International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 3 Issue 11 ǁ Nov. 2015 ǁ PP.85-91

Big Data Analytics Frameworks

Tanmaya Kumar Pattnaik, Pratyushbhanu Khuntia, Biraja Prasad Nayak
Department of Computer science and Engineering, NM Institute of Engineering and Technology,Bhubaneswar , Odisha
Department of Computer science and Engineering ,Aryan Institute of Engineering and Technology Bhubnaeswar , Odisha
Department of Computer science and Engineering,Capital Engineering College,odisha

ABSTRACT— *Big Data concerns massive, heterogeneous, autonomous sources with distributed and decentralized control. These characteristics make it an extreme challenge for organizations using traditional data management mechanism to store and process these huge datasets. It is required to define a new paradigm and re-evaluate current system to manage and process Big Data. In this paper, the important characteristics, issues and challenges related to Big Data management has been explored. Various open source Big Data analytics frameworks that deal with Big Data analytics workloads have been discussed. Comparative study between the given frameworks and suitability of the same has been proposed.*

Keywords—*Big Data Analytics; Big Data Issues and Challenges; Apache Hadoop; Apache Drill; Project Storm*

I. INTRODUCTION

Digital universe is flooded with huge amount of data generated by number of users worldwide. These data are of diverse in nature, come from various sources and in many forms. To keep with the desire to store and analyze ever larger volumes of complex data, relational databases vendors have delivered specialized analytical platforms that come in many shapes and sizes from software only to analytical services that run in third party hosted environments. In addition new technologies have emerged to address exploding volumes of complex data, including web traffic, social media content and machine generated data including sensor data, global positioning system data. New non-relational database vendors combine text indexing and natural language processing techniques with traditional database technologies to optimize ad-hoc queries against semi-structured data. Number of analytical platform are available in the market for analysis of complex structured and unstructured data, each of which is designed to handle specific type of data/workload. In this paper we will discuss three open source Big Data Analytics frameworks suitable for different types of workload.

This paper is organized as follows: Section 2 we will discuss characteristics of Big Data, motivation for adapting Big Data and analytics platform in organization. Section 3 will discuss issues and challenges organization facing with Big Data and Analytics. Section 4 discusses three open source Big Data analytics frameworks. Comparison between these 3 frameworks and suitability of the framework is suggested in section 5.

II. BIG DATA AND ANALYTICS CHARACTESISTICS

The term Big Data covers diverse technologies same as cloud computing. Input to Big Data systems comes from web server logs, social networks, satellite imagery, traffic flow sensors, broadcast audio sensors, banking transaction etc. This data is called Big Data. To identify data as Big Data should be analyzed from different dimensions.

A. Characteristics of Big Data

Big Data can be characterized by different aspects. The commonly used aspects are Volume, Velocity and Variety. Veracity and Value are also used to characterize Big Data. They are helpful lens through which we can understand the nature of Big Data and the platform available to exploit them.

Volume - As infrastructure becomes increasingly available and affordable, data generated by different sources is very huge in size; petabytes or zettabytes. This huge amount of data is called Big Data.

Velocity - The sheer velocity at which we are creating data is huge cause of Big Data. Digital universe expands from 130 million to 40 trillion in 8 years (2005-2013). The data generated from various sources range from Batch to Real time. So this high velocity data defines new term called “Big Data” .

Variety - The representation of data generated by various sources are diverse in nature; for example ecommerce web sites deal with structured data[12], web server logs deal with semi structured data[13] and social websites deal with unstructured data like audio, video, images etc… Hence big data can be categorized into structured, unstructured and semi structured types and digital universe deals with combination of all.

Veracity - Duo to sheer velocity of some data we cannot spend time in cleans the data before using it. Compiling multisource data and use it for decision making for business requires mechanism that deals with imprecise data. Hence combination of precise, imprecise, accurate, data can be called big data.

Value - By processing huge volume, high velocity, variety and veracity of data, presents a new dimension for analyzing big data called "value". Collaborating different types of data, putting them all together in order to extract hidden knowledge for business and getting competitive advantage from it represents value of big data.

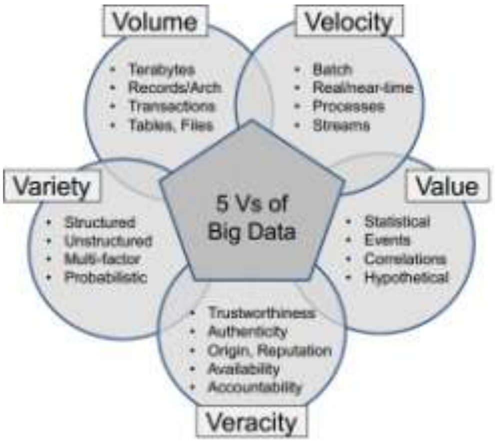

Fig. 1. Characteristics of Big Data.[11]

B. Motivaton for Big Data and Analytics

Statistics [18] shows that rate of data generated on digital universe is increasing exponentially. Current tools and technologies are not up to the mark to store and process huge amount of data. They are also unable to extract value from these data which is most important. When an enterprise can leverage all the information available with large data rather than just a subset of its data then it has a powerful advantage over the market competitors. Big Data can help to gain insights and make better decisions. In order to handle big data modified paradigms are required. Following are some areas where Big Data can play important role.

Big Data Analytics and Health Care

Medical practitioners store huge amount of data about patients' medical history, medication and other details. Huge amount of data are being stored by drug manufacturing company. These data are very complex in nature and sometimes practitioners cannot correlate with other information, thus results in important information remains hidden. By applying advance analytics techniques, this hidden information can be extracted, which results in personalized medication. Advance analytics techniques can also gain insight into genetic and environmental causes of diseases.

Big Data Analytics and intelligence agencies

Intelligence agencies collect huge amount of data from different sources like satellite imagery, signal intercepts, and publicly available sources. Connecting dots, by linking all the information, possible threats can be found, thefts can be prevented or detected. All of these requires robust analytics technique that handles large amount of complex data.

Big Data Analytics and Environment

Understanding environment requires huge amount of data collected from various sources like sensors monitoring air and water quality, metrological conditions, proportion of $CO2$ and other gas in air etc. By linking all information together important trending such as increased $CO2$ emission, increase or decrease of greenhouse effect can be found out.

All above example shows that adaption of new frameworks, tools and technologies result into extraction of valuable information which remains hidden previously.

III. BIG DATA ANALYTICS ISSUES AND CHALLENGES

Organization dealing with Big Data facing numerous challenges. System working with Big Data need to understand the need of technology and need of user. Meeting challenges presented by Big Data will be difficult; volume of data increasing every day, velocity of its generation is increasing faster than ever; variety of data is also expanding.

Current tools, technologies, architecture, management, and analysis approaches are unable to cop up with complexity of data presented. Some challenges are presented below.

Privacy, Security and Trust - Organization using Big Data, committed to protect the privacy, security of its users and should ensure that the organization must comply all privacy and security related act to enhance the protection of and set clear boundaries for usage of personal information.

Trust in the organization needs to be maintained as the volume of data holding increases. The trust that users have in these agencies and their abilities to securely hold information of a personal can easily be affected by leakage of data or information into public domain.

Data Management and Sharing - Agencies realize that for data to have any value, it needs to be discoverable, accessible and usable. Agencies must achieve these requirements but still adhering to privacy laws. Current trends towards open data has seen a focus on making datasets available to the public. Agencies must put focus on making data available, open and standardize within and between agencies in such a way that allows agencies to use and collaborate to the extent made possible by privacy laws.

Technology and Analytical skills - Big Data and Analytics put lot of stress on ICT providers for developing new tools and technology to handle complex data. Current tools and technologies are unable to store, process and analyze massive amount of diverse data. Vendors and developers of Big Data systems and solutions including open source software are developing more capable tools to simplify the challenges of Big Data Analytics.

Some specific challenges related to Big Data and Analytics are:

Data Storage and Retrieval - Current available technologies are able to handle data entry and data storage. But the tools designed for transaction processing which will add, update, search for small to huge amount of data is not be able to handle big data. How to handle semi or unstructured data for processing it is yet unknown [2].

Quality vs. quantity - When dealing with huge amount of data, sometime it is difficult to decide:

- Which data is inappropriate and how do we select most appropriate data?
- How do ensure authenticity of the data?
- How to estimate the value of data?

Data Growth and Expansion - As the organizations increases their services, their data is also expected to grow. Few organization also consider data expansion because of data grow in richness, data evolved with new techniques [2].

Speed and scale - When volume of data grows, it is difficult to gain insight into data within time period. Gaining insight into data is more important than processing complete set of data. Processing near real time data will always require processing interval in order to produce satisfactory output [2].

Structured and unstructured data - Transition between structured data- stored in well-defined tables and unstructured data (images, videos, text) required for analysis will affect end to end processing of data. Invention of new non-relational technologies will provide some flexibility in data representation and processing [2].

Data ownership - Very huge amount of data resides in the servers of social media service providers. These data is not really owned by them but they store data of their users. Actual owner of the page is one who has created the page or account. This is ongoing and big challenge in area of social media [2].

IV. BIG DATA ANALYTICS FRAMEWORKS

Different types of data when we consider Big Data.

Different types of framework required to run different types of analytics. A variety of workloads present in large-scale data processing enterprise. In order to achieve a business goal, we often see a combination of said workloads deployed:

- Batch-oriented processing, for example, Map Reduce based frameworks like Hadoop, for recurring tasks such as large-scale data mining or aggregation [8].
- OLTP, such as user-facing e-commerce transactions, with Apache HBase [14]

- Stream processing, to handle stream sources such as social media feeds or sensor data, with Storm being a representative framework [9].
- Interactive ad-hoc query and analysis with Apache Drill [5].

A. Apache Hadoop

Apache Hadoop is open source software library which includes framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It has a variety of options ranging from single computer to thousands of computers, each of which offering local computation and storage.

Instead of depending on hardware, library itself designed to detect and handle failure and assure high-availability at application layer [7].

Apache Hadoop include following modules:

a) Hadoop core: Common utilities that support other modules
b) Hadoop distributed file system: Provide high throughput access to application data.
c) Hadoop YARN: Framework for job scheduling and resource management
d) Hadoop Map Reduce: Framework for parallel processing of large data set.

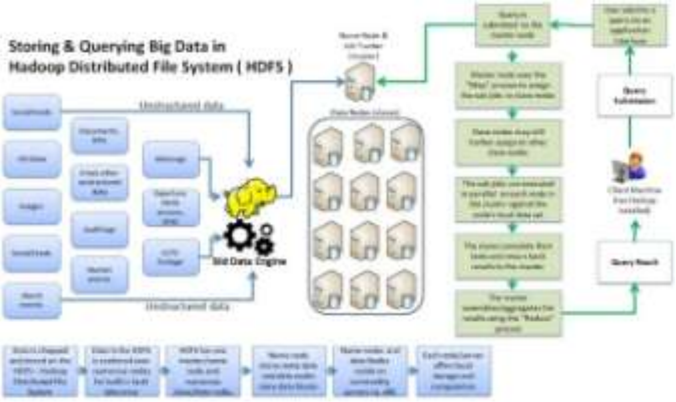

Fig. 2. Data store and retrival in Apache Hadoop system [15]

Data management with Apache Hadoop is shown in Fig.2. Here query is submitted by user to Hadoop Engine which will take input data from HDFS. Data is spread across number of data nodes. There is one name node or Job Tracker which will take care of assigning the work among data nodes and producing the result and responding back to user. Architecture of Apache Hadoop is very robust and fault-tolerant. Job Tracker is continuously tracing the status of data node and if data node remains silent for more than predefined time, task of that data node is given to another data node.

B. Project Storm

Hadoop and related technologies have made it possible to store and process data at scales previously unthinkable.

Unfortunately, these data processing technologies are not real-time systems. However, real-time data processing at massive scale is becoming more and more of a requirement for businesses. Storm exposes a set of primitives for doing real- time computation. Like how Map Reduce greatly eases the writing of parallel batch processing, Storm's primitives greatly ease the writing of parallel real-time computation [9].

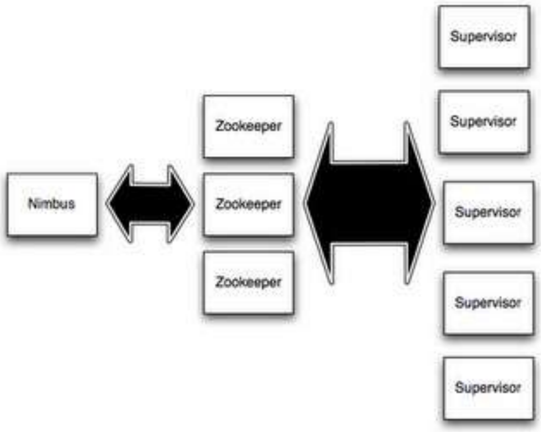

Fig. 3. Architecture of Storm cluster [16]

Architecture of storm cluster is shown in Fig 3. A Storm cluster is superficially similar to a Hadoop cluster. Whereas on Hadoop you run "Map-Reduce jobs", on Storm you run "topologies". "Jobs" and "topologies" themselves are very different -- one key difference is that a Map-Reduce job eventually finishes, whereas a topology processes messages forever (or until you kill it) [9].

There are two kinds of nodes on a Storm cluster: the master node and the worker nodes. The master node runs a daemon called "Nimbus" that is similar to Hadoop's "Job-Tracker". Nimbus is responsible for distributing code around the cluster, assigning tasks to machines, and monitoring for failures. Each worker node runs a daemon called the "Supervisor". The supervisor listens for work assigned to its machine and starts and stops worker processes as necessary based on what Nimbus has assigned to it. Each worker process executes a subset of a topology; a running topology consists of many worker processes spread across many machines. All coordination between Nimbus and the Supervisors is done through a Zookeeper [17] cluster- a coordinating service in distributed environment which will take care of naming, configuration management, synchronization etc. Additionally, the Nimbus daemon and Supervisor daemons are fail-fast and stateless; all state is kept in Zookeeper or on local disk. This means you can kill -9 Nimbus or the Supervisors and they'll start back up like nothing happened. This design leads to Storm clusters being incredibly stable.

C. Apache Drill

Apache Drill is a distributed system for interactive ad-hoc analysis of large-scale datasets. Designed to handle up to petabytes of data spread across thousands of servers, the goal of Drill is to respond to ad-hoc queries in a low latency manner.

Many a times it happens that human sits in front of business application and need to execute ad-hoc queries as per business needs. Query should not need more then few seconds to execute even at scale; some time user do not know which query to fire in advance; also, user need to react to changing circumstances. Apache drill will provide the solution for all above issues [5].

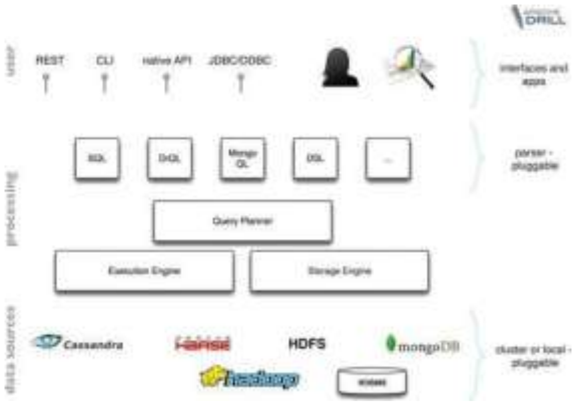

Fig. 4. Architecture block diagram of Apache Drill [5]

High level architecture of Apache Drill is shown in Fig 4. At high level Apache Drill's architecture contains following layers:

User - providing interfaces such as a command line interface (CLI), a REST interface, JDBC/ODBC, etc., for human or application driven interaction.

Processing - allowing for pluggable query languages as well as the query planner, execution, and storage engines.

Data sources - pluggable data sources either local or in a cluster setup, providing in-situ data processing.

Apache Drill is not a database but rather a query layer that works with a number of underlying data sources. It is primarily designed to do full table scans of relevant data as opposed to, say, maintaining indices. Apache Drill provides for a flexible query execution framework, enabling a number of use cases from quick aggregation of statistics to explorative data analysis.

The workers in Apache Drill, suitably called drill-bits, run on each processing node in order to maximize data locality. The coordination of the drill-bits, the query planning, as well as the optimization, scheduling, and execution are performed and distributed.

V. CONCLUSION

In this work a detailed study of Big Data and analytics has been performed and comparison between different frameworks is given below:

TABLE 1 COMPARISION BETWEEN BIG DATA ANALYTICS FRAMEWORKS

Features	**Apache Hadoop**	**Project Storm**	**Apache Drill**
Owner	Community	Community	Community
Workload	Batch processing	Real time computation / stream analysis	Interactive and Ad-hoc analysis
Source code	Open	Open	Open
Low Latency	No	Yes	Yes
Complexity	Easy	Easy	Complex

As shown in above table, Apache Hadoop is suited for workload where time is not critical factor whereas Project storm is well suited for data stream analysis in which analysis performed is real time and Apache drill is best for interactive and ad-hoc analysis. Following points related to Big Data and Analytics are worth noted.

- There is a requirement of Big Data Analytics frameworks for the organization that deal with different types of Big Data workloads. In addition a middleware architecture is also required to integrate and process all Big Data related workloads.
- Organization dealing with Big Data and Analytics need to deal with challenges like privacy, security, data management and sharing, technology, skills and other specific challenges related to workload present in the organization.

REFERENCES

[1] Katal, A., Wazid, M., Goudar, R.H., "Big data: Issues, challenges, tools and Good practices", Sixth International Conference on Contemporary Computing (IC3) 2013.

[2] Stephen K, Frank A, J. Alberto E, William M, "Big Data: Issues and Challenges Moving Forward", IEEE, 46th Hawaii International Conference on System Sciences, 2013.

[3] Sachchidanand S, Nirmala S, "Big Data Analytics", IEEE, International Conference on Communication, Information & Computing Technology (ICCICT),Oct. 19-20, 2012.

[4] Katina Michael, Keith W. Miller, "Big Data: New Opportunities and New Challenges", IEEE Technology and Society Magazine, vol 13.

[5] Michael Hausenblas, Jacques Nadeau, "Apache Drill Ad-hoc interactive analysis at scale", June 2013.

[6] Sergey M, Andrey G, Jing Jing L, Geoffrey R, Shiva S, Matt T, Theo V, "Dremel: Interactive Analysis of Web-Scale Datasets", Google 2013.

[7] Jeffrey Dean and Sanjay Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters", Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

[8] "Apache-Hadoop"- http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F

[9] "Project Storm" - http://storm-project.net/

[10] Apache-Drill"- https://cwiki.apache.org/confluence/display/DRILL/Apache+Drill+Wiki

[11] Characteristics of Big Data - http://www.datatechnocrats.com/tag/big- data/
[12] “Structured Data” - http://www.webopedia.com/TERM/S/structured_data.html
[13] “Semi structured Data” – http://en.wikipedia.org/wiki/Semi- structured_data
[14] Apache HBase - http://hbase.apache.org/
[15] Storing and querying data Big Data in HDFS - http://ecomcanada.wordpress.com/2012/11/14/storing-and-querying-big- data-in-hadoop-hdfs/
[16] Storm cluster - https://github.com/nathanmarz/storm/wiki/Tutorial
[17] Apache Zookeeper - http://zookeeper.apache.org/
[18] Big Data statistics - wikibon.org/blog/big-data-statistics/