Academic Guru Publication House

Advances in Cloud Computing Security
Techniques and Applications
Volume 1
Year: 2021

AG PH Books

A Survey on Big Data: Technologies, Trends and Tools

Dr. Sarika A. Panwar[1*], Dr. Pallavi S. Deshpande[2]

[1]*Assistant Professor, Department of Electronics and Telecommunication Engineering, AISSMS Institute of Information Technology, Pune-01*

[2]*Department of Electronics and Telecommunication Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune-43*

Abstract

Data sets which are too huge or complicated for typical data processing tools, such as relational databases, are referred to as "big data." From the outset, big data has been at the core of companies such as Ebay, Google, LinkedIn, and Fb. Massive as well as complicated data sets, including social media analytics as well as data management skills, as well as real-time data, are included within the collection. The complexity of big data necessitates the development of new methods, algorithms, and analytics for their management and analysis, as well as for their value creation and information extraction. The primary goal of this article is to provide an overview of the current status of Big Data research. In addition, we'll talk about the latest in technology as well as tools, as well as potential problems and emerging trends.

Keywords: Big data, Big Data Quality, Big Data Quality Dimensions, Big Data Analysis.

1. INTRODUCTION

Semi-structured, Structured, as well as unstructured data all fall under the umbrella of "big data." A structured form of such data refers to data that has been properly formatted or tabulated [1]. The data that includes both text and pictures fall underneath the semi-structured category. Unstructured data is any data that does not follow a predetermined structure, such as text, photos, or videos. Databases

* ISBN No. 978-81-955340-6-7

designed for the relational data cannot handle the billions of records that this sort of data includes. As a result, Big Data Analysts must use additional tools and methods to do this [2]. As a result, dealing with tools as well as approaches is a highly challenging work for the Big Data Analysts.

Instances of the Big Data in an action may be seen here:

- In order to get a better understanding of their customers' habits, preferences, as well as perceptions, firms and retailers are using social media such as Facebook as well as Twitter.
- As equipment degrades, manufacturers may track the tiniest variations in the vibration data to determine when to replace or repair it. If it's replaced too early, money is wasted, and if it's replaced too late, a costly work halt result.
- Social media is being used by manufacturers in such a different way from marketers: They are looking for difficulties regarding warranty support prior they can become public.
- Governments around the country, states, and cities are making data available to public so that citizens may build innovative apps which can better serve the people.
- In order to produce more relevant as well as intelligent offerings, financial institutions are using data extracted from client interactions to divide their consumers into the finely calibrated categories.
- Advertisers as well as marketing firms are keeping tabs on social media to assess that house insurance applications can indeed be handled promptly, as well as which ones require a face-to-face visit to verify their authenticity.
- Retail companies are well engaging brand champions, altering the perspective of brand adversaries, or even allowing passionate consumers to pitch their items. •. Using social media has made all of this possible.
- Hospitals use medical data as well as patient records to forecast which patients seem to be likely to return within the next few months following release. So, the hospital is able to avoid another expensive hospital stay.
- Web-based firms are creating information products which aggregate client data in order to provide more enticing suggestions and more effective discount programmes.
- In addition to analysing sales of ticket , sports clubs are now employing big data to monitor team strategy.

Fig. 1 shows the four stages of such big data processing.

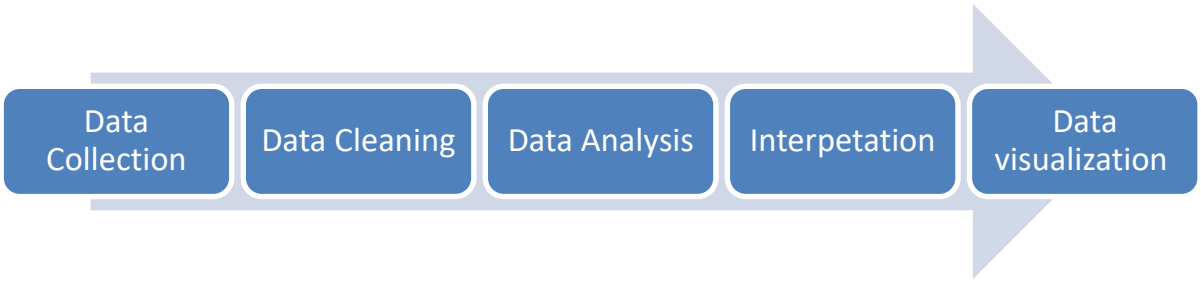

Figure1: Four stage process of Data Mining

- **Data Collection**: Collecting data from various sources should be based on the specifications you've established. This information may be gleaned through a wide range of resources, such as surveys as well as interviews as well as direct observation. For the analysis, it is important to arrange the data you have gathered.
- **Data Cleaning:** It's time to go through the data you've gathered to see what you can utilise. During this step, you'll be removing any empty spaces, duplicated records, as well as other typographical problems. The data must be cleaned up before it can be used for any further investigation.
- **Data Analysis.** Data analysis software as well as some other tools are being used to assist you comprehend the data as well as the draw conclusions. Xls, Python language, R,Rapid Miner, Looker,Chartio, Redash, Metabase,and Microsoft Power BI are among the data analysis tools.
- **Data Interpretation:** Now that you've gotten your outcomes, it's time to analyse them and choose the best next steps depending on what you've learned.
- **Data Visualization:** The term "data visualisation" is indeed a fancy way of stating "graphically present the information in a manner that others can read as well as comprehend." "A variety of tools are at your disposal, including charts, maps, graphs and bullet points. In order to get significant insights, visualisation lets you to compare datasets as well as discover the links between them.

2. BIG DATA MANAGEMENT

Inside the Big Data-related project, how should it be managed and developed? What kind of design should we use to keep track of all the different parts of Big Data? The structure of Big Data should be linked with firm's support infrastructure. Data is being generated by a variety of sources, many of which are unreliable, loud, and dirty. Here, we'll quickly touch on Hadoop, as well as the other management tools, in just this part.

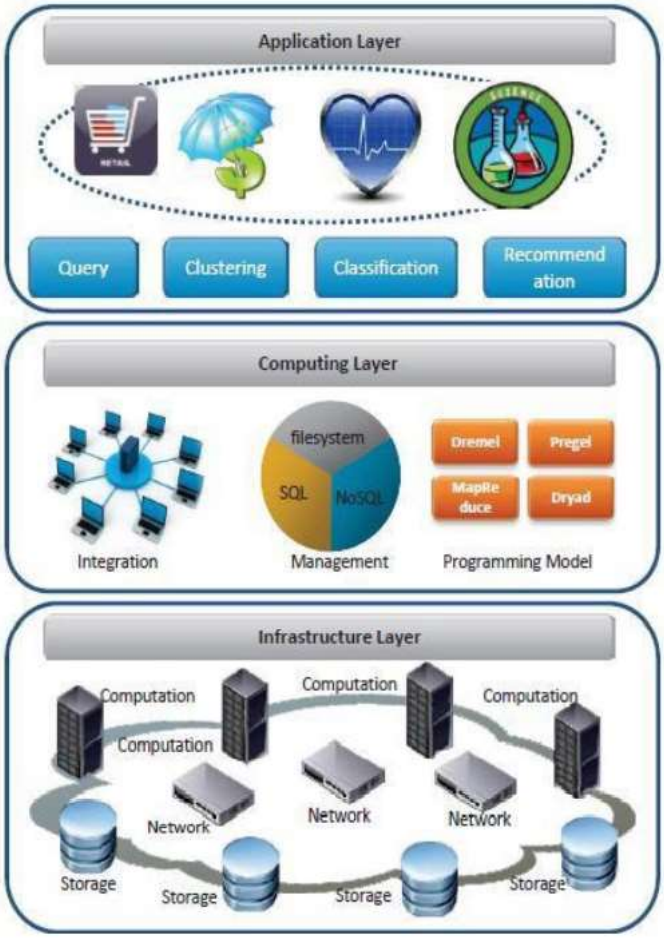

Figure 2: A layered Big Data Architecture

Interactions between systems, data storage, including network devices external towards the system are all handled by the infrastructure layer, which also responds to requests for the data retrieval through other levels like computing. The physical layer of the Big Data is another name for this layer. An abstraction of such underlying data layer is provided through this Mid Layer, which facilitates access to as well as retrieval of data. Distributed storage devices are organised by indexing data in just this layer. Multiple repositories are used to arrange data into chunks. The analytics or even application layer includes the necessary tools and methodologies, as well as the logic needed to provide domain-specific analytics. Another name for this layer might be "Logical layer."

Storage management has a wide variety of tools as well as approaches to choose from. Simple DB, Google Big Table, MemcacheDB, NoSQL are some of the options. [4].

2.1. Hadoop

For the search engine initiative, Doug Cutting and Mike Cafarella initiated a project that would index approximately 1 billion pages. A Google File System, or GFS, was first established in 2003 by the search giant. Later that year, Google released Map Reduce architecture that served as the basis for the Hadoop platform. MapReduce as well as HDFC are at the heart of the (Hadoop Distributed File System) Hadoop system. Let us take a quick look at this Hadoop component in this part.

Hadoop's Design Principles It comprises mostly of four parts.

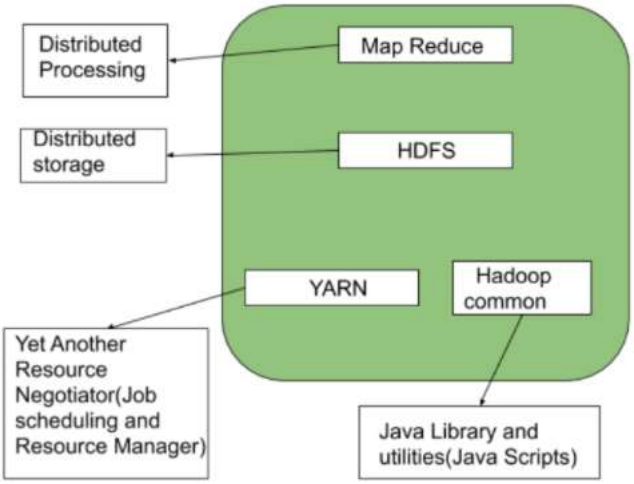

Figure 3. Hadoop Architecture with 4 components

- HDFS (Hadoop distributed File System)
- MapReduce
- YARN (Yet Another Resource Framework)
- Common Utilities or Hadoop Common

A. HDFS:

HDFS is indeed the Java-based file system for large-scale commodity server clusters which enables scalable as well as dependable data storage. There are two sorts of nodes within a cluster. There is indeed a master node just at top of the tree, which is the name node. It is also possible to have a data node acting as a slave node. The default block size for HDFS is 64MB. In order to analyse enormous volumes of data simultaneously, these files are duplicated in multiples.

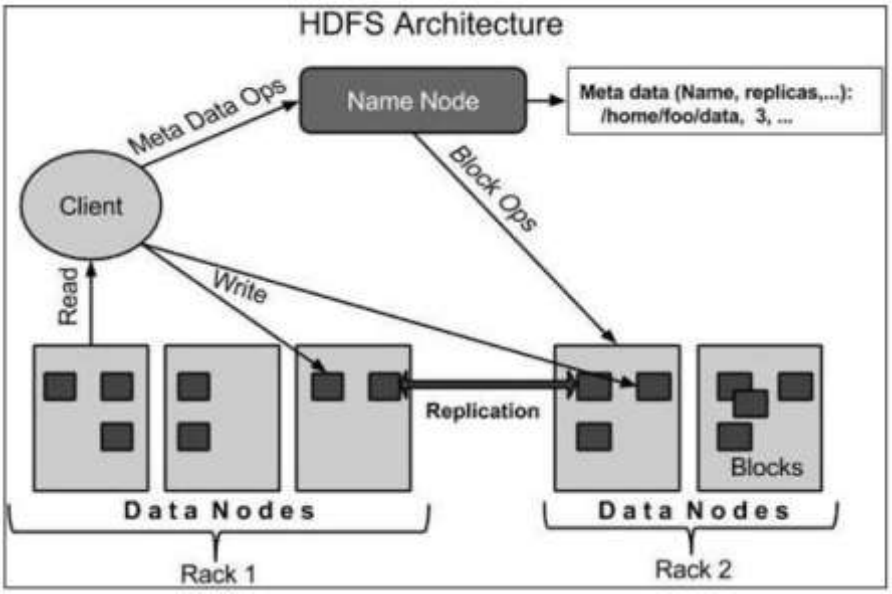

Figure 4. HDFS Architecture [5]

B. MapReduce

Java-based Map Reduce is indeed a distributed computing programming methodology. It's a method of preparing something for consumption or use. Map as well as Reduce are key components of Map Reduce algorithm. The Hadoop program's Map that Reduces function really consists of two discrete and different operations. In the first place, there's the map task, that takes a collection of data as well as turns it into the new set of data under which such individual pieces are broken down further in the tuples (key or the value pairs). The result of a map is used as input for a reduction operation, which merges the tuples into the smaller set. Map and reduce jobs are always executed sequentially, in accordance with name Map Reduce.

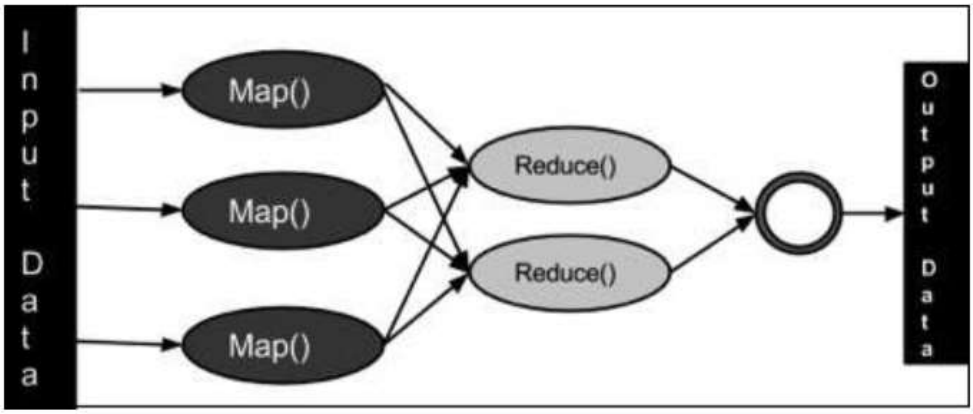

Figure 5. MapReduce Architecture [6]

C. YARN

It stands such Yet another Resource Negotiator, and it was introduced within Hadoop version 2 as a way to manage clusters. Big data applications use YARN, which is today regarded as a large-scale, widely dispersed operating system. It is referred to through Apache as a new resource management. Apps run on resources provided by YARN Infrastructure, such as CPUs and RAM that may be accessed by the applications themselves. A resource manager as well as the node manager are two sub-parts of the manager. The master is indeed Resource Manager (that is one for each cluster). They understand where slaves are (Rack Awareness) as well as how much they have at their disposal. The Resource Scheduler, which determines how resources are assigned, is perhaps the most critical of the many services it provides. The infrastructure is the master of the Node Managers (of which there are several in a cluster). Whenever it begins, the Resource Manager is notified. The Resource Manager receives a pulse from it on a regular basis.

Features of YARN

- Multi-Tenancy
- Scalability
- Cluster-Utilization
- Compatibility

D. Common Utilities or Hadoop Common

Inside the context of Hadoop cluster, "common utilities" refers to the java library as well as Java files, or even "java scripts," which we require to run all of the other's components in the system. YARN, HDFS, as well as Map Reduce all rely on these tools to operate the cluster. Because hardware failure is so prevalent among Hadoop clusters, the problem must be primarily managed in software through the Hadoop Framework, according to Hadoop Common.

3. BIG DATA APPLICATION

These properties may be derived from the vast amounts of data that can be analysed via Big Data Analysis. An overview of Big Data use cases is provided in the section.

a) **Text Data Analysis**

Text mining is yet another term for the text analytics. To get more information about clients fro the unstructured data sources, text analytics may be used. Computational linguistics, Machine learning and statistics all play a role in the text analytic projects. With the use of text analytics, companies can turn massive amounts of the human-generated text into digestible summaries that aid in fact-based

decision-making. If you want to anticipate the stock market, for instance, you may use the text analytics for extracting information from the financial news. (NLP) Natural Language Processing is used in the majority of text mining approaches (NLP). Text analysis, interpretation, and generation are all possible with NLP. Certain NLP-based text mining approaches have been implemented, such as the extraction of data, the creation of topic models, the summarising of text, the clustering, classification the answering of questions, and the gathering of opinions.

b) Social Media Analysis

There are a variety of approaches used to study social networks, that are networks which are made up of persons or organisations that are interdependent in some way, whether it's via friendship, a shared interest or monetary transactions. Nodes and linkages connect each other within a social network. A network of nodes (actors) as well as connections (relationships between nodes) forms the basis of the social structure, which may be shown as a network diagram. The actor, relation, as well as the network are indeed the three main building blocks of social network systems. Categories of social media services include link-based as well as content-based analyses [8]. It's always been the goal of link-based structural analysis to focus on link the prediction, community discovery, the development of social networks, and social impact analysis and many more.

c) Mobile Data Analysis

It is the process of analysing and acting on user behaviour data in order to increase, engagement, user retention and conversions in mobile applications. A new field of study has opened up as a result of advancements in the wireless sensor, or even mobile communication, and the stream processing technologies. In addition to health as well as business-related applications, there are a number of additional Big Data application areas, like surveillance monitoring, weather forecasting, as well as the multimedia data analysis.

4. BIGDATA CHALLENGES

Big Data's heterogeneity, size, complexity, timeliness, and privacy issues inhibit development at all stages of such data pipeline which might provide value. Currently, we are forced to make judgments on the ad hoc basis about which data to preserve and which to discard, as well as on how to consistently save data we do keep with appropriate metadata, due to data tsunami. For example, unlike pictures as well as video, which are stored and shown in organised formats, most data today also isn't inherently structured. For semantic material, though, you should instead search for keywords. The major challenge is to organise this data so that it can be analysed in the future. In the context of other data, the value of a piece of information increases exponentially. A key source of added worth comes from data integration. Today, the vast majority of data is produced digitally, giving us the possibility and the task of both influencing the development of new data as well as automatically linking previously-made data to make new connections. Other core issues include data analysis, organisation, recovery, and modelling. Due to

the intricacy of such data which has to be processed, as well as the limited scale of an original method, data analysis is often a bottleneck in many applications. Finally, non-technical domain specialists are essential to deriving actionable Knowledge from the data.

CONCLUSION

Various technologies for dealing with huge data have been examined in such research. Utilizing HDFS Hadoop distributed data storage as a framework, this article explores Big Data architecture and explains its many components. The mining sector, for example, finds a lot of value in the big data. Big data isn't a luxury for an industry that conducts billions of dollars in commerce every year, it's a need. Overall system as well as application performance was the primary focus of this article, that included a review of several Big Data architectures and their handling approaches for dealing with large amounts of data from multiple sources. Big data has had a significant impact on the corporate sector. It is possible to exploit big data in a variety of ways, and in ways that many people have never considered before. The mining sector, for example, finds a lot of value in the big data. Big data isn't a luxury for the industry that conducts billions of dollars in commerce every year, it's a need. The algorithms for effectively and swiftly mining large amounts of data are always being improved by researchers. In addition, a look at some of the analytics as well as management tools being provided.

References

1. Arti Chandani, M. M. (2015). Banking On Big Data: A Case Study. ARPN Journal Of Engineering And Applied Sciences , 4.
2. Basvanth Reddy, P. B. (2016). Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique. International Journal of Advanced Research in Computer and Communication Engineering , 5.
3. Keshav Sanse, Meena Sharma, (2015).Clustering methods for Big data analysis,(IJARCET) Volume 4 Issue 3, March.
4. Min Chen Shiwen Mao Yunhao Liu, (2014). Big Data: A Survey,Mobile Networks and Application 19:171:209.
5. Sharma, Sugam, et al. (2014): "A brief review on leading big data models." Data Science Journal 14-041.
6. Jena, Bibhudutta, et al. (2016) "Improvising name node performance by aggregator aided HADOOP framework." 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). IEEE,

7. Chung, W. (2014). BizPro: Extracting and categorizing business intelligence fac-tors from textual news articles. International Journal of Information Management,34(2), 272284.
8. Aggarwal CC (2011) An introduction to social network data analytics. Springer.