ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

AGNI-PANKH 16

Jawaharlal Darda Institute of Engineering and Technology, Yavatmal

Vol. 4, Special Issue 3, January 2017

Overview and Analysis of Big Data and Hadoop

Ms. Ashwini K. Patriwar[1], Ms. Gauri P. Rathod[2], Ms. Ashwini L. Shrirang[3], Ms. Karishma P. Jaiswal[4]

Student, CSE Department, JDIET, Yavatmal, India[1, 2, 4]

Student, IT Department, JDIET, Yavatmal, India[3]

Abstract: The term “Big Data” is an evolving term that describes any amount of structured, semistructured and unstructured data that has the potential to be mined for information. Relational database management systems and desktop have difficult in handling big data. To handle these large amounts of data may require software running on tens, hundreds, or even thousands of servers". To solve these problems of making it useful for analytics purposes “Hadoop” technology is used. Hadoop technology is the core platform and for structuring Big Data. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Hadoop is written in Java. Hadoop programs can be written using a small API in Java or Python. Hadoop provides to the application programmer the abstraction of map and reduce. Map and reduce are available in many languages, such as Lisp and Python.

Keywords: Big data, Hadoop, API, RFID.

I. INTRODUCTION

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set."There is little doubt that the quantities of data now available are indeed large, but that’s not the most relevant characteristic of this new data ecosystem."

Data sets grow rapidly - in part because they are increasingly gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte’s (2.5×1018) of data is generated. One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.

Relational database management systems and desktop statistics- and visualization-packages often have difficulty handling big data. The work may require "massively parallel software running on tens, hundreds, or even thousands of servers". What counts as "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target.

Big data can be described by the following characteristics:

Volume

The quantity of generated and stored data. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Variety

The type and nature of the data. This helps people who analyse it to effectively use the resulting insight.

Velocity

In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development.

Variability

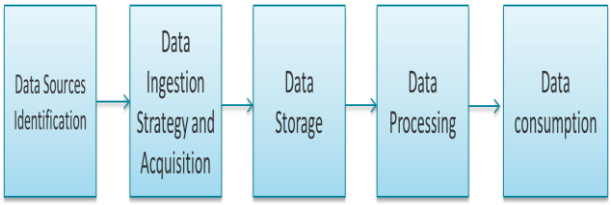
Inconsistency of the data set can hamper processes to handle and manage it.

Veracity

The quality of captured data can vary greatly, affecting accurate analysis.

II. ARCHITECTURE

The following diagram shows the architecture of big data:



Get to the Source!

Source profiling is one of the most important steps in deciding the architecture. It involves identifying the different source systems and categorizing them based on their nature and type.

Points to be considered while profiling the data sources:

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
AGNI-PANKH 16
Jawaharlal Darda Institute of Engineering and Technology, Yavatmal
Vol. 4, Special Issue 3, January 2017

- Identify the internal and external sources systems
- High Level assumption for the amount of data ingested from each source
- Identify the mechanism used to get data – push or pull
- Determine the type of data source – Database, File, web service, streams etc.
- Determine the type of data – structured, semi structured or unstructured

Ingestion Strategy and Acquisition

Data ingestion in the Hadoop world means ELT (Extract, Load and Transform) as opposed to ETL (Extract, Transform and Load) in case of traditional warehouses.

Points to be considered:

- Determine the frequency at which data would be ingested from each source
- Is there a need to change the semantics of the data append replace etc?
- Is there any data validation or transformation required before ingestion (Pre-processing)?
- Segregate the data sources based on mode of ingestion – Batch or real-time

Storage

One should be able to store large amounts of data of any type and should be able to scale on need basis. We should also consider the number of IOPS (Input output operations per second) that it can provide. Hadoop distributed file system is the most commonly used storage framework in BigData world, others are the NoSql data stores – MongoDB, HBase, Cassandra etc. One of the salient features of Hadoop storage is its capability to scale, self-manage and self-heal.

There are 2 kinds of analytical requirements that storage can support:

- Synchronous – Data is analysed in real-time or near real-time, the storage should be optimized for low latency.
- Asynchronous – Data is captured, recorded and analysed in batch.

And Now We Process

Not only the amount of data being stored but the processing also has increased multifold.

Earlier frequently accessed data was stored in Dynamic RAMs but now due to the sheer volume, it is been stored on multiple disks on a number of machines connected via the network. Instead of bringing the data to processing, in the new way, processing is taken closer to data which significantly reduce the network I/O.

III. HADOOP ARCHITECTURE

Hadoop framework includes following four modules:

 Hadoop Common: These are Java libraries and utilities required by other Hadoop modules. These libraries provide filesystem and OS level abstractions and contains the necessary Java files and scripts required to start Hadoop.

 Hadoop YARN: This is a framework for job scheduling and cluster resource management.

 Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.

 Hadoop MapReduce: This is YARN-based system for parallel processing of large data sets.

How Does Hadoop Work?

Stage I

A user/application can submit a job to the Hadoop (a hadoop job client) for required process by specifying the following items:

1. The location of the input and output files in the distributed file system. 2. The java classes in the form of jar file containing the implementation of map and reduce functions. 3. The job configuration by setting different parameters specific to the job.

Stage II

The Hadoop job client then submits the job (jar/executable etc) and configuration to the JobTracker which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client.

Stage III

The TaskTrackers on different nodes execute the task as per MapReduce implementation and output of the reduce function is stored into the output files on the file system.

IV. APPLICATIONS

- ❑Monitor premature infants to alert when interventions is needed
- ❑Predict machine failures in manufacturing
- ❑Prevent traffic jams, save fuel, reduce pollution

V. ROLE OF BIG DATA

1. In BDA: Big Data Analytics Applications (BDA Apps) are a new type of software applications, which analyse big data using massive parallel processing frameworks (e.g., Hadoop). Developers of such applications typically develop them using a small sample of data in a pseudo-cloud environment. the applications in a large-scale cloud environment with considerably. more processing power and larger The big data can come from sources such as runtime information about traffic, tweets during the Olympic Games, stock market updates, usage information of an online game, or the data from any other rapidly growing data-intensive software system.

3. In Clustering: Using clustering (K-means algorithm) through a simple point and click dialog, users can automatically find groups within data based on specific data dimensions. With clustering, it is then simple to identify and address groups by customer type, text

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

AGNI-PANKH 16

Jawaharlal Darda Institute of Engineering and Technology, Yavatmal

Vol. 4, Special Issue 3, January 2017

documents, products, patient records, click path, behaviour, purchasing patterns, etc.

4. In Data Mining: Decision Tree--Datameer's decision trees automatically help users understand what combination of data attributes result in a desired outcome. Decision trees illustrate the strengths of relationships and dependencies within data and are often used to determine what common attributes influence outcomes such as disease risk, fraud risk, purchases and online signups. The structure of the decision tree reflects the structure that is possibly hidden in your data.

VI. ADVANTAGES

1. Derives innovative solutions.
2. Accesses vast information's via surveys.
3. Delivers answer of any query.
4. Every second, additions are made.
5. One platform carries unlimited information.

VII. DISADVATAGES

1. It can be misleading.
2. Complex task to gain insight.
3. Bespoke solution to query not possible.
4. Updates can mismatch real figures.
5. Battle to fetch relevant information.
6. input data (reminiscent of the mainframe days) Big Data Analytics Applications (BDA Apps) are a new category of software applications that leverage largescale data, which is typically too large to fit in memory or even on one hard drive,-

VIII. CONCLUSION

The main aim of this paper is to explore the role of Big Data in various fields. Big Data is a powerful tool that makes things ease in various fields as said above. Big data used in so many applications they are banking, agriculture, data mining, finance, marketing, stocks, BDA, customer goods, credit cards etc.

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. Hadoop is designed to scale up from single server to thousands of machine.

ACKNOWLEDGEMENT

We would like to thank, Respected Sir/Mam for giving us such a wonderful opportunity to expand our knowledge. It helped us lot to realize of what we study for. Secondly, we would like to thanks our Teachers who helped us. Thirdly, we would like to thanks our parents who patiently helped us. Next, we would thank Microsoft for developing such a wonderful tool like MS Word. It helped our work a lot to remain error-free.

REFERENCES

[1] Big Data at Work: Dispelling the Myths, Uncovering the Opportunities
[2] www.google.com
[3] www.wikipedia.org
[4] www.studymafia.org
[5] Hadoop: The Definitive Guide
[6] HBase: The Definitive Guide Lars George (Author) O'Reilly Media
[7] Hadoop Operations Eric Sammer