Big Data and Apache Spark: A Review

Abhishek Bhattacharya[1*], Shefali Bhatnagar[2]

Deptt. Of CSE, ASET, Amity University, Noida, India

Abstract— *Big Data is currently a very burning topic in the fields of Computer Science and Business Intelligence, and with such a scenario at our doorstep, a humungous amount of information waits to be documented properly with emphasis on the market. By market, we mean the current technologies in use, the current prevalent tools, and the companies playing an imperative role in taming the data with such a colossal outreach.*

Keywords— Big Data, Cloud, Apache, Hadoop, Spark, Analytics.

I. INTRODUCTION

This paper details the concept of big data, the nitty-gritty of the field, and the associated analytics. The paper is divided into seven sections. We start by introducing the concept of Big Data. We move forward towards sections, one of which explains a very important slice of big data, the three V's. Subsequently, we have sections on big data analytics and security issues in big data analytics. This is followed by an entire section that gives a very streamlined idea of the enormity of the extent to which data is generated in this world – what are the sources, what are the sinks, and how we go about transforming them to develop lineages or provenances – following the ETL.

Then we go about discussing the variety of excellent tools available in the market coming from big names such as Apache, on which note, we have considered writing two sections – one on Apache Hadoop, and other on Apache Spark. The prime difference we put a spotlight on is the use of disk by Hadoop'sMapReduce as compared to Apache Spark's use of memory, making Spark a more competitive product which has established quite a few benchmark records by now.

II. THE 3V'S OF BIG DATA

A. **Volume**:Volume refers to amount of data. Datavolumecontinues to increase at an unprecedentedrate.Volume of data stored in enterprise repositories have grown from megabytes and gigabytes to petabytes.

B. **Variety**:There are many different types of data as text,sensordata,audio, video, graph and more.Variety is about managing the complexity of multiple data types,includingstructured,semi-structured and unstructured data.

C. **Velocity**: Velocity refers to the speed of data processing.Data is arriving continuously as streams ofdata, and we are interested in obtaining useful informationfrom it in real time. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.

III. STATISTICS OF BIG DATA

In today's scenario, almost everything happens online. Every click, every play/pause while you watch a video, every server request, every network message, every fault and any diversion from a particular service at any point of time by you, can be recorded.

For Example - Log Files (apache server). Every click and request is recorded.

For Example - Machine Syslog file (millions of machines available and actions are recorded.)

This scenario can very easily be thought of as analogous to an iceberg. A humungous amount of data is or can be indeed collected every second, but the analysis of any of the data is on a very minuscule level.

Let us discuss a few ways that contribute to the enormity:

Every post on Facebook is data, and every day it's in the order of several TBs. Can you imagine that? Your general hard disk drive is usually a TB. Every minute YouTube observes an upload of over 300 hours of video, which subsequently generate billions of views. That is big data.Reviews on Yelp or Zomato generate a lot of big data too; and, so do tweets on Twitter and the billion searches on Google. Doesn't this seems to be a gigantic amount of data coming in every moment, making it quite a bit difficult to be tamed?A lot of dense big data is present in the form of graphs. For graphs, we can discuss the Facebook user graph, which is an example of a very dense graph. Then there are telecommunication networks, road and similar route

maps, consumer relationship maps et al., where dealing with the data becomes more essential without which the data might stand to be not of much use.

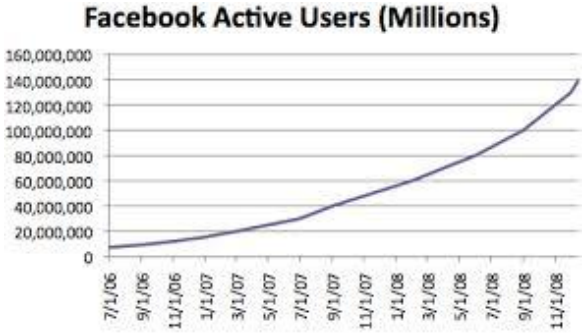

FIG 1. FACEBOOK USER GRAPH

Products such as Nest constantly monitor room temperature. They get a lot of data collected every minute and function accordingly. Also, traffic responders collect huge amounts of data while used for paying tolls or to get a traffic density overview. That is exactly where Internet of Things is heading us. It is going to be a future with the data of every moment of our day, stored and connected, and it is going to be big.

What can we do with Big Data?

Let's say we focus on crowdsourcing, that is, considering people from a wide region, say, all over the world, contributing to a cause. There can be inputs from people all over the world which combined with physical modeling sensing and data assimiliation, can generate results which can map anything ranging from traffic at general geographic locations, temperature rise or fall over areas with similar land features et al.

We fit big data and its analytics in *three primary models* by a three different pioneers of the field:

TABLE 1

BIG DATA AND ITS ANALYTICS IN THREE PRIMARY MODELS

Jim Gray's model of data science	Ben Fry's model of data science	Jeff Hammerbacher's model
Capture	Acquire	Identify Problem
Curate	Parse	Instrument data sources
Communicate	Filter	Collect Data
	Mine	Prepare Data (integrate, transform, clean, filter, aggregate)
	Represent	Build Model
	Refine	Evaluate Model
	Interact	Communicate Results

IV. SECURITY ISSUES IN BIG DATA

The protection of information is another enormous concern, and one that increments in the connection of Big Data. Unlike customary security technique, security in huge information is fundamentally in the type of how to process information mining without uncovering delicate data of clients. Additionally, current innovations of security insurance are primarily in view of static information set, while information is dependably alertly changed, including information design, variety of trait and expansion of new information. In this way, it is a test to execute successful protection insurance in this mind boggling situation. Also, legitimate and administrative issues likewise require consideration. For electronic wellbeing records, there are strict laws overseeing what should and can't be possible. For other information, regulations, especially in the US, are less strong. Overseeing security is successfully both a specialized and a sociological issue, which must be tended to mutually

from both points of view to understand the guarantee of enormous information. Knowledge driven security depends on huge information investigation. By keeping information in one spot, it happens to be an objective for assailants to harm the association. It obliged that huge information stores are rightly controlled. To guarantee confirmation a cryptographically secure correspondence system must be executed. Controls ought to be utilizing standard of lessened benefits, particularly for access rights, with the exception of a head who have authorization information to physical access. For viable access controls, they ought to be ceaselessly watched and exchanged as change workers association parts so representatives don't total radical rights that could be abused. Other security strategies are expected to catch and dissect system movement, for example, metadata, bundle catch, stream and log data. Associations ought to ensure interests in security items utilizing nimble advancements based examination not static supplies.

Another issue is connected with arranging consistence of information security laws. Associations need to consider lawful fanning for putting away information. Nonetheless, enormous information has security points of interest. At the point when associations order information, they control information as indicated by determined by the regulations, for example, forcing store periods. This permits associations to choose information that has neither little esteem nor any should be kept so it is no more accessible for robbery. Another advantage is enormous information can be dug for dangers, for example, confirmation of malware, inconsistencies, or phishing. The generally less organized and casual nature of numerous Big Data methodologies is their quality, however it additionally represents an issue: if the information included is touchy for reasons of security, endeavor security, or administrative necessity, then utilizing such methodologies may speak to a genuine security rupture. Database administration frameworks bolster security strategies that are truly granular, ensuring information at both coarse and fine grain level from wrong get to. Huge Data programming for the most part has no such protects. Ventures that incorporate any touchy information in Big Data operations must guarantee that the information itself is secure, and that the same information security approaches that apply to the information when it exists in databases or documents are likewise authorized in the Big Data connection. Inability to do as such can have genuine negative outcomes.

V. TOOLS FOR BIG DATA ANALYTICS

A. **Hadoop**: Hadoop is a tool that needs to be discussed whenever big data is talked about. It has emerged to be a very intrinsic to big data, and is a framework for distributed processing of large data sets across clusters of computers using simple programming models. It can scale up to a number of machines and provide local storage and computation. This works on all the three popular operating systems.

B. **MapReduce**: Google developed MapReduce. It uses a parallel and distributed algorithm on a cluster to process and generate large data sets. MapReduce is a framework that is used by Hadoop as well as other data processing applications. It is completely OS independent.

C. **Storm**: Apache made this highly appreciated product that is now a part of Twitter. Storm makes it easy for unbounded streams of data to be reliably processes, and that too real-time. It is indeed called the "Hadoop of real-time", and is compatible with several platforms and several languages while being scalable, robust, and fault-tolerant. It is used by big names that have large and active datasets, such as Twitter, Yelp, Spotify, Alibaba.com, Baidu, Flipboard, Groupon, The Weather Channel et al. The use cases range from distributed ETL, online machine learning, continuous computation, real-time analytics, and more.

D. **Cassandra**: Managed by the Apache Foundation, Cassandra was originally developed by Facebook, and is a NoSQL database. It provides at par performance with scalability and high availability. Fault-tolerance on cloud infrastructure ensures it to be a great platform for "mission-critical" data. Cassandra supports replication on multiple datacenters with lower latency for the users and ability to survive regional outages.

E. **Spark**: Apache Spark, as defined by its website, is a fast and general engine for large-scale data processing. It has tremendous speed as compared to HadoopMapReduce and is up to 100 times faster in memory, and 10 times on disk. It can be interactively used from the Python or Scala shells, making it easy to build parallel apps. It has a powerful stack of high-level tools for streaming, SQL, and complex analytics. Running everywhere from Hadoop, Mesos, in the cloud, or standalone, it can access diverse data sources including S3, HDFS, Cassandra et al.

These were five of the very popular and efficient tools currently available on the market for big data analytics. We would now proceed to discuss Hadoop and later, Spark, in a bit of detail emphasizing their importance.

VI. APACHE SPARK

Getting a good hold on the huge data, it needs to be pretty fast. For example, if we talk about the data being generated by Walmart stores all over the world, that would be millions of sale entries every hour, right? So, it would be really bad to have data scientists to provide insights on sales during a particular time of day only if the computation takes a day. Also, the data scientists should be able to process it in entirety at once. Hence, it is required of Spark to be available on clusters, rather than insisting the requirement of a single machine.

So, Spark boasts a world record in large scale sorting by Databricks. That was made possible because of the two features discussed above. Spark stores data sets in memory, which makes it a 100x faster than Hadoop, which does so on disk. Also, allowing user programs to load data to a cluster's memory and allowing repeated querying, it is a framework well suited to machine learning algorithms.

Components of Spark:

A. Resilient Distributed Datasets and the Spark Core: The Spark Core is the foundation and provides basic I/O functionalities, task dispatching and scheduling.

RDDs are basically a collection of partitioned data. These are generally created by referencing datasets in storages such as Cassandra, HBase et al., or by applying transformations such as map, reduce, filter etc. on existing RDDs.

B.**Spark SQL:** Spark SQL, a component on the Core, introduces a new data abstraction called DataFrame, for providing support for structured data. It provides a language to manipulate DataFrames in Java, Python or Scala.

C.**Spark Streaming:** Spark streaming rests on the Core as well, and levera on top of the Core which is proven to be ten times faster than Hadoop's disk-based Apache Mahout due to the distributed memory-based Spark architecture. It implements common algorithms to simplify large scale machine learning pipelines, like logistic or linear regression, decision trees or k-means clustering.

D.**MLlib Machine Learning Library:** This is a machine learning framework on top of the Core which is proven to be ten times faster than Hadoop's disk-based Apache Mahout due to the distributed memory-based Spark architecture. It implements common algorithms to simplify large scale machine learning pipelines, like logistic or linear regression, decision trees or k-means clustering.

E.**GraphX:** It is a graph-processing framework on the Core, and provides an API for graph computation that can model the Pregel abstraction, providing an optimized runtime.

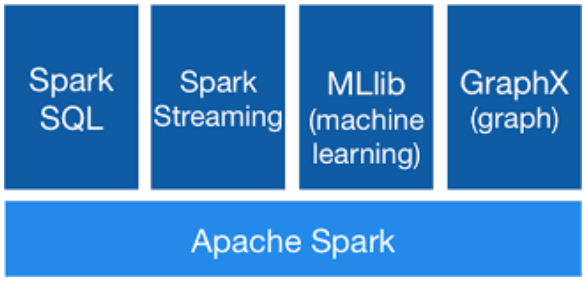

FIG. 2 APACHE SPARK STACK

VII. WORLD RECORD SET BY APACHE SPARK

This is a comparison between Hadoop Map Reduce and Apache Spark for sorting data and setting world record:

TABLE 2
COMPARISON BETWEEN HADOOP MAP REDUCE AND APACHE SPARK

	Hadoop MR Record	Spark Record	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
Cluster disk throughput	3150 GB/s	618 GB/s	570 GB/s
Sort Rate	1.42TB/min	4.27TB/min	4.27TB/min

VIII. CONCLUSION

The paper concludes with the proposition that Big Data is a booming field at the current moment, and the immense amount of data that gets generated every moment calls for a very effective management and analysis system that can deal with the magnitude.

Furthermore, this paper seeks to justify the characteristics of Apache Spark and its standing as very efficient software pertaining to the current scenario of Big Data. In October 2014, Databricks took an interest in the Sort Benchmark and set another world record for sorting 100 terabytes (TB) of information, or 1 trillion 100-byte records. The group utilized Apache Spark on 207 EC2 virtual machines and sorted 100 TB of information in 23 minutes. In comparison, the past world record set by Hadoop MapReduce utilized 2100 machines as a part of a private data center and took 72 minutes. This section tied with a UCSD research group constructing high performance frameworks.

REFERENCES

[1] Harshawardhan S. Bhosale,Prof. Devendra P. Gadekar, "A review, paper on big data and hadoop", International Journal of Scientific Research and Publications, Vol. 4, October 2014

[2] BhartiThakur,ManishMann, "Data Mining for Big Data:AReview",International Journal of Advanced Research in Computer Science and Software Engineering,Vol. 4,May 2015

[3] Wei Fan,AlbertWeifet,"Mining Big Data:CurrentStatus,and Forecast to the Future",SIGKDDExplorations,Vol. 14,February 2012

[4] Hsinchunchen,Roger H.L chiang,Veda C. storey, "Business Intelligence and Analytics: From Big Data to Big Impact",MIS Quarterly Special Isuue: Business Intelligence Research,Vol. 36,December 2012

[5] Sanjay Rathee, "Big Data and Hadoop with components like Flume, Pig, Hive and Jaql",International Conference on Cloud,Big Data and Trust 2013,Vol. 15,November 2013

[6] A. Vailaya, "What's All the Buzz Around"Big Data?"", IEEE Women in Engineering Magazine, December 2012, pp. 24-31,

[7] E. Begoli and J. Horey, "Design Principles for Effective Knowledge Discovery from Big Data", Software Architecture (WICSA) and EuropeanConference on Software Architecture (ECSA) Joint Working IEEE/IFIP Conference on, Helsinki, August 2012