Applied Science & Engineering Journal for Advanced Research
ISSN (Online): 2583-2468
Volume-1 Issue-3 || May 2022 || PP. 19-22 DOI: 10.54741/asejar.1.3.3

Current Challenges and Future Research Perspectives on Big Data

Vibha Gadage
M.E Scholar, Department of Computer Science and Engineering, GHRCEM, Pune, India

Corresponding Author: deeptirani2008@gmail.com

Received: 24-04-2022 Revised: 08-05-2022 Accepted: 18-05-2022

ABSTRACT
Data is being gathered on a scale never before seen in a wide range of application fields. Based on the actual data, rather than guesswork, decisions can now be made that were previously impossible. Mobile services, their manufacturing, retail, financial amenities, and life and physical sciences are just a few of the industries now benefiting from such Big Data analysis. Data heterogeneity, scalability, timeliness, and complexity are just a few of the issues this paper tackles when it comes to big data.

Keywords: *sql, analytics, big data*

I. INTRODUCTION

As the benefits of data ambitious conclusion making become more widely recognised, so does interest in Big Data. The word "Big Data" refers to the collection, processing, analysis, and visualisation of potentially large datasets in a reasonable amount of time using techniques that are not typically available through standard IT systems. As a result, "Big Data technologies" refers to the platform, tools, and software used to accomplish this.

In order to extract value from data, every stage of the data pipeline is hampered by the heterogeneity, timeliness, involvedness, scale and privacy issues that come with Big Data. Other fundamental challenges include data analysis, organisation, retrieval, and modelling. It took a long time to find relevant data scattered across multiple databases, database tables, and/or files. For the most part, organisations lacked the documentation and search tools necessary to locate the information they seek quickly. Analysts, on the other hand, frequently sought assistance from their coworkers, such as database administrators.

II. SURVEY OF LITERATURE

2.1 Elucidations for Supervision Big Data

Since commodity hardware and Cloud-based storage solutions are now widely available, the cost of storing data has plummeted. When production with large measurements of data, it is necessary to allot data and capability across multiple servers. Virtual file systems whichever open source or wholesaler specific, helped conversion from a managed infrastructure to a service-based approach; A new generation of merchandises such as noSQL databases and the Hadoop map-reduce platform have resulted from new database designs and resourceful ways to maintenance extremely parallel dispensation.

2.2 Big Data Opportunities

We are overprovided in a flood of data today.

Methodical Research

Big Data has completely altered the game. Researchers from around the globe now use the Sloan Digital Sky Survey as their primary source of information. Today's astronomer's job entails searching through a database for interesting objects and phenomena that have already been photographed and catalogued.

Teaching

This data could be used to design the most effective approaches to education, starting from reading, writing, college-level, math, to advanced, courses. Nowadays there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about student's performance. In the world of education there is access to a huge database and collection of every detailed measure of every student's academic performance.

Supplementary Fields

Metropolitan planning, intelligent transportation, environmental modelling, energy conservation, smart materials, computational social sciences a new approach fast becoming popular because of the dramatically lower cost of data collection and analysis genomics and payer–provider data (electronic health records, insurance records, prescriptions written by pharmacies, and patient feedback and responses) are the two main sources of health big data.

2.3 Pipeline Phases

The investigation of Big Data implicates multiple distinct phases as shown in the figure below, every of which familiarizes contests.

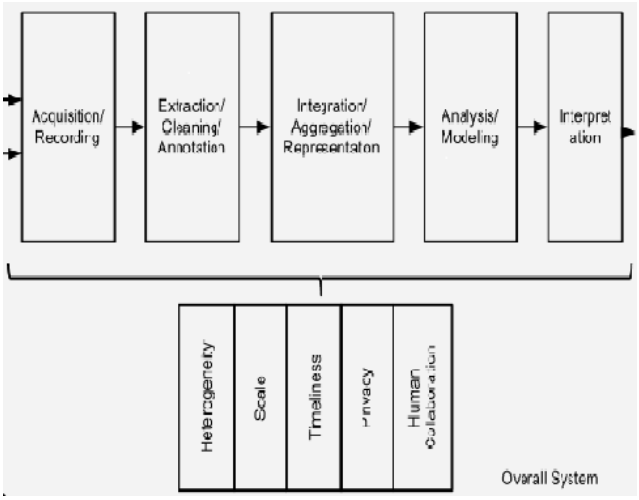


Figure.1: Phases of Big data analysis

a. Data Procurement and Cassette

A data generating source is the source of Big Data. It is possible to filter and compress this data by orders of magnitude, and much of it is irrelevant. Defining these filters in such a way that they don't omit useful information is a major challenge. In addition, creating accurate metadata to describe what data is being collected and how it is being collected and measured is a significant challenge.

b. Information Scrubbing and Withdrawal

The Information extraction is the process of extracting data from various sources and transforming it into a structured format that can then be analysed.

c. Data Integration, Aggregation, and Representation

There are two ways to create effective database designs: either by developing tools to assist in the design process or by abandoning it entirely and developing techniques that allow databases to be used effectively even if they are not designed intelligently.

d. Query Dispensation, Analysis and Data Exhibiting

It is fundamentally different from traditional statistical analysis to query and mine Big Data (noisy, dynamic and heterogeneous, inter-related and untrustworthy). Big-data computing environments, declarative querying and mining interfaces, scalable mining algorithms, and integrated, cleaned, trustworthy and easily accessible data are all required for mining. Aside from providing intelligent querying, data mining can help improve the data's quality and reliability, as well as uncover its semantics. As a result of Big Data, interactive data analysis with real-time answers is also possible. An issue with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying capabilities, and analytics packages that perform various non-SQL processing activities, including data mining and statistical analyses. " Both the expressiveness and the performance of the analysis will benefit from a close coupling between declarative query languages and such packages' functions.

e. Interpretation

A decision-maker, provided with the result of analysis, has to interpret these results. Interpretation involves examining all the assumptions made and retracing the analysis.

Applied Science & Engineering Journal for Advanced Research
ISSN (Online): 2583-2468
Volume-1 Issue-3 || May 2022 || PP. 19-22 DOI: 10.54741/asejar.1.3.3

III. PROGRESSIVE ANALYTICS OF BIG DATA

"Big data" has been used to describe data sets that are more than terabytes in size. Automated censors in the office, social media, websites, and robotics are all contributing to the proliferation of structured, unstructured, and semi-structured data in the workplace. With big data analytics, it is possible to put all these fragmented and often disconnected pieces together in order to generate actionable insights for your organisation. Since data is now being generated and captured across multiple channels, KloudData recognises the importance of adjusting business analytics in order to accommodate the various formats in which it is being generated and captured. Using Sybase IQ and Hadoop, KloudData's big data offering aims to speed up the development and adoption of big data analytics solutions. Big data mining, extraction, analysis, and presentation are all critical components of KloudData's mission to empower business users with the information they need to make better-informed decisions. For today's fast-moving enterprises, our goal is to provide high-availability and scalable analytics solutions.

IV. BIG DATA EXPLORATION IN EXPERIMENTS

4.1 Incompleteness and Heterogeneity

It comes to information consumption, most people don't mind a lot of variety. Because of its depth and variety, natural language is a useful tool for conveying complex ideas. Computers can only process homogeneous data, so they don't have the ability to discern subtlety. As a result, data must be organised in a precise manner before it can be analysed. e.g. A hospitalised patient who undergoes numerous medical procedures. It's possible that we'll keep a single record for each procedure or test the patient has, or that we'll keep a single record for their entire hospital stay or for all of their hospital interactions throughout their life. Computer systems are most effective when they can store a large number of identically sized and shaped items in a single location.

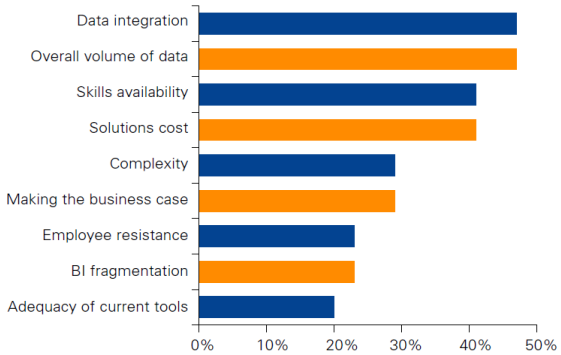
Incompleteness and errors in data are likely to persist even after cleaning and fixing them. During data analysis, these inconsistencies and errors must be taken into consideration.

4.2 Scale

Data volume is increasing at a faster rate than computing resources, but CPU speeds remain the same. As a result of the vastly increased amount of processor, cache, and processor memory channels shared among cores in a single node, the traditional parallel data processing techniques that were used to process data across nodes no longer apply directly to intra-node parallel processing. Rethinking the way we design, build, and operate data processing components is necessary to address this issue.

Cloud computing, which now aggregates multiple disparate workloads with varying performance goals (e.g. interactive services demand that the data processing engine return back an answer within a fixed response time cap), is the second major shift that is currently taking place. It is necessary to develop new methods for determining how to run and execute data processing jobs in order to meet the goals of each workload while also dealing with system failures, which become more frequent as clusters grow in size and cost (that are required to deal with the rapid growth in data volumes).

The traditional I/O subsystem is also undergoing significant transformation, which is a third major change. Increasingly HDDs (persistent data storage and slower random I/O performance) are being replaced by solid state drives and Phase Change Memory. It's necessary to rethink how storage subsystems for data processing systems are designed because the performance spread between sequential and random I/O isn't as wide with these newer storage technologies. According to Fig. 2, big data analytics faces a variety of challenges.



Figurer 2: Big Data and Analytics Success

4.3 Timeliness

The slower it takes to analyse large data sets, the more common it is. There are instances in which the findings of an investigation must be made available right away. e.g. A credit card transaction that appears to be fraudulent should be flagged before it is completed. It's obviously impractical to search through a large data set in search of appropriate elements in data analysis. There is an index structure, but it only supports a few class criteria. There are new types of criteria specified and new index structures needed to support such criteria, as new analyses using Big Data are desired. When data volume grows quickly and queries have short response times, creating such structures becomes even more difficult to do well.

4.4 Privacy

There is a great deal of public concern about the misuse of personal data, particularly through the linking of data from various sources. When it comes to big data, privacy management is both a technical and social issue that must be addressed simultaneously in order to realise its full potential.

There are a plethora of other intriguing research questions to be addressed. For example, no one knows yet how to share private data while restricting disclosure and ensuring that the shared data has sufficient data utility, security for information sharing in Big Data use cases, etc.

V. TOOLS: OPEN SOURCE REVOLUTION

The open source software revolution is inextricably linked to the Big Data phenomenon. It's a win-win situation for companies like Facebook, Twitter and LinkedIn when they work on open source projects. Hadoop and other related software are used in the Big Data infrastructure.

a. Apache Hadoop

MapReduce programming model and a distributed file system called Hadoop Distributed File system are the foundations of this software for data-intensive distributed applications (HDFS). A large number of compute nodes in a large cluster can be used to write applications that rapidly process large amounts of data in parallel with Hadoop. The input dataset is divided into independent subsets for parallel processing by map tasks in a job called MapReduce. After mapping, the next step is to reduce the number of tasks. These reduced tasks rely on the maps' output to get the job's final result.

b. Apache Hadoop related projects

Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra,Cascading, Scribe and many others.

c. Apache S4

Streaming data is processed on this platform. S4 is a data stream management platform. Streaming and processing elements are combined in real time in S4 apps.

d. Storm

At Twitter, Nathan Marz developed the software for streaming data-intensive distributed applications, which is similar to S4.

VI. CONCLUSION

This paper, we discuss the various ways in which big data can be put to good use. Many technical issues are discussed in this paper, from data collection to result interpretation, including: scale, heterogeneity, lack of structure, privacy, provenance, timeliness, visualization and error handling. For Big Data to deliver on its promises and deliver on the promised benefits, it requires support and encouragement of fundamental research.

REFERENCES

1. Ji, Changqing, et al. (2012). Big data processing in cloud computing environments. *Pervasive Systems, Algorithms and Networks (ISPAN), 12th International Symposium on. IEEE*.
2. Keim, Daniel, Huamin Qu, & Kwan-Liu Ma. 92013). Big-data visualization. *Computer Graphics and Applications, 33*(4), 20-21.
3. Madden, Sam. 92012). From databases to big data. Internet Computing, *IEEE 16*(3), 4-6.
4. Chittaranjan, Gokul, Jan Blom, & Daniel Gatica-Perez. (2013). Mining large-scale smartphone data for personality studies. *Personal and Ubiquitous Computing*, *17*(3), 433-450.