schemas. Typically, a DW is a purpose-built relational database running on specialized hardware either on the premises or in the cloud [13]. DWs have been used widely for storing enterprise data and fueling business intelligence and analytics applications [14–16].

Data lakes (DLs) have emerged as big data repositories that store raw data and provide a rich list of functionalities with the help of metadata descriptions [10]. Although the DL is also a form of enterprise data storage, it does not inherently include the same analytics features commonly associated with data warehouses. Instead, they are repositories storing raw data in their original formats and providing a common access interface. From the lake, data may flow downstream to a DW to get processed, packaged, and become ready for consumption. As a relatively new concept, there has been very limited research discussing various aspects of data lakes, especially in Internet articles or blogs.

Although data warehouses and data lakes share some overlapping features and use cases, there are fundamental differences in the data management philosophies, design characteristics, and ideal use conditions for each of these technologies. In this context, we provide a detailed overview and differences between both the DW and DL data management schemes in this survey paper. Furthermore, we consolidate the concepts and give a detailed analysis of different design aspects, various tools and utilities, etc., along with recent developments that have come into existence.

The remainder of this paper is organized as follows. In Section 2, the terminology and basic definitions of big data analytics and the data management schemes are analyzed. Furthermore, the related works in the field are also summarized in this section. In Section 3, the architectures of both the data warehouse and data lake are presented. Next, in Section 4, the key design aspects of the DW and DL models along with their practical aspects are presented at length. Section 5 summarizes the various popular tools and services available for enterprise data management. In Sections 6 and 7, the open challenges and promising directions are explained, respectively. In particular, the pros and cons of various methods are critically discussed, and the observations are presented. Finally, Section 8 concludes this survey paper.

2. Definition: Big Data Analytics, Data Warehouses, and Data Lakes

The definitions and fundamental notions of various data management schemes are provided in this section. Furthermore, related works and review papers on this topic are also summarized.

2.1. Big Data Analytics

With significant advancements in technology, unprecedented usage of computer networks, multimedia, the Internet of Things, social media, and cloud computing has occured [17]. As a result, a huge amount of data, known as "big data", has been generated. It is required to collect, manage, and analyze these data efficiently via big data processing. The process of big data processing is aimed at data mining (i.e., extracting knowledge from large amounts of data), leveraging data management, machine learning, high-performance computing, statistics, pattern recognition, etc. The important characteristics of big data (known as the seven Vs of big data) (https://impact.com/marketing-intelligence/7-vs-big-data/, accessed on 25 September 2022) are as follows:

- Volume, or the available amount of data;
- Velocity, or the speed of data processing;
- Variety, or the different types of big data;
- Volatility, or the variability of the data;
- Veracity, or the accuracy of the data;
- Visualization, or the depiction of big data-generated insights through visual representation;
- Value, or the benefits organizations derive from the data.

Typically, there are mainly three kinds of big data processing possible: batch processing, stream processing, and hybrid processing [18]. In batch processing, data stored in the non-volatile memory will be processed, and the probability and temporal characteristics of data conversion processes will be decided by the requirements of the problems. In stream processing, the collected data will be processed without storing them in non-volatile media, and the temporal characteristics of data conversion processes will mainly be determined by the incoming data rate. This is suitable for domains that require low response times. Another kind of big data processing, known as hybrid processing, combines both the batch and stream processing techniques to achieve high accuracy and a low processing time [19]. Some examples of hybrid big data processing are Lambda and Kappa Architecture [20]. The Lambda Architecture processes huge quantities of data, enabling the batch-processing and stream-processing methods with a hybrid approach. The Kappa Architecture is a simpler alternative to the Lambda Architecture, since it leverages the same technology stack to handle both real-time stream processing and historical batch processing. However, it avoids maintaining two different code bases for the batch and speed layers. The major notion is to facilitate real-time data processing using a single stream-processing engine, thus bypassing the multi-layered Lambda Architecture without compromising the standard quality of service.

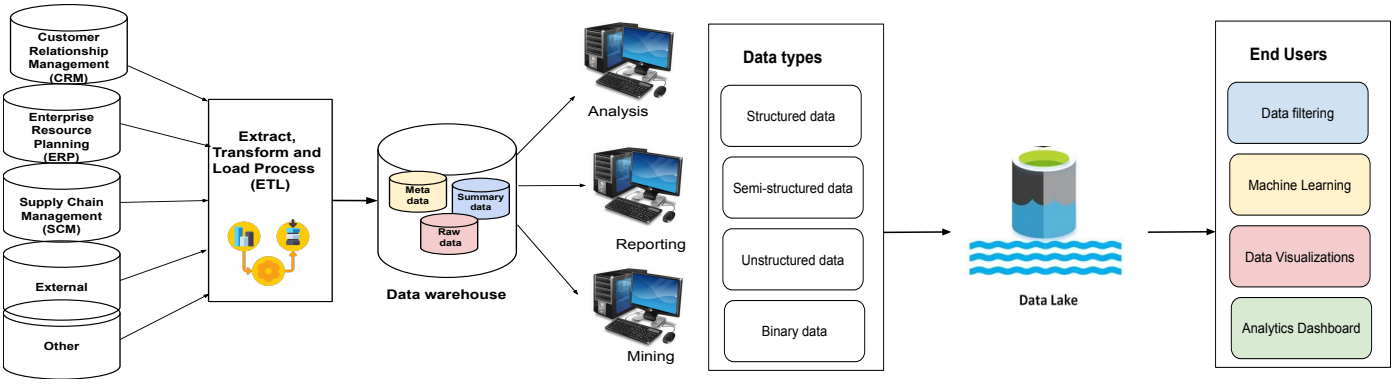
2.2. Data Warehouses

The concept of data warehouses (DWs) was introduced in the late 1980s by IBM researchers Barry Devlin and Paul Murphy with the aim to deliver an architectural model to solve the flow of data to decision support environments [21]. According to the definition by Inmon, "*a data warehouse is a subject-oriented, nonvolatile, integrated, time-variant collection of data in support of management decisions*" [22]. Formally, a data warehouse (DW) is a large data repository wherein data can be stored and integrated from various sources in a well-structured manner and help in the decision-making process via proper data analytics [23]. The process of compiling information into a data warehouse is known as data warehousing.

In enterprise data management, data warehousing is referred to as a set of decision-making systems targeted toward empowering the information specialist (leader, administrator, or analyst) to improve decision making and make decisions quicker. Hence, DW systems act as an important tool of business intelligence, being used in enterprise data management by most medium and large organizations [24,25]. The past decade has seen unprecedented development both in the number of products and services offered and in the wide-scale adoption of these advancements by the industry. According to a comprehensive research report by Market Research Future (MRFR) titled "*Data Warehouse as a Service Market*

information by Usage, by Deployment, by Application and Organization Size—forecast to 2028", the market size will reach USD 7.69 billion, growing at a compound annual growth rate of 24.5%, by 2028 [26].

In the data warehouse framework, data are periodically extracted from programs that aid in business operations and duplicated onto specialized processing units. They may then be approved, converted, reconstructed, and augmented with input from various options. The developed data warehouse then becomes a primary origin of data for the production, analysis, and presentation of reports via instantaneous reports, e-portals, and digital readouts. It employs "online analytical processing" (OLAP), whose utility and execution needs differ from those of the "online transaction processing" (OLTP) implementations typically backed up by functional databases [27,28]. OLTP programs often computerize the handling of administrative data processes, such as order entry and banking transactions, which are an organization's necessary activities. Data warehouses, on the other hand, are primarily concerned with decision assistance. As shown in Figure 1a, a data warehouse integrates data from various sources and helps with analysis, data mining, and reporting. A detailed description of a DW's architecture is presented in Section 3.1.



(**a**) Data warehouse (**b**) Data lake

Figure 1. Data warehouse architecture vs. data lake architecture.

Data warehousing advancements have benefited various sectors, including production (for supply shipment and client assistance), business (for profiling of clients and stock governance), monetary administrations (for claims investigation, risk assessment, billing examination, and detecting fraud), logistics (for vehicle administration), broadcast communications (in order to analyze calls), utility companies (in order to analyze power use), and medical services [29]. The field of data warehousing has seen immense research and developments over the last two decades in various research categories such as data warehouse architecture, data warehouse design, and data warehouse evolution.

2.3. Data Lake

By the beginning of the 21st century, new types of diverse data were emerging in ever-increasing volumes on the Internet and at its interface to the enterprise (e.g., web-based business transactions, real-time streaming, sensor data, and social media). With the huge amount of data around, the need to have better solutions for storing and analyzing large amounts of semi-structured and unstructured data to gain relevant information and valuable insight became apparent. Traditional schema-on-write approaches such as the extract, transform, and load (ETL) process are too inefficient for such data management requirements. This gave rise to another popular modern enterprise data management scheme known as data lakes [30–32].

Data lakes are centralized storage repositories that enable users to store raw, unprocessed data in their original format, including unstructured, semi-structured, or structured data, at scale. These help enterprises to make better business decisions via visualizations or

dashboards from big data analysis, machine learning, and real-time analytics. A pictorial representation of a data lake is given in Figure 1b.

According to Dixon, "*whilst a data warehouse seems to be a bottle of water cleaned and ready for consumption, then "Data Lake" is considered as a whole lake of data in a more natural state*" [33]. Another definition for the data lake is provided in [34], and it is as follows: "*a data lake stores disparate information while ignoring almost everything*". The explanation of data lakes from an architectural viewpoint is given in [35], and it is as follows: "*A data lake uses a flat architecture to store data in its raw format. Each data entity in the lake is associated with a unique, i.e.tifier and a set of extended metadata, and consumers can use purpose-built schemas to query relevant data, which will result in a smaller set of data that can be analyzed to help answer a consumer's question*". A data lake houses data in its original raw form. The data in data lakes can vary drastically in size and structure, and they lack any specific organizational structure. A data lake can accommodate either very small or huge amounts of data as required. All of these features provide flexibility and scalability to data lakes. At the same time, challenges related to its implementation and data analytics also arise.

Data lakes are becoming increasingly popular for organizations to store their data in a centralized manner. A data lake may contain unstructured or multi-structured data, where most of them may have unrealized value for the enterprise. This allows organizations to store their data from different sources without any overhead related to the transformation of the data [30]. This also allows ad hoc data analyses to be performed on this data, which can then be used by organizations to drive key insights and data-driven decision making. DLs replace the previous way of organizing and processing data from various sources with a centralized, efficient, and flexible repository that allows organizations to maximize their gains from a data-driven ecosystem. Data lakes also allow organizations to scale them to their needs. This is achieved by separating storage from the computational part. Complex transformation and preprocessing of data in the case of data warehouses is eliminated. The upfront financial overhead of data ingestion is also reduced. Once data are collated in the lake or hub, it is available for analysis for the organization.

2.4. The Difference between Data Warehouses and Data Lakes

Although data warehouses and data lakes are used as two interchangeable terms, they are not the same [21]. One of the major differences between them is the different structures (i.e., processed vs. raw data). A data warehouse stores data in processed and filtered form, whereas data lakes store raw or unprocessed data. Specifically, data are processed and organized into a single schema before being put into the warehouse, whereas raw and unstructured data are fed into a data lake. Analysis is performed on the cleansed data in the warehouse. On the contrary, in a data lake, data are selected and organized as and when needed.

As for storing processed data, a data warehouse is economic. On the contrary, data lakes have a comparatively larger capacity than the data warehouse and are ideal for raw and unprocessed data analysis and employing machine learning. Another key difference is the objective or purpose of use. Typically, processed data that flow into data warehouses are used for specific purposes, and hence the storage space will not be wasted, whereas the purpose of usage for the data lake is not defined and can ideally be used for any purpose. To use processed or filtered data, no specialized expertise is required, as merely familiarization with the presentation of data (e.g., charts, sheets, tables, and presentations) will do. Hence, DWs can be used by any business or individual. On the contrary, it is comparatively difficult to analyze DLs without familiarity with unprocessed data, hence requiring data scientists with appropriate skills or tools to comprehend them for specific business use. Accessibility or ease of use of data repositories is yet another aspect that differentiates data warehouses and data lakes. Since the architecture of a data lake has no proper structure, it has flexibility of use. Instead, the structure of a DW makes sure that no foreign particles invade it, and it is very costly to manipulate. This feature makes it very

secure, too. A detailed analysis of the differences between data warehouses and data lakes is given in Table 1.

Table 1. Differences between data warehouses and data lakes.

Parameters	Data Warehouse	Data Lake
Data	Data warehouse focuses only on business processes	Data lakes store everything
Processing	Highly processed data	Data are mainly unprocessed
Type of Data	They are mostly in the tabular form and structure	They can be unstructured, semi-structured, or structured
Task	Optimized for data retrieval	Share data stewardship
Agility	Less agile and has fixed configuration compared with data lakes	Highly agile and can configure and reconfigure as needed
Users	Widely used by business professionals and business analysts	Data lakes are used by data scientists, data developers, and business analysts
Storage	Expensive storage that gives fast response times is used	Data lakes are designed for low-cost storage
Security	Allows better control of the data	Offers less control
Schema	Schema on writing (predefined schemas)	Schema on reading (no predefined schemas)
Data Processing	Time-consuming to introduce new content	Helps with fast ingestion of new data
Data Granularity	Data at the summary or aggregated level of detail	Data at a low level of detail or granularity
Tools	Mostly commercial tools	Can use open-source tools such as Hadoop or Map Reduce

2.5. Literature Review

A summary of various research works in the field of data warehouses and data lakes is presented here. A list of various survey articles on data warehouses and data lakes is depicted in Table 2. Mainly, data warehouse review works address architecture modeling and its comparisons [36,37], the evolution of the DW concept [38], real-time data warehousing and ETL [39], etc. Compared with the data warehouse literature reviews, data lake papers are relatively fewer in number. Data lake review works summarize recent approaches and the architecture of DLs [31,32] as well as the design and implementation aspects [30]. To the best of our knowledge, only one work on comparing data warehouses and data lakes was found in the literature [12]. In contrast to that article, our work provides a comprehensive analysis of both data management schemes by addressing various aspects such as, definitions, architecture, practical design considerations, tools and services, challenges, and opportunities in detail. In addition to the survey papers, we also consolidate various works on data warehouses and data lakes in the reported literature and classify them in Table 3 based on their functions and utility.

Table 2. Summary of existing survey articles on data warehouses and data lakes.

Topic	Survey Papers	Contributions
Data warehouse	[28]	Data warehouse concepts, multilingualism issues in data warehouse design and solutions
Data warehouse	[36]	Data warehouse architecture modeling and classifications
Data warehouse and big data	[40]	A comprehensive survey on big data, big data analytics, augmentation, and big data warehouses
Data warehouse	[11]	Data warehouse survey
Data warehouse	[39]	Real-time data warehouse and ETL
Data warehouse	[41]	Architectures of data warehouses (DWs) and their selection

Table 2. *Cont.*

Topic	Survey Papers	Contributions
Data warehouse	[38]	Data warehouse (DW) evolution
Data warehouse	[42]	Data warehouse modeling and design
Data warehouse	[37]	Comparative study on data warehouse architectures
Data lake	[30]	A survey on designing, implementing, and applying data lakes
Data lake	[31]	Recent approaches and architectures using data lakes
Data lake	[32]	Overview of data lake definitions, architectures, and technologies
Data lake vs. data warehouse	[12]	Explores the two architectures of data warehouses and data lakes

Table 3. Related works: classification of data warehouse and data lake solutions.

Systems or Topic Area	Data Warehouse	Data Lake	Function or Work Performed	Reference
OLAP	✓		Online analytical processing (OLAP)	Providing OLAP to User-Analysts: an IT Mandate [28]
GEMMS		✓	Metadata extraction, Metadata modeling	Metadata Extraction and Management in Data Lakes with GEMMS [30]
KAYAK		✓	Dataset preparation and organization	KAYAK: a Framework for Just-in-Time Data Preparation in a Data Lake [43]
DWHA	✓		Modeling and classification of DW	Analysis of Data Warehouse Architectures: Modelling and Classification [36]
DATAMARAN		✓	Metadata extraction	Navigating the Data Lake with DATAMARAN: Automatically Extracting Structure from Log Datasets [44]
Geokettle	✓		Data warehouse architecture, design, and testing	Extraction, Transformation, and Loading (ETL) Module for Hotspot Spatial Data Warehouse Using Geokettle [45]
GOODS		✓	Dataset preparation and organization, metadata enrichment	Managing Google's data lake: an overview of the Goods system [46]
VOLAP	✓		OLAP, query processing, and optimization	VOLAP: a Scalable Distributed System for Real-Time OLAP with High-Velocity Data [47]
Dimension constraints	✓		Multidimensional data modeling, OLAP, query processing, and optimization	Capturing summarizability with integrity constraints in OLAP [48]
CLAMS		✓	Data quality improvement	CLAMS: Bringing Quality to Data Lakes [49]
Juneau		✓	Dataset preparation and organization, discover related data sets, and query-driven data discovery	Juneau: Data Lake Management for Jupyter [50]
JOSIE		✓	Discover related data sets and query-driven data discovery	Josie: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes [51]
CoreDB		✓	Metadata enrichment and query heterogeneous data	CoreDB: a Data Lake Service [52]
Constance		✓	Unified interface for query processing and data exploration	Constance: An Intelligent Data Lake System [53]
ODS	✓		Operational data store	Combining the Data Warehouse and Operational Data Store [54]

3. Architecture

In this section, the architectures of the data warehouse and data lake schemes are described in detail. Furthermore, the classification of data warehouse and data lake solutions based on function is carried out and summarized as a table.

3.1. Data Warehouse Architecture

The data warehouse architecture contains historical and commutative data from multiple sources. Basically, there are three kinds of architectures [55]:

- *Single-tier architecture*: This kind of single-layer model minimizes the amount of data stored. It helps remove data redundancy. However, its disadvantage is the lack of a component that separates analytical and transactional processing. This kind of architecture is not frequently used in practice.
- *Two-tier architecture*: This model separates physically available sources and the data warehouse by means of a staging area. Such an architecture makes sure that all data loaded into the warehouse are in an appropriate cleansed format. Nevertheless, this architecture is not expandable nor can it support many end users. Additionally, it has connectivity problems due to network limitations.
- *Three-tier architecture*: This is the most widely used architecture for data warehouses [56,57]. It consists of a top, middle, and bottom tier. In the bottom tier, data are cleansed, transformed, and loaded via backend tools. This tier serves as the database of the data warehouse. The middle tier is an OLAP server that presents an abstract view of the database by acting as a mediator between the end user and the database. The top tier, the front-end client layer, consists of the tools and an API that are used to connect and get data out from the data warehouse (e.g., query tools, reporting tools, managed query tools, analysis tools, and data mining tools).

The architecture of a data warehouse is shown in Figure 2. It consists of a central information repository that is surrounded by some key DW components, making the entire environment functional, manageable, and accessible.

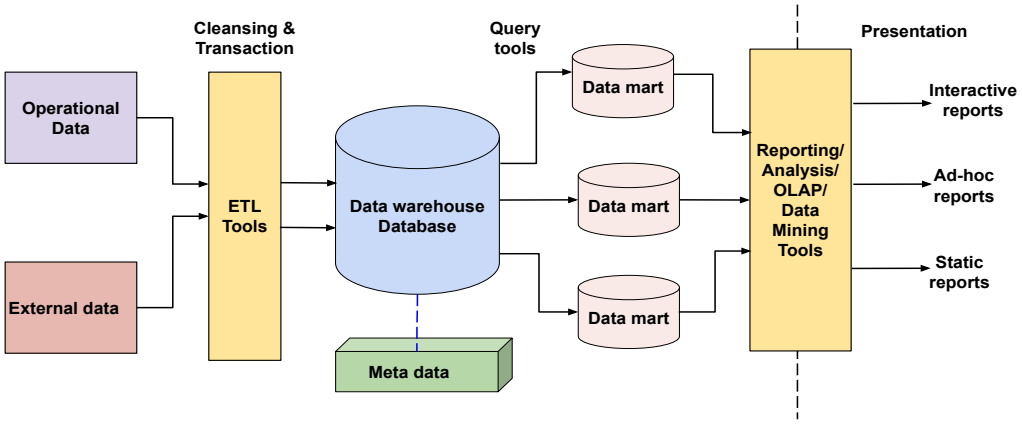


Figure 2. Data warehouse architecture.

- **Data warehouse database**: The core foundation of the data warehouse environment is its central database. This is implemented using RDBMS technology [58]. However, there is a limitation to such implementations, since the traditional RDBMS system is optimized for transactional database processing and not for data warehousing. In this regard, the alternative means are (1) the usage of relational databases in parallel, which enables shared memory on various multiprocessor configurations or parallel processors, (2) new index structures to get rid of relational table scanning and improve the speed, and (3) multidimensional databases (MDDBs) used to circumvent the limitations caused by the relational data warehouse models.
- **Extract, transform, and load (ETL) tools**: All the conversions, summarizations, and changes required to transform data into a unified format in the data warehouse are

carried out via extract, transform, and load (ETL) tools [59]. This ETL process helps the data warehouse achieve enhanced system performance and business intelligence, timely access to data, and a high return on investment:

- *Extraction*: This involves connecting systems and collecting the data needed for analytical processing;
- *Transformation*: The extracted data are converted into a standard format;
- *Loading*: The transformed data are imported into a large data warehouse.

ETL anonymizes data as per regulatory stipulations, thereby anonymizing confidential and sensitive information before loading it into the target data store [60]. ETL eliminates unwanted data in operational databases from loading into DWs. ETL tools carry out amendments to the data arriving from different sources and calculate summaries and derived data. Such ETL tools generate background jobs, Cobol programs, shell scripts, etc. that regularly update the data in the data warehouse. ETL tools also help with maintaining the metadata.

- **Metadata**: Metadata is the data about the data that define the data warehouse [61]. It deals with some high-level technological concepts and helps with building, maintaining, and managing the data warehouse. Metadata plays an important role in transforming data into knowledge, since it defines the source, usage, values, and features of the data warehouse and how to update and process the data in a data warehouse. This is the most difficult tool to choose due to the lack of a clear standard. Efforts are being made among data warehousing tool vendors to unify a metadata model. One category of metadata known as *technical metadata* contains information about the warehouse that is used by its designers and administrators, whereas another category called *business metadata* contains details that enable end users to understand the information stored in the data warehouse.
- **Query Tools**: Query tools allow users to interact with the DW system and collect information relevant to businesses to make strategic decisions. Such tools can be of different types:
  - **Query and reporting tools**: Such tools help organizations generate regular operational reports and support high-volume batch jobs such as printing and calculating. Some popular reporting tools are Brio, Oracle, Powersoft, and SAS Institute. Similarly, query tools help end users to resolve pitfalls in SQL and database structure by inserting a meta-layer between the users and the database.
  - **Application development tools**: In addition to the built-in graphical and analytical tools, application development tools are leveraged to satisfy the analytical needs of an organization.
  - **Data mining tools**: This tool helps in automating the process of discovering meaningful new correlations and structures by mining large amounts of data.
  - **OLAP tools**: Online analytical processing (OLAP) tools exploit the concepts of a multidimensional database and help analyze the data using complex multidimensional views [28,62]. There are two types of OLAP tools: multidimensional OLAP (MOLAP) and relational OLAP (ROLAP) [63]:
    - **MOLAP**: In such an OLAP tool, a cube is aggregated from the relational data source. Based on the user report request, the MOLAP tool generates a prompt result, since all the data are already pre-aggregated within the cube [64].
    - **ROLAP**: The ROLAP engine acts as a smart SQL generator. It comes with a "designer" piece, wherein the administrator specifies the association between the relational tables, attributes, and hierarchy map and the underlying database tables [65].

3.2. Data Lake Architecture

The architecture of a business data lake is depicted in Figure 3. Although it is treated as a single repository, it can be distinguished as separate layers in most cases.

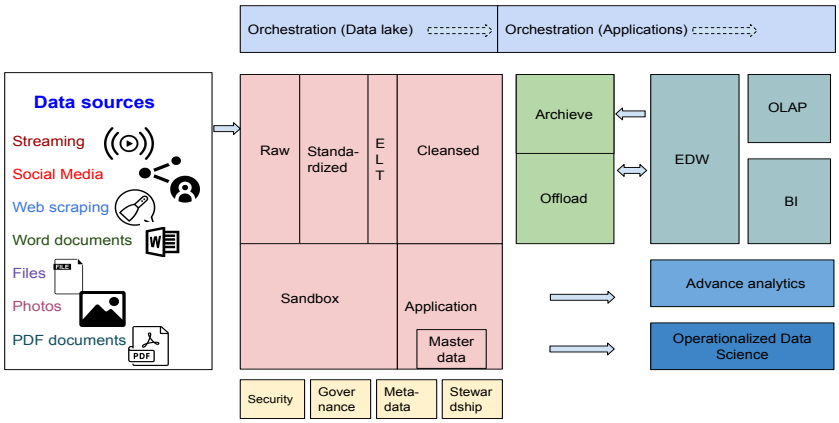


Figure 3. Data lake building blocks.

- **Raw data layer**: This layer is also known as the ingestion layer or landing area because it acts as the sink of the data lake. The prime goal is to ingest raw data as quickly and as efficiently as possible. No transformations are allowed at this stage. With the help of the archive, it is possible to get back to a point in time with raw data. Overriding (i.e., handling duplicate versions of the same data) is not permitted. End users are not granted access to this layer. These are not ready-to-use data, and they need a lot of knowledge in terms of relevant consumption.
- **Standardized data layer**: This is optional in most implementations. If one expects fast growth for his or her data lake architecture, then this is a good option. The prime objective of the standardized layer is to boost the performance of the data transfer from the raw layer to the curated layer. In the raw layer, data are stored in their native format, whereas in the standardized layer, the appropriate format that fits best for cleansing is selected.
- **Cleansed layer or curated layer**: In this layer, data are transformed into consumable data sets and stored in files or tables. This is one of the most complex parts of the whole data lake solution since it requires cleansing, transformation, denormalization, and consolidation of different objects. Furthermore, the data are organized by purpose, type, and file structure. Usually, end users are granted access only to this layer.
- **Application layer**: This is also known as the trusted layer, secure layer, or production layer. This is sourced from the cleansed layer and enforced with requisite business logic. In case the applications use machine learning models on the data lake, they are obtained from here. The structure of the data is the same as in the cleansed layer.
- **Sandbox data layer**: This is also another optional layer that is meant for analysts' and data scientists' work to carry out experiments and search for patterns or correlations. The sandbox data layer is the proper place to enrich the data with any source from the Internet.
- **Security**: While data lakes are not exposed to a broad audience, the security aspects are of great importance, especially during the initial phase and architecture. These are not like relational databases, which have an artillery of security mechanisms.
- **Governance**: Monitoring and logging operations become crucial at some point while performing analysis.
- **Metadata**: This is the data about data. Most of the schemas reload additional details of the purpose of data, with descriptions on how they are meant to be exploited.
- **Stewardship**: Based on the scale that is required, either the creation of a separate role or delegation of this responsibility to the users will be carried out, possibly through some metadata solutions.

- **Master Data**: This is an essential part of serving ready-to-use data. It can be either stored on the data lake or referenced while executing ELT processes.
- **Archive**: Data lakes keep some archive data that come from data warehousing. Otherwise, performance and storage-related problems may occur.
- **Offload**: This area helps to offload some time- and resource-consuming ETL processes to a data lake in case of relational data warehousing solutions.
- **Orchestration and ELT processes**: Once the data are pushed from the raw layer through the cleansed layer and to the sandbox and application layers, a tool is required to orchestrate the flow. Either an orchestration tool or some additional resources to execute them are leveraged in this regard.

Many implementations of a data lakes are originally based on Apache Hadoop. The Highly Available Object Oriented Data Platform (Hadoop) is a widely popular big data tool especially suitable for batch processing workloads of big data [66]. It uses HDFS as its core storage and MapReduce (MR) as the basic computing model. Novel computing models are constantly proposed to cope with the increasing needs for batch processing performance (e.g., Tez, Spark, and Presto) [67,68]. The MR model has also been replaced with the directed acyclic graph (DAG) model, which improves computing models' abstract concurrency. The second phase of data lake evolution has happened with the arrival of the *Lambda Architecture* [69,70], owing to the constant changes in data processing capabilities and processing demand. It presents stream computing engines, such as Storm, Spark Streaming, and Flink [71]. In such a framework, batch processing is combined with stream computing to meet the needs of many emerging applications. Yet another advanced phase is for the *Kappa Architecture* [72]. The two models of batch processing and stream computing are unified by improving the stream computing concurrency and increasing the time window of streaming data. In this regard, stream computing is used that features an inherent and scalable distributed architecture.

4. Design Aspects

The design aspects and practical implementation constraints are to be studied in detail to develop a suitable data management solution. This section presents the design aspects to be considered in data warehouse- and data lake-based enterprise data management.

4.1. Data Warehouse Design Considerations for Business Needs

To design a successful data warehouse, one should also realize the requirements of an organization and develop a framework for them. Some of the key criteria to keep in mind when choosing a data warehouse are as follows:

- **User needs and appropriate data model**: The very first design consideration in a data warehouse is the business and user needs. Hence, during the designing phase, the integration of the data warehouse with existing business processes and compatibility checks with long-term strategies have to be ensured. Enterprises have to clearly comprehend the purpose of their data warehouse, any technical requirements, benefits of end users from the system, improved means of reporting for business intelligence (BI), and analytics. In this regard, finding the notion of what information is important to the business is quintessential to the success of the data warehouse. To facilitate this, creating an appropriate data model of the business is a key aspect when designing DWs (e.g., SQL Developer Data Modeler (SDDM)). Furthermore, a data flow diagram can also help in depicting the data flow within the company in diagram format.
- **Adopting a standard data warehouse architecture and methodology**: While designing a DW, yet another important practical consideration is to leverage a recognized DW modeling standard (e.g., 3NF, star schema (dimensional), and Data Vault) [73]. Selecting such a standard architecture and sticking to the same one can augment the efficiency within a data warehouse development approach. Similarly, an agile data warehouse methodology is also an important practical aspect. With proper planning,

DW projects can be compartmentalized to smaller pieces capable of delivering faster. This design trick helps to prioritize the DW as a business's needs change.

- **Cloud vs. on-premise storage**: Enterprises can opt for either on-premises architecture or a cloud data warehouse [13]. The former category requires setting up the physical environment, including all the servers necessary to power ETL processes, storage, and analytic operations, whereas the latter can skip this step. However, a few circumstances exist where it still makes sense to consider an on-premises approach. For example, if most of the critical databases are on-premises and are old enough, they will not work well with cloud-based data warehouses. Furthermore, if the organization has to deal with strict regulatory requirements, which might include no offshore data storage, an on-premise setting might be the better choice. Nevertheless, cloud-based services provide the most flexible data warehousing service in the market in terms of storage and the pay-as-you-go nature.
- **Data tool ecosystem and data modeling**: The organization's ecosystem plays a key role. Adopting a DW automation tool ensures the efficient usage of IT resources, faster implementation through projects, and better support by enforcing coding standards (Wherescape (https://www.wherescape.com, accessed on 25 September 2022), AnalytixDS, Ajilius (https://tracxn.com/d/companies/ajilius.com, accessed on 25 September 2022), etc.).The data modeling planning step imparts detailed, reusable documentation of a data warehouse's implementation. Specifically, it assesses the data structures, investigates how to efficiently represent these sources in the data warehouse, specifies OLAP requirements, etc.
- **ETL or ELT design**: Selection of the appropriate ETL or ELT solution is yet another design concern [39]. When businesses use expensive in-house analytics systems, much prep work including transformations can be conducted, as in the ETL scheme. However, ELT is a better approach when the destination is a cloud data warehouse. Once data are colocated, the power of a single cloud engine can be leveraged to perform integrations and transformations efficiently. Organizations can transform their raw data at any time according to their use case, rather than a step in the data pipeline.
- **Semantic and reporting layers**: Based on previously documented data models, the OLAP server is implemented to facilitate the analytical queries of the users and to empower BI systems. In this regard, data engineers should carefully consider time-to-analysis and latency requirements to assess the analytical processing capabilities of the data warehouse. Similarly, while designing the reporting layer, the implementation of reporting interfaces or delivery methods as well as permissible access have to be set by the administrator.
- **Ease of scalability**: Understanding current business needs is critical to business intelligence and decision making. This includes how much data the organization currently has and how quickly its needs are likely to grow. Staffing and vendor costs need to be taken into consideration while deciding the scale of growth.

4.2. Data Lake Design Aspects for Enterprise Data Management

At a high level, the concept of a data lake seems to be simple. Irrespective of the format, it stores data from multiple sources in one place, leverages big data technologies, and deploys on a commodity infrastructure. However, many a time, reality may fail due to various practical constraints. Hence, it is quite important to consider several key criteria while designing an enterprise data lake:

- **Focus on business objectives rather than technology**: By anchoring the business objectives, a data lake can prioritize the efforts and outcomes accordingly. For instance, for a particular business objective, there may be some data that are more valuable than others. This kind of comprehension and analysis is the key to an enterprise's data lake success. With such an oriented goal, data lakes can start small and then accordingly learn, adapt, and produce accelerated outcomes for a business. In particular, some

key factors in this regard are (1) whether it solves an actual business problem, (2) if it imparts new capabilities, and (3) the access or ownership of data, among others.

- **Scalability and durability** are two more major criteria [74]. Scalability enables scaling to any size of data while importing them in real time. This is an essential criterion for a data lake since it is a centralized data repository for an entire organization. Another important aspect (i.e., durability) deals with providing consistent uptime while ensuring no loss or corruption of data.
- Another key design aspect in a data lake is its **capability to store unstructured, semi-structured, and structured data**, which helps organizations to transfer anything from raw, unprocessed data to fully aggregated analytical outcomes [75]. In particular, the data lake has to deliver business-ready data. Practically speaking, data by themselves have no meaning. Although file formats and schemas can parse the data (e.g., JSON and XML), they fail at delivering insight into their meaning. To circumvent such a limitation, a critical component of any data lake technical design is the incorporation of a knowledge catalog. Such a catalog helps in finding and understanding information assets. The knowledge catalog's contents include the semantic meaning of the data, format and ownership of data, and data policies, among other elements.
- **Security** considerations are also of prime importance in a data lake in the cloud. The three domains of security are encryption, network-level security, and access control. Network-level security imparts a robust defense strategy by denying inappropriate access at the network level, whereas encryption ensures security at least for those types of data that are not publicly available. Security should be part of data lake design from the beginning. Compliance standards that regulate data protection and privacy are incorporated in many industries, such as the Payment Card Industry Data Security Standard (PCI DSS) for financial services and Health Insurance Portability and Accountability Act (HIPAA) for healthcare [76]. Furthermore, two of the biggest regulations regarding consumer privacy (i.e., California's Consumer Privacy Act (CCPA) and the European Union's General Data Protection Regulation (GDPR)) restrict the ownership, use, and management of personal and private data.
- A data lake design must include **metadata storage functionality** to help users to search and learn about the data sets in the lake [77]. A data lake allows the storage of all data that are **independent of the fixed schema**. Instead, data are read at the time of processing, should they be parsed and adapted into a schema, only as necessary. This feature saves plenty of time for enterprises.
- **Architecture in motion** is another interesting concept (i.e., the architecture will likely include more than one data lake and must be adaptable to address changing requirements). For instance, on-premises work with Hadoop could be moved to the cloud or a hybrid platform in the future. By facilitating the innovation of multi-cloud storage, a data lake can be easily upgraded to be used across data centers, on premises, and in private clouds. In addition, machine learning and automation can augment the data flow capabilities of an enterprise's data lake design.

5. Tools and Utilities

In this section, we categorize and detail the popular data warehouse and data lake tools and services in Sections 5.1 and 5.2, respectively.

5.1. Popular Data Warehouse Tools and Services

An enterprise data warehouse is one of the primary components of business intelligence [14,16]. It stores data from one or more heterogeneous sources and then analyzes and extracts insights from them to support decision making. Some of the popular top data warehousing tools are explained below:

- **Amazon Web Services (AWS) data warehouse tools**: AWS is one of the major leaders in data warehousing solutions [78] (https://aws.amazon.com/training/classroom/data-warehousing-on-aws/, accessed on 25 September 2022). AWS has many services,

such as AWS Redshift, AWS S3, and Amazon RDS, making it a very cost-effective and highly scalable platform. **AWS Redshift** is a suitable platform for businesses that require very advanced capabilities that exploit high-end tools [79]. It consists of an in-house team that organizes AWS's extensive menu of services. **Amazon Simple Storage Service (AWS S3)** is a low-cost storage solution with industry-leading scalability, performance, and security features. **Amazon Relational Database Service (Amazon RDS)** is an AWS cloud data storage service that runs and scales a relational database. It has resizable and cost-effective technology that facilitates an industry-standard relational database and manages all database management activities.

- **Google data warehouse tools**: Google is highly acclaimed for its data management skills along with its dominance as a search engine (https://cloud.google.com, accessed on 25 September 2022). Google's data warehouse tools (https://research.google/research-areas/data-management/, accessed on 25 September 2022) excel in cutting-edge data management and analytics by incorporating machine intelligence. **Google BigQuery** is a business-level cloud-based data warehousing solution platform specially designed to save time by storing and querying large data sets through using super-fast SQL searches against multi-terabyte data sets in seconds, offering customers real-time data insights. **Google Cloud Data Fusion** is a cloud ETL solution which is entirely managed and allows data integration at any size with a visual point-and-click interface. **Dataflow** is another cloud-based data-processing service that can be used to stream data in batches or in real time. **Google Data Studio** enables turning the data into entirely customizable, easy-to-read reports and dashboards.
- **Microsoft Azure Data Warehouse tools**: Microsoft Azure is a recent cloud computing platform that provides Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) as well as 200+ products and cloud services [80] (https://azure.microsoft.com/en-in/, accessed on 25 September 2022). **Azure SQL Database** is suitable for data warehousing applications with up to 8 TB of data volume and a large number of active users, facilitating advanced query processing. **Azure Synapse Analytics** consists of data integration, big data analytics, and enterprise data warehousing capabilities by also integrating machine learning technologies.
- **Oracle Autonomous Data Warehouse**: Oracle Autonomous Data Warehouse [81] is a cloud-based data warehouse service that manages the complexities associated with data warehouse development, data protection, data application development, etc. The setting, safeguarding, regulating, and backing up of data are all automated using this technology. This cloud computing solution is easy to use, secure, quick to respond, as well as scalable.
- **Snowflake**: Snowflake [82] is a cloud-based data warehouse tool offering a quick, easy-to-use, and adaptable data warehouse platform (https://www.snowflake.com, accessed on 25 September 2022). It has a comprehensive Software as a Service (SaaS) architecture since it runs entirely in the cloud. This makes data processing easier by permitting users to work with a single language, SQL for data blending, analysis, and transformations on a variety of data types. Snowflake's multi-tenant design enables real-time data exchange throughout the enterprise without relocating data.
- **IBM Data Warehouse tools**: IBM is a preferred choice for large business clients due to its huge install base, vertical data models, various data management solutions, and real-time analytics (https://www.ibm.com/in-en/analytics, accessed on 25 September 2022). One DW tool (i.e., **IBM DB2 Warehouse**) is a cloud DW that enables self-scaling data storage and processing and deployment flexibility. Another tool is **IBM Datastage**, which can take data from a source system, transform it, and feed it into a target system. This enables the users to merge data from several corporate systems using either an on-premises or cloud-based parallel architecture.

5.2. Popular Data Lake Tools and Services

A data lake stores structured data from relational databases, where semi-structured data, unstructured data, and binary data and can be set up "on the premises" or in the "cloud" [83,84]. Some of the most popular data lake tools and services are analyzed below:

- **Azure Data Lake**: Azure Data Lake makes it easy for developers and data scientists to store data of any size, shape, and speed and conduct all types of processing and analytics across platforms and languages (https://azure.microsoft.com/en-in/solutions/data-lake/, accessed on 25 September 2022). It removes the complexities associated with ingesting and storing the data and makes it faster to bring up and execute with batch, streaming, and interactive analytics [85]. Some of the key features of Azure Data Lake include unlimited scale and data durability, on-par performance even with demanding workloads, high security with flexible mechanisms, and cost optimization through independent scaling of storage.
- **AWS**: Amazon Web Services claims to provide "the most secure, scalable, comprehensive, and cost-effective portfolio of services for customers to build their data lake in the cloud" (https://aws.amazon.com/lake-formation/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc, accessed on 25 September 2022). AWS Lake Formation helps to set up a secure data lake that can collect and catalog data from databases and object storage, move the data into the new Amazon Simple Storage Service (S3) data lake, and clean and classify the data using ML algorithms. It offers various aspects of scalability, agility, and flexibility that are required by the companies to fuse data and analytics approaches. AWS customers include NETFLIX, Zillow, NASDAQ, Yelp, and iRobot.
- **Google BigLake**: BigLake is a storage engine that unifies data warehouses and lakes (https://cloud.google.com/biglake, accessed on 25 September 2022). It removes the need to duplicate or move data, thus making the system efficient and cost-effective. BigLake provides detailed access controls and performance acceleration across BigQuery and multi-cloud data lakes, with open formats to ensure a unified, flexible, and cost-effective lakehouse architecture. The top features of BigLake include (1) users being able to enforce consistent access controls across most analytics engines with a single copy of data and (2) unified governance and management at scale. Users can extend BigQuery to multi-cloud data lakes and open formats with fine-grained security controls without setting up a new infrastructure.
- **Cloudera**: Cloudera SDX is a data lake service for creating safe, secure, and governed data lakes with protective rings around the data wherever they stored, from object stores to the Hadoop Distributed File System (HDFS) (https://www.cloudera.com, accessed on 25 September 2022). It provides the capabilities needed for (1) data schema and metadata information, (2) metadata governance and management, (3) data access authorization and authentication, and (4) compliance-ready access auditing.
- **Snowflake**: Snowflake's cross-cloud platform breaks down silos and enables a data lake strategy (https://www.snowflake.com/workloads/data-lake/, accessed on 25 September 2022). Data scientists, analysts, and developers can seamlessly leverage governed data self-service for a variety of workloads. The key features of Snowflake include (1) all data on one platform that combines structured, semi-structured, and unstructured data of any format across clouds and regions, (2) fast, reliable processing and querying, simplifying the architecture with an elastic engine to power many workloads, and (3) secure collaboration via easy integration of external data without ETL.

6. Challenges

This section addresses some of the key challenges in big data analytics problems. In addition, the implementation challenges encountered in data warehouses and data lake paradigms are also critically analyzed.

6.1. Challenges in Big Data Analytics

In the past few years, big data have been accumulated in every walk of human life, including healthcare, retail, public administration, and research. Web-based applications have to deal with big data frequently, such as Internet text and documents (corpus, etc.), social network analysis, prediction markets, and Internet search indexing [86]. Although we can clearly observe the potential and current advantages of big data, there are some inherent challenges also present that have to be tackled to achieve the full potential of big data analytics [87].

The first hurdle for big data analytics is the **storage mediums and higher I/O speed** [88]. Storage of big data causes a financial overhead which is not affordable or profitable for many enterprises. Furthermore, this also results in slower processes [89]. In decades gone by, analysts made use of hard disk drives for data storage purposes, but this is slower in terms of random I/O performance compared with sequential I/O. To overcome this limitation, the concept of solid-state drives (SSDs) and phase change memory were introduced. However, the currently available storage tech simply does not possess the required performance for processing big data and delivering insights in a timely fashion. Companies opt for various modern techniques to handle large data sets, such as compression (reducing the number of bits within the data), data tiering (storing data in several storage tiers), and deduplication (the process of removing duplicates and unwanted data).

Anther challenge is **the lack of proper understanding of big data and the lack of knowledge professionals**. Due to insufficient understanding, organizations may fail in big data initiatives. This may be due to the absence of skilled data professionals, the lack of a transparent picture for employees, or improper usage of data repositories, among other reasons. It is highly encouraged to conduct big data workshops and seminars at companies to enable every level of the organization to inculcate a basic understanding of knowledge concepts. Furthermore, companies should invest in recruiting skilled professionals, supplying training programs to the staff, as well as purchasing knowledge analytics solutions powered by advanced artificial intelligence or machine learning tools.

Yet another challenge in big data analytics is the **confusion with suitable tool selection**. For instance, many a time, it is not so clear whether Hadoop or Spark is a better option for data analytics and storage. Sometimes, the wrong selection may result in poor decisions and the selection of inappropriate technology. Hence, money, time, effort, and work hours are wasted. The best solution would be to make use of experienced professionals or data consulting to obtain a recommendation for the tools that can support a company based on its scenario.

Data in a corporation come from various sources, such as customer logs, financial reports, social media platforms, e-mails, and reports created by employees. **Integrating data from such a huge spread of sources** is another challenging task [90]. This consolidation task, known as data integration, is crucial for business intelligence. Hence, enterprises purchase proper tools for data integration purposes. Talend Data Integration, IBM InfoSphere Xplenty, Informatica PowerCenter, and Microsoft SQL QlikView are some of the popular data integration tools [91].

Security of huge sets of knowledge, especially ones that involve many confidential details of customers, is one of the, inevitable challenges in big data analytics [92,93]. The careless treatment of data repositories may invite malicious hackers, which can cost millions for a stolen record or a knowledge breach. The remedy would be to foster a cybersecurity division of a company to guard their data and to implement various security actions such as data encryption, data segregation, identity and access control, implementation of endpoint security, real-time security monitoring, and using big data security tools (e.g., IBM Guardian).

6.2. Data Warehouse Implementation Challenges

Implementation of a data warehouse requires proper planning and execution based on proper methods. Some of the major challenging considerations that arise with data warehousing are design, construction, and implementation [94,95].

The efficiency and working of a warehouse are **dependent on the data** that support its operations. With incorrect or redundant data, warehouse managers cannot accurately measure the exact costs. A key solution is to automate the system to improve the lead data quality and make sure that the sales team receives complete, correct, and consistent lead information. Another major concern in a data warehouse is the **quality control of data (i.e., quality and consistency of data)** [96]. The business intelligence process can be fine-tuned by incorporating flexibility to accept and integrate analytics as well as update the warehouse's schema to handle evolutions.

Another major challenge is **differences in naming, domain definitions, and identification numbers from heterogeneous sources**. The data warehouse has to be designed in such a way that it can accommodate the addition and attrition of data sources and the evolution of the sources and source data, thus avoiding major redesign. Yet another challenge is **customizing the available source data into the data model of the warehouse** because the capabilities of a DW may change over time based on the change in technology [97]. Further, **broader skills** are required for the administration of data warehouses in traditional database administration. Hence, managing the data warehouse in a large organization, the design of the management function, and selecting the management team for a database warehouse are some of the important aspects of a data warehouse.

Data security is another critical requirement in DWs, given that business data are extremely sensitive and can be easily obtained [98]. Unfortunately, the typical security paradigm—based on tables, lines, and characteristics—is incompatible with DWs. Following that, the model should be changed to one that is firmly integrated with the applicable model and is focused on the key notions of multidimensional display, such as facts, aspects, and measures. Furthermore, as is frequently advised in computer programming, **information security** should be considered at all stages of the improvement process, from prerequisite analysis to execution and upkeep. In addition, **data warehouse governance** is yet another important consideration, which includes approval of the data modeling standards and metadata standards, the design of a data access policy, and a data backup strategy [99].

6.3. Data Lake Implementation Challenges

The data lake is relatively novel technology and has not matured yet. Hence, there are many challenges in its implementation, including many of the same challenges that early data warehouses confronted [75,100]. The first challenge is the **high cost of data lakes**. They are expensive to implement and maintain. Data lake platforms that exploit the cloud may be easier to deploy, but they may also come with high fees. Some of the platforms such as Hadoop are open source and hence free of cost. Nevertheless, the implementation and management may take more time and more expert staff. **Management difficulty** is another issue [75]. The management of the DL involves various complex tasks, such as ensuring the capacity of the host infrastructure to cope with the growth of the DL and dealing with data redundancy and data security. This puts forth challenges even to skilled engineers. Furthermore, it is required to have more domain experts and engineers with real expertise in setting up and managing data lakes. In the current scenario, there is a shortage of both data scientists and data engineers in the field. This **lack of skills** is yet another challenge.

Another aspect for consideration is the **long time to value** (i.e., it takes years to become full-fledged and to be integrated well with the workflow and analytics tools to impart real value to the enterprise) [101]. As mentioned in the case of data warehouses, in the case of DLs, **data security** is also a major concern. It requires special security measures to be considered to enforce data governance rules and to secure the data in the DL with the help of cyber security specialists and security tools. Another critical challenge is the **computation resources and increase in computing power**. This is due to the fact that data are growing unprecedentedly faster than computing power. At the same rate, the existing computers are not well equipped to host and manage them at the same rate due to a lack of power. Similarly, open-source data platforms also find many core problems surrounding data lakes which are too costly to manage. This also requires massive computing power to overcome such serious skill gaps.

To build a better data lake, it is required to modernize the way businesses build and manage data lakes. One key takeaway is to take full **advantage of the cloud**, as opposed to building cumbersome data lakes on a tailor-made infrastructure [102]. It helps to get rid of data silos and to build data lakes that are applicable to various use cases, rather than only fitting them to a certain range of needs.

7. Opportunities and Future Directions

Based on our survey, we discuss novel trends in modern enterprise data management and point out some promising directions for future research in this section.

7.1. Data Warehouses: Opportunities and Future Directions

The business management landscape has witnessed a massive change with the emergence of the data warehouse. The **advancements in cloud technology, the Internet of Things, and big data analytics** have brought effective data solutions in modern data warehouses [77,103]. With the rapid evolution of technology, many enterprises have migrated their data to the cloud to expand their networks and markets. **Cloud data warehouses** help to overcome the huge costs of purchasing, infrastructure, installation, etc. [104]. Hence, in the coming years, more sophisticated technology in cloud DWs is envisaged to enhance intense, easy-to-use, and economical data clouds as well. The long-term gains for the adoption of cloud warehousing are mainly data availability and scalability. The flexibility to store a variety of data formats—not just relational—combined with the intrinsic flexibility of cloud-based services enables a very broad distribution of cloud services.

Another massive change is in the means of **data analytics**. In contrast to the older times, wherein data analytics and business intelligence occurred in two different divisions, which delayed the overall efficiency of the system, the modern data warehouse provides an advanced structure for storage and faster data flow, thus making them easily accessible for business users. Such an agility model is powered by data fragmentation, allowing access to and the analysis of data across the enterprise in real time.

Another big advancement is in the **Internet of Things (IoT)** platforms for sharing and storing data. This has changed the face of data streaming by enabling users to store and access data across multiple devices. The concept of the IoT is more pertinent to the real world due to the increasing popularity of mobile devices, embedded and ubiquitous communication technologies, cloud computing, and data analytics. In a broader sense, as with the Internet, the IoT enables devices to exist in many places and facilitates applications from trivial to the most crucial. Several technologies such as computational intelligence and big data can be incorporated together with the IoT to improve data management and knowledge discovery on a large scale. Much research in this sense has been carried out by Mishra et al. [105].

In summary, the future of data warehouses comprises features that enable the following:

- All the data are accessible from a single location;
- The capability to outsource the task of maintaining that service's high availability to all customers;
- Governance based on policies;
- Platforms with high user experience (UX) discoverability;
- Platforms that cater to all customers.

7.2. Data Lakes: Opportunities and Future Directions

One of the core capabilities of a data lake architecture is its ability to quickly and easily ingest multiple types of data (e.g., real-time streaming data from on-premises storage platforms, structured data generated and processed by mainframes and data warehouses, and unstructured or semi-structured data). The ingestion process makes use of a high degree of parallelism and low latency since it requires interfacing with external data sources with limited bandwidth. Hence, ingestion will not carry out any deep analysis of the downloaded data. However, there are possibilities for **applying shallow data sketches**

on the downloaded contents and their metadata to maintain a basic organization of the ingested data sets.

In another phase of data lake management (i.e., **the data extraction stage**), the raw data are transformed into a predetermined data model. Although various studies have been conducted on this topic, there still remains room for improvement. Rather than conducting extraction on one file at a time, one can take advantage of the knowledge from the history of extractions. Similarly, in the cleaning phase of the data lake, not much work has not been performed in the literature other than some approaches such as CLAMS [49]. One opportunity in this regard will be to make use of the lake's wisdom and perform collective data cleaning. In addition, it is important to investigate the possible means of errors in the lake and to get rid of them efficiently to obtain a clean data lake.

The common methods to retrieve the data from the data lake are query-based retrieval (a user starts a search with a query for data retrieval) and data-based retrieval (a user navigates a data lake as a linkage graph or a hierarchical structure to find data of interest) [75]. A new direction may be to incorporate **analysis-driven or context-driven** approaches (i.e., augmenting a data set with relevant data and some contextual information to facilitate learning tasks).

Another direction of research is related to the exploration of **machine learning in data lakes**. Specifically, many studies are underway focusing on ML application toward data set organization and discovery. The data set discovery task is often associated with finding "similar" attributes extracted from the data, metadata, etc. which could be further coupled with classification or clustering tasks. Some recent works have leveraged ML techniques, such as the KNN classifier [106] and a logistic regression model for optimizing feature coefficients [107]. More advanced deep learning and similar sophisticated ML techniques are envisaged to augment the data set discovery process in the coming years.

Metadata management is an important task in a data lake, since a DL does not come with descriptive data catalogs [75,77]. Due to the lack of such explicit metadata of data sets, especially during the discovery and cleaning of data, there is a chance for a data lake to become a data swamp. Hence, it is quite necessary to extract meaningful metadata from data sources and to support efficient storage and query answering of metadata. In this field of metadata management, there remain more topics to explore further in extracting knowledge from lake data and incorporating them into existing knowledge bases. Yet another key aspect is **data versioning**, wherein new versions of the already existing files enter into a dynamic data lake [77]. Since versioning-related operations can affect all stages of a data lake, it is a very crucial aspect to address. There are some large-scale data set version control tools, such as DataHub (https://datahubproject.io, accessed on 25 September 2022), that provide a git-like interface to handle version creation, branching, and merging operations. Nevertheless, more research and development may be carried out further to deal with schema evolution.

As a final note, there is an emerging data management architecture trend called the *data lakehouse* that couples the flexibility of a data lake with the data management capabilities of a data warehouse. Specifically, it is considered a unique data storage solution for all data—unstructured, semi-structured, and structured—while providing the data quality and data governance standards of a data warehouse [108]. Such a data lakehouse would be capable of imparting better data governance, reduced data movement and redundancy, efficient use time, etc., even with a simplified schema. This topic of the *data lakehouse* is envisaged to be an excellent research area of data management in the future.

8. Conclusions

Enterprises and business organizations exploit a huge volume of data to understand their customers and to make informed business decisions to stay competitive in the field. However, big data come in a variety of formats and types (e.g., structured, semi-structured and unstructured data), making it difficult for businesses to manage and use them effectively. Based on the structure of the data, typically, two types of data storage are utilized in enterprise data management: the data warehouse (DW) and data lake (DL). Although being

used as interchangeable terms, they are two distinct storage forms with unique characteristics that serve different purposes.

In this review, a comparative analysis of data warehouses and data lakes by highlighting the key differences between the two data management approaches was envisaged. In particular, the definitions of the data warehouse and data lake, highlighting their characteristics and key differences, were detailed. Furthermore, the architecture and design aspects of both DWs and DLs are clearly discussed. In addition, a detailed overview of the popular DW and DL tools and services was also provided. The key challenges of big data analytics in general, as well as the challenges of implementation of DWs and DLs, were also critically analyzed in this survey. Finally, the opportunities and future research directions were contemplated. We hope that the thorough comparison of existing data warehouses vs. data lakes and the discussion of open research challenges in this survey will motivate the future development of enterprise data management and benefit the research community significantly.

Author Contributions: Conceptualization, A.N. and D.M.; methodology, A.N. and D.M.; validation, A.N.; formal analysis, D.M.; investigation, A.N.; data curation, A.N. and D.M.; writing—original draft preparation, A.N. and D.M.; writing—review and editing, A.N. and D.M.; visualization, A.N.; supervision, A.N.; project administration, A.N. All authors have read and agreed to the published version of the manuscript.

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 authors declare no conflict of interest.

References

1. Tsai, C.W.; Lai, C.F.; Chao, H.C.; Vasilakos, A.V. Big data analytics: A survey. *J. Big Data* **2015**, *2*, 21. [CrossRef]
2. Big Data—Statistics & Facts. Available online: https://www.statista.com/topics/1464/big-data/ (accessed on 27 October 2022).
3. Wise, J. Big Data Statistics 2022: Facts, Market Size & Industry Growth. Available online: https://earthweb.com/big-data-statistics/ (accessed on 27 October 2022).
4. Jain, A. The 5 V's of Big Data. 2016. Available online: https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/ (accessed on 27 October 2022).
5. Gandomi, A.; Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manag.* **2015**, *35*, 137–144. [CrossRef]
6. Sun, Z.; Zou, H.; Strang, K. Big Data Analytics as a Service for Business Intelligence. In *Open and Big Data Management and Innovation*; Springer International Publishing: Cham, Switzerland, 2015; Volume 9373, pp. 200–211. [CrossRef]
7. Big Data and Analytics Services Global Market Report. Available online: https://www.reportlinker.com/p06246484/Big-Data-and-Analytics-Services-Global-Market-Report.html (accessed on 27 October 2022).
8. BI & Analytics Software Market Value Worldwide 2019–2025. Available online: https://www.statista.com/statistics/590054/worldwide-business-analytics-software-vendor-market/ (accessed on 27 October 2022).
9. Kumar, S. What Is a Data Repository and What Is it Used for? 2019. Available online: https://stealthbits.com/blog/what-is-a-data-repository-and-what-is-it-used-for/ (accessed on 27 October 2022).
10. Khine, P.P.; Wang, Z.S. Data lake: A new, ideology in big data era. *ITM Web Conf.* **2018**, *17*, 03025. [CrossRef]
11. Arif, M.; Mujtaba, G. A Survey: Data Warehouse Architecture. *Int. J. Hybrid Inf. Technol.* **2015**, *8*, 349–356. [CrossRef]
12. El Aissi, M.E.M.; Benjelloun, S.; Loukili, Y.; Lakhrissi, Y.; Boushaki, A.E.; Chougrad, H.; Elhaj Ben Ali, S. Data Lake Versus Data Warehouse Architecture: A Comparative Study. In *WITS 2020*; Bennani, S., Lakhrissi, Y., Khaissidi, G., Mansouri, A., Khamlichi, Y., Eds.; Springer: Singapore, 2022; Volume 745, pp. 201–210. [CrossRef]
13. Rehman, K.U.u.; Ahmad, U.; Mahmood, S. A Comparative Analysis of Traditional and Cloud Data Warehouse. *VAWKUM Trans. Comput. Sci.* **2018**, *6*, 34–40. [CrossRef]
14. Devlin, B.A.; Murphy, P.T. An architecture for a business and information system. *IBM Syst. J.* **1988**, *27*, 60–80. [CrossRef]
15. Garani, G.; Chernov, A.; Savvas, I.; Butakova, M. A Data Warehouse Approach for Business Intelligence. In Proceedings of the 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Napoli, Italy, 12–14 June 2019; pp. 70–75. [CrossRef]

16. Gupta, V.; Singh, J. A Review of Data Warehousing and Business Intelligence in different perspective. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 8263–8268.
17. Sagiroglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47. [CrossRef]
18. Miloslavskaya, N.; Tolstoy, A. Application of Big Data, Fast Data, and Data Lake Concepts to Information Security Issues. In Proceedings of the 2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), Vienna, Austria, 22–24 August 2016; pp. 148–153. [CrossRef]
19. 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, Porto, Portugal, 26–28 July 2018; pp. 294–301. [CrossRef]
20. Lin, J. The Lambda and the Kappa. *IEEE Internet Comput.* **2017**, *21*, 60–66. [CrossRef]
21. Devlin, B. Thirty Years of Data Warehousing—Part 1. 2020. Available online: https://www.irmconnects.com/thirty-years-of-data-warehousing-part-1/ (accessed on 27 October 2022).
22. Inmon, W.H. *Building the Data Warehouse*, 4th ed.; Wiley Publishing: Indianapolis, IN, USA, 2005.
23. Chandra, P.; Gupta, M.K. Comprehensive survey on data warehousing research. *Int. J. Inf. Technol.* **2018**, *10*, 217–224. [CrossRef]
24. Simões, D.M. Enterprise Data Warehouses: A conceptual framework for a successful implementation. In Proceedings of the Canadian Council for Small Business & Entrepreneurship Annual Conference, Calgary, AL, Canada, 28–30 October 2010.
25. Al-Debei, M.M. Data Warehouse as a Backbone for Business Intelligence: Issues and Challenges. *Eur. J. Econ. Financ. Adm. Sci.* **2011**, *33*, 153–166.
26. Report by Market Research Future (MRFR). Available online: https://finance.yahoo.com/news/data-warehouse-dwaas-market-predicted-153000649.html (accessed on 27 October 2022).
27. Chaudhuri, S.; Dayal, U. An overview of data warehousing and OLAP technology. *ACM Sigmod Rec.* **1997**, *26*, 65–74. [CrossRef]
28. Codd, E.F.; Codd, S.B.; Salley, C.T. In *Providing OLAP to User-Analysts: An IT Mandate*; Codd & Associates: Ladera Ranch, CA, USA, 1993; pp. 1–26.
29. The Best Applications of Data Warehousing. 2020. Available online: https://datachannel.co/blogs/best-applications-of-data-warehousing/ (accessed on 27 October 2022).
30. Hai, R.; Quix, C.; Jarke, M. Data lake concept and systems: A survey. *arXiv* **2021**, arXiv:2106.09592.
31. Zagan, E.; Danubianu, M. Data Lake Approaches: A Survey. In Proceedings of the 2020 International Conference on Development and Application Systems (DAS), Suceava, Romania, 21–23 May 2020; pp. 189–193. [CrossRef]
32. Cherradi, M.; El Haddadi, A. Data Lakes: A Survey Paper. In *Innovations in Smart Cities Applications*; Ben Ahmed, M., Boudhir, A.A., Karaș, R., Jain, V., Mellouli, S., Eds.; Lecture Notes in Networks and Systems; Springer International Publishing: Cham, Switzerland, 2022; Volume 5, pp. 823–835. [CrossRef]
33. Dixon, J. Pentaho, Hadoop, and Data Lakes. 2010. Available online: https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/ (accessed on 27 October 2022).
34. King, T. The Emergence of Data Lake: Pros and Cons. 2016. Available online: https://solutionsreview.com/data-integration/the-emergence-of-data-lake-pros-and-cons/ (accessed on 27 October 2022).
35. Alrehamy, H.; Walker, C. Personal Data Lake with Data Gravity Pull. In Proceedings of the IEEE Fifth International Conference on Big Data and Cloud Computing 2015, Beijing, China, 26–28 August 2015. [CrossRef]
36. Yang, Q.; Ge, M.; Helfert, M. Analysis of Data Warehouse Architectures: Modeling and Classification. In Proceedings of the 21st International Conference on Enterprise Information Systems, Heraklion, Greece, 3–5 May 2019; pp. 604–611.
37. Yessad, L.; Labiod, A. Comparative study of data warehouses modeling approaches: Inmon, Kimball and Data Vault. In Proceedings of the 2016 International Conference on System Reliability and Science (ICSRS), Paris, France, 15–18 November 2016; pp. 95–99. [CrossRef]
38. Oueslati, W.; Akaichi, J. A Survey on Data Warehouse Evolution. *Int. J. Database Manag. Syst.* **2010**, *2*, 11–24. [CrossRef]
39. Ali, F.S.E. A Survey of Real-Time Data Warehouse and ETL. *Int. J. Sci. Eng. Res.* **2014**, *5*, 3–9.
40. 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; pp. 2785–2794. [CrossRef]
41. Alsqour, M.; Matouk, K.; Owoc, M. A survey of data warehouse architectures—Preliminary results. In Proceedings of the Federated Conference on Computer Science and Information Systems, Wroclaw, Poland, 9–12 September 2012; pp. 1121–1126.
42. Rizzi, S.; Abelló, A.; Lechtenbörger, J.; Trujillo, J. Research in data warehouse modeling and design: Dead or alive? In Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, DOLAP '06, Arlington, VA, USA, 10 November 2006; Association for Computing Machinery: New York, NY, USA, 2006; pp. 3–10. [CrossRef]
43. Maccioni, A.; Torlone, R. KAYAK: A Framework for Just-in-Time Data Preparation in a Data Lake. In *Advanced Information Systems Engineering*; Krogstie, J., Reijers, H.A., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; pp. 474–489. [CrossRef]
44. Gao, Y.; Huang, S.; Parameswaran, A. Navigating the Data Lake with DATAMARAN: Automatically Extracting Structure from Log Datasets. In Proceedings of the 2018 International Conference on Management of Data, Houston, TX, USA, 10–15 June 2018; ACM: Houston, TX, USA, 2018; pp. 943–958. [CrossRef]

45. Astriani, W.; Trisminingsih, R. Extraction, Transformation, and Loading (ETL) Module for Hotspot Spatial Data Warehouse Using Geokettle. *Procedia Environ. Sci.* **2016**, *33*, 626–634. [CrossRef]
46. Halevy, A.V.; Korn, F.; Noy, N.F.; Olston, C.; Polyzotis, N.; Roy, S.; Whang, S.E. Managing Google's data lake: An overview of the Goods system. *IEEE Data Eng. Bull.* **2016**, *39*, 5–14.
47. Dehne, F.; Robillard, D.; Rau-Chaplin, A.; Burke, N. VOLAP: A Scalable Distributed System for Real-Time OLAP with High Velocity Data. In Proceedings of the 2016 IEEE International Conference on Cluster Computing (CLUSTER), Taipei, Taiwan, 13–15 September 2016; pp. 354–363. [CrossRef]
48. Hurtado, C.A.; Gutierrez, C.; Mendelzon, A.O. Capturing summarizability with integrity constraints in OLAP. *ACM Trans. Database Syst.* **2005**, *30*, 854–886. [CrossRef]
49. 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 '16, San Francisco, CA, USA, 26 June–1 July 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 2089–2092. [CrossRef]
50. Zhang, Y.; Ives, Z.G. Juneau: Data lake management for Jupyter. *Proc. VLDB Endow.* **2019**, *12*, 1902–1905. [CrossRef]
51. Zhu, E.; Deng, D.; Nargesian, F.; Miller, R.J. JOSIE: Overlap Set Similarity Search for Finding Joinable Tables in Data Lakes. In Proceedings of the 2019 International Conference on Management of Data, SIGMOD '19, Amsterdam, The Netherlands, 30 June–5 July 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 847–864. [CrossRef]
52. Beheshti, A.; Benatallah, B.; Nouri, R.; Chhieng, V.M.; Xiong, H.; Zhao, X. CoreDB: A Data Lake Service. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17, Singapore, 6–10 November 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 2451–2454. [CrossRef]
53. Hai, R.; Geisler, S.; Quix, C. Constance: An Intelligent Data Lake System. In Proceedings of the 2016 International Conference on Management of Data, SIGMOD '16, San Francisco, CA, USA, 26 June–1 July 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 2097–2100. [CrossRef]
54. Ahmed, A.S.; Salem, A.M.; Alhabibi, Y.A. Combining the Data Warehouse and Operational Data Store. In Proceedings of the Eighth International Conference on Enterprise Information Systems, Paphos, Cyprus, 23–27 May 2006; pp. 282–288. [CrossRef]
55. Software Architecture: N Tier, 3 Tier, 1 Tier, 2 Tier Architecture. Available online: https://www.appsierra.com/blog/url (accessed on 27 October 2022).
56. Han, S.W. Three-Tier Architecture for Sentinel Applications and Tools: Separating Presentation from Functionality. Ph.D. Thesis, University of Florida, Gainesville, FL, USA, 1997.
57. What Is Three-Tier Architecture. Available online: https://www.ibm.com/in-en/cloud/learn/three-tier-architecture (accessed on 27 October 2022).
58. Phaneendra, S.V.; Reddy, E.M. Big Data—Solutions for RDBMS Problems—A Survey. *Int. J. Adv. Res. Comput. Commun. Eng.* **2013**, *2*, 3686–3691.
59. Simitsis, A.; Vassiliadis, P.; Sellis, T. Optimizing ETL processes in data warehouses. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; pp. 564–575. [CrossRef]
60. Prasser, F.; Spengler, H.; Bild, R.; Eicher, J.; Kuhn, K.A. Privacy-enhancing ETL-processes for biomedical data. *Int. J. Med. Inform.* **2019**, *126*, 72–81. [CrossRef]
61. Rousidis, D.; Garoufallou, E.; Balatsoukas, P.; Sicilia, M.A. Metadata for Big Data: A preliminary investigation of metadata quality issues in research data repositories. *Inf. Serv. Use* **2014**, *34*, 279–286. [CrossRef]
62. Mailvaganam, H. Introduction to OLAP—Slice, Dice and Drill! 2007. Data Warehousing Review. Retrieved on 18 March 2008. Available online: https://web.archive.org/web/20180928201202/http://dwreview.com/OLAP/Introduction_OLAP.html (accessed on 25 September 2022).
63. Pendse, N. What is OLAP? Available online: https://dssresources.com/papers/features/pendse04072002.htm (accessed on 27 October 2022).
64. Xu, J.; Luo, Y.Q.; Zhou, X.X. Solution for Data Growth Problem of MOLAP. *Appl. Mech. Mater.* **2013**, *321–324*, 2551–2556. [CrossRef]
65. Dehne, F.; Eavis, T.; Rau-Chaplin, A. Parallel multi-dimensional ROLAP indexing. In Proceedings of the CCGrid 2003. 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid, Tokyo, Japan, 12–15 May 2003; pp. 86–93. [CrossRef]
66. Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R. The Hadoop Distributed File System. In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), Incline Village, NV, USA, 3–7 May 2010; pp. 1–10. [CrossRef]

67. Luo, Z.; Niu, L.; Korukanti, V.; Sun, Y.; Basmanova, M.; He, Y.; Wang, B.; Agrawal, D.; Luo, H.; Tang, C.; et al. From Batch Processing to Real Time Analytics: Running Presto® at Scale. In Proceedings of the 2022 IEEE 38th International Conference on Data Engineering (ICDE), Kuala Lumpur, Malaysia, 9–12 May 2022; pp. 1598–1609. [CrossRef]
68. Sethi, R.; Traverso, M.; Sundstrom, D.; Phillips, D.; Xie, W.; Sun, Y.; Yegitbasi, N.; Jin, H.; Hwang, E.; Shingte, N.; et al. Presto: SQL on Everything. In Proceedings of the 2019 IEEE 35th International Conference on Data Engineering (ICDE), Macao, China, 8–1 April 2019; pp. 1802–1813. [CrossRef]
69. Kinley, J. The Lambda Architecture: Principles for Architecting Realtime Big Data Systems. 2013. Available online: http://jameskinley.tumblr.1084com/post/37398560534/thelambda-architecture-principles-for (accessed on 27 October 2022).
70. Ferrera Bertran, P. Lambda Architecture: A state-of-the-Art. Datasalt. 17 January 2014. Available online: https://github.com/pereferrera/trident-lambda-splout (accessed on 25 September 2022).
71. Carbone, P.; Katsifodimos, A.; Ewen, S.; Markl, V.; Haridi, S.; Tzoumas, K. Apache Flink™: Stream and Batch Processing in a Single Engine. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.* **2015**, *36*, 28–38.
72. Kreps, J. Questioning the Lambda Architecture. 2014. Available online: https://www.oreilly.com/radar/questioning-the-lambda-architecture/ (accessed on 27 October 2022).
73. Data Vault vs Star Schema vs Third Normal Form: Which Data Model to Use? Available online: https://www.matillion.com/resources/blog/data-vault-vs-star-schema-vs-third-normal-form-which-data-model-to-use (accessed on 27 October 2022).
74. Patranabish, D. Data Lakes: The New Enabler of Scalability in Cross Channel Analytics—Tech-Talk by Durjoy Patranabish | ET CIO. Available online: http://cio.economictimes.indiatimes.com/tech-talk/data-lakes-the-new-enabler-of-scalability-in-cross-channel-analytics/585 (accessed on 27 October 2022).
75. Nargesian, F.; Zhu, E.; Miller, R.J.; Pu, K.Q.; Arocena, P.C. Data lake management: Challenges and opportunities. *Proc. VLDB Endow.* **2019**, *12*, 1986–1989. [CrossRef]
76. A Brief Look at 4 Major Data Compliance Standards: GDPR, HIPAA, PCI DSS, CCPA. Available online: https://www.pentasecurity.com/blog/4-data-compliance-standards-gdpr-hipaa-pci-dss-ccpa/ (accessed on 27 October 2022).
77. Sawadogo, P.; Darmont, J. On data lake architectures and metadata management. *J. Intell. Inf. Syst.* **2021**, *56*, 97–120. [CrossRef]
78. Overview of Amazon Web Services: AWS Whitepaper. 2022. Available online: https://d1.awsstatic.com/whitepapers/aws-overview.pdf (accessed on 27 October 2022).
79. Pandis, I. The evolution of Amazon redshift. *Proc. VLDB Endow.* **2021**, *14*, 3162–3174. [CrossRef]
80. Microsoft Azure Documentation. Available online: http://azure.microsoft.com/en-us/documentation/ (accessed on 27 October 2022).
81. Automate Your Data Warehouse. Available online: https://www.oracle.com/autonomous-database/autonomous-data-warehouse/ (accessed on 27 October 2022).
82. Dageville, B.; Cruanes, T.; Zukowski, M.; Antonov, V.; Avanes, A.; Bock, J.; Claybaugh, J.; Engovatov, D.; Hentschel, M.; Huang, J.; et al. The Snowflake Elastic Data Warehouse. In Proceedings of the 2016 International Conference on Management of Data, San Francisco, CA, USA, 26 June–1 July 2016; ACM: San Francisco, CA, USA, 2016; pp. 215–226. [CrossRef]
83. Mathis, C. Data Lakes. *Datenbank-Spektrum* **2017**, *17*, 289–293. [CrossRef]
84. Zagan, E.; Danubianu, M. Cloud DATA LAKE: The new trend of data storage. In Proceedings of the 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Online, 11–13 June 2021; IEEE: Ankara, Turkey, 2021; pp. 1–4. [CrossRef]
85. Ramakrishnan, R.; Sridharan, B.; Douceur, J.R.; Kasturi, P.; Krishnamachari-Sampath, B.; Krishnamoorthy, K.; Li, P.; Manu, M.; Michaylov, S.; Ramos, R.; et al. Azure Data Lake Store: A Hyperscale Distributed File Service for Big Data Analytics. In Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD '17, Chicago, IL, USA, 14–19 May 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 51–63. [CrossRef]
86. Elgendy, N.; Elragal, A. Big Data Analytics: A Literature Review Paper. In *Advances in Data Mining. Applications and Theoretical Aspects*; Perner, P., Ed.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2014; pp. 214–227. [CrossRef]
87. Jin, X.; Wah, B.W.; Cheng, X.; Wang, Y. Significance and Challenges of Big Data Research. *Big Data Res.* **2015**, *2*, 59–64. [CrossRef]
88. Agrawal, R.; Nyamful, C. Challenges of big data storage and management. *Glob. J. Inf. Technol. Emerg. Technol.* **2016**, *6*, 1–10. [CrossRef]
89. Padgavankar, M.H.; Gupta, S.R. Big Data Storage and Challenges. *Int. J. Comput. Sci. Inf. Technol.* **2014**, *5*, 2218–2223.
90. Kadadi, A.; Agrawal, R.; Nyamful, C.; Atiq, R. Challenges of data integration and interoperability in big data. In Proceedings of the 2014 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 27–30 October 2014; IEEE: Washington, DC, USA, 2014; pp. 38–40. [CrossRef]
91. Best Data Integration Tools. Available online: https://www.peerspot.com/categories/data-integration-tools (accessed on 27 October 2022).
92. Toshniwal, R.; Dastidar, K.G.; Nath, A. Big Data Security Issues and Challenges. *Int. J. Innov. Res. Adv. Eng.* **2014**, *2*, 15–20.
93. Demchenko, Y.; Ngo, C.; de Laat, C.; Membrey, P.; Gordijenko, D. Big Security for Big Data: Addressing Security Challenges for the Big Data Infrastructure. In *Secure Data Management*; Jonker, W., Petković, M., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2014; pp. 76–94. [CrossRef]
94. Chen, E.T. Implementation issues of enterprise data warehousing and business intelligence in the healthcare industry. *Commun. IIMA* **2012**, *12*, 3.

95. Cuzzocrea, A.; Bellatreche, L.; Song, I.Y. Data warehousing and OLAP over big data: Current challenges and future research directions. In Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP, DOLAP '13, San Francisco, CA, USA, 28 October 2013; Association for Computing Machinery: New York, NY, USA, 2013; pp. 67–70. [CrossRef]
96. Singh, R.; Singh, K. A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. *Int. J. Comput. Sci. Issues* **2010**, *7*, 41.
97. Longbottom, C.; Bamforth, R. Optimising the Data Warehouse. 2013. Available online: https://www.it-daily.net/downloads/WP_Optimising-the-data-warehouse.pdf (accessed on 27 October 2022).
98. Santos, R.J.; Bernardino, J.; Vieira, M. A survey on data security in data warehousing: Issues, challenges and opportunities. In Proceedings of the 2011 IEEE EUROCON—International Conference on Computer as a Tool, Lisbon, Portugal, 27–29 April 2011, pp. 1–4. [CrossRef]
99. Responsibilities of a Data Warehouse Governance Committee. Available online: https://docs.oracle.com/cd/E29633_01/CDMOG/GUID-7E43F311-4510-4F1E-A17E-693F94BD0EC7.htm (accessed on 28 October 2022).
100. Gupta, S.; Giri, V. *Practical Enterprise Data Lake Insights: Handle Data-Driven Challenges in an Enterprise Big Data Lake*, 1st ed.; Apress: Berkeley, CA, USA, 2018.
101. Giebler, C.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B. Leveraging the Data Lake: Current State and Challenges. In *Big Data Analytics and Knowledge Discovery*; Ordonez, C., Song, I.Y., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; pp. 179–188. [CrossRef]
102. Lock, M. Maximizing Your Data Lake with a Cloud or Hybrid Approach. 2016. Available online: https://technology-signals.com/wp-content/uploads/download-manager-files/maximizingyourdatalake.pdf (accessed on 27 October 2022).
103. Kumar, N. Cloud Data Warehouse Is the Future of Data Storage. 2020. Available online: https://www.sigmoid.com/blogs/cloud-data-warehouse-is-the-future-of-data-storage/ (accessed on 27 October 2022).
104. Kahn, M.G.; Mui, J.Y.; Ames, M.J.; Yamsani, A.K.; Pozdeyev, N.; Rafaels, N.; Brooks, I.M. Migrating a research data warehouse to a public cloud: Challenges and opportunities. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 592–600. [CrossRef]
105. Mishra, N.; Lin, C.C.; Chang, H.T. A Cognitive Adopted Framework for IoT Big-Data Management and Knowledge Discovery Prospective. *Int. J. Distrib. Sens. Netw.* **2015**, *2015*, 1–12. [CrossRef]
106. Alserafi, A.; Abelló, A.; Romero, O.; Calders, T. Keeping the Data Lake in Form: DS-kNN Datasets Categorization Using Proximity Mining. In *Model and Data Engineering*; Schewe, K.D., Singh, N.K., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; pp. 35–49. [CrossRef]
107. 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: Dallas, TX, USA, 2020; pp. 709–720. [CrossRef]
108. Armbrust, M.; Ghodsi, A.; Xin, R.; Zaharia, M. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. In Proceedings of the Conference on Innovative Data Systems Research, Virtual Event, 11–15 January 2021.