The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2659 – 2666

The Need of a Unified File Format in Big Data Analysis

***Srihari Desai**
Dept. Of CSE
Bangalore Institute of Technology
Bangalore, India
*srihari.desai12@gmail.com
Tushar Kumar Chopra
Dept. Of CSE
Bangalore Institute of Technology
Bangalore, India
tusharkumarchopra@gmail.com
Yashas M B
Dept. Of CSE
Bangalore Institute of Technology
Bangalore, India
yashasmb2003@gmail.com
Suneetha K R
Professor, Dept. Of CSE
Bangalore Institute of Technology
Bangalore, India
suneetha.bit@gmail.com

Abstract— Big data is a larger, more complex data set extracted from different data sources. These enormous amounts of data may be utilized to solve several issues in industries like business, health, and technology that weren't previously solvable. Utilizing such a large and varied amount of data requires an effective management system. It consists of extraction of data, processing it to meet the requirements and providing required storage. Big data preprocessing constitutes a challenging task, as the previous existent approaches cannot be directly applied, since the size of the data sets or data streams make them unfeasible. As a result, data preparation has become more popular in cloud computing, and its contributions to the big data framework have been upgraded to include techniques like feature selection, imperfect data, imbalances learning, and instance reduction. The rise of technologies like machine learning, data analytics, and artificial intelligence, is altering the big data technology landscape. The use of these technologies in conjunction with big data allows businesses to improve their visualization capabilities, and make complex data more usable, and more accessible through visual representation. Big data framework is used to work with real-time data. It is crucial to maintain proper data file formats in to enable effective big data storage and exploitation. For data to be shared between different settings, file formats are crucial. Information, received from different sources, use different file formats. Hence, this paper tries to provide an idea to form a single platform in order to process variety of data.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2659 – 2666
Keywords— Big data,Data Abstraction,Data Management.

Introduction

The capacity to analyse massive datasets has become essential in a wide range of academic areas in an information technology era. Data created by machines, gadgets, cloud-based solutions, management consulting, and so on has reached massive proportions and is predicted to more than double in the future. Formally, big data is defined from 3Vs to 4Vs.The 3 V's are – Volume,Velocity and Variety.Volume refers to the huge amount of data.Velocity refers to the growth of data and how fast the data moves.Variety refers to all the structures,semi-structured and unstructured data generated. The fourth V, or veracity, stands for the reliability, accuracy, excellence, and accountability of the data.

Big data refers to massive volumes of information obtained from millions of individuals and kept throughout the internet in numerous locations. By relaxing theoretical model assumptions, controlling noisy data, preventing overfitting, and providing appropriate test data to validate models, big data helps data scientists overcome the obstacles associated with dealing with small data samples.

Data in a Data Warehouse system is loaded only when all three ETL procedures have been completed (Extract, Transform, Load). ETL pulls data from operational systems, changes it by executing data cleansing procedures, and then puts it into the Data Warehouse. ETL is not confined to data warehousing settings; it is a vital component of handling huge volumes of data, maintaining data quality and consistency, and allowing data-driven judgement in any company's IT system that incorporates proprietary applications and database systems.

Big data structure can be divided into three group - *structure,unstructured and semi-structure*.

Structured Data is a standardised format that has a well-defined structure. Structured data is organised in a table with column and row relationships. Structure data can be found in Excel files or SQL databases, as an example. A data model - an idea of how data is stored, retrieved, and processed — is required for structured data. Each field has its own identity and may be viewed independently or in combination with data from those other sections.

Unstructured data has an ambiguous form and cannot be stored or processed in traditional ways until it is converted to a structured format. Big data does not have a definite shape or organisation. Multimedia material such as audios, videos, and images can be used to represent unstructured data. It is essential to recognize that unstructured data is currently overtaking other types of big data. Every day, companies gather massive amounts of unstructured data and more, which are then utilised to train massive models using Deep Learning to solve some of the most complex real-world problems

Semi- structured data is information which falls somewhere between structured and unstructured data. It has some organisational structure, but not as much as structured data. In the digital age, this type of data, such as social network data, log files, and email communications, is becoming increasingly common. Semi-structured data might be difficult to interpret due to the various degrees of structure, yet it provides critical insights for businesses. Semi-structured data is frequently stored and processed using NoSQL databases,

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2659 – 2666
Hadoop, and Spark, and machine learning algorithms can be applied to analyse and provide information.

There are 3 main stages in Big Data Processing(ETL)

1. *Data Extraction:* At this early step, data is collected from a variety of source materials, such as enterprise , web sites, sensors, marketing tools, and transactional records. Data analysis specialists use both structured and unstructured data sources to extract information. During the extraction process, information from several sources may be merged, and the information's accuracy may also be checked. Verifying that the data has been properly sorted and gathered is the final step in this process, which provides a standard for future decision-making that will be improved.
2. *Data Transformation:* In second stage of big data processing, data is transformed or changed into the required format for additional analysis and display. The techniques used in this process include aggregation, normalisation, feature selection, binning, clustering, and idea hierarchy construction. These methods aid in the transformation of unstructured data into structured data and structured data into a more comprehensible format.
3. *Data Loading:* At this stage, the collected data is sent to a big data platform, such as a Hadoop Distributed File System (HDFS)(for structured data) or a NoSQL database(for non-structured data). In order to handle and analyse the data for important information, Big Data's data loading technique (ETL) aims to quickly, reliably, and scalably move data into the platform as the ETL process makes the process of loading automates.

Related Works

Big data is a category of information that is becoming more and more prominent in modern days. It is different from traditional data. The factors that determine big data are volume, velocity, variety (the 3V's), along with veracity and value. Their primary sources are mainly decentralized entities such as individuals and IoT devices [1].

Big data has become a part and parcel of multiple fields of science and technology like machine learning, artificial intelligence, data science [2] and even medicine. Big data analytics is also implemented in business which helps in improved agility and industrialization performance [3]. In the educational field, online learning and teaching activities produce huge amounts of data. Having found that only 10% of studies were done to include big data into the curriculum, many changes have been proposed to implement it.[4]. The medical sector makes use of the unstructured and structured data that is received from databases, emails, documents, transactions, devices, and sensors. Medical facilities are availing the perks of data-based health care. [5].

Big data, being a complex data set, requires proper management. This comes with its own set of problems. Their challenges are mainly seen in storage, data integration and analysis. Big data analytics involves the process of examining information to unravel hidden patterns [6]. A process called ETL is implemented to deal with big data. It involves 3 parts. A) Data extraction, B) Transformation techniques like normalizing, filtering, and sorting to clean the data and C) Migrating or loading data into a data warehouse [7].

Due to the lack of coordination between database systems and the analysis tools, many of the approaches in data mining that have been offered are typically unable to handle huge datasets

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2659 – 2666
satisfactorily. It is seen that every big data platform has its unique target, and different analytical methods, such as, quantum computing, data mining, future streaming, cloud computing, data mining and statistical analysis are used, as per the needs [8].

Big data preprocessing is a cumbersome task as the prevalent techniques cannot be used directly, owing to the size of the data sets or data streams. This has led to an increase of preprocessing of information in cloud computing along with an updated classification of data preprocessing variants under the big data framework which involves approaches like feature selection, imperfect data, imbalances learning and instance reduction [9].

It is laborious for the existing batch-processing techniques to change according to expanding data volume and significant real-time requirements, requiring the need for a framework to deal with real time data. Universal solutions with fair performance are generally economical than customized solutions [10].

Big data preprocessing involves data reduction, filtering noisy data and dealing with missing data. A similar concept is applied in discovering Smart Data, which is information of high quality, that is mined, which ensures sustainable storage. A technique, k-NN, is used to mine smart data. K-NN based preprocessing models have been carried out under Apache Spark, along with empirical analysis of its behaviors on numerous big data sets. [11].

Preprocessing can also be improved by choosing the appropriate File Format. Big data file formats consist of five categories, that are, text-based, row-based, column-based, in-memory and data storage services. In many cases, column-based file formats are seen to work better than the row-based ones, as only a few columns are retrieved during most queries. A combination of these formats is also used, based on the requirements [12].

File formats are also essential for creation of data marts. With big data, it is common to come across data having their own structure requiring the need of tools to understand and implement it. During data analysis, the availability of precise data becomes essential to tackle an issue. For this, data marts are used. The examination of file formats show that the Apache Parquet format is preferrable to store data in data marts.[13]

Certain evaluations on the file formats such as json, csv, orc, avro and parquet have also been conducted. These are used in the Hadoop framework when implementing the data storage and processing. The study consisted of stages of comparative and experimental evaluations. It also consisted of the development of an algorithm to choose the appropriate file format using tropical optimization methods.[14]

Text mining has also become an integral part of big data analytics, enabling users to derive patterns and knowledge from unstructured textual data. It is different from data mining as it doesn't deal with structured data. It plays a vital role in collecting and interpreting large scale, complex data that is present in a wide variety of disciplines and cultures.[15]

Big Data Frameworks

Big data frameworks deals with dataset sizes that are too large for use with conventional data preparation technologies. The necessity for the frameworks, required to handle huge amounts of data has increased along with the usage of big data.

Some of the majorly used frameworks are:

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2659 – 2666
1. *Hadoop*—Large data sets and distributed storage may both be handled by this open-source batch processing solution. The Hadoop framework relies on computer clusters and modules that were developed with the understanding that hardware malfunctions are inevitable and should be handled as such by the system.
2. *Storm*—Big data processing framework Apache Storm enables applications to be organised as directed acyclic networks. It can handle more than 10,00,000 tuples per second per node and is efficient at managing unbounded data streams regardless of the language used. It is quite adaptable.
3. *Spark*—A well-known and rising big data framework is Apache Spark. It has an expressive and user-friendly application programming interface and is a quick, in-memory data processing engine. With rapid access to datasets, data workers may readily carry out operations like structured query language, machine learning, or streaming.
4. *Hive*—Facebook created Apache Hive to combine the features of well-known large data frameworks. SQL queries are transformed into MapReduce jobs using this engine. Executor, Optimizer, and Parser are a few of the components that make up the Apache Hive engine. In order to analyse massive amounts of data, it may be connected with Hadoop.
5. *MapReduce*—Google initially launched MapReduce, a web engine for the Hadoop framework, in 2004 as a method for handling enormous amounts of data. Since then, it has transformed into the MapReduce data processing technology that we are familiar with. Map, Shuffle, and Reduce are the three stages of the engine's data processing.
6. *Presto*—Presto is an open platform and distributed SQL engine that enables interactive analytics on gigabyte- to petabyte-scale data sources. It enables data searching in relational databases, Cassandra, and Hive in addition to private data storage.
7. *Heron*—Twitter created Apache Heron, a large data tool, as a modern alternative to Storm. It is designed to be used for ETL, trend analysis, and continuous spam detection operations.
8. *Flink*—A leading open-source big data framework for stream processing massive data is Apache Flink. It is a dependable, unceasingly accessible, and powerful platform for data streaming applications. It has excellent throughput and latency, is fault-tolerant and stateful, and can recover from errors.
9. *Kudu*—An innovative new storage component called Apache Kudu was created to improve intricate pipelines in the Hadoop ecosystem. It is a system that resembles SQL and supports both sequential and random writes and reads.
10. *Samza*—Samza is a LinkedIn-developed fully accessible streaming data processing tool. Streaming, Execution, and Processing are its three tiers. Samza offers batch processing, high performance, horizontal scalability, ease of use, and a pluggable architecture for streaming data. ADP, VMWare, Expedia, and Optimizely are a few businesses that use Samza.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 11 Issue 1, 2023
pp: 2659 – 2666

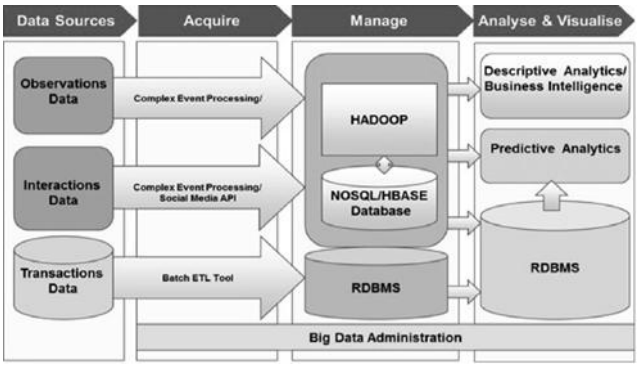

Fig 1. Big Data Architecture

File Formats

Data may be present in various forms. Some are structured, while others may be semi-structured or unstructured. In order to efficiently store such varieties of data, file formats are used. This helps us to obtain a proper understanding of the data sets. For Big Data, many file formats are used based on the requirements. The following are the majorly used ones:

CSV—A good option for compatibility, spreadsheet processing, and data that is readable by users. The data must be flat. It is ineffective and unable to manage nested data. There could be problems with the separator, which could affect the quality of the data. This is used t for small data sets, Custom software, or exploratory research.

JavaScript Object Notation (JSON)—It is a nested design which is frequently used in APIs. It is suitable for landing data, tiny data collections, or API integration. It can be converted to a more efficient file format to process significant amount of data.

Avro—It is effective at storing row data. Both evolution and a schema are supported. Kafka integration is excellent. Allows for file splitting. Used in Kafka or for row-level operations. Faster to write data than read it.

Parquet—Schemas are supported by Parquet. It is a great solution to store columnar data in deep storage that can be accessed via SQL, and it works well with Hive and Spark. As data is stored in columns, query engines read files that contain the chosen columns, thereby reducing query time.

Optimized Row Columnar (ORC)—It offers better compression and is similar to Parquet as shown in the below Table 1. It is less popular but offers stronger support for schema evolution as well.

Table 1.Comparison of File Formats in Big Data

	Avro	***Parquet***	***ORC***
Schema Evolution	Best	Good	Better
Compression	Good	Better	Best
Splittability	Good	Good	Best
Row or Column	Row	Column	Column
Read or Write	Write	Read	Read

Although the mentioned file formats help in processing and organizing data, choosing the right file format is not easy. As seen in Table 1., the file formats vary in their fields of expertise and don't provide a complete solution. This leads to a compromise, which might not bring about the complete essence of the data. Hence, a new file format is to be developed, that allows the implementation of a variety of data under a single model.

Conclusion

This paper provides the details about big data and the various frameworks along with the file formats needed to process these huge datasets. Big Data was created as a result of the emergence of novel technologies and the rapid data collection that resulted from them. In the age of information technology, the ability to process huge data effectively has become vital for a variety of academic and scientific sectors. Given that many technologies in a diverse technological environment depend on data sharing in order to function, it is easy to understand the significance of big data file formats. The usage of standard data file formats is the way for such data exchange that is most optimal. There are a wide range of tools and frameworks that indeed help to process this massive amount of data. It is difficult to manage data from different file formats, hence it is necessary to find a solution to integrate or merge the data into one file format. Future work in this area involves finding a solution to unify or find a unique file format in which data can be represented.

References

[1] Hiba Alsghaier, Mohammed Akour, Issa Shehabat, Samah Aldiabat. The Importance of Big Data Analytics in Business: A Case Study. American Journal of Software Engineering and Applications. Vol. 6, No. 4, 2017, pp. 111-115. doi: 10.11648/j.ajsea.20170604.12

[2] Sun Z, & Huo Y (2021) The spectrum of big data analytics. Journal of Computer Information Systems 61(2): 154-162. DOI. 10.1080/08874417.2019.1571456.

[3] Hiba Alsghaier, Mohammed Akour, Issa Shehabat, Samah Aldiabat. The Importance of Big Data Analytics in Business: A Case Study. American Journal of Software Engineering and Applications. Vol. 6, No. 4, 2017, pp. 111-115. doi: 10.11648/j.ajsea.20170604.12

[4] Baig, M.I., Shuib, L. & Yadegaridehkordi, E. Big data in education: a state of the art, limitations, and future research directions. Int J Educ Technol High Educ 17, 44 (2020). https://doi.org/10.1186/s41239-020-00223-0

[5] Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. J Big Data. 2022;9(1):3. doi: 10.1186/s40537-021-00553-4. Epub 2022 Jan 6. PMID: 35013701; PMCID: PMC8733917.

[6] Aburawi, Yousef & Albaour, Abdulbaset. (2021). Big Data: Review Paper. International Journal Of Advance Research And Innovative Ideas In Education. 7. 2021.pp 729-734

[7] Aqlan, Faisal & Nwokeji, Joshua. (2018). Big Data ETL Implementation Approaches: A Systematic Literature Review. Conference of 2018 Software Engineering and Knowledge Engineering, 10.18293/SEKE2018-152.

[8] D. P. Acharjya, Kauser Ahmed P, ” A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools”, February 2016 International Journal of Advanced Computer Science and Applications 7(2):511-518 DOI:10.14569/IJACSA.2016.070267.

[9] Salvador García , Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, Francisco Herrera, ” Big data preprocessing: methods and prospects”, November 2016. Big Data Analytics 1(1) DOI:10.1186/s41044-016-0014-0.

[10] Sowmya R, Suneetha K R, ”Data Mining with Big Data”, January 2017 DOI:10.1109/ISCO.2017.7855990 Conference: 2017 11th International Conference on Intelligent Systems and Control (ISCO).

[11] Triguero, Isaac & García-Gil, Diego & Maillo, Jesús & Luengo, Julián & García, Salvador & Herrera, Francisco. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 9. e1289. 10.1002/widm.1289.

[12] Samiya Khan, Mansaf Alam, ” File Formats for Big Data Storage Systems”, October 2019. PP: 2906-2912 /Volume-9 Issue-1, October 2019 /DOI: 10.35940/ijeat.A1196.109119

[13] Belov, Vladimir & Kosenkov, Alexander & Nikulchev, Evgeny. (2021). Experimental Characteristics Study of Data Storage Formats for Data Marts Development within Data Lakes. Applied Sciences. 11. 8651. 10.3390/app11188651.

[14] Belov, Vladimir & Tatarintsev, Andrey & Nikulchev, Evgeny. (2021). Comparative Characteristics of Big Data Storage Formats. Journal of Physics Conference Series. 1727. 012005. 10.1088/1742-6596/1727/1/012005.

[15] Hassani, Hossein & Beneki, Christina & Unger, Stephan & Mazinani, Maedeh & Yeganegi, Mohammad. (2020). Text Mining in Big Data Analytics. Big Data and Cognitive Computing. 4. 1. 10.3390/bdcc4010001