International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 06 | June-2016 www.irjet.net p-ISSN: 2395-0072

Study And Analysis Of Big Data And It's Framework(Apache Hadoop versus Apache spark)

Vivek Kushwaha[1]

[1]Department Of Computer Science and Engineering,KIIT University, Bhubaneshwar, Orissa,India.

Abstract - Big data is the term used for the data set which are very large and complex and cannot be handled by the traditional data processing applications The working of Big Data includes data analysis, capture, data curation, searching, data sharing, storage, transfer, visualization, querying, updating and information privacy. Big Data refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

Key Words: Big Data , Apache Hadoop, Apache Spark, Data Curation, Data Analysis.

1.INTRODUCTION *(Size 11 , cambria font)*

Big data is the term used for the data set which are very large and complex and cannot be handled by the traditional data processing applications The working of Big Data includes data analysis, capture, data curation, searching, data sharing, storage, transfer, visualization, querying, updating and information privacy. Big Data refers simply to the use of predictive analytics or certain other advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk.

Data Analysis is the process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information suggesting conclusions, and supporting decision-making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.

Data curation is a term used to indicate management activities related to organization and integration of data collected from various sources, annotation of the data, and publication and presentation of the data such that the value of the data is maintained over time and the data remains available for reuse and preservation. Data curation includes "all the processes needed for principled and controlled data creation, maintenance, and management, together with the capacity to add value to data

Data sharing is the practice of making data used for scholarly research available to other investigators.

Data Storing often called storage or memory is a technology consisting of computer components and recording media used to retain digital data. It is a core function and fundamental component of computers.

Data transfer is the physical transfer of data over a point to point or point to multipoint communication channel.

1.1 Purpose

Big Data have many applications. The most common purpose of the Big Data is to produce small data. There is almost always a drastic filtering process that reduces big data into smaller data.

	BIG DATA
GOAL	There is a vague goal, but there really is no way to completely specify what the big data resource will contain and how the various types of data held in the resource will be organized, connected to other data resources, or usefully analyzed
LOCATION	Typically spread throughout electronic space, typically parceled onto multiple Internet servers, located anywhere on earth.
DATA STRUCTURE AND CONTENT	Must be capable of absorbing unstructured data The subject matter of the resource may cross multiple disciplines, and the individual data objects in the resource may link to data contained in other, seemingly unrelated, big data resources.
DATA PREPARATION	The data comes from many diverse sources, and it is prepared by many people. People who use the data are seldom the people who have prepared the data.
LONGEVITY	Big data projects typically contain data that must be stored in perpetuity. Ideally, data stored in a big data resource will be absorbed into another resource when the original resource terminates. Many big data projects extend into the future and the past (e.g., legacy data), accruing data prospectively and retrospectively.
MEASUREMENT	Many different types of data are delivered in many different electronic formats. Measurements, when present, may be obtained by many different protocols. Verifying the quality of big data is one of the most difficult tasks for data managers.
REPRODUCIBILITY	Replication of a big data project is seldom feasible. In most instances, all that anyone can hope for is that bad data in a big data resource will be found and flagged as such.
STAKES	Big data projects can be obscenely expensive. A failed big data effort can lead to bankruptcy, institutional collapse, mass firings, and the sudden disintegration of all the data held in the resource. Though the costs of failure can be high in terms of money, time, and labour, big data failures may have some redeeming value. Each failed effort lives on as intellectual remnants consumed by the next big data effort.
INTROSPECTION	Unless the big data resource is exceptionally well designed, the contents and organization of the resource can be inscrutable, even to the data managers. Complete access to data, information about the data values, and information about the organization of the data is achieved through a technique herein referred to as introspection.
ANALYSIS	With few exceptions, such as those conducted on supercomputers or in parallel on multiple computers, big data is ordinarily analysed in incremental steps. The data are extracted, reviewed, reduced, normalized, transformed, visualized, interpreted, and reanalysed with different methods .

Table1: Showing goals of big data.

1.2 CHARACTERISTICS OF BIG DATA

- Support for multiple data types
- Handle batch processing and/or real time data streams
- Utilize what already exists in your environment:
- Support NoSQL and other newer forms of accessing data:
- Overcome low latency:
- Provide cheap storage:
- Integrate with cloud deployments:

1.3. The original three 'V' Dimension Characteristics of Big Data identified in 2001 are

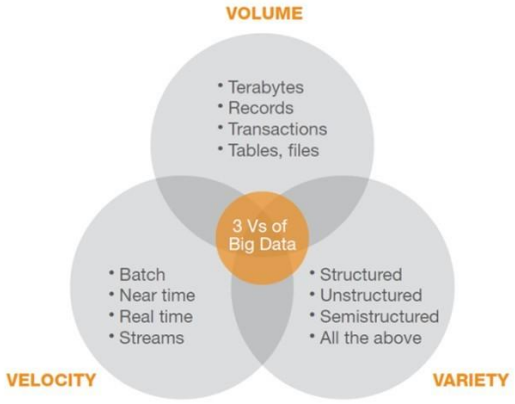

Fig1: The 3 V's of the Big Data.

1.4 .BIG DATA FRAMEWORK

→Hadoop

It is an open source software framework for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.

→Spark

It is an open source software web application framework and domain specific language written in

java. It is an alternative to other java web application frameworks for instance JAX-RS, Play framework etc.

→Flink

It is a streaming dataflow engine that provides data distribution, communication, and fault tolerance for distributed computations over data streams.

→Storm

It is a free and open source distributed real time computation system It is simple, can be used with any programming language.

→Samza

It is a distributed stream processing framework , uses Apache Kafka for messaging, and Apache Hadoop Yarn to provide fault tolerance, processor isolation, security, and resource management.

Table -2: . Comparison between Apache Hadoop and Apache Spark

ASPECT	APACHE HADOOP	APACHE SPARK
Difficulty	Difficult to program and needs abstraction	Easy to program and does not need abstraction
Interactive mode	No in-built interactive mode except pig and hive	It has interactive mode
Streaming		
Performance		
Latency	Completely disk oriented	Lower latency
Easy of code	Writing pipeline is complex and lengthy process.	Always compact than Hadoop MR

Table -3: Hardware Requirements

	Apache Spark	Apache Hadoop
Cores	8–16	4
Memory	8 GB to hundreds of gigabytes	24 GB
Disk	4–8	4–6 one-TB disks
Network	10 GB or more	1 GB Ethernet all-to-all

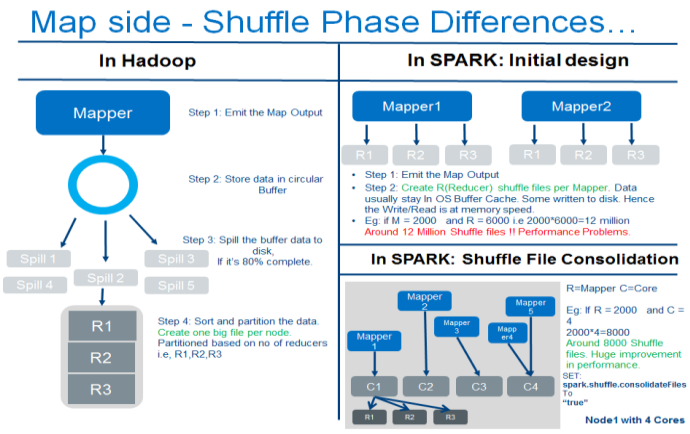

Fig -2: Map Side.

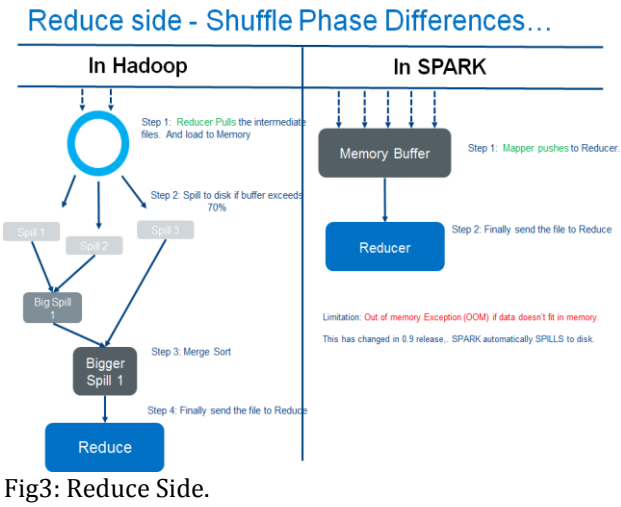

Fig3: Reduce Side.

3. CONCLUSIONS

Spark has excellent performance and is highly cost-effective due to in-memory data processing. It's compatible with all of Hadoop's data sources and file

formats, and thanks to friendly APIs that are available in several languages; it also has a faster learning curve. Spark even includes graph processing and machine-learning capabilities.

Hadoop MapReduce is a more mature platform and it was built for batch processing. It can be more cost-effective than Spark for truly Big Data that doesn't fit in memory and also due to the greater availability of experienced staff. Furthermore, the Hadoop MapReduce ecosystem is currently bigger thanks to many supporting projects, tools and cloud services.

REFERENCES

[1] → Qura.com
https://www.quora.com/What-is-the-difference-between-Apache-Spark-and-Apache-Hadoop-Map-Reduce

[2] Xplenty
https://www.xplenty.com/blog/2014/11/apache-spark-vs-hadoop-mapreduce/