International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 7 | ISSN : 2456-3307

Big Data, Technologies and Trends A Study

Jayamma Rodda[1], R. VijayaKumari[2]

[1]Assistant Professor, Department of Master of Computer Science, KBN College, Vijayawada, Andhra Pradesh, India

[2]Assistant Professor, Department of Computer Science and Applications, Krishna University, Machilipatnam, Andhra Pradesh, India

ABSTRACT

Big Data is the most droned terms among Scholars and trade. Sprouting Big Data applications has become gradually more significant in the last few years. In actuality, quite a few organizations from dissimilar sectors depend all the time more on knowledge extracted from giant volume of data. The different types of users produce massive quantity of data which is not alike in features. The term of big data is referring big data set coming from various sources in the form of structure and non-structured data. They need to cater to find out the useful information from the massive, noisy data. However, in Big Data context, traditional data techniques and platforms are less efficient. They show a slow responsiveness and lack of scalability, performance and accuracy. Big Data aims to help to select and adopt the right combination of different Big Data technologies according to their technological needs and specific applications' requirements. The objective of this paper is to provide a simple, comprehensive and brief introduction of Big Data and its technologies. This paper may not cover every dimension of Big Data only few of its essential aspects are covered.

Keywords : Big Data, Hadoop, HDFS, Map Reduce, HD Insight, No SQL, Poly base, Presto, PIG, HIVE, and R

I. INTRODUCTION

The word "Big data" is used by sociologist Mr. Charles Tilly in his article. Later the word "Big Data" is used by CNN in year of 2001 in news article. Now a days data has become the principal demand in the industry. A huge amount of data is produced every day by the extremeuse of internet. This data can be generated in either of the two ways by human or machine. Human produces data is in the form of emails, videos, documents, images, videos and the posts they post in the twitter and face book or any other websites etc., The machine generated data is logs data (which is from web logs, email logs, click logs) and sensor data. This category of data is larger in size than the data generated by the human.With the advent of the Big Data technologies the processing of this machine generated data has been the task. The Key basis for the Big Data are Transaction data that is produced by the purchases made in online , data from Web, data from Social media, data generated by click stream, the data from sensors and GPS signal data from Mobile phones[1-2]. Major amount of social media data is produced from the social networking sites like LinkedIn, Twitter, face book. Always the E-commerce or online ad companies observe for the users steering data from their click stream to a website. Usually the machines have sensors embedded in themthese sensors produces enormous data.If we look into the sources of Big Data face book has at the most 40PB of data and daily it seizes more than 100 TB of data, whereas The Twitter seizes 8 TB data daily and Yahoo has 60 PB data[1]. The processing, analysing and knowledge mining from this outsized data has always been challenging. As the Classic data analytical tools do not make a provision for the large scale data therefore other distributed data analytical tools are to be used for data analytical tools cannot support large scale data. Nowadays, the web is burdened with a proponent production of excessive amount of data by individuals and organizations. It is doubling in size every two years.

II. METHODS AND MATERIAL

1. BIG DATA

Big Data as the term implies the data in the aspect of storing and processing which is difficult to be handled by a single system. This Big Data is the kind of data which is not only big in size but also cleaning, storing, managing and manipulating and processing.
Big Data encompasses three types of data[42].

1. **Structured Data**Data that has a well-defined structure falls under Structured Data. This data has records divided into rows and columns. As a result, it is easy to read and manage this data. Usually, data in databases falls under Structured Data. Well defined software are present to read structured data. Furthermore, it is easy to set constrains for this type of data.

   Structured Data accounts for only 20% of the total data available. Specialized programs can be written to read this data and write into databases. On the other hand, users can manually enter data into databases.
2. **Semi-structured data** The main difference between Structured and Semi-structured is that Structured data cannot be stored in traditional databases. However, Semi-Structured still retains some organizational properties which makes it easier to process than Unstructured data. Some common examples of Semi-structured data is data stores in CSV files and No SQL documents.
3. **Unstructured data**Unstructured data, no format is present. Hence, unstructured data can be used to store any kind of data to it. The downside of using unstructured data is that the data does not have any constraints to it. Hence storing this data and managing and manipulating this data proves to be a difficult task. Unstructured data provides a way to store data like images and videos which cannot be processed using traditional databases.

Definition

Most well-known definition of Big Data is a five Vs[42] concept: Volume, Velocity, Variety, Veracity and Value. So data possesses large volume, comes with high velocity, from variety of sources and formats with great value and having great uncertainty is referred as Big Data. Volume- represents scale of data i.e. Big Data has massive volume. Velocity- refers speed of generation and processing of data i.e. rate of entering streaming data in the system is really fast. Variety- refers different form of data i.e. unstructured or semi-structured data (text, sensor data, audio, video, click stream, log file, XML) originated from different sources. Veracity- refers uncertainty of data i.e. quality of data being captured. Data like posts on social networking sites are imprecise [5-6]. Value represents that big data must have value.

To extract knowledge from Big Data, various models, programs, softwares, hard wares and technologies have been designed and proposed. They try to ensure more accurate and reliable results for Big Data applications. In this paper, we present a survey on recent technologies developed for Big Data.

2. Big Data Technologies:

Top big data technologies used to store and analyse data are:

Apache Hadoop. Apache Hadoop is a java based free software framework that can effectively store large amount of data in a cluster.

Hadoop[41] is a large-scale distributed batch processing infrastructure for parallel processing of big data on large cluster of commodity computers [16]. Hadoop consists of three core components: HDFS, MapReduce and YARN. HDFS and MapReduce design are based on Google's File System and MapReduce. YARN framework is a NextGenMapReduce also called MapReduce 2.0, was added in Hadoop-2.x version for job scheduling and resource management of Hadoop cluster. Hadoop is extremely scalable distributed system and requires minimum networks bandwidth. Hadoop infrastructure automatically handles fault tolerance, data distribution, parallelization and load balancing tasks. In traditional parallel and distributed system, data are moved to the node for computation which can never be feasible in case of Big Data. Hadoop is a joint system providing computational power i.e. MapReduce and distributed storage i.e. HDFS at one place. Its design is based on distributing computational power to where the data is; instead of moving data [17].

HDFS Architecture

HDFS[41] is a distributed file system, which provides unlimited storage, scalable and fast access to stored data. It supports horizontal scalability. Thousands of nodes in a cluster hold petabyte scale of data and if there is a requirement of more storage, one needs to just add more nodes only [1]. It uses block-structured file system and stores the files in a replicated manner after breaking the file into fixed size blocks. Default block size is 64 MB and each block is replicated at three nodes by default. Storing data in this way provides high fault tolerance and availability during execution of Big Data applications on Hadoop cluster [16-17].

Hadoop is designed on Master-Slave architecture. There is single master node known as NameNode and multiple slave nodes known as DataNodes. Master node coordinates all slave nodes. DataNodes are the workhorses and stores all data. NameNode is the administrator of file system operations i.e. file creation, permissions etc. Without NameNode no one can operate cluster and write/read data. NameNode is called a single point failure [1]. Fig. 1 shows the functionality of NameNode and DataNode in HDFS. NameNode assigns a block id to each block of a file and stores all the metadata of the files in its memory in order to be fast accessed. Metadata are the file name, permission, replication and location of each block of the file. DataNodes store all the files as replicated blocks and retrieve them whenever required.

MapReduce Framework

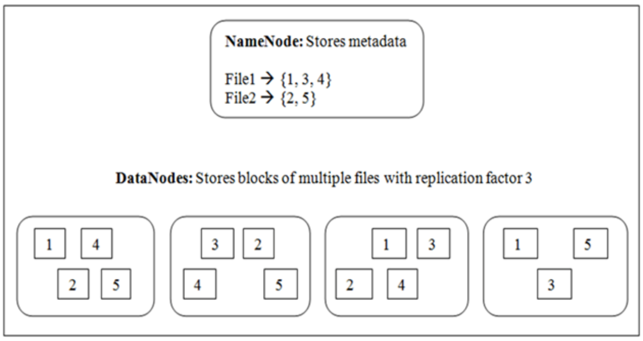
NameNode: Stores metadata
File1 → {1, 3, 4}
File2 → {2, 5}
DataNodes: Stores blocks of multiple files with replication factor 3
1 4 2 5
3 2 4 5
1 3 2 4
1 5 3

Figure 1. NameNode and DataNodes in HDFS [14][41]

MapReduce is an efficient, scalable and simplified programming model for large scale distributed data processing on a large cluster of commodity computers [16] [18-19]. It works on the data residing in HDFS. Map Reduce is a programming framework, which provides generic templates that can be customized by programmer's requirements. It process large volumes of data in parallel by breaking the computation job into independent tasks across a large number of machines. It distributes the tasks across machines in Hadoop cluster and put together the results of computations from each machine. It takes care of the hardware and network failure. A failed task is assigned to other node to re-execute itself without re-executing other tasks. It balances the workload and increase the throughput by assigning work of slower or busy nodes to idle nodes [1] [20].

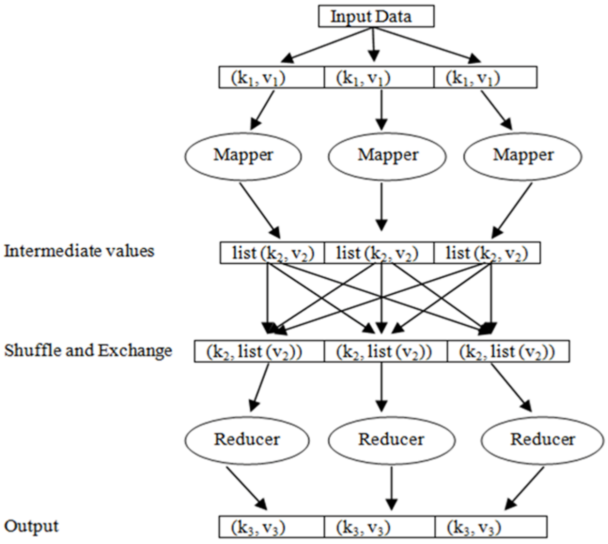

Figure 2. Work flow in MapReduce framework [18][41]

Map Reduce programs can run on Hadoop in multiple languages like Java, Python, Ruby and C++ [21]. MapReduce program consists of two functions Mapper and Reducer which runs on all machines in parallel fashion. The input and output of these functions must be in form of (key, value) pairs. Fig. 2 illustrates the work flow in MapReduce framework.

Map Function: Mapper is applied in parallel on input data set. The Mapper takes the input (k1, v1) pairs from HDFS and produces a list of intermediate (k2, v2) pairs. Mapper output are partitioned per reducer i.e. the number of reduce task for that job.

Reduce Function: The Reducer takes (k2, list (v2)) values as input, make sum of the values in list (v2) and produce new pairs (k3, v3) as final result. *Combiner Function:* It is optional and also known as MiniReducer. It is applied to reduce the communication cost of transferring intermediate outputs of mappers to reducers. Shuffle and exchange is the single point of communication in MapReduce. MapReduce framework shuffle theintermediate output pairs of mappers and exchange them between reducers to send all pairs with the same key to a single reducer [16].

Daemon Processes in Hadoop

Hadoop has five daemons that are the processes running in background. These are NameNode (NN), Secondary NameNode (SNN), DataNodes (DN), JobTracker and TaskTrackers and described as follows [16] [23-24].

NameNode: Each Hadoop cluster has exactly one NameNode which runs on master machine. NameNode manages metadata and access control of the file system.

Secondary NameNode: There is also a backup NameNode named as Secondary NameNode which periodically wakes up and process check points and downloads updates from NameNode. It can be used latter to restore failed NameNode, providing fault tolerance.

DataNodes: DataNode runs on each slave machines in cluster and holds file system. Each DataNode manages blocks of the file system assigned to it. *JobTracker:* Exactly one JobTracker runs in a cluster. All running tasks are halted if JobTracker goes down. Initially jobs are submitted to JobTracker. Then it talks to the NameNode to determine the location of data and talks to TaskTrackers to submit the tasks. *TaskTrackers:*TaskTracker runs on each slave node and accepts map & reduce tasks and shuffle operations from JobTracker.

HDInsight

There are two flavors of HDInsight: Windows Azure HDInsight Service and Microsoft HDInsight Server for Windows (recently quietly killed but lives on in a different form). Both were developed in partnership with Hadoop software developer and distributor

Hortonworks and were made generally available in October, 2013.

Windows Azure HDInsight Service (try) is a service that deploys and provisions Apache Hadoop clusters in the Azure cloud, providing a software framework designed to manage, analyze and report on big data. It makes the HDFS/MapReduce software framework and related projects such as Pig, Sqoop and Hive available in a simpler, more scalable, and cost-efficient environment.

Windows Azure HDInsight Service uses Azure Blob Storage as the default file system (or you can store it in the native Hadoop Distributed File System (HDFS) file system that is local to the compute nodes but would lose the data if you deleted your cluster). There is a thin layer over Azure Blob Storage that exposes it as an HDFS file system called Windows Azure Storage-Blob or WASB Microsoft HDInsight Server for Windows was killed shortly after it was released but lives on in two flavors: Hortonworks Data Platform (HDP) (try) and Microsoft's Parallel Data Warehouse (PDW). Both are on-premise solutions. With HDP, it includes core Hadoop (meaning the HDFS and MapReduce), plus Pig for MapReduce programming, Hive data query infrastructure, Hortonworks' recently introduced HCatalog table management service for access to Hadoop data, Scoop for data movement, and the Ambari monitoring and management console. All of the above have been reengineered to run on Windows and all are open-source components that are compatible with Apache Hadoop and are being contributed back to the community. With PDW, you can add an HDInsight region into the appliance, and this region includes HDP and can be accessed via Polybase.

The HDInsight Server is designed to work with (but does not include) Windows Server and Microsoft SQL Server. In the case of Windows, HDInsight is integrated with Microsoft System Center for administrative control and Active Directory for access control and security.

Types in HDInsight

HDInsight includes specific cluster types and cluster customization capabilities, such as the capability to add components, utilities, and languages. HDInsight offers the following cluster types:

Apache Hadoop: A framework that uses HDFS, YARN resource management, and a simple MapReduce programming model to process and analyze batch data in parallel.

Apache Spark: An open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications. See What is Apache Spark in HDInsight?.

Apache HBase: A NoSQL database built on Hadoop that provides random access and strong consistency for large amounts of unstructured and semi-structured data--potentially billions of rows times millions of columns. See What is HBase on HDInsight?

ML Services: A server for hosting and managing parallel, distributed R processes. It provides data scientists, statisticians, and R programmers with on-demand access to scalable, distributed methods of analytics on HDInsight. See Overview of ML Services on HDInsight.

Apache Storm: A distributed, real-time computation system for processing large streams of data fast. Storm is offered as a managed cluster in HDInsight. See Analyze real-time sensor data using Storm and Hadoop.

Apache Interactive Query preview (AKA: Live Long and Process): In-memory caching for interactive and faster Hive queries. See Use Interactive Query in HDInsight.

Apache Kafka: An open-source platform that's used for building streaming data pipelines and applications. Kafka also provides message-queue functionality that allows you to publish and subscribe to data streams. See Introduction to Apache Kafka on HDInsight.

NoSQL

NoSQL, which stand for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built. NoSQLdatabases are especially useful for working with large sets of distributed data.[27] NoSQL encompasses a wide variety of different database technologies that were developed in response to the demands presented in building modern applications.

Developers are working with applications that create massive volumes of new, rapidly changing data types structured, semi-structured, unstructured and polymorphic data.

- Long gone is the twelve-to-eighteen month waterfall development cycle. Now small teams work in agile sprints, iterating quickly and pushing code every week or two, some even multiple times every day.
- Applications that once served a finite audience are now delivered as services that must be always-on, accessible from many different devices and scaled globally to millions of users.
- Organizations are now turning to scale-out architectures using open source software, commodity servers and cloud computing instead of large monolithic servers and storage infrastructure.
- Relational databases were not designed to cope with the scale and agility challenges that face modern applications, nor were they built to take advantage of the commodity storage and processing power available today.

Sqoop

ApachSqoop [38] is an open source software.It provides a command-line interface (CLI) that ensures an efficient transfer of bulk data between Apache Hadoop and structured data- stores (such as relational databases, enterprise data warehouses and NoSQL databases). Sqoop offers many advantages. For instance,it provides fast performance, fault tolerance and optimal system utilization to reduce processing loads to external systems. The transformation of the imported data is done using MapReduce or any other high-level language like Pig, Hive or JAQL [39]. It allows easy integration with HBase, Hive and Oozie. When Sqoop imports data from HDFS, the output will be in multiple files. These files may be delimited text files, binary Avro or SequenceFiles con- taining serialized data. The process of Sqoop Export will read a set of delimited text files from HDFS in parallel, parse them into records, and insert them as new rows in a target database table.

PolyBase

PolyBase enables your SQL Server 2016 instance to process Transact-SQL queries that read data from Hadoop. The same query can also access relational tables in your SQL Server. PolyBase enables the same query to also join the data from Hadoop and SQL Server. In SQL Server, an external tableor external data source provides the connection to Hadoop.

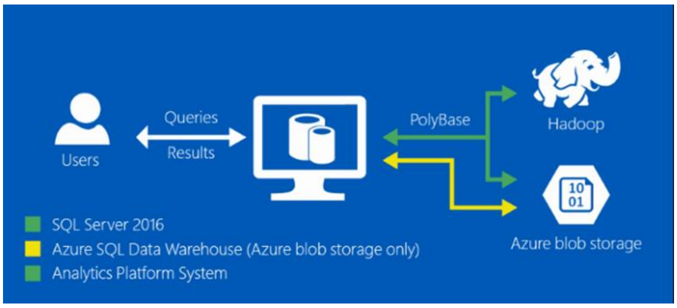
Users
Queries
Results
PolyBase
Hadoop
10
01
Azure blob storage
🟩 SQL Server 2016
🟨 Azure SQL Data Warehouse (Azure blob storage only)
🟩 Analytics Platform System

Figure 3.

PolyBase pushes some computations to the Hadoop node to optimize the overall query. However, PolyBase external access is not limited to Hadoop. Other unstructured non-relational tables are also supported, such as delimited text files.

Supported SQL products and services

PolyBase provides these same functionalities for the following SQL products from Microsoft:

- SQL Server 2016 and later versions (Windows only)
- Analytics Platform System (formerly Parallel Data Warehouse)
- Azure SQL Data Warehouse

Azure integration

With the underlying help of PolyBase, T-SQL queries can also import and export data from Azure Blob Storage. Further, PolyBase enables Azure SQL Data Warehouse to import and export data from Azure Data Lake Store, and from Azure Blob Storage.

In the past it was more difficult to join your SQL Server data with external data.[36] You had the two following unpleasant options:

- Transfer half your data so that all your data was in one format or the other.
- Query both sources of data, then write custom query logic to join and integrate the data at the client level.

PolyBase avoids those unpleasant options by using T-SQL to join the data.

To keep things simple, PolyBase does not require you to install additional software to your Hadoop environment. You query external data by using the same T-SQL syntax used to query a database table. The support actions implemented by PolyBase all happen transparently. The query author does not need any knowledge about Hadoop.

PolyBase enables the following scenarios in SQL Server:

Query data stored in Hadoop from SQL Server or PDW. Users are storing data in cost-effective distributed and scalable systems, such as Hadoop. PolyBase makes it easy to query the data by using T-SQL.

Query data stored in Azure Blob Storage. Azure blob storage is a convenient place to store data for use by Azure services. PolyBase makes it easy to access the data by using T-SQL.

Import data from Hadoop, Azure Blob Storage, or Azure Data Lake Store. Leverage the speed of Microsoft SQL's columnstore technology and analysis capabilities by importing data from Hadoop, Azure Blob Storage, or Azure Data Lake Store into relational tables. There is no need for a separate ETL or import tool.

Export data to Hadoop, Azure Blob Storage, or Azure Data Lake Store. Archive data to Hadoop, Azure Blob Storage, or Azure Data Lake Store to achieve cost-effective storage and keep it online for easy access.

Integrate with BI tools. Use PolyBase with Microsoft's business intelligence and analysis stack, or use any third party tools that are compatible with SQL Server.

Big data in EXCEL

Excel is very commonly used for dataprocessing. With a Table configuration, the column filter reflects on the whole data. With the Sort feature I can order any column numerically, alphabetically or even by date. Pivot Tables – A useful and powerful tool included in Excel to summarize a big table, by the quantities that you need and want.

Excel's role in big data

There are a variety of different technology demands for dealing with big data: storage and infrastructure, capture and processing of data, ad-hoc and exploratory analysis, pre-built vertical solutions, and operational analytics baked into custom applications.[37]The sweet spot for Excel in the big data scenario categories is exploratory/ad hoc analysis. Here, business analysts want to use their favorite analysis tool against new data stores to get unprecedented richness of insight. They expect the tools to go beyond embracing the "volume, velocity and variety" aspects of big data by also allowing them to ask new types of questions they weren't able to ask earlier: including more predictive and prescriptive experiences and the ability to include more unstructured data (like social feeds) as first-class input into their analytic workflow.

Broadly speaking, there are three patterns of using Excel with external data, each with its own set of dependencies and use cases. These can be combined together in a single workbook to meet appropriate needs.

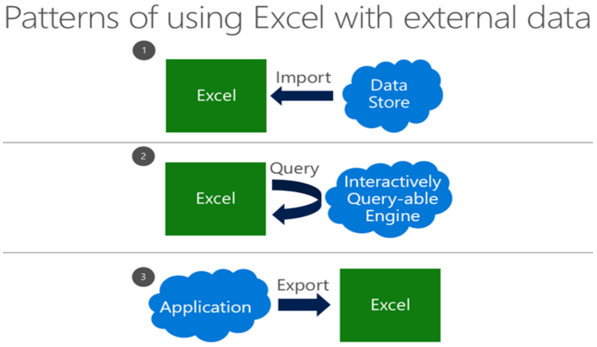
Patterns of using Excel with external data
1
Excel
Import
Data Store
2
Excel
Query
Interactively Query-able Engine
3
Application
Export
Excel

Figure 4

When working with big data, there are a number of technologies and techniques that can be applied to make these three patterns successful.

Import data into Excel

Many customers use a connection to bring external data into Excel as a refreshable snapshot. The advantage here is that it creates a self-contained document that can be used for working offline, but refreshed with new data when online. Since the data is contained in Excel, customers can also transform it to reflect their own personal context or analytics needs.

When importing big data into Excel, there are a few key challenges that need to be accounted for:

Querying big data—Data sources designed for big data, such as SaaS, HDFS and large relational sources, can sometimes require specialized tools. Thankfully, Excel has a solution: Power Query, which is built into Excel 2016and available separately as a download for earlier versions. Power Query provides several modern sets of connectors for Excel customers, including connectors for relational, HDFS, SaaS (Dynamics CRM, SalesForce), etc. We're constantly adding to this list and welcome your feedback on new connectors we should provide out of the box at our UserVoice.

Transforming data—Big data, like all data, is rarely perfectly clean. Power Query provides the ability to create a coherent, repeatable and auditable set of data transformation steps. By combining simple actions into a series of applied steps, you can create a reliably clean and transformed set of data to work with.

Handling large data sources—Power Query is designed to only pull down the "head" of the data set to give you a live preview of the data that is fast and fluid, without requiring the entire set to be loaded into memory. Then you can work with the queries, filter down to just the subset of data you wish to work with, and import that.

Handling semi-structured data—A frequent need we see, especially in big data cases, is reading data that's not as cleanly structured as traditional relational database data. It may be spread out across several files in a folder or very hierarchical in nature. Power Query provides elegant ways of treating both of these cases. All files in a folder can be processed as a unit in Power Query so you can write powerful transforms that work on groups (even filtered groups!) of files in a folder. In addition, several data stores as well as SaaS offerings embrace the JSON data format as a way of dealing with complex, nested and hierarchical data. Power Query has a built-in support for extracting structure out of JSON-formatted data, making it much easier to take advantage of this complex data within Excel.

Handling large volumes of data in Excel—Since Excel 2013, the "Data Model" feature in Excel has provided support for larger volumes of data than the 1M row limit per worksheet. Data Model also embraces the Tables, Columns, Relationships representation as first-class objects, as well as delivering pre-built commonly used business scenarios like year-over-year growth or working with organizational hierarchies. For several customers, the headroom Data Model is sufficient for dealing with their own large data volumes. In addition to the product documentation, several of our MVPs have provided great content on Power Pivot and the Data Model.

Presto (SQL query engine)[34]

Facebook commenced development efforts on Presto in 2012, and announced its release as open source for Apache Hadoop in 2013.[29-30]. In 2014, Netflix disclosed they used Presto on 10 petabytes of data stored in the Amazon Simple Storage Service (S3).[31]Airbnb released the source to web interface software called Airpal for Presto in March, 2015.[32-33]

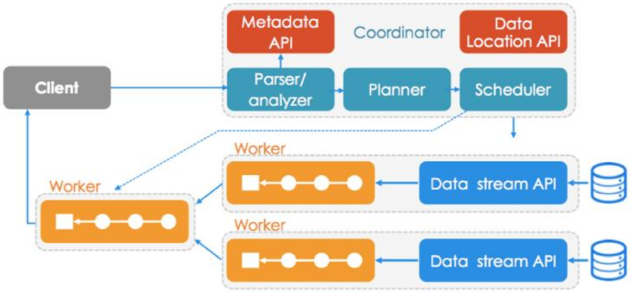

Figure 5. Presto Architecture

PIG

Apache [42]Pig [21] is a higher level of abstraction over MapReduce. Founded in 2006 at Yahoo, Pig was mover into the Apache Software Foundation in 2007. Pig has a few distinctive features that makes it an ideal choice for big data processing.

- Uses a simple language called Pig Latin, for writing code to process data.
- Since it is a higher level of abstraction over MapReduce, Pig code gets converted into MapReduce code internally.
- Users can use UDF (User Defined Functions) that can be written in other programming languages likeJava, etc.

- Compared toMapReduce, Pig requires a smaller amount of code to accomplish the same task.
- User can choose between two execution modes, Local mode and Map Reduce mode.
- Pig uses multiple-query approach, thus reducing the size of the code.
- There are many built in operations and nested data types present in Pig.
- The tasks in Pig are automatically optimized by thePig Engine.

Perhaps the biggest advantage that Pig offers is its ease to handle different types of data. Pig can handle all three types of data, structured, semi structured and unstructured with the same ease and using the same tools.

Pig consists of two parts, Latin Pig and Pig Engine. Latin Pig [20] is the high-level language in which all the Pig scripts are written. Pig Engine is the execution engine that takes in3 Pig script and outputs MapReduce code. This code then runs on the Hadoop platform on the data stored in HDFS and produces the information.

The various parts of Pig Engine are

- Grunt shell
- Pig Server
- Parser
- Optimizer
- Compiler
- Execution Engine

However, since the code needs to be converted into MapReduce code, the time taken by the system to produce the results is high in Pig. Still, owing to its simplicity and ease to learn, Pig has been used by a large number of users as their primary Big Data analysis tool.

HIVE

[42]Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive.
The features of Hive are -

- It stores schema in a database and processed data intoHDFS.
- It is designed for OLAP.
- It provides SQL type language for querying calledHiveQL [19] or HQL.
- It is familiar, fast, scalable, and extensible.

The working of Hive is as follows:

1. Execute Query: The Hive interface such as Command Line or Web UI sends query to Driver (any database driver such as JDBC, ODBC, etc.) to execute.
2. Get Plan: The driver takes the help of query compiler that parses the query to check the syntax and query plan or the requirement of query.
3. Get Metadata: The compiler sends metadata request toMetastore (any database).
4. Send Metadata: Metastore sends metadata as a response to the compiler.
5. Send Plan: The compiler checks the requirement and resends the plan to the driver. Up to here, the parsing andcompiling of a query is complete.
6. Execute Plan: The driver sends the execute plan to the execution engine.
7. Execute Job: Internally, the process of execution job is a MapReduce job. The execution engine sends the job to JobTracker, which is in Name node and it assigns this.

8. Metadata Ops: Meanwhile in execution, the executionengine can execute metadata operations with Metastore.
9. Fetch and send Result: The execution engine receives the results from Data nodes.

The execution engine sends those resultant values to the driver the driver sends the results

R

[42]R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering and many more) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is a programming language and software environment for statistical analysis, graphics representation and reporting. The following are the important features of R:

- R is a well-developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.
- R has an effective data handling and storage facility,
- R provides a suite of operators for calculations on arrays, lists, vectors and matrices.
- R provides a large, coherent and integrated collection of tools for data analysis.
- R provides graphical facilities for data analysis and display either directly at the computer or printing at the papersto Hive Interfaces.

III. RESULTS AND DISCUSSION

Big Data Trends

[41]Big Data opens new opportunities in research and development and is not only limited to Hadoop and its eco-system. A number of tools and projects dedicated to customized requirements are being developed to deploy on top of Hadoop. Many enterprises are launching their own Hadoop distributions. Cloud computing is using Hadoop to provide data processing and storage services. Computation framework of Hadoop is being efficient and flexible. This section gives a brief description of some trends of Big Data.

Big Data Eco-System

Big Data Eco-system[41] is even bigger than Hadoop Eco- System and growing rapidly. We can categorize the projects and tools of Big Data Eco-System on the basis of their core functionality for which they are developed. Table 1 summarizes the Big Data related projects.

Sl. No.	Core	Tools/Projects
1	Getting	Flume, Chukwa, Scoop,
2	Compute	MapReduce, YARN,
3	Querying	Pig, Hive, Cascading

4	Real Time	HBase, Apache Drill, Citus
5	Big Data	HBase, Cassandra, Amazon
6	Hadoop in	Amazon Elastic MapReduce
7	Work Flow	Oozie, Cascading, Scalding,
8	Serializatio	Avro, Protobuf, Trevni
9	Monitoring	Hue, Ganglia, Open, Nagios
10	Application	Mahout, Giraph
11	Stream	Storm, Apache S4, Samza,
12	Business	Datameer, Tableau,

Applications of Big Data

Let us consider some of the Big Data applications[40]: *Smart Grid case:* it is crucial to manage in real time the national electronic power consumption and to monitor Smart grids operations. This is achieved through multiple connections among smart meters, sensors, control centers and other infrastructures. Big Data analytics helps to identify at-risk transformers and to detect abnormal behaviorsof the connected devices. Grid Utilities can thus choose the best treatment or action. The real-time analysis of the generated Big Data allow to model incident scenarios. This enables to establish strategic preventive plans in order to decrease the cor- rective costs. In addition, Energy-forecasting analytics help to bet- ter manage power demand load, to plan resources, and hence to maximize profits [7]

E-health: connected health platforms are already used to personalize health services (e.g., CISCO solution) [8] . Big Data is generated from different heterogeneous sources (e.g., laboratory and clinical data, patients symptoms uploaded from distant sensors, hospitals operations, pharmaceutical data). The advanced analysis of medical data sets has many beneficial applications. It enables to personalize health services (e.g., doctors can monitor online patients symptoms in order to adjust prescription); to adapt public health plans according to population symp- toms, disease evolution and other parameters.It is also useful to optimize hospital operations and to decrease health cost expenditure.

*Internet of Things (IoT):*IoT [9]) represents one of the main markets of big data applications. Because of the high vari- ety of objects, the applications of IoT are continuously evolving. Nowadays, there are various Big Data applications supporting for logistic enterprises. In fact, it is possible to track vehicles positions with sensors, wireless adapters, and GPS. Thus, such data driven applications enable companies not only to supervise and manage employees but also to optimize delivery routes. This is by exploit- ingand combining various information including past driving experience.Smart city is also a hot research area based on the application of IoT data.

Openbenefits: Utilities such as water supply organizations are placing sensors in the pipelines to monitor flow of water in the complex water supply networks. It is reported in the Press that Bangalore Water Supply and Sewage Board is implementing a real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure equitable supply of water to different areas of the city. It helpstp reduce the need for valve operators and to timely identifying and fixing water pipes that are leaking.

Transportation and logistics: [10] Many public road transport companies are using RFID (Radiofrequency Identifi- cation) and GPS to track buses, taxis, autos and explore interesting data to improve there services. In India the call cab or auto services is one of the most implemented services. For instance, data collected about the number of passengers using the buses in different routes are used to optimize bus routes and the frequency of trips. various real-time system has been implemented not only to provide passengers with recommendations but also to offer valuable information on when to expect the next bus which will take him to the desired destination. Mining Big Data helps also to improve travelling business by predicting demand about public or private networks. For instance, in India that has one of the largest railway networks in the world, the total number of reserved seats issued every day is around 250,000 and reservation can be made 60 days in advance. Making predictions from such data is a complicated issue because it depends on several factors such as weekends, festivals, night train, starting or intermediate station. By using the machine learning algorithms, it is possible to mine and apply advanced analytics on past and new big data collection. In fact advanced analytics can ensure high accuracy of results regarding many issues. *Government services and monitoring:* Many government such as India and United States are mining data to monitor political trends and analyze population sentiments. There are many applications that combine many data sources: social network communications, personal interviews, and voter compositions. Such systems enable also to detect local issues in addition to national issues. Furthermore, governments may use Big Data systems to optimize the use of valuable resources and utilities. For instance, sensors can be placed in the pipelines of water supply chains to monitor water flow in large networks. So it is possible for many countries to rely on real-time monitoring system to detect leakages, illegal connections and remotely control valves to ensure equitable supply of water to different areas of the city.

IV. CONCLUSION

There are many researches that have been proposed in this domain and also so many are being going on to make the people aware of any new and unseen facts of this technology.In this paper, we have focused on the components and technologies used in Big Data platforms. Different technologies have been discussed. Also we have discussed in brief about the big data Trends and its applications.Various technologies, that many short comings exist. Most of the time, they are related to adopted architectures and techniques. Thus, further work needs to be carried out in several areas such as data organization, domain specific tools and platform tools in order to create next generation Big Data infrastructures. Hence, technological issues in many Big Data areas can be further studied and constitute an important research topic.

V. REFERENCES

[1]. Mark Kerzner and SujeeManiyam, "Hadoop Illuminated," https://github.com/hadoop-illuminated/hadoop-book, 2013, Accessed on Sept. 20, 2015.

[2]. L Douglas, "3d data management: Controlling data volume, velocity and variety," Gartner, Retrieved 6 (2001).

[3]. IBM What is big data? - Bringing big data to the enterprise. http://www-01.ibm.com/software/in/data/bigdata/, Accessed on Sept. 20, 2015.

[4]. M A. Beyer and L. Douglas, "The importance of big data: A definition," Stamford, CT: Gartner, 2012.

[5]. IBM Big Data &Analytics Hub, http://www.ibmbigdatahub.com/infographic/four-vs-big-data, Accessed on Sept. 20, 2015.

[6]. J S. Ward and A. Barker, "Undefined By Data: A Survey of BigData Definitions," http://arxiv.org/abs/1309.5821v1.

[7]. Stimmel, C.L., 2014. Big Data Analytics Strategies for the Smart Grid. CRC Press.Stoianov, N., Uruea, M., Niemiec, M., Machnik, P., Maestro, G., 2013. Integrated security infrastructures for law enforcement agencies. Multimedia Tools App.,1-16

[8]. Nambiar, R., Bhardwaj, R., Sethi, A., Vargheese, R., 2013. A look at challenges and

opportunities of Big Data analytics in healthcare. In: In: 2013 IEEE International Conference on Big Data. IEEE, pp.17-22.

[9]. Chen, M., Mao, S., Zhang, Y., Leung, V.C., 2014b. Big Data: Related Technologies, Challenges and Future Prospects. Springer.

[10]. Rajaraman, V., 2016. Big data analytics. Resonance 21, 695-716.

[11]. Tom White, "Hadoop: The definitive guide," O'Reilly Media, Inc.,2012.

[12]. S. Ghemawat, H. Gobioff and ST Leung, "The Google file system," in ACM SIGOPS operating systems review, vol. 37, no. 5, ACM,2003.

[13]. J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in Proc. 6th Symposium on Opearting Systems Design & Implementation, 2004.

[14]. ApacheHadoop, http://hadoop.apache.org

[15]. Sort Benchmark, http://sortbenchmark.org/

[16]. Yahoo! Hadoop Tutorial, http://developer.yahoo.com/hadoop/tutorial/index.html, Accessed onSept. 20, 2015.

[17]. HDFS Architecture Guide, https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, Accessedon Sept. 20, 2015.

[18]. Sudhakar Singh, RakhiGarg and P K Mishra, "Review of Apriori Based Algorithms on MapReduce Framework," in Proc. International Conference on Communication and Computing (ICC - 2014), Elsevier Science and Technology Publications, 2014.

[19]. MapReduce Tutorial, http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html, Accessed on Sept. 20, 2015.

[20]. K-H. Lee, Y-J. Lee, H. Choi, Y. D. Chung and B. Moon, "Parallel Data Processing with MapReduce: A Survey," in ACM SIGMOD Record, vol. 40, no. 4, pp. 11-20, (2011).

[21]. Hadoop Tutorials, http://hadooptutorials.co.in/tutorials/hadoop/understanding-hadoop- ecosystem.html, Accessed on Sept. 20, 2015.

[22]. Hadoop Architecture Overview, http://ercoppa.github.io/HadoopInternals/HadoopArchitectureOverview.html, Accessed on Sept. 20, 2015.

[23]. IBM developer Works, http://www.ibm.com/developerworks/library/l-hadoop-1/,Accessed on Sept. 20, 2015.

[24]. R. P. Padhy, "Big Data Processing with Hadoop-MapReduce in Cloud Systems," in International Journal of Cloud Computing and Services Science (IJ-CLOSER), vol. 2, no. 1, pp.16-27, 2013.+.

[25]. https://customers.microsoft.com/en-us/story/pros

[26]. https://azure.microsoft.com/en-us/blog/announcing-public-preview-of-apache-kafka-on-hdinsight-with-azure-managed-disks/

[27]. https://www.mongodb.com/nosql-explained

[28]. https://en.wikipedia.org/wiki/Presto_(SQL_query_engine)

[29]. Joab Jackson (November 6, 2013). "Facebook goes open source with query engine for big data". Computer World.Retrieved April 26, 2017.

[30]. Jordan Novet (June 6, 2013). "Facebook unveils Presto engine for querying 250 PB data warehouse". Giga Om.Retrieved April 26, 2017.

[31]. Eva Tse, ZhenxiaoLuo, NezihYigitbasi (October 7, 2014). "Using Presto in our Big Data Platform on AWS".Netflix technical blog.Retrieved April 26, 2017.

[32]. Doug Henschen (March 5, 2015). "Airbnb Boosts Presto SQL Query Engine ForHadoop". Information Week.Retrieved April 26, 2017.

[33]. James Mayfield (March 4, 2015). "Airpal: a Web UI for PrestoDB". Airbnb blog post.Archived

from the original on March 6, 2015.Retrieved April 26, 2017.

[34]. https://en.wikipedia.org/wiki/Presto_(SQL_query_engine)

[35]. https://en.wikipedia.org/wiki/NoSQL

[36]. https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-guide?view=sql-server-2017

[37]. https://www.microsoft.com/en-us/microsoft-365/blog/2016/06/23/excel-and-big-data/

[38]. Vohra, D., 2016. Using apache sqoop. In: Pro Docker. Springer, pp.151-183.

[39]. Jain, A., 2013. Instant Apache Sqoop.Packt Publishing Ltd..

[40]. Ahmed Oussous , Fatima-Zahra Benjelloun , Ayoub Ait Lahcen , Samir Belfkih "Big Data technologies: A survey"LGS, National School of Applied Sciences (ENSA), Ibn Tofail University, Kenitra, Morocco LRIT, Unit associe au CNRST URAC 29, Mohammed V University in Rabat, Morocco Journal of King Saud University - Computer and Information Sciences

[41]. Sudhakar Singh a,*, Pankaj Singh b, Rakhi Garg c, P K Mishra a "Big Data: Technologies, Trends and Applications" / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4633-4639

[42]. Rachit Singhal, Mehak Jain and Shilpa Gupta " Comparative Analysis of Big Data Technologies" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 3822-3830 Research India Publications. http://www.ripublication.com 3822