GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

TOOLS AND TECHNIQUES FOR BIG DATA

Anil B Jangid[*1] **& Arshiya Aman**[2]
[*1&2]Student, Dept. of BCA, Dayananda Sagar Institution, Bangalore, India

ABSTRACT

"Big Data" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Data is growing at a huge speed making it difficult to handle such large amount of data (exabytes). The term "Big Data" is used to describe the collection of complex and large data sets such that it's difficult to capture, process, store, search and analyse this kind of data using conventional data base management tools and traditional databases management system This paper analyses current technologies which deals with variety, velocity, and large volume of data.

Keywords*- Big Data, NoSQL, Data Wrangling*

I. INTRODUCTION

Knowledge sharing sites like www.epinions.com allow users to offer advice or rate products to help others. Users can rate the usefulness or "trustworthiness" of a review, and may possibly rate other reviewers as well. In this way, a network of trust relationships between users evolves, representing a social network for mining. Such tracking and analysis can provide critical information for decision making in various domains. But these data can be structured data - which can be represented in a well designed and rigidly defined tabular format, semi-structured data - which does not have a formal data model like XML and unstructured data - which does not have a pre-defined data model like the data found in log files, blogs, mails etc.

Over 2.5 Exabyte of data is generated every day. This data can be accumulated from various sources like transactions from a large stock exchange captures more than 1 TB of data every day, data transmitted from about 1.75 billion smart phones, more than 48 hours of video every minute from YouTube, data from social media such as Twitter and Facebook captures more than 10 TB of data daily etc.

This paper aims at analyzing various tools and techniques to store, manage, analyze and extract values and hidden knowledge from large volume of rapidly changing big data.
The organization of the paper is as follows. The concept of Big Data is presented in section II, the challenges of traditional technologies are discussed in Section III. Section IV presents the prominent tools and techniques needed to handle Big Data and finally conclusions are drawn in section V.

II. BIG DATA

Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. The four dimensions of big data are Volume, Variety, Velocity and Veracity.

Volume defines large amount of data. The benefit gained from the ability to process large amounts of information is the main attraction of big data analytics. More data leads to more accurate analysis. Velocity refers to speed of generating and processing data. Variety defines both structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analysing these data types together. Big Data Veracity refers to the biases, noise and abnormality in data which deals with is the data that is being stored, and mined meaningful to the problem being analyzed.

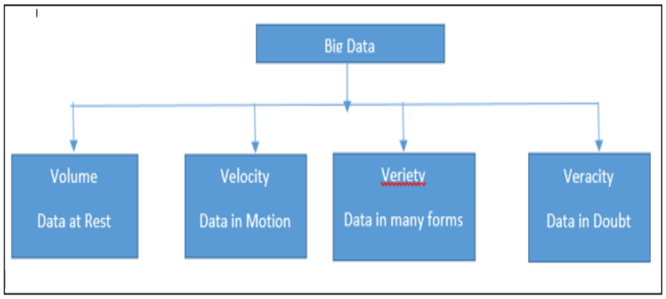

Fig.1 Classification of Big Data

III. CHALLENGES OF TRATIONAL TECHNOLOGIES

- The traditional databases are not designed to handle database insert/update rates required to support the Velocity at which Big Data arrives or needs to be analysed.
- The traditional databases require the database schema to be created in advance to define the data how it would look like which makes it harder to handle Variety
- Traditional databases can't analyze data from social media, data from videos, data from sensors as this type of data grows at very fast speed and also this is unstructured data.
- RDBMS's are normally architected in a centralized, scale up, master-slave fashion where All writes are written to the master and All reads performed against the replicated slave databases, but Big data need distributed, scale-out technique to handle large volume of data.
- RDBMS databases which scale vertically, But Big data needs horizontal scaling by adding more servers.

IV. TOOLS AND TECHNIQUES FOR BIG DATA

There is no single tool, or choice of a platform that will remedy all of the challenges of Big Data business. The technological paradigms of Big Data are distributed data storage and parallel processing or distributed computing, machine learning and data mining, or information extraction. Some of the tools are,

- NoSQL Databases : MongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- Processing - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop
- MapReduce, Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine,S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- Storage - S3, Hadoop Distributed File System
- Servers - EC2, Google App Engine, Elastic, Beanstalk, Heroku

1. Hadoop

Hadoop designed specifically for information that comes in many forms, such as server log files or personal productivity documents. It is an open source, java-based programming framework. Supports the processing of large data sets in a distributed computing environment. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Hadoop uses map-reduce to spread analytical processing across armies of commodity servers. Hadoop components are

a. **HDFS:** A highly faults tolerant distributed file system that is responsible for storing data on the clusters.
b. **Mapreduce:** A powerful parallel programming technique for distributed processing of vast amount of data on clusters
c. **Hbase:** A column oriented distributed NoSQL database for random read /write access.

d. **Pig**: A high level data programming language / Script for analyzing data of hadoop computation.
e. **Hive**: A data warehousing application that provides a SQL like access and relational model.
f. **YARN** Performs job scheduling and cluster resource management
g. **Amazon Web Services -** Elastic MapReduce : Amazon Web Services (AWS) Elastic MapReduce (EMR) is Amazon's packaged Hadoop offering. Rather than building Hadoop deployments manually on EC2 (Elastic Compute Cloud) clusters, users can spin up fully configured Hadoop installations using simple invocation commands, either through the AWS Web Console or through command-line tools.

2. NoSQL

As the term says NoSQL, it means non relational or Non SQL database, refer to Hbase, Cassandra, MongoDb, Riak, CouchDB. It does not have a schema. NoSQL deals with the unstructured, unpredictable kind of data according to the system requirement and is vertically scalable. Features of NoSQL are

a. Dynamic Schema
b. Linear Scalability
c. Continuous Data Availability
d. Running well on clusters
e. Eventually consistent/BASE
f. Auto Sharding
g. Mostly open-source
h. Allows Distributed and Parallel Processing

3. MongoDB

MongoDB is an agile NoSQL document database, unlike the traditional database, MongoDB stores the document data in binary form of JSON document which is also known as BSON format. It is used for high scalability, availability and performance. In MongoDB dynamic schemas are the unit of database, which found in document where set of documents are found in collection while set of collection makes the database. It supports horizontal scaling and auto sharding. MongoDB supports map reduce processing.

4. Cassandra

This NoSQL database is used to handle the large set of data when we need to scale the database with high performance. Cassandra deals with the fault tolerance and replication of the data. It follows the column family model. It supports best query capability and don't have joins feature.

5. IPython

The architecture of Ipython provides parallel and distributed computing. IPython enables parallel applications to be developed, executed, debugged and monitored interactively. IPython supports both Task parallelism and data parallelism. IPython Notebook is a web-based interactive computational environment. An IPython notebook is a JSON document containing an ordered list of input/output cells which can contain code, text, mathematics, plots and rich media. IPython notebooks can be converted to a number of open standard output formats (HTML, presentation slides, LaTeX, PDF, ReStructuredText, Markdown, Python) through 'Download As' in the web interface and 'ipython nbconvert' in a shell.

6. RCloud

RCloud was developed in AT&T Labs, was created to address the need for a collaborative data analysis environment for R. Similar to IPython, RCloud allows researchers to analyze large data sets and share their results across an organization. Conceptually, RCloud is similar to an R CRAN (Comprehensive R Archive Network) package, but augmented by wiki-like collaboration features. Notebooks and code are stored in GitHub. Although a relative newcomer with little documentation outside of AT&T, RCloud shows a lot of promise.

7. Storm

It is a distributed, fault-tolerant, and high-performance real time computation system that provides strong guarantees on the processing of data. Similar to how Hadoop provides a set of general primitives for doing batch processing, Storm provides a set of general primitives for doing real-time computation. Its use cases span stream processing, distributed RPC, continuous computation, and more.

8. Lambda

A data-processing architecture designed to handle massive quantities of Big Data by taking advantage of both batch- and stream-processing methods. Its architecture attempts to balance latency, throughput, and fault-tolerance by using batch processing to provide comprehensive and accurate views of batch data, while simultaneously using real-time stream processing to provide views of online data. The two view outputs are joined before presentation.

V.CONCLUSIONS

Given the continuing trend of data growth in online, a new generation of solutions are required to tailor them. Data analytics over humongous data sets seem possible only by using distributed computation framework. A new generation of information management systems, termed Big Data, caters to this trend and is apt for businesses that are planning to migrate existing applications to adapt to new trends of data growth, develop new applications involving large volume of unstructured data.

REFERENCES

1. *Leavitt, N.," Will NoSQL Databases Live Up to Their Promise?," IEEE Volume: 43 , Issue: 2 Publication Year: 2010 , Page(s): 12 – 14.*
2. *Sebnem Rusitschka, Alejandro Ramirez, "Big Data Technologies and Infrastructures", http://byte-project.eu/wp-content/uploads/2014/09/BYTE_D1-4_BigDataTechnologiesInfrastructures_FINAL.compressed-1.pdf*
3. *Gudivada, V.N. ; Rao, D. ; Raghavan, V.V.,"NoSQL Systems for Big Data," Management in Proceedings of IEEE World Congress , Publication Year: 2014 , Page(s): 190 – 197.*
4. *Bednar, P. ; Sarnovsky, M. ; Demko, V.,"RDF vs. NoSQL, databases for the semantic web applications Applied Machine Intelligence and Informatics (SAMI)," IEEE 12th International Symposium, Publication Year: 2014 , Page(s): 361 – 364.*
5. *XiaomingGao ;Qiu, J. "Supporting Queries and Analyses of Large-Scale Social Media Data with Customizable and Scalable Indexing Techniques over NoSQL databases," Cluster, Cloud and Grid Computing (CCGrid), 14th IEEE/ACM International Symposium, Publication Year: 2014 , Page(s): 587 – 590.*
6. *Man Qi," Digital forensics and NoSQL databases", Fuzzy Systems and Knowledge Discovery (FSKD), 11th IEEE International ConferencePublication Year: 2014, Page(s): 734 - 739*
7. *Brewer, E. A. (July 2000). Towards Robust Distributed Systems. ACM Symposium on the Principles of Distributed Computing. Retrieved from http://www.cs.berkeley.edu/~brewer/cs262b-2004/PODC-keynote.pdf.*