MDPI

Article

A Survey of Big Data Archives in Time-domain Astronomy

Manoj Poudel *, Rashmi P. Sarode *, Yutaka Watanobe, Maxim Mozgovoy, and Subhash Bhalla

Graduate Department of Computer and Information Systems, The University of Aizu, Aizu-Wakamatsu, Fukushima 965-8580, Japan; yutaka@u-aizu.ac.jp, mozgovoy@u-aizu.ac.jp; bhalla.subhash@gmail.com
* Correspondence: pmanoj0091@gmail.com (M.P.); rashmipsarode@gmail.com (R.P.S.)

Abstract: The rise of Big Data has resulted in the proliferation of numerous heterogeneous data stores. Even though multiple models are for integrating these data, combining such huge amounts of data into a single model remains challenging. There is a need in the database management archives to manage such huge volumes of data without any particular structure which comes from unconnected and unrelated sources. This data is growing in size and thus demands special attention. The speed with which this data is growing as well as the varied data types involved and stored in scientific archives is posing further challenges. Astronomy is also increasingly becoming a science which is now based on a lot of data processing and involves assorted data. This data is now stored in domain-specific archives. Many astronomical studies are producing large scale archives of data and these archives are then published in the form of data repositories. These mainly consist of images and text without any structure in addition to data with some structure such as relations with key values. When the archives are published as remote data repositories, it is challenging work to organize the data against its increased diversity and to meet the information request from users. To address this problem, Polystore Systems present a new model of data integration and have been proposed to access unrelated data repositories using an uniform single query language. This article highlights the Polystore system for integrating large scale heterogeneous data in the astronomy domain.

Keywords: Big Data; Data Integretion; Astronomy; Polystore;

1. Introduction

Due to an abundance of data sources, the amount of data available for analysis is increasing rapidly, and there has been a great deal of research into how to manage heterogeneous data. Big data refers to large or complex information that cannot be processed using traditional methods. For a long time, people have been storing and accessing huge amounts of data for analytics [1]. In today's world big data both structured as well as unstructured data is available. Databases and spreadsheets that have been used in the past include structured data that is often numerical. An unstructured data set is a collection of unrelated pieces of information that does not follow a predetermined structure. It is common practice to store and process large amounts of data using specialized computer databases and applications [2].

The Semantic Web, often known as the Web of Data, is based on the concept of Linked Data [3]. The Semantic Web's emphasis on Linked Data's best practices for creating meaningful links between resources will benefit humans and robots. It is a collection of design concepts for sharing machine-readable and cross-referenced data [4]. Linked Data is a collection of best practices for publishing and interconnecting structured data on the Internet. Uniform Resource Identifiers (URIs, are a common means of identifying concepts or entities), Hypertext Transfer Protocol (HTTP, a protocol used to retrieve resources or descriptions of resources, which is fundamental and universal) and Resource Description Framework (RDF, a data model based on graphs to organize data describing things in the world by structuring and linking it) are some technologies that support Linked Data [5]. Linked Data is a way to identify entities or concepts in the world and retrieve resources or

Citation: Poudel, M.; Sarode, R.P.; Watanobe, Y; Mozgovoy, M; Bhalla, S. A Survey on Time-domain Astronomy in Big Data Archives. *Journal Not Specified* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

descriptions of resources [6]. Open-linked data is known as linked open data (LOD) [7].The Tim Berners-Lee definition of linked open data compared to linked data is the most precise: "Linked Open Data is Linked Data provided under an open license that does not hinder its free usage " [6].

The current model that focuses on data integration is open data integration. This feature is coherent worldwide discovery, or data influenced by the need for data analysis. The latest integration techniques require extensive data analysis to pinpoint the data to be used for extracting information with respect to big data [8]. Stores with copious amounts of varied data sets containing diverse data, called data lakes, may contain data in raw format, or may be organized in rich data schemas. The challenge presented to scientists is to match the pace of growth of data repositories and archives [9]. Other problems arise when storing heterogeneous data in native data stores and creating a communication protocol in data stores to query for information retrieval.

Celestial phenomena that arise outside our atmosphere are studied by astronomers in their quest to better understand the universe. Similar to other scientific disciplines, astronomy deals with data tsunamis that require new methods and techniques for conducting scientific research [10]. Advances in telescopes, detector technology, and exponential increases in processing power have led to an avalanche of data in astronomy and other data-heavy sciences such as physics, biology, and geology. This data avalanche is expected to continue for several years. An efficient federation of database technologies is required to properly handle this avalanche and process enormous amounts of data. Ultimately, data-mining tools for analysis are crucial for extracting knowledge from enormous data volumes. In Section 3, we discuss mining in astronomy in detail [11].

The goal of this research is to survey big data archives in time-domain astronomy and find a solution to efficiently manage large repositories of data. Modern astronomical studies are heavily focused on undertaking large sky surveys and archiving the results. Prototypes and demonstrations of the classification and mining of distributed data are necessary to investigate user interfaces and interaction models. For instance, AstroDAS [12] is a system in which scientists can record and share their assertions regarding the data integration process. Web services, linked databases and grids need to be investigated to dispense the method of mapping entities obtained from their research over accepted scientific databases. The identification of similar celestial objects within the present network of diverse catalogs is what astronomers want to share in the future.

The rest of the manuscript is as follows: We discuss big data in astronomical domain in Section 2, Section 3 explores time domain astronomy, and Section 4 analyzes astronomical data mining. In Section 5, we explore Scientific Archive Services and Query Languages, Section 6 emphasizes on Polystore system, Section 7 deals with the Challenge of the future management in astronomical big open linked data, and Section 8 presents the summary and conclusions.

2. Big Data in Astronomical Domain

The term "big data" refers to the huge amount of data that can be found in various forms and formats. Traditional relational data are not the only source of unstructured data, which are continuously expanding [13]. It is typical for big data to be available from various sources. Data derived from machines, for example, grows exponentially and holds a wealth of information that will be uncovered in the future. However, even though the data gathered by humans are more textual, valuable insights can still be retrieved from them [14].

For computer scientists interested in astronomy, most surveys make their entire data collection and any derived parameters readily available online, in the form of enormous databases. In today's world, data are becoming increasingly large, unstructured, and fast-moving, making traditional data management methods ineffective. Over a billion stars have been studied, millions of objects have their spectra taken, and hundreds of new planets are still being identified. The Sloan Digital Sky Survey (SDSS) [15] is one

of the largest astronomical surveys. More than 200 million galaxies and more stars have been discovered in the almost a million field images capture by the SDSS telescope per night. Future surveys are expected to yield significantly more information. The Legacy Survey of Space and Time (LSST), now known as the Vera C. Rubin Observatory, is another prospective future survey that will offer wide-field images of the sky and expose galaxies that are currently too faint to be noticed.

2.1. The four V's of Big Data

The four V's of big data (volume, velocity, variety, and value) are the pillars of this concept [16]. They are explained as follows:

(i) **Volume:** The quantity of data is known as the volume. The data are described in terms of terabytes, petabytes, and even exabytes. Consequently, the collection, cleaning, curation, integration, storage, processing, indexing, search, sharing, transfer, mining, analysis, and visualization of large amounts of data are complicated by the quantity of data [10]. There is too much data for current technology to handle effectively. There is an avalanche of data in astronomy generated by a number of earth- and space-based broad astronomical observations.

(ii) **Variety:** Variety is an indicator of data complexity. Structured, semi-structured, unstructured, and mixed data are all types of data [17]. Images, spectra, time series, and simulations constitute a large amount of astronomical data. Catalogs and databases contain a vast majority of information. This complicates data integration from different telescopes and projects because the data are stored in different formats. The high dimensionality problem is exacerbated because each piece of data has thousands or more features.

(iii) **Velocity:** The term "velocity" represents the rate of creation, communication, and analysis of data. For ten years, The Rubin Observatory Legacy Survey of Space and Time (LSST) [18] generated one Sloan Digital Sky Survey (SDSS) [15] of data volume one night [19]. It is necessary to analyze data in a bundle, stream, near-instantaneous, or instantaneous setting. The LSST expects to discover a thousand additional supernova explosions every night for the next ten years. Scientists have a significant challenge in determining how to mine, correctly identify, and target supernova prospects in ten years.

(iv) **Value:** Discovering new and unusual astronomical objects and events is a challenge that has inspired and exhilarated scientists. Therefore, spotting a new pattern or law in the data distribution is valuable. The term "value" refers to the enormous astronomical worth of the data [10].

Optical time-domain astronomy is nearing a tipping point in terms of the data rate and volume. By 2023, the amount of data is expected to grow by a factor of three. As the number of recognized sources increases, it is be necessary to create efficient and well-designed databases. To successfully manage these data, highly efficient machine learning algorithms for categorizing source types are required [20].

3. Time-domain Astronomy

Time-domain astronomy focuses on studying systems that change over time. In recent years, large-scale surveys of the sky have made it more important to study this topic because it is possible to detect changes that were previously too small to detect [21]. Time-domain astronomy is a branch of astronomy and astrophysics that explores lifetime evolution and changes in a wide variety of cosmic objects, particularly when these changes occur on short cosmic time scales (hours or days to a year) [22]. Novae, supernovae, gamma-ray bursts, active galactic nuclei, binary stars, and pulsars are the specific objects of interest. These are appropriately referred to as transients because the electromagnetic signature radiated by an event such as an explosion is transient. It briefly appears as a flash in the sky before fading away gradually. By capturing these electromagnetic signatures, astronomers can learn about cosmic objects and the physical processes that govern their evolution [23].

Sequences of observations or data points grouped chronologically are called "time series" [24]. It is not typical for researchers to use time series in their work, such as meteorology, electroencephalography, and financial markets. Many characteristics of time series data can be observed, such as their non-isolated generation, their temporal variation, and the presence of a trend or cyclic components. Multiple goals can be achieved by studying time series data, such as gaining insight into the mechanism that generates the data or predicting future outcomes and behaviors [25].

To depict the brightness change of an object over time in time-domain astronomy, light curves [26], are commonly used to represent data collected by telescopes (for a visual representation). Based on the variability features of light curves, astronomical objects can be categorized into several groups (quasars, long-period variables, and eclipsing binaries) and consequently can be researched in-depth. There are a variety of approaches for classifying data into groups based on light curve data, the most common of which is the use of machine learning algorithms to extract features from the light curve data and then arrange the features into categories. Variability classes can be characterized and differentiated using these light-curve traits. There are a wide variety of features, ranging from simple statistical properties, such as the mean or standard deviation to more complicated time series properties, such as auto correlation function. Machine learning and other algorithms can use these properties to distinguish between different types of light curves [25].

3.1. Science Project and Virtual Observatory in the Era of Big Data in Astronomy

Many scientific operations, including astronomy, have become data-intensive in the age of big data and archives. The rapid advancement of technology, particularly in computer hardware (with low-cost, high-capacity storage, and processing) and microelectronics (such as: charge-coupled devices (CCD)) devices [27], has revolutionized the majority of natural science through an explosion in the number of measurements and simulation data [28].

Astronomers use a blink comparator to compare two-night sky images and identify differences. It was possible to "blink" back and forth between two images of the same part of the sky obtained at various points in time using this technique. Asteroids and comets could be differentiated in images taken a few day apart since they would appear to oscillate between two positions. On the contrary, all the other stars stood still. Photographs separated by a longer period of time, can be used to discover stars with large proper motion; to distinguish binary stars from optical doubles, and variable stars [29]. Clyde Tombaugh used the blink comparator to discover Pluto. On loan from the Lowell Observatory and on display in the Museum of Washington, DC's Exploring the Planets Gallery [30].

Modern astronomy has come a long way since Galileo made his initial views of stars in 1609 with his refracting telescope [31]. Astronomy advanced significantly in the first decade of the 1600s by discovering an optical telescope and its use to study the night sky. Global astronomical research projects aim to meet the data volume and computational challenges associated with tackling the forefront research problems. The virtual observatory has been proposed as the response of the astronomical community to the challenges passed by the new massive and complex data sets [32]. The following example of astronomical projects are data intensive, and need virtual observatory for world wide scientific collaboration.

Using A 7.2 deg2 camera mounted on the Palomar Samuel Oschin 48-inch (1.2 meters) Schmidt telescope, Palomar Transient Factory (PTF), has been in service since 2009 [33] [34]. These telescopes observe the night sky in the visible and infrared spectra. It is a fully-automated, wide-field survey aimed at the systematic exploration of the optical transient sky [35]. Two automatic reduction pipelines received data from the camera. Lawrence Berkeley National Laboratory (LBNL) [36] runs a near-real-time image subtraction pipeline to recognize optical transients minutes after the images are recorded. To arrive at a collection of probabilistic claims concerning the scientific classification of the transients, the output of this pipeline is submitted to UC Berkeley, where a source classifier analyzes it [33].

The images were also entered into a database at the Infrared Processing and Analysis Center (IPAC) within a few days of capture. Calibration and object detection are performed on each incoming frame before they are combined into a database. Using the P48 photometric follow-up telescope, the P60 automatically generates colors and light curves for intriguing transients identified [37]. Fifteen more telescopes were used for photometric and spectroscopic follow-up as part of the PTF cooperation. The coordination of observations and reporting is handled by an automated system that collects data from the Berkeley classification engine and distribute it to various follow-up facilities [33].

The intermediate Palomar Transient Factory (iPTF) [38], which began operation in 2012, is the successor of the PTF. The image processing and differencing pipeline innovations have made it possible to receive transient candidates significantly faster (from 30-60 minutes to 10 minutes in iPTF) than before [38] [39]. The PTF/iPTF generated 1 gigabyte data per 90 seconds, and which was approximately 0.05 petabytes per year. The iPTF has conducted a series of fast-cadence studies to identify and characterize young supernovae and rapidly changing transients. A follow-up study is needed to gather more information on the detected transients, such as ultraviolet–optical–infrared light curves and hues, spectroscopic categorization, X-ray and radio observations for non-thermal emission, and a complete multi-wavelength follow-up survey [39].

After years of development, the Zwicky Transient Facility (ZTF) was launched in 2018 with the largest instantaneous field of view for any camera on a telescope with an aperture bigger than 0.5 m using a 48-inch Schmidt Telescope [40]. The ZTF observing system provides high-speed, wide-field-of-view, multi-band optical imagery for time-domain astrophysics analysis [20] [41]. The work of ZTF expands our understanding of the temporal and dynamic sky. This category includes Near-Earth Asteroids (NEAs), unusual and rapidly developing flux transients, and all sorts of galactic variable sources. Managing data transfer from the P48, raw data ingestion, all processing pipelines, long-term archiving and curation of data products, user interfaces for data retrieval and access management, near-real-time distribution of flux-transient alerts and potentially new solar system objects (SSOs), generation of quality assurance (QA) metrics for the project, analysis and trending, and maintenance of all software, hardware fundamental areas of ZTF [42].

Astronomical observatories in the modern era are building work on the Legacy Survey of Space and Time (LSST), which is now taking place in conjunction with the Vera C. Rubin Observatory. Images and data items of approximately 500 petabytes are sent via the LSST [18]. The Rubin Observatory includes an 8.4-meter primary mirror, the world's largest digital camera, a complex data processing system, and an online education platform. The 8.4-meter Simonyi Survey Telescope, which boasts a unique three-mirror construction and an incredibly wide field of vision, takes only three nights to survey the entire sky [43]. LSST will shed light on the unseen components of the universe by tracking the motion of billions of galaxies and analyzing how they distort space and time. For example, variable stars, supernovas, and black holes will be studied in unprecedented detail owing to the LSST. New classes of transient occurrences were discovered. The telescope will reveal the motions of millions of stars and provide a three-dimensional image of our galaxy, which is 1,000 times larger than that of earlier surveys. Over 90% of the potentially hazardous asteroids larger than 140 meters in diameter were investigated using LSST. Beyond Neptune, it should be able to pick up an additional 40,000 bodies [44].

4. Astronomical Data Mining

Due to the rapid development in data volume from various sky surveys, data repositories have grown in size from gigabytes to terabytes and petabytes as discussed in brief in Section 2. The term "big data analysis" refers to analyzing large amounts of data. Astronomy is now a data-intensive science and is expected to become even more data-intensive in the next decade. Clustering and classification problems have traditionally been the norm for astronomers. For astronomers who use observational (experimental) methods to gather data on celestial objects and then analyze that data to determine the objects' physical prop-

erties and the underlying physics that underlies those properties, this is especially true [45]. Data mining is a collection of techniques used to reduce, improve, and clean a large amount of data. These techniques include summarization, classification, regression, clustering, association, time series analysis, and outlier/anomaly identification. The most important source of astronomical data is systematic observation of the sky over a wide range of wavelengths. Scientific data mining has ensured the effectiveness and completeness of these data, leading to new astronomical research. In addition, numerous simulations generate large amounts of data[14]. Distributed data mining (DDM), is becoming more common as astronomical data sets (from many large sky surveys) grow too large to be downloaded to a single location for analysis. The discovery of hidden knowledge in geographically distributed heterogeneous databases is made possible by DDM algorithms [45].

Knowledge discovery in databases (KDD) is the primary emphasis of the data mining overview. However, the concept of a database encompasses machine-readable astronomical information [46]. The KDD focuses on extracting knowledge (high-level information) from low-level data (usually stored in large databases). KDD involves many steps such as data preparation and cleaning, data selection, sampling, preprocessing and transformation, data mining to extract patterns and models, interpretation and evaluation of extracted information, and evaluation, rendering, or use of the final extracted knowledge. However, it is important to keep in mind that data mining is only one part of the KDD process [47]. Many terms are related to data mining, and we begin by introducing some of them:

- **Data Collection:** All actions necessary to gather the desired data in digital form were included in data collection. As a part of the research process, data collection methods include acquiring fresh observations, querying existing databases, and completing data mergers (data fusion). An enormous cross-matching dataset can introduce confusing matches, discrepancies in the point spread function (object resolution) inside or between data sets, adequate processing time, and data transit needs. A few arcseconds of astrometric tolerance are typically utilized when each database item lacks exact identification [48]. If the researcher chooses a method of collecting data based on a legitimate premise, he or she must weigh the method's strengths and weaknesses when analyzing their results. In qualitative research, it is critical that participants be recruited in a transparent manner [49].
- **Processing of data:** Data preprocessing, such as sample cuts in database searches, may be required during the data collection process. It is essential to use caution when preprocessing data because the input data can significantly affect many data mining approaches. For a specific algorithm, preprocessing can be divided into two types: procedures that make it meaningful for reading and processes that alter the data in some manner [48]. Data preprocessing includes the preparation and transformation of data so that it may be used in the mining process. Data preprocessing attempts to minimize data size, identify the relationships between data, normalize data, remove outliers, and extract characteristics from the data. Numerous methods have been proposed, such as cleaning, integration, transformation, and reduction of data sets [50].
- **Selection of Attributes:** Some properties of an object are not necessary for its proper functioning. To maximize performance, it is possible to use all the qualities of the object. Several low-density habitats and gaps have been created because of this. It is difficult to extract new ideas from data. As a result, dimension reduction is essential for retaining as much information as possible while using fewer attributes. Several algorithms are hampered by the presence of unnecessary, redundant, or otherwise unimportant features [48]. Filters and wrappers are prominent phrases used to describe the nature of the metric used to evaluate the value of attributes in a categorization. The accuracy estimates produced by the real target learning method are used by wrappers to evaluate attributes. Filters, on the other hand, work independently of any learning process and use generic properties of the data to evaluate attributes [51].

- **Use of Machine Learning Algorithms:** Machine learning algorithms are usually classified into supervised, semi-supervised, and unsupervised methods. Semi-supervised approaches use two sets of objects for which the target property, such as classification, is known with confidence. The algorithm was trained on these objects and then applied to others without the target characteristics. The test set included these additional items. In most astronomy cases, a photometric sample of objects can predict qualities that ordinarily require a spectroscopic sample. The parameter space spanned by the input attributes must span the time for which the algorithm is employed. This may appear restricted initially, but may often be overcome by merging data sets [48]. The research, development, and validation of algorithms for web service-based (possibly grid-based or peer-to-peer) classification and mining of distributed data are required. A combined text-numeric data mining algorithm may be the most effective, and thus has to be explored for these algorithms to be successful [45].

We explored the vast survey databases produced by numerous NASA missions while discussing the possibility of distributed data mining to assist astronomical research. GALEX conducts all-sky surveys in the near-UV and far-UV regions at various depths. Large-scale infrared survey are being carried out by the Spitzer Space Telescope, which includes areas of the sky that have already been extensively investigated by the Hubble Space Telescope (optical), Chandra X-ray Observatory, and other observatories [52]. The WISE mission, launched in 2009, conducted an infrared survey of the sky. Millions of stars and galaxies in the near-infrared region have been cataloged by the 2-Micron All-Sky Survey (2MASS). Numerous classes of astrophysical specimens can be studied with the using these wavebands [45].

5. Scientific Archives Services

The archives listed below are examples of time-domain astronomy. They map a large portion of the sky from a variety of observatories. Transient occurrences from more than a century of observations can be detected, owing to their extensive data library and more current images.

(i) **SIMBAD (Set of Identification, Measurements, and Bibliography for Astronomical Data):** SIMBAD [53][54] is the principal database for astronomical object identification and bibliography. e Centre de Donn´ees astronomiques de Strasbourg (CDS) developed and maintained the SIMBAD. Many astronomical objects are included in the database, bibliography, and selected observational measurements. Priority is given to catalogs and tables that span a wide range of wavelengths and serve large-scale research efforts. Meanwhile, systematic scanning of the bibliography provides an overview of current astronomical studies, including their diversity and broader trends. The World Wide Web (WWW) interface for Simbad is available at:http://simbad.u-strasbg.fr/Simbad.

(ii) **SMOKA (Subaru-Mitaka-Okayama-Kiso-Archive):** Multiple telescope data can be found in the SMOKA [55] science archive system. More than 20 million astronomical images are currently stored on the server, totaling more than 150 gigabytes. Additionally, the search interface can be used to perform searches based on various search restrictions and flexible image transport system (FITS)-Header keyword values for specific data sets. Data from telescopes and observatories from Subaru (Subaru), OAO (Okayama), Kiso (Kiso) and MITSuME (MITSuME) instruments and reduction tools can be accessed via the search interface.

(iii) **IRSA (NASA/IPAC Infrared Science Archive):** Several National Aeronautics and Space Act (NASA) [56] programs are supported by the Infrared Processing and Analysis Center (IPAC) [57], including Spitzer, the (NEO)WISE and 2MASS satellites, and the IRAS. IPAC also manages NASA's data archives. The IRSA also provides access to data from ESA missions, including Herschel and Planck, in collaboration with NASA. IRTF and SOFIA data will soon be archived at IRSA. IPAC's non-NASA or non-infrared initiatives, such as the Palomar Transient Factory (PTF),

Zwicky Transient Facility (ZTF), and Vera C. Rubin Observatory (VCRO), benefit from IPAC's archiving technology (formerly known as LSST) [58]. One petabyte of data from more than 15 projects can be found at IRSA. More than 100 billion astronomical measurements can be accessed through IRSA, including all-sky coverage.

An important aspect of astronomical research is the design, implementation, and archiving of large-scale surveys. Projects such as the LSST and PTF are expected to yield huge catalogs of stars and galaxies in the future. By integrating and cross-correlating data across these various survey dimensions in a virtual collection, these catalogs will dramatically boost science returns and enable new discoveries. A new paradigm for informatics is required to mine the riches of this data loop, which will require distributed database queries and data mining across virtual tables of decentralized, linked, and integrated sky survey catalogs. The problems that this subject presents are challenging, as they are in the majority of today's fields that produce large amounts of data at rapid rates [45].

Data from numerous celestial bodies can only be accessed by astronomers using query tools. Images and image-related information are the most frequently researched topics on the internet. Depending on these conditions, the user may have a wide range of requirements. Querying a single or group of items may be desired by users to retrieve information. To obtain precise information in the astronomical domain, users must develop extensive programs or formulate intricate queries. However, this has also resulted in a wide range of data storage interfaces and the loss of a uniform programming paradigm. Consequently, it is extremely difficult for a user to design applications that employ several data stores. Heterogeneous data sources, such as relational database management systems (RDBMS) or extensible markup language (XML), and database management system (DBMS) databases, have been studied extensively in the context of multi-database systems for many years in the astronomical domain [59].

It is possible to access numerous data stores in the cloud using state-of-the-art multi-database query processing systems. However, cloud operations differ when accessing data sources over a wide area network or the Internet. First, there are a variety of queries, including web data integration queries, such as those from price comparison sites, which can access many similar web data sources. By contrast, cloud data integration queries must access several different data stores. The users must be granted permission to access each data store. There are only a limited number of locations where mediators and data source wrappers can be placed [59].

For decades, large-scale data management has relied on parallel DBMS. In addition to typical relational DBMSs such as MySQL and Oracle, new data stores based on the (atomicity, consistency, isolation, durability (ACID)) [60] principles have recently been proposed to handle huge amounts of data. Large-scale data storage and warehousing systems such as Megastore, Mesa, and Spanner, have been developed with SQL-based query languages in mind. In addition, NewSQL databases are built for high-throughput online transaction processing (OLTP) while retaining ACID features. Many big data applications do not require rigorous ACID compliance and prioritize performance over consistency and reliability [61].

A wide variety of databases, data, and storage options are available to businesses today. The incompatibility of systems or the difficulty of developing new connectors and translators between them can impede the development of analytics and applications that work across these modalities [62]. This has led to the development of specialized multistore systems (also known as polystores) that enable integrated access to a number of cloud data stores via one or more query languages. It is difficult to evaluate different multistore systems because of their varied goals, topologies, and query processing methodologies [59]. Thus, the database community has come up with "Polystores" as a solution to the huge amount of data [63].

6. Polystores

A polystore system is a DBMS constructed on top of numerous integrated heterogeneous storage engines [64]. In federated databases, polystore manages various data models in many stores. It offers a single query language that can be used in various data models. Because they are built using numerous heterogeneous and interconnected storage engines, Polystore data management systems may work with various databases. It is possible to query various data models consistently with polystores. Polystores are required to quickly and efficiently manage information across multiple data models. Therefore, polystores are utilized for huge datasets or data models to handle data management solutions [63]. The polystore system is depicted in Figure 1.

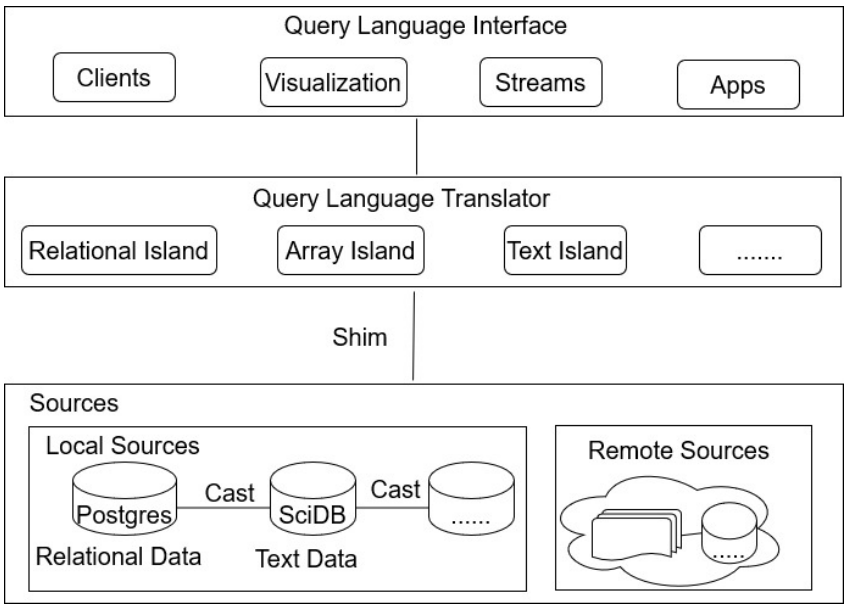

Figure 1. Polystore System

6.1. Existing Architecture of a Polystore Database System

The MIT BigDAWG (Big Data Analytics Working Group) Architecture contains a query mechanism for enormous multiple data sets in the MIMIC II medical domain. It comprises four layers: Database and Storage Engines, Islands, Middleware, API, and Applications. The initial releases of BigDAWG supported the open-source database engines PostgreSQL (SQL), Apache Accumulo (NoSQL), and SciDB (NEWSQL). Additionally, relational, array, and text islands were also supported. Figure 2 illustrates the architecture of the BigDAWG. The client is connected to the API through the middleware. The middleware receives a client query and passes it to the appropriate execution island(s). Shim translates and passes queries from each island to an appropriate database. Casts are used to migrate data across heterogeneous databases [65].

6.2. Major BigDAWG Components

The middleware or API comprises four parts: Planner, monitor,executor, and migrator. The planner element parses the entering query into an array of objects and creates query plan trees. Additionally, the planner component highlights potential data engines for each group of objects. These trees are then communicated to the monitor elements, which identify the optimal tree for each object group. The trees are then sent to the executor elements that assemble the collection of objects necessary to execute the query. Depending on the requirements of the query plan, the Executor element can use the migrator element to move objects across islands and engines [65]. The components of BigDAWG are illustrated in Figure 3.

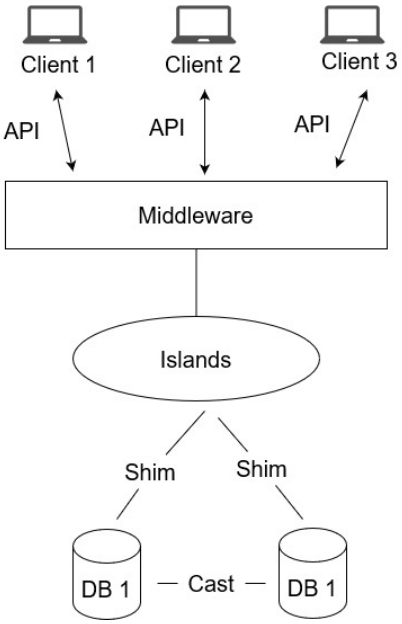

Figure 2. BigDAWG Architecture [65]

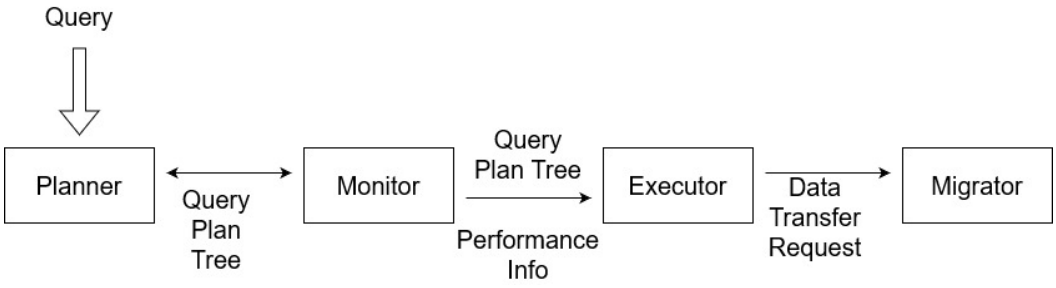

Figure 3. BigDAWG Components [63]

Query Endpoints communicates with users at a fundamental level, accepts queries, directs them to aggregation middleware, and returns the results. The catalog is a PostgreSQL engine that stores metadata about other engines, islands, datasets, and connectors. Interface middleware manages all these components. The initial release relies on Docker to simplify the installation and startup. The interface middleware can run on a server and connect to current database engines [63].

7. Challenge of the Future Management in Astronomical Data Archives

The current database management trends require the use of many models and data repositories. Previous models such as federated database systems (FDBS) and Data warehouses function well with relational data but are incapable of storing a large variety of data types (arrays, graphs, and images). Different data stores that manage various types of data have their own local languages.

Cooperative but autonomous databases comprise the FDBS [66]. With FDBS, local databases with decentralized control can gain more control over the information that they can exchange. Federated query agents (FQA) are used by FDBS to process queries. To store and execute queries, these agents can serve as intermediaries (mediators) between them two. FDBS data are stored in a relational database, which is a single data model supported by the system [67].

A data warehouse is a relational database built for analysis rather than for transaction processing. It normally comprises historical data collected from transactions, but can also contain data from other sources. It allows a company to combine data from multiple sources while separating the analytical and transaction workloads. Other programs that manage

the process of obtaining data and distributing it to business users can be found in a data warehouse environment. A central data store or warehouse controls the data warehouses [68].

Due to the rise of big data, the models such as FDBS and Data Warehouse seem to be inefficient as they can integrate only databases with a single data model that is no longer relevant. In addition, the volume and velocity of the data growth cannot be accommodated. These models also fall short in terms of their cost and performance. It is now safe to conclude that the numerous methods provided by earlier data integration models for managing heterogeneous data have failed. The database community has become increasingly interested in managing huge amounts of unstructured data from numerous heterogeneous data repositories. Due to the development in data size, rate of data incre-mentation, and appearance of new data types in various scientific data archives, this issue has received increased attention. The "one size fits all" approach [69] is no longer appropriate for modern database engineering. The underlying DBMS must also have complete autonomy to optimize queries. A model that can span heterogeneous data sources using a unified query language is required. Data virtualization via mediation is essential for meeting these needs.

Polystore can span many data management systems without requiring an underlying data location or storage engines, and it can be queried using a single language [63]. Polystore facilitate many-to-many interactions between information islands and data management systems across numerous distributed data models and query languages [65]. Polystores also provide seamless access to cloud data stores. The CloudMdsQL Polystore provides a functional SQL-like query language to access many data sources (relational, NoSQL, and HDFS) [70].

Polystore systems were recently proposed as a novel data integration approach that provides integrated access to heterogeneous data stores via a unified single query language. Moreover, Polystore Systems eliminate heterogeneity issues by implementing a communication protocol within the underlying database management systems via islands/shims, mediation, or APIs. Different types of data, such as text, graph, image, log data require different user interfaces. However, Polystore helps integrate all these data into a single query interface so that multiple models can be uniformly spanned.

8. Summary and Conclusions

We discussed all types of data, including big data, astronomical data, open data, and linked open data. These data are vast, heterogeneous, and complex. We also explored big data in astronomy and existing archives. In addition, large astronomical archives have also been investigated. This huge amount of data must be managed efficiently. These data can also be utilized for data mining to ensure accuracy and completeness. We also reviewed a few scientific archive services from various observatories, where transient events have been observed for over a century. Single-data-model models from the past, such as FDBS and data warehouses, are inefficient for managing huge volumes of data. Thus, to manage this huge amount of data with different goals, architectures, and query languages, the concept of Polystore was elaborated. Polystores utilizes a uniform query language to span several heterogeneous data models efficiently.

Most single data models or multi-data stores fail to handle huge amounts of data efficiently. Various scientific projects and scientific archives are available for research. The huge amount of data available in these resources can be mined, although they cannot be integrated. Polystores help integrate different types of data and query multiple models effectively for information retrieval, data visualization, and the development of useful online applications, thus solving the problem of heterogeneous data integration.

Author Contributions:

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. SAS. *Big Data*, 2022. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html.
2. Segal, T. *Big Data*, 2022. https://www.investopedia.com/terms/b/big-data.asp.
3. Tillett, B. RDA and the semantic web, linked data environment. *RDA and the Semantic Web, Linked Data Environment* **2013**, pp. 139–146.
4. Heath, T.; Bizer, C. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology* **2011**, *1*, 1–136.
5. Eibeck, A.; Zhang, S.; Lim, M.Q.; Kraft, M. A Simple and Effective Approach to Unsupervised Instance Matching and its Application to Linked Data of Power Plants.
6. Portal, L. *What is Linked Open Data?*, 2022. https://landportal.org/developers/what-is-linked-open-data.
7. Monaco, D.; Pellegrino, M.A.; Scarano, V.; Vicidomini, L. Linked open data in authoring virtual exhibitions. *Journal of Cultural Heritage* **2022**, *53*, 127–142.
8. Beno, M.; Figl, K.; Umbrich, J.; Polleres, A. Open data hopes and fears: determining the barriers of open data. In Proceedings of the 2017 Conference for E-Democracy and Open Government (CeDEM). IEEE, 2017, pp. 69–81.
9. of Public Expenditure, D. *What is open data?*, 2021. https://data.gov.ie/edpelearning/en/module1/#/id/co-01.
10. Zhang, Y.; Zhao, Y. Astronomy in the big data era. *Data Science Journal* **2015**, *14*.
11. Zhang, Y.; Zhao, Y. *Data mining in astronomy*, 2008. https://spie.org/news/1283-data-mining-in-astronomy?SSO=1.
12. Bose, R.; Mann, R.G.; Prina-Ricotti, D. Astrodas: Sharing assertions across astronomy catalogues through distributed annotation. In Proceedings of the International Provenance and Annotation Workshop. Springer, 2006, pp. 193–202.
13. Zakir, J.; Seymour, T.; Berg, K. Big Data Analytics. *Issues in Information Systems* **2015**, *16*.
14. Chathuranga, K. *Big Data in Astronomy*, 2018. https://doi.org/10.13140/RG.2.2.31794.96962.
15. York, D.G.; Adelman, J.; Anderson Jr, J.E.; Anderson, S.F.; Annis, J.; Bahcall, N.A.; Bakken, J.; Barkhouser, R.; Bastian, S.; Berman, E.; et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal* **2000**, *120*, 1579.
16. Bryant, A.; Raja, U. In the realm of Big Data. *First monday* **2014**, *19*.
17. Jena, M.; Behera, R.K.; Dehuri, S. Hybrid Decision Tree for Machine Learning: A Big Data Perspective. In *Advances in Machine Learning for Big Data Analysis*; Springer, 2022; pp. 223–239.
18. Schmidt, S.; Malz, A.; Soo, J.; Almosallam, I.; Brescia, M.; Cavuoti, S.; Cohen-Tanugi, J.; Connolly, A.; DeRose, J.; Freeman, P.; et al. Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). *Monthly Notices of the Royal Astronomical Society* **2020**, *499*, 1587–1606.
19. Robertson, B.E.; Banerji, M.; Brough, S.; Davies, R.L.; Ferguson, H.C.; Hausen, R.; Kaviraj, S.; Newman, J.A.; Schmidt, S.J.; Tyson, J.A.; et al. Galaxy formation and evolution science in the era of the Large Synoptic Survey Telescope. *Nature Reviews Physics* **2019**, *1*, 450–462.
20. Poudel, M.; Sarode, R.P.; Shrestha, S.; Chu, W.; Bhalla, S. Development of a Polystore Data Management System for an Evolving Big Scientific Data Archive. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*; Springer, 2019; pp. 167–182.
21. University, H. *Time Domain Astronomy*, 2021. https://www.cfa.harvard.edu/research/topic/time-domain-astronomy.
22. Unsöld, A.; Baschek, B. *The new cosmos: an introduction to astronomy and astrophysics*; Springer Science & Business Media, 2013.
23. of Technology, C.I. *Time Domain Astronomy*, 2021. https://www.growth.caltech.edu/tda.html.
24. Vaughan, S. Random time series in astronomy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2013**, *371*, 20110549.
25. Isadora Nun, P.P. *Feature Analysis for Time Series*, 2021. https://isadoranun.github.io/tsfeat/FeaturesDocumentation.html.
26. Kasliwal, M.; Cannella, C.; Bagdasaryan, A.; Hung, T.; Feindt, U.; Singer, L.; Coughlin, M.; Fremling, C.; Walters, R.; Duev, D.; et al. The growth marshal: a dynamic science portal for time-domain astronomy. *Publications of the Astronomical Society of the Pacific* **2019**, *131*, 038003.
27. Janesick, J.R.; Elliott, T.; Collins, S.; Blouke, M.M.; Freeman, J. Scientific charge-coupled devices. *Optical Engineering* **1987**, *26*, 268692.
28. Szalay, A.; Gray, J. Science in an exponential world. *Nature* **2006**, *440*, 413–414.
29. Projects, S. *Blink Comparator*, 2021. https://science-projects.org/portfolios/blink-comparator/.

30. Institution, S. *Blink Comparator*, 2022. https://airandspace.si.edu/multimedia-gallery/11363hjpg.
31. Sheehan, W. *Planets & perception: telescopic views and interpretations, 1609-1909*; University of Arizona Press, 1988.
32. Ragagnin, A.; Dolag, K.; Biffi, V.; Bel, M.C.; Hammer, N.J.; Krukau, A.; Petkova, D.S.M.; Steinborn, D. An online theoretical virtual observatory for hydrodynamical, cosmological simulations. *ArXiv e-prints* **2016**.
33. Law, N.M.; Kulkarni, S.R.; Dekany, R.G.; Ofek, E.O.; Quimby, R.M.; Nugent, P.E.; Surace, J.; Grillmair, C.C.; Bloom, J.S.; Kasliwal, M.M.; et al. The Palomar Transient Factory: system overview, performance, and first results. *Publications of the Astronomical Society of the Pacific* **2009**, *121*, 1395.
34. Stritzinger, M.; Leibundgut, B.; Walch, S.; Contardo, G. Constraints on the progenitor systems of type Ia supernovae. *Astronomy & Astrophysics* **2006**, *450*, 241–251.
35. Shrestha, S.; Poudel, M.; Wu, Y.; Chu, W.; Bhalla, S.; Kupfer, T.; Kulkarni, S. PDSPTF: polystore database system for scalability and access to PTF time-domain astronomy data archives. In *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*; Springer, 2018; pp. 78–92.
36. Bebek, C.; Coles, R.; Denes, P.; Dion, F.; Emes, J.; Frost, R.; Groom, D.; Groulx, R.; Haque, S.; Holland, S.; et al. CCD research and development at Lawrence Berkeley National Laboratory. In Proceedings of the High Energy, Optical, and Infrared Detectors for Astronomy V. International Society for Optics and Photonics, 2012, Vol. 8453, p. 845305.
37. Grillmair, C.; Laher, R.; Surace, J.; Mattingly, S.; Hacopians, E.; Jackson, E.; van Eyken, J.; McCollum, B.; Groom, S.; Mi, W.; et al. An overview of the palomar transient factory pipeline and archive at the infrared processing and analysis center. *Astronomical data analysis software and systems XIX* **2010**, *434*, 28.
38. Kulkarni, S. The intermediate palomar transient factory (iptf) begins. *The Astronomer's Telegram* **2013**, *4807*, 1.
39. Cao, Y.; Nugent, P.E.; Kasliwal, M.M. Intermediate palomar transient factory: Realtime image subtraction pipeline. *Publications of the Astronomical Society of the Pacific* **2016**, *128*, 114502.
40. Bellm, E. The Zwicky transient facility. In Proceedings of the The Third Hot-wiring the Transient Universe Workshop, 2014, Vol. 27.
41. Bellm, E.C.; Kulkarni, S.R.; Graham, M.J.; Dekany, R.; Smith, R.M.; Riddle, R.; Masci, F.J.; Helou, G.; Prince, T.A.; Adams, S.M.; et al. The Zwicky Transient Facility: system overview, performance, and first results. *Publications of the Astronomical Society of the Pacific* **2018**, *131*, 018002.
42. Masci, F.J.; Laher, R.R.; Rusholme, B.; Shupe, D.L.; Groom, S.; Surace, J.; Jackson, E.; Monkewitz, S.; Beck, R.; Flynn, D.; et al. The zwicky transient facility: Data processing, products, and archive. *Publications of the Astronomical Society of the Pacific* **2018**, *131*, 018003.
43. Raiteri, C.M.; Carnerero, M.I.; Balmaverde, B.; Bellm, E.C.; Clarkson, W.; D'Ammando, F.; Paolillo, M.; Richards, G.T.; Villata, M.; Yoachim, P.; et al. Blazar Variability with the Vera C. Rubin Legacy Survey of Space and Time. *The Astrophysical Journal Supplement Series* **2021**, *258*, 3.
44. Xi, S. *Large Synoptic Survey Telescope*, 2022. https://www.americanscientist.org/article/large-synoptic-survey-telescope.
45. Borne, K.D. Scientific data mining in astronomy. In *Next Generation of Data Mining*; Chapman and Hall/CRC, 2008; pp. 115–138.
46. Frawley, W.J.; Piatetsky-Shapiro, G.; Matheus, C.J. Knowledge discovery in databases: An overview. *AI magazine* **1992**, *13*, 57–57.
47. Fayyad, U.M. Data mining and Knowledge discovery in databases: Applications in Astronomy and Planetary Science. Technical report, American Association for Artificial Intelligence, Menlo Park, CA (United States), 1996.
48. Brunner, N.M.B..R.J. *Data Mining and Machine Learning in Astronomy*, 2010. https://ned.ipac.caltech.edu/level5/March11/Ball/Ball2.html.
49. Kairuz, T.; Crump, K.; O'Brien, A. Tools for data collection and analysis. *Pharmaceutical Journal (Vol 278)* **2007**.
50. Alasadi, S.A.; Bhaya, W.S. Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences* **2017**, *12*, 4102–4107.
51. Hall, M.A.; Holmes, G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data engineering* **2003**, *15*, 1437–1447.
52. Werner, M.W.; Roellig, T.; Low, F.; Rieke, G.H.; Rieke, M.; Hoffmann, W.; Young, E.; Houck, J.; Brandl, B.; Fazio, G.; et al. The Spitzer space telescope mission. *The Astrophysical Journal Supplement Series* **2004**, *154*, 1.
53. Shaw, R.A.; Hill, F.; Bell, D.J. Astronomical Data Analysis Software and Systems XVI. *Astronomical Data Analysis Software and Systems XVI* **2007**, *376*.
54. Wenger, M.; Ochsenbein, F.; Egret, D.; Dubois, P.; Bonnarel, F.; Borde, S.; Genova, F.; Jasniewicz, G.; Laloë, S.; Lesteven, S.; et al. The SIMBAD astronomical database-The CDS reference database for astronomical objects. *Astronomy and Astrophysics Supplement Series* **2000**, *143*, 9–22.
55. *SMOKA Science Archive*, 2022. https://smoka.nao.ac.jp/.
56. Kurtz, M.J.; Eichhorn, G.; Accomazzi, A.; Grant, C.S.; Murray, S.S.; Watson, J.M. The NASA astrophysics data system: Overview. *Astronomy and astrophysics supplement series* **2000**, *143*, 41–59.
57. Laher, R.R.; Surace, J.; Grillmair, C.J.; Ofek, E.O.; Levitan, D.; Sesar, B.; van Eyken, J.C.; Law, N.M.; Helou, G.; Hamam, N.; et al. IPAC image processing and data archiving for the Palomar Transient Factory. *Publications of the Astronomical Society of the Pacific* **2014**, *126*, 674.
58. *Science Data Center for Astrophysics Planetary Sciences*. https://www.ipac.caltech.edu/.
59. Bondiombouy, C.; Valduriez, P. Query processing in multistore systems: an overview **2016**.

60. Xia, Y.; Yu, X.; Butrovich, M.; Pavlo, A.; Devadas, S. Litmus: Towards a Practical Database Management System with Verifiable ACID Properties and Transaction Correctness.
61. Han, R.; John, L.K.; Zhan, J. Benchmarking big data systems: A review. *IEEE Transactions on Services Computing* **2017**, *11*, 580–597.
62. Gadepally, V.; Chen, P.; Duggan, J.; Elmore, A.; Haynes, B.; Kepner, J.; Madden, S.; Mattson, T.; Stonebraker, M. The BigDAWG polystore system and architecture. In Proceedings of the 2016 IEEE High Performance Extreme Computing Conference (HPEC). IEEE, 2016, pp. 1–6.
63. Patidar, R.G.; Shrestha, S.; Bhalla, S. Polystore Data Management Systems for Managing Scientific Data-sets in Big Data Archives. In Proceedings of the International Conference on Big Data Analytics. Springer, 2018, pp. 217–227.
64. MIT, B. *BigDAWG - Introduction and Overview*, 2022. https://bigdawg-documentation.readthedocs.io/en/latest/intro.htm.
65. Duggan, J.; Elmore, A.J.; Stonebraker, M.; Balazinska, M.; Howe, B.; Kepner, J.; Madden, S.; Maier, D.; Mattson, T.; Zdonik, S. The bigdawg polystore system. *ACM Sigmod Record* **2015**, *44*, 11–16.
66. Shrestha, S.; Bhalla, S. A Survey on the Evolution of Models of Data Integration. *International Journal of Knowledge Based Computer Systems8 (1 & 2), June & December* **2020**, pp. 11–16.
67. Poudel, M.; Shrestha, S.; Sarode, R.P.; Chu, W.; Bhalla, S. Query Languages for Polystore Databases for Large Scientific Data Archives. In Proceedings of the 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019, pp. 185–190.
68. Oracle. *Data Warehousing Concepts*, 1999. https://docs.oracle.com/cd/A84870_01/doc/server.816/a76994/concept.htm.
69. Stonebraker, M.; Çetintemel, U. " One size fits all" an idea whose time has come and gone. In *Making Databases Work: the Pragmatic Wisdom of Michael Stonebraker*; 2018; pp. 441–462.
70. Kranas, P.; Kolev, B.; Levchenko, O.; Pacitti, E.; Valduriez, P.; Jiménez-Peris, R.; Patiño-Martinez, M. Parallel query processing in a polystore. *Distributed and Parallel Databases* **2021**, *39*, 939–977.