International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 2, No 02, 2024. ISSN: 2348-6600
Reference ID: IJCS-506 PAGE NO: 3507-3514

CHALLENGES, RESEARCH ISSUES AND TOOLS INVOLVED IN BIG DATA AND HADOOP

Dr.K.S.GOMATHI
Principal & Head,
Department of Computer Science & Computer Applications,
Madurai Gandhi NMR Subbaraman College for Women,
Madurai, Tamil Nadu, India.

R.MAHALAKSHMI
Assistant Professor,
Department of Computer Science,
Madurai Gandhi N.M.R.Subbaraman College for Women,
Madurai, Tamil Nadu, India.

ABSTRACT

A huge repository of terabytes of data is generated each day from modern information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these massive data requires a lot of efforts at multiple levels to extract knowledge for decision making. Therefore, big data analysis is a current area of research and development. The basic objective of this paper is to explore the potential impact of big data challenges, open research issues, and various tools associated with it. As a result, this article provides a platform to explore big data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.

Keywords: Big Data Analytics; Hadoop; Massive Data; Structured Data; Unstructured Data.

I. INTRODUCTION

In digital world, data are generated from various sources and the fast transition from digital technologies has led to growth of big data. It provides evolutionary breakthroughs in many fields with collection of large datasets. In general, it refers to the collection of large and complex datasets which are difficult to process using traditional database management tools or data processing

All Rights Reserved ©2024 International Journal of Computer Science (IJCS Journal)
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 2, No 02, 2024. ISSN: 2348-6600
Reference ID: IJCS-506 PAGE NO: 3507-3514

applications. These are available in structured, semi-structured, and unstructured format in petabytes and beyond. Formally, it is defined from 3Vs to 4Vs. 3Vs refers to volume, velocity, and variety. Volume refers to the huge amount of data that are being generated everyday whereas velocity is the rate of growth and how fast the data are gathered for being analysis. Variety provides information about the types of data such as structured, unstructured, semi- structured etc. The fourth V refers to veracity that includes availability and accountability.

I. CHALLENGES IN BIG DATA ANALYTICS

Recent years big data has been accumulated in several domains like health care, public administration, retail, bio-chemistry, and other interdisciplinary scientific researches. Web-based applications encounter big data frequently, such as social computing, internet text and documents, and inter- net search indexing.

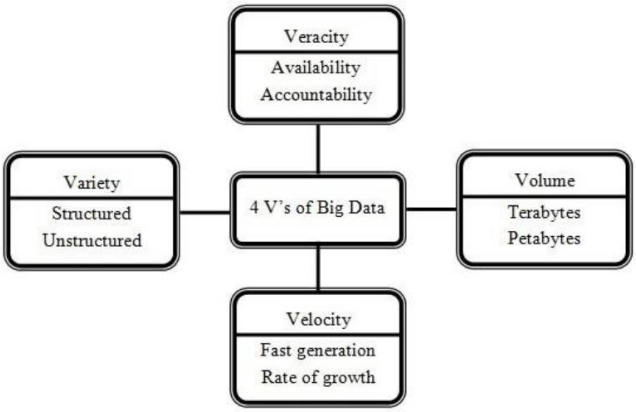


Fig. 1: Characteristics of Big Data

A. DATA STORAGE AND ANALYSIS

In recent years the size of data has grown exponentially by various means such as mobile devices, aerial sensory technologies, remote sensing, radio frequency identification readers etc. These data are stored on spending much cost whereas they ignored or deleted finally becuase there is no enough space to store them. Therefore, the first challenge for big data analysis is storage mediums and higher input/output speed. In such cases, the data accessibility must be on the top priority for the knowledge discovery and representation. In past decades, analyst use hard disk drives to store data but, it slower random input/output performance than sequential input/output. To overcome this limitation, the concept of solid state drive (SSD) and phrase change memory (PCM) was introduced. However the available storage

All Rights Reserved ©2024 International Journal of Computer Science (IJCS Journal)
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 2, No 02, 2024. ISSN: 2348-6600
Reference ID: IJCS-506 PAGE NO: 3507-3514

technologies cannot possess the required performance for processing big data.

Another challenge with Big Data analysis is attributed to diversity of data. With the ever growing of datasets, data mining tasks has significantly increased. Additionally data reduction, data selection, feature selection is an essential task especially when dealing with large datasets.

B. KNOWLEDGE DISCOVERY AND COMPUTATIONAL COMPLEXITIES

Knowledge discovery and representation is a prime issue in big data. It includes a number of sub fields such as authentication, archiving, management, preservation, information retrieval, and representation. There are several tools for knowledge discovery and representation such as fuzzy set , rough set , soft set , near set etc The most popular approach in case of large dataset management is data warehouses and data marts. Data warehouse is mainly responsible to store data that are sourced from operational systems whereas data mart is based on a data warehouse and facilitates analysis.

C. SCALABILITY AND VISUALIZATION OF DATA

The objective of visualizing data is to present them more adequately using some techniques of graph theory. Graphical visualization provides the link between data with proper interpretation. However, online marketplace like flipkart, amazon, e-bay have millions of users and billions of goods to sold each month. This generates a lot of data.

Information Security

In big data analysis massive amount of data are correlated, analyzed, and mined for meaningful patterns. All organizations have different policies to safe guard their sensitive information. Preserving sensitive information is a major issue in big data analysis. There is a huge security risk associated with big data . Therefore, information security is becoming a big data analytics problem. Security of big data can be enhanced by using the techniques of authentication, authorization, and encryption.

Various security measures that big data applications face are scale of network, variety of different devices, real time security monitoring, and lack of intrusion system.

II. OPEN RESEARCH ISSUES IN BIG DATA ANALYTICS

Data science aims at researching big data and knowledge extraction from data. Applications of big data and data science include information science, uncertainty modeling, uncertain data analysis, machine learning, statistical learning, pattern recognition, data warehousing, and signal

All Rights Reserved ©2024 International Journal of Computer Science (IJCS Journal)
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 2, No 02, 2024. ISSN: 2348-6600
Reference ID: IJCS-506 PAGE NO: 3507-3514

processing. The research issues pertaining to big data analysis are classified into three broad categories namely internet of things (IoT), cloud computing, bio inspired computing, and quantum computing.

A. IoT FOR BIG DATA ANALYTICS

Internet has restructured global interrelations, the art of businesses, cultural revolutions and an unbelievable number of personal characteristics. Internet of Things is attracting the attention of recent researchers for its most promising opportunities and challenges.

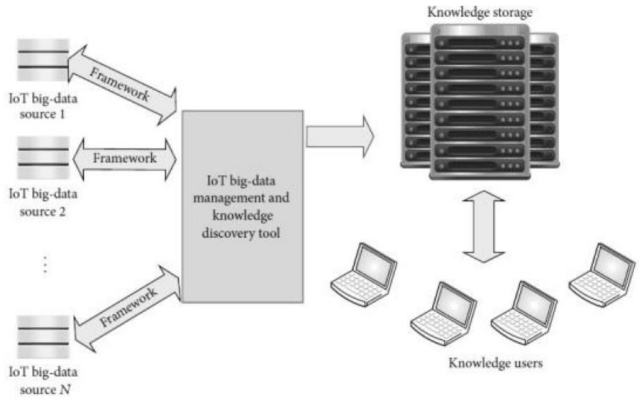


Fig. 2: IoT Big Data Knowledge Discovery

The discovered knowledge is stored in knowledge bases and expert systems are generally designed based on the discovered knowledge. The final phase is to apply discovered knowledge in various applications.

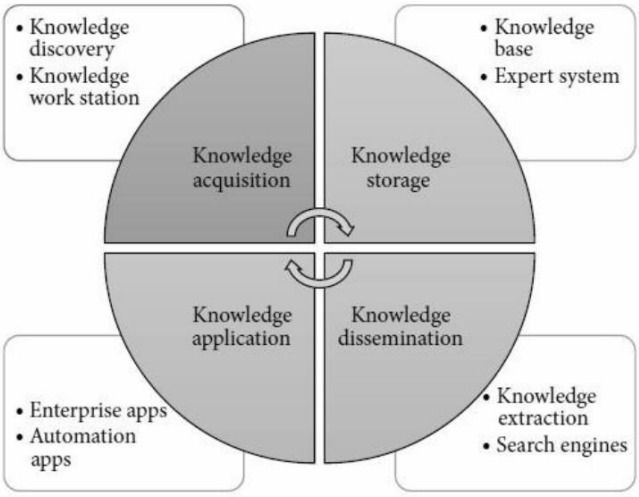


Fig. 3: IoT Knowledge Exploration System

B. CLOUD COMPUTING FOR BIG DATA ANALYTICS

Computing infrastructures that are hidden in virtualization software make systems to behave like a true computer, but with the flexibility of specification details such as number of processors, disk space, memory, and operating system Cloud computing harmonize massive data by on- demand access to configurable computing resources through virtualization techniques.

Big data forms a framework for discussing cloud computing options. Depending on special need, user can go to the marketplace and buy infrastructure services from cloud service providers such as Google, Amazon, IBM, software as a service (SaaS)

All Rights Reserved ©2024 International Journal of Computer Science (IJCS Journal)
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 2, No 02, 2024. ISSN: 2348-6600
Reference ID: IJCS-506 PAGE NO: 3507-3514

from a whole crew of companies such as NetSuite, Cloud9, Job science etc

C. BIO-INSPIRED COMPUTING FOR BIG DATA ANALYTICS

Bio-inspired computing is a technique inspired ny nature to address complex real world problems. Biological systems are self-organized without a central control. These techniques are developed by biological molecules such as DNA and proteins to conduct computational calculations involving storing, retrieving, and processing of data.

D. QUANTUM COMPUTING FOR BIG DATA ANALYSIS

A quantum computer has memory that is exponentially larger than its physical size and can manipulate an exponential set of inputs simultaneously This exponential improvement in computer systems might be possible.

III. TOOLS FOR BIG DATA PROCESSING

Most batch processing tools are based on the Apache Hadoop infrastructure such as Mahout and Dryad. Stream data applications are mostly used for real time analytic. Some examples of large scale streaming platform are Strom and Splunk.

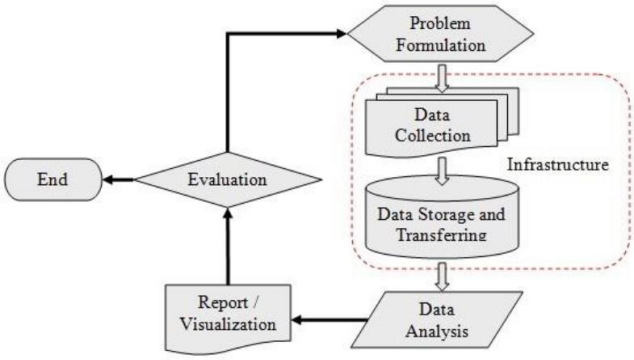


Fig. 4: Workflow of Big Data Project

A. APACHE HADOOP AND MAPREDUCE

The most established software platform for big data analysis is Apache Hadoop and Mapreduce. It consists of Hadoop kernel, MapReduce, Hadoop distributed file system (HDFS) and apache hive etc. Map reduce is a programming model for processing large datasets is based on divide and conquer method. The divide and conquer method is implemented in two steps such as Map step and Reduce Step. Hadoop works on two kinds of nodes such as master node and worker node. The master node divides the input into smaller sub problems and then distributes them to worker nodes in map step. Thereafter the master node combines the outputs for all the subproblems in reduce step. Moreover, Hadoop and MapReduce works as a powerful software framework for solving big data problems.

All Rights Reserved ©2024 International Journal of Computer Science (IJCS Journal)
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 2, No 02, 2024. ISSN: 2348-6600
Reference ID: IJCS-506 PAGE NO: 3507-3514

B. APACHE MAHOUT

Apache mahout aims to provide scalable and commercial machine learning techniques for large scale and intelligent data analysis applications. Core algorithms of mahout including clustering, classification, pattern mining, regression, dimensionalty reduction, evolutionary algorithms, and batch based collaborative filtering run on top of Hadoop platform through map reduce framework

C. APACHE SPARK

Apache spark is an open source big data processing frame- work built for speed processing, and sophisticated analytics. Spark runs on top of existing Hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster.

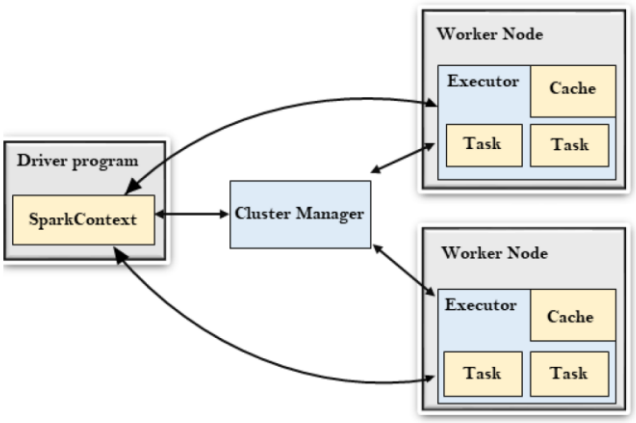


Fig. 5: Architecture of Apache Spark

D. DRYAD

It is another popular programming model for implementing parallel and distributed programs for handling large context bases on dataflow graph. It consists of a cluster of computing nodes, and an user use the resources of a computer cluster to run their program in a distributed way. A dryad user use thousands of machines, each of them with multiple processors or cores.

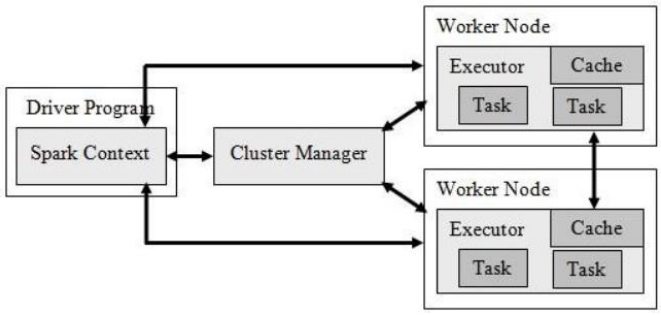
Storm

Storm is a distributed and fault tolerant real time computation system for processing large streaming data. It is specially designed for real time processing in contrasts with hadoop which is for batch processing. On storm cluster users run different topologies for different storm tasks whereas hadoop platform implements map reduce jobs for

All Rights Reserved ©2024 International Journal of Computer Science (IJCS Journal)
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 2, No 02, 2024. ISSN: 2348-6600
Reference ID: IJCS-506 PAGE NO: 3507-3514

corresponding applications. A storm cluster consists of two kinds of nodes such as master node and worker node. The master node and worker node implement two kinds of roles such as nimbus and supervisor respectively.



E. APACHE DRILL

Apache drill is another distributed system for interactive analysis of big data. It has more flexibility to support many types of query languages, data formats, and data sources. It is also specially designed to exploit nested data.

F. JASPERSOFT

The Jasper soft package is an open source software that produce reports from database columns. It is a scalable bigdata analytical platform and has a capability of fast data visualization on popular storage platforms, including MangoDB, Cassandra, Redis etc. One important property of Jaspersoft is that it can quickly explore big data without extraction, transformation, and loading (ETL).

G. SPLUNK

In recent years a lot of data are generated through machine from business industries. Splunk is a real-time and intelligent platform developed for exploiting machine generated big data. It combines the up-to-the-moment cloud technologies and big data. In turn it helps user to search, monitor, and analyze their machine generated data through web interface.

IV. CONCLUSION

In recent years data are generated at a dramatic pace. Analyzing these data is challenging for a general man. To this end in this paper, we survey the various research issues, challenges, and tools used to analyze these big data. From this survey, it is understood that every big data platform has its individual focus. Some of them are designed for batch processing whereas some are good at real-time analytic. Each big data platform also has specific functionality. Different techniques used for the analysis include statistical analysis, machine learning, data mining, intelligent analysis, cloud computing, quantum computing, and data stream processing.

All Rights Reserved ©2024 International Journal of Computer Science (IJCS Journal)
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 12, Issue 2, No 02, 2024. ISSN: 2348-6600
Reference ID: IJCS-506 PAGE NO: 3507-3514

REFERENCES

1) M. K.Kakhani, S. Kakhani and S. R.Biradar, Research issues in big data analytics, International Journal of Application or Innovation in Engineering & Management, 2(8) (2015), pp.228-232.
2) A. Gandomi and M. Haider, Beyond the hype: Big data concepts, meth- ods, and analytics, International
3) Journal of Information Management, 35(2) (2015), pp.137-144.
4) X. Jin, B. W.Wah, X. Cheng and Y. Wang, Significance and challenges of big data research, Big Data Research, 2(2) (2015), pp.59-64.
5) R. Kitchin, Big Data, new epistemologies and paradigm shifts, Big Data Society, 1(1) (2014), pp.1-12.
6) C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Infor- mation Sciences, 275 (2014), pp.314-347.
7) K. Kambatla, G. Kollias, V. Kumar and A. Gram, Trends in big data analytics, Journal of Parallel and Distributed Computing, 74(7) (2014), pp.2561-2573.

All Rights Reserved ©2024 International Journal of Computer Science (IJCS Journal)
Published by SK Research Group of Companies (SKRGC) - Scholarly Peer Reviewed Research Journals