A STUDY ON BIGDATA: CHALLENGES, TECHNOLOGIES AND TOOLS

Abstract

In modern days data is generated at a substantial pace. Analyzing these data is difficult for the common man. Meantime the data may be varied and also it'll growing up every second because these data are generated from the web, socialmedia, etc... So here we need heterogeneous support and huge space storing or occupying applications for data analysis. Big data is a better solution for this problem. It has large collections of the dataset and supports different kinds of data (Structured, Unstructured and Semi-Structured data). This paper impacts the overview of big data and also focuses on the importance, challenges, tools and technologies that will be worthwhile in future research.

Keywords: Bigdata, Data growth, Data analysis, Hadoop

Authors

S. Priya
Research Scholar
Department of Computer and Information Science
Annamalai University
Tamil Nadu, India.
priyas2vr@gmail.com

Dr. A. Subashini
Assistant Professor
Department of Computer Application
Government Arts College
C. Mutlur, Chidambaram, Tamil Nadu, India.

Copyright © 2024 Authors

I. INTRODUCTION

Big data is a large collection of datasets in different formats. Like data in structured, unstructured and semi-structured. These data are not processed, which uses traditional techniques for computing. The main difficulty in handling such massive data because that the volume is increasing rapidly in comparison to the computing resources. The Big data term which is being used nowadays is kind of mistermed as it points out only the size of the data not putting too much attention to its other existing properties. Big data can be defined with the following characteristics,

1. **Variety:** It provides the accessing services to the different formats of data like structure and structured and semi-structure

2. **Velocity:** , Today world internet users may increase the speed of data access is most important in big data provide a better speed of data in data delivery

3. **Volume:** It is an important feature for accessing data storage or the space conception which has huge storage for a large data set.

This paper focuses on challenges in big data and its available techniques. So, to elaborate on this, the paper is sliced into the following sections. Sections 2 deals with challenges that originate during fine-tuning of big data. Section 3 furnishes the emerging technologies that will help us to process big data. Section 4 provides insight into big data tools. Conclusion remarks are provided in section 5 to summarize outcomes.

II. CHALLENGES OF BIG DATA

1. **The quick Growth of Data:** Every day every search more data are growing up and data are relevant and useful for further analysis.

2. **Storage:** An organization, side have a large amount of data those data are complex and also operational data there are difficult to store preserve and manage or handle without any technologies and supporting tools.

3. **Syncing Across Data Source:** To import data from different data sources that time we compare those data because some data or not up to date so comparison of different data sources is an important task that is done by big data easily.

4. **Security:** We have a huge amount of data that can be easily targeted for suspicious activities or persistent thread for the challenges of this issue big data provide better data security through proper encryption and authentication services.

5. **Unreliable Data:** Bigdata support different structured data so it cannot be 100% accurate it might have redundant or inconsistent data.

Copyright © 2024 Authors

6. **Skill and Professional Availability:** The date of growth is inverse the big data concept has also been implemented in many industries but the technical person available is less in the present situation because the skilled person only can recover if any problem occurred in Big Data System.

III. EMERGING TECHNOLOGIES FOR BIGDATA

Providing more accurate analysis is the most important technology of big data intelligence these techniques are as follows,

1. **Operational Big Data:** This included system provides operation capabilities and interactive workload for real-time problems. Which data are previously captured and stored? NOSQL data system or design to take features of new downloadcomputing architecture. Some systems of NOSQL provide new view patterns and friends inreal-time and they do not need the date of a scientist or additional features.

2. **Analytical Bigdata:** These systems are massive parallel processing (MPP) database systems and MapReduce which provide analytical capabilities. For complex analysis, a new method of analytical data used in maps reduces these capabilities or is provided by SQL. MapReduce can be from a single server to thousands of high and low-end machines.

Table 1 : Comparison of Operational Vs Analytical Bigdata

	Operational Big data	**Analytical Big data**
Delay	1ms-100ms	1min-100min
Concurrency	1000-100000	1-10
Access Pattern	Write and Read	Read-only
Queries	Selective	Un-Selective
Data Scope	Operational	Retrospective
End-user	Customer	Data Scientist
Technology	NOSQL	Map reduce and MPP Database

3. **Data Science:** A combination of statistics, mathematics, programming, problem-solving and ingeniously capturing data is called data science and also it provides the services of cleansing, preparing and alignment of the data.

- **Applications of Data Science**

  - **Internet Search:** Search engines make use of data science algorithms to deliver the best result of search queries in a fraction of a second.

  - **Digital Advertisement:** The entire digital marketing spectrum uses the data science algorithm from digital banners to digital billboards. This is the main reason for Digital ads getting higher CTR than traditional advertisements.

Copyright © 2024 Authors

- **Recommendation System:** In recommendation system not only make it easy to find a relevant product from the billion product available but also add a lot to use experience a lot of companies use this system to promote their product and suggestion by the user's demand and relevant information. The recommendations systems are based on uses previous search results.

4. **Hadoop:** Apache open source Framework written by Java that allows distributed processing of large data sets across clusters of computers using programming models. The Apache Framework application works in employment that provides distributed storage and computation across a group of computers how does this design scale from a single server to thousands of machines each offering local competition and storage?

- **Benefits of Hadoop**

  - **Scalable :** Hadoop is an extremely scalable storage technology since it can distribute and store a huge data sets across thousands of inexpensive servers and operates in parallel like traditional database systems that can't scale the process of a large amount of database business to run applications on thousands of node or elements involving many thousands of terabyte of data.

  - **Cost-effective:** It offers here cost-effective storage solution for business exploding data set this is the important advantage of Hadoop in this problem with traditional relational database management system that is expensive and extremely cost of a beat to scale to process such massive volume of data reduce cost many companies in the past would have had to down sample data and clarified based on certain assumption as to which data was the most valuable.

  - **Flexible:** It enables businesses to easily Access new data sources and tap into different types of data to generate volume from the data this means the businesses can use Hadoop to drive valuable business inside from data sources, such as social media and email conversations.

  - **Immediate:** Distributed file system-based method is a unique feature of Hadoop. Faster data processing is how does can efficiently process and also it is dealing with a large volume of unstructured data. It efficiently processes terabytes of data in just a minute and petabytes in hours.

  - **Resilient to Failure:** Fault tolerance is the main key advantage of using the individual note that data is also replicable to the other note in the cluster which means that in the event of failure, there is another copy available for use.

- **Hadoop Ecosystem:** Hadoop is an open source framework designed to make working with big data easier. It has found a home in businesses and industries that need working with enormous data sets. They are delicate and require handling that is effective or efficient. As a framework, Hadoop is made up of a number of modules that are supported by a sizable ecosystem of technologies, allowing it to process enormous data sets that are located in or belong to clusters. Apache Hadoop consists

Copyright © 2024 Authors

of other commercial tools and solutions in addition to Apache initiatives. Hadoop is made up of four main parts: HDFS, Map Reduce, YARN, and Hadoop Common.

Most tools are utilized as a supplement to or as a support for these important components. All of these tools function properly to deliver services.

- HDFS
- Map Reduce
- YARN and
- Hadoop Common

Most of the tools are used to supplementary or support these major elements all these tools work correctly to provide services observation storage and maintenance of data.
Collective components of Hadoop ecosystems are,

- HDFS
- Map Reduce
- YARN
- Spark
- PIG and HIVE
- HBASE
- Oozie
- Mahout and SparkMLLib
- Zookeeper
- Solar,Lucene

❖ **HDFS [** Hadoop Distributed File System**]:** The main part of the Hadoop ecosystem is responsible for storing large data sets of structured unstructured data across various nodes and maintaining the master data in the form of log files.

HDF consists of two components
- Name node
- Data node

- **Name Node:** It is the main node which consists of metadata (data about data) requiring fewer resources than the data nodes that store the actual data.

- **Data Node:** In Hadoop Distributed File System the data nodes maintain all the coordination between the clusters and the hardware does working at the heart of the system.

❖ **YARN:** Yet another resource negotiator, Yarn is the one who aids in managing the resources throughout the cluster, as the name suggests. It manages resource allocation and scheduling for the Hadoop system.
Three major components of YARN,
- Resource manager
- Node manager and
- Application manager

Copyright © 2024 Authors

- The **resource manager** has the right of allocating resources for the application in your system.

- The **node manager** works on the allocation of resources such as CPU memory bandwidth machines and later on acknowledgement the resource manager.

- The application manager functions as an interface, negotiating on behalf of the resource manager and node manager.MapReduce

- Map reduce uses the distributed and parallel algorithm map-reduce to enable the transfer of processing logic and support the creation of applications that reduce a large amount of data to a manageable amount. Map reduce uses two functions.,

The first one is **Map** and the second one is **Reduce**,

- **MAP() :** Map performance of data and organizing them in the form of a group map generates key-valuepair-based results, which are later on processed by the reduced method. **Reduce()**

  Reduce as the name suggest does the summarization by aggregating the map data in simple reduce takes the output generated by the map as input that combines those couples into the smallest set of tuples.

- **HBase:** HBase is a NoSQL database or non-relational database, which is important and mainly used when you need random, real-time, read, or write access to your Big Data. It provides support to a high volume of data and high throughput. In an HBase, a table can have thousands of columns.

- **Pig:** It is a procedural language platform used to develop a script for Map Reduce operations. It is a high-level data flow platform for executing Map Reduce programs of Hadoop. The language used for Pig is Pig Latin, this is a data flow language used by Apache Pig to analyze the data in Hadoop. It is a text-based language that transforms the Java Map Reduce programming paradigm into a notation. The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Apart from that, Pig can also execute its job in Apache Tez or Apache Spark. A pig can handle multiple types of data, i.e., structured, semi-structured or unstructured and stores the corresponding results in the Hadoop Data File System.

- **Hive:** It is a tool for creating SQL-like scripts that perform Map Reduce processes. A Hadoop infrastructure utility for processing structured data is called Hive. It is built on top of Hadoop to condense Big Data and facilitate easy querying and analysis. Facebook originally created Hive; afterwards, the Apache Software Foundation picked it up and continued to work on it as an open source project under the moniker Apache Hive. It is utilized by various businesses. Amazon utilizes it, as an illustration, in Amazon Elastic Map Reduce. A relational database, an OLTP system, and a language for real-time queries and row-level changes are not all part of the Hive design.

Copyright © 2024 Authors

IV. TOOLS FOR BIGDATA

We have huge varieties of information acid from creative data processing and are cost-effective which make better decision making rather than traditional processing method. Software utilities are big data Technologies that are included to analyze process and extract information from large data sets. The top technologies are,

1. **Apache Hadoop:** Created by Apache Software Foundation which is a Java-based open source Framework for analyzing and storing large items of the dataset. It provides your distributed stories in prospector by using map-reduce programming methodology.

2. **MongoDB:** MongoDB is an open-source document-oriented and cross-platform database. It provides basic services of storing and accessing Huge data and especially maintaining high performance available and scalability of data MangoDB is classified as a NoSQL database because which do not store under the drive data in table format.

3. **Rain Store:** This is the database management system that handles massive data and also eliminates duplicate files.

4. **Apache Cassandra:** Which is an open-source distributed with no SQL database apache cassandra allows in-depth analysis of real-time data with no compromises for high scalability and availability of performance.

5. **Presto:** Facebook created Presto with an open-source SQL query engine. It allows interactive query analysis on large amounts of data. This supports quick analysis because it is a distributed search engine. The range of search is gigabytes to petabytes.

6. **Rapid Miner:** It is a predictive analytical data mining application and also it is an open source. This enables data scientists and big data analysts to quickly analyze their data because it has a strong data science platform. In addition to data mining, it supports model operations and deployment of modelling.

7. **Elastic Search:** This is one of the open-source analytical model search engines. It supports distributed processing and also it is based on Apache Lucene. Elastic search allows index searching and analyzes all types or kinds of data. The most common uses of Elastic searches are log analytical, full-text search, business analytics operation intelligence and security intelligence.

8. **Kafka :** Apache Kafka is a popular open-source event Store streaming technology in Java and Scala by Apache Foundation. Thousands of industries rely on the streaming analytical platform, mission-critical data integration application.

9. **Plunk:** Sophisticated and scalable software platform which get the data from website sensor and other sources to perform analysis and visualize machine generator data. It offers matrices under-diagnosis problems to the cooperated processing. It has a searchable repository so they can easily generate an alert, report graph and visualisation of the dashboard which indexes and correlated the real-time data.

Copyright © 2024 Authors

10. **KNIME:** Generally referred to as **Konstanz Information Minor** is a platform for data analysis reports and integration data in open access.

V. CONCLUSION

Big data plays an essential role in data analysis and is a major part of bigdata analytics to assure perfect results from complex datasets. In this paper, we shortly reviewed various technologies and tools from its inception to the future. This review puts focus on the tropical and auspicious areas of big data. This paper provides a new viewpoint on research regarding big data.

REFERENCES

[1] D. P. Acharjya Kauser Ahmed P, A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016

[2] Nada Elgendy and Ahmed Elragal, Big Data Analytics: A Literature Review Paper(https://www.researchgate.net/profile/Ahmed-Elragal / publication / 264555968_Big_Data_Analytics_A_Literature_Review_Paper / links / 541e9b9a0cf203f155c0655a/ Big-Data-Analytics-A-Literature-Review-Paper.pdf)

[3] Jasmine Zakir, Tom Seymour, Kristi Berg, BIG DATA ANALYTICS, Issues in Information SystemsVolume 16, Issue II, pp. 81-90, 2015 (https://doi.org/10.48009/2_iis_2015_81-90)

[4] T. K. Das and P. M. Kumar, Big data analytics: A framework for unstructured data analysis, International Journal of Engineering and Technology, 5(1) (2013), pp.153-156.

[5] Sofiya Mujawar, Aishwarya Joshi.” Data Analytics Types, Tools and their Comparison” IIJARCE 2015Vol. 4, Issue 2, pp. 488-491

[6] Sofiya Mujawar, Aishwarya Joshi.” Data Analytics Types, Tools and their Comparison” IIJARCE 2015Vol. 4, Issue 2, pp. 488-491

[7] https://www.tutorialspoint.com/hive/hive_introduction.htm

Copyright © 2024 Authors