Big Data Analysis

R. Dharshini[1] C. Deivasikamani[2] V. Kavipriya[3] K. Kousalyaa[4]
[1,2,3,4]UG Student
[1,2,3,4]Department of Computer Science
[1,2,3,4]Sri Krishna Arts and Science, India

Abstract— The Journal of Big Data publishes high-quality, scholarly research papers, methodologies and case studies covering a broad range of topics, from big data analytics to data-intensive computing and all applications of big data research. The journal examines the challenges facing big data today and going forward including, but not limited to: data capture and storage; search, sharing, and analytics; big data technologies; data visualization; architectures for massively parallel processing; data mining tools and techniques; machine learning algorithms for big data; cloud computing platforms; distributed file systems and databases; and scalable storage systems. Academic researchers and practitioners will find the journal of big data to be a seminal source of innovative material.

Key words: Methodologies, Big Data Challenges, Data Visualization, Case Studies & Data Analytics

I. INTRODUCTION

The term Big Data refers to all the data that is being generated across the globe at an unprecedented rate. This data could be either structured or unstructured. Today's business enterprises owe a huge part of their success to an economy that is firmly knowledge-oriented. Data drives the modern organizations of the world and hence making sense of this data and unravelling the various patterns and revealing unseen connections within the vast sea of data becomes critical and a hugely rewarding end eavour indeed. There is a need to convert Big Data into Business Intelligence that enterprises can readily deploy. Better data leads to better decision making and an improved way to strategize for organizations regardless of their size, geography, market share, customer segmentation and such other categorizations. Hadoop is the platform of choice for working with extremely large volumes of data. The most successful enterprises of tomorrow will be the ones that can make sense of all that data at extremely high volumes and speeds in order to capture newer markets and customer base.

II. BIG DATA CHALLENGES

A. Volume

The first characteristic of Big Data, which is "Volume", refers to the quantity of data that is being manipulated and analyzed in order to obtain the desired results. It represents a challenge because in order to manipulate and analyze a big volume of data requires a lot of resources that will eventually materialize in displaying the requested results. For example a computer system is limited by current technology regarding the speed of processing operations. The size of the data that is being processed can be unlimited, but the speed of processing operations is constant. To achieve higher processing speeds more computer power is needed and so, the infrastructure must be developed, but at higher costs. By trying to compress huge volumes of data and then analyze it, is a tedious process which will ultimately prove more ineffective. To compress data it takes time, almost the same amount of time to decompress it in order to analyze it so it can be displayed, by doing this, displaying the results will be highly delayed. One of the methods of mining through large amount of data is with OLAP solutions (Online Analytical Processing).An OLAP solution consists of tools and multidimensional databases that allow users to easily navigate and extract data from different points of view. Therefore, it identifies relations between elements in the database so it can be reached in a more intuitive way. For obtaining results various OLAP tools are used in order for the data to be mined and analyzed.

B. Velocity

"Velocity" is all about the speed that data travels from point A, which can be an end user interface or a server, to point B, which can have the same characteristics as point A is described. This is a key issue as well due to high requests that end users have for streamed data over numerous devices (laptops, mobile phones, tablets etc.). For companies this is a challenge that most of them can't keep up to. Usually data transfer is done at less than the capacity of the systems. Transfer rates are limited but requests are unlimited, so streaming data in real-time or close to real-time is a big challenge. The only solution at this point is to shrink the data that is being sent. A good example is Twitter. Interaction on Twitter consists of text, which can be easily compressed at high rates. But, as in the case of "Volume" challenge, this operation is still time-consuming and there will still be delay in sending-receiving data. The only solution to this right now is to invest in infrastructure.

C. Variety

"Variety" is the third characteristic of Big Data. It represents the type of data that is stored, analyzed and used. The type of data stored and analyzed varies and it can consist of location coordinates, video files, data sent from browsers, simulations etc. The challenge is how to sort all this data so it can be "readable" by all users that access it and does not create ambiguous results. The mechanics of sorting has two key variables at the beginning: the system that transmits data and the system that receives it and interpret it so that can be later displayed.[1]

III. DISTRIBUTED SYSTEM CHALLENGES

- Programming Complexity
- Finite
- The data bottleneck

IV. HADOOP INTRODUCTION

A scalable fault-tolerant distributed system for data storage and processing. Distribute data when the data is stored.

Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.[2]
Some of the characteristics:

- Open source
- Distributed processing
- Distributed storage
- Scalable
- Reliable
- Fault-tolerant
- Economical
- Flexible

A. History

Originally built as an Infrastructure for the "Nutch" project. Based on Google's map reduce and google File System. Created by Doug Cutting in 2005 at Yahoo Named after his son's toy yellow elephant.

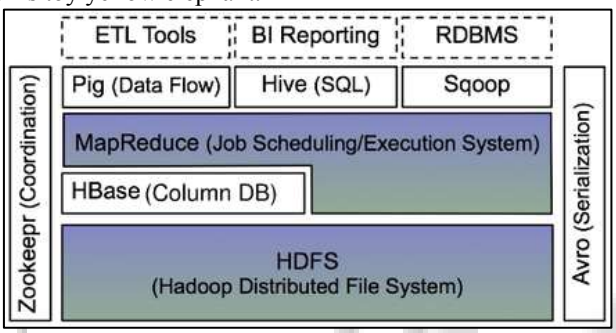

Fig. 1: Hadoop Ecosystem

V. HADOOP COMPONENTS

A. HDFS & MAP Reduce

Hadoop Distributed File System (HDFS) is designed to reliably store very large files across machines in a large cluster. It is inspired by the Google File System .Distribute large data file into blocks. Blocks are managed by different nodes in the cluster. Each block is replicated on multiple nodes. Name node stored metadata information about files and blocks [3].

1) The Mapper

Each block is processed in isolation by a map task called mapper. Map task runs on the node where the block is stored.

2) The Reducer:

Consolidate result from different mappers. Produce final output.

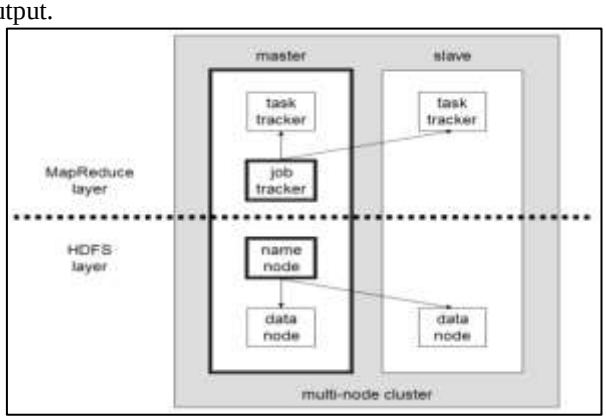

Fig. 2: HDFS & MAP Reduce

B. HBASE

HBase is an open source, non-relational, distributed database modeled after Google's Big Table. It runs on top of Hadoop and HDFS, providing Big Table-like capabilities for Hadoop. When there is real big data: millions or billions of rows, in other way data cannot store in a single node. When random read/write access to big data. When require to do thousands of operations on big data. When there is no need of extra features of RDMS like typed columns, secondary indexes, transactions, advanced query languages, etc. When there is enough hardware.[4]

C. HIVE

An SQL like interface to Hadoop. Data warehouse infrastructure built on top of Hadoop. Provide data summarization, query and analysis. Query execution via MapReduce. Hive interpreter convert the query to Map reduce format. Open source project. Developed by Facebook. Also used by Netflix, Cnet, Digg, eHarmony etc.

1) HiveQL Example

SELECT customerId, max(total_cost) from hive_purchases GROUP BY customerId HAVING count(*) > 3;

D. PIG

A scripting platform for processing and analyzing large data sets. Apache Pig allows to write complex MapReduce programs using a simple scripting language. High level language. Pig Latin. Pig Latin is data flow language. Pig translate Pig Latin script into MapReduce to execute within Hadoop. Open source project. Developed by Yahoo[5].
Pig Latin example:

```
A = LOAD 'student' USING PigStorage() AS (name:chararray, age:int, gpa:float);

X = FOREACH A GENERATE name,$2;

DUMP X;
```

E. SQOOP

Command-line interface for transforming data between relational database and Hadoop. Support incremental imports. Imports use to populate tables in Hadoop. Exports use to put data from Hadoop into relational database such as SQL server.

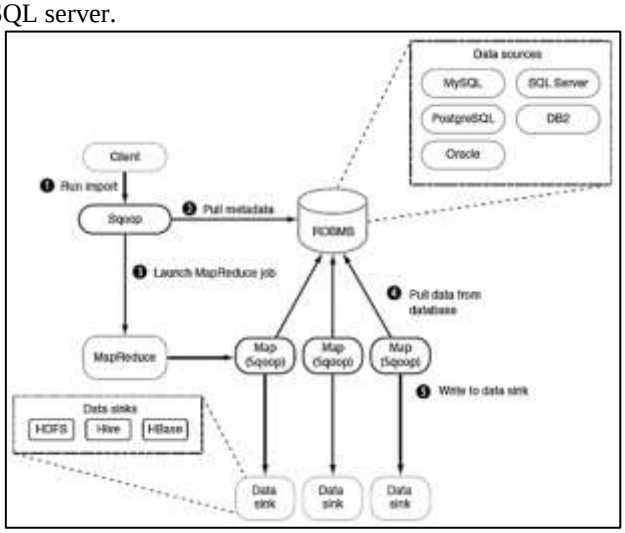

Fig. 3: Working of SQOOP

F. FLUME

Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).[6]

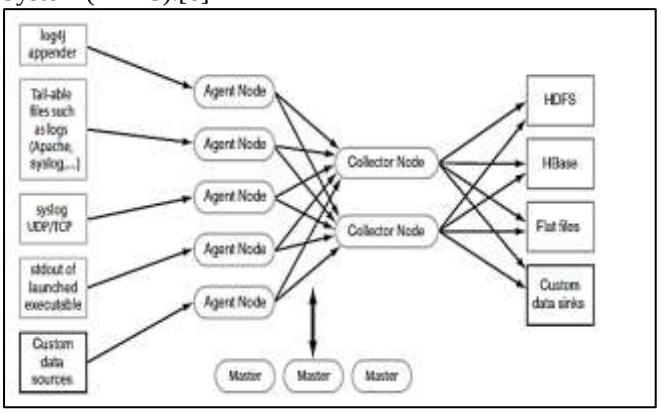

Fig. 4: Working of Flume

Data flows like: Agent tier -> Collector tier -> Storage tier

Agent nodes are typically installed on the machines that generate the logs and are data's initial point of contact with Flume. They forward data to the next tier of collector nodes, which aggregate the separate data flows and forward them to the final storage tier.

G. HUE

Graphical front end to the cluster. Open source web interface. Makes Hadoop platform (HDFS, Map reduce, oozie, Hive, etc.) easy to use.

H. ZOOKEEPER

Because coordinating distributed systems is a Zoo. Zoo Keeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.[7]

HBase	Hadoop database for random read/write access
Hive	SQL-like queries and tables on large datasets
Pig	Data flow language and compiler
Oozie	Workflow for interdependent Hadoop jobs
Sqoop	Integration of databases and data warehouses with Hadoop
Flume	Configurable streaming data collection
ZooKeeper	Coordination service for distributed applications

Table 1:

VI. CONCLUSION

The availability of Big Data, low-cost commodity hardware, & new information management and analytic software have produced a unique moment in the history of data analysis. The convergence of these trends means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability.

The Age of Big Data is here, and these are truly revolutionary times if both business and technology professionals continue to work together and deliver on the promise.

REFERENCES

[1] Global data center traffic – Cisco Forecast Overview - http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html
[2] http://training.cloudera.com/essentials.pdf
[3] http://en.wikipedia.org/wiki/Apache_Hadoop
[4] http://practicalanalytics.wordpress.com/2011/11/06/explaining-hadoop-to-management-whats-the-big-data-deal/
[5] https://developer.yahoo.com/hadoop/tutorial/module1.html
[6] http://hadoop.apache.org/
[7] http://wiki.apache.org/hadoop/FrontPage/

