JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BIG DATA ANALYTICS: MAKING SENSE OF BIG DATA

[1]VIKAS SHUKLA, [2]SANJEEV SHUKLA, [3]PAWAN KUMAR

[1]Axis Business School (Uttar Pradesh), INDIA, [2]Axis Business School (Uttar Pradesh), INDIA, [3]Axis Institute of HigherEducation (Uttar Pradesh), INDIA

Abstract:. Today, the enterprise collects more data than ever before, from a wide variety of sources and in a wide variety of formats. Along with traditional transactional and analytics data stores, we now collect additional data across social media activity, web server log files, financial transactions and sensor data from equipment in the field. The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. In this paper, I am presenting basic analysis workflow of big data with six step implementation strategy to build predictive models.

Index terms: Big Data, NoSQL, Hadoop, Hive, Map Reduce, In-memory processing, Data mining, Data analysis, Data transformation, parallel processing

I. INTRODUCTION

Dataset whose volume, velocity, variety and complexity are beyond the ability of commonly used tools to capture, process, store, manage and analyze them can be termed as "BIGDATA".

The tremendous increase in volume of data which is being generated from different sources is increasing day by day. In the year 2000, (800,000 petabytes (PB)) of data were stored in the world. Of course, a lot of the data that's being created today isn't analyzed at all and that's another problem that I am trying to address in my paper. We expect this number to reach 35 ZETTABYTES (ZB) by 2020.

It's not that we have started generating data in a day or two. It was there from long, but what has changed is its exponential growth with the growth of technologies. On an average per minute we are creating.

- 98000+ tweets
- 695,000 status updates
- 11,000,000 chat messages
- 700,000+ Google searches
- 168,000,000+ emails sent
- 217 new mobile web
- A total of 1820 TB data gets created.

Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains. Efforts are needed to manage such large and unstructured data it's a challenge for any data scientist to apply complex mathematical algorithms time and again and implement machine learning applications for big data.

Big data analytics is the process of examining large amounts of data of a variety of types (big data) to uncover hidden patterns, unknown correlations and other useful information. Such information can provide competitive advantages over rival organizations and result in business benefits, such as more effective marketing and increased revenue. Big data analytics can be done with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics and data mining. But the unstructured data sources used for big data analytics may not fit in traditional data warehouses.

Furthermore, traditional data warehouses may not be able to handle the processing demands posed by big data. As a result, a new class of big data technology has emerged and is being used in many big data analytics environments.

The technologies associated with big data analytics include NoSQL databases and Map Reduce. These technologies form the core of an open source software framework that supports the processing of large data sets across clustered systems.

Data mining is a particular data analysis technique that focuses on modeling and knowledge discovery

for predictive rather than purely descriptive purposes. Business intelligence covers data analysis that relies heavily on aggregation, focusing on business information. In statistical applications, some people divide data analysis into descriptive statistics, exploratory data analysis (EDA),and confirmatory data analysis (CDA).EDA focuses on discovering new features in the data and CDA on confirming or falsifying existing hypotheses. Predictive analytics focuses on application of statistical or structural models for predictive forecasting or classification, while text analytics in the realm of big data applies statistical, linguistic, and structural techniques to extract and classify information from textual sources, a species of unstructured data. All are varieties ofdata analysis.

II. BIG DATA CHALLENGES:

In last few years, internet age firms like Google, Amazon, Yahoo and Facebook started collection of peta, zeta bytes of unstructured and structured data. Efforts needed to manage such large and unstructured data exposed the limitations of in-memory data processing on costly hardware. These firms led the innovation to create new parallel processing technologies such as Map Reduce using frameworks such as Hadoop, Hive, HBase and Cassandra. Vendors such as Hortonworks and Cloudera offer easy installation, infrastructure services and support for these distributed computing technologies.

Together, these technologies offer excellent storage and querying capabilities irrespective of the data formats, size and memory restrictions and utilize combined computing power of cluster of utility machines. But it still leaves businesses and data scientists with lots of challenges to learn and to implement efficient, end-to-end data analysis workflows to make sense of big data. It's a challenge for any data scientist to apply complex mathematical algorithms time and again and implement machine learning applications for big data. Today, businesses in every industry domain are struggling to create new analysis workflows from scratch and design complex applications for each stage of big data analysis: inspection, cleaning, transformation, and modeling to derive business value.

Now we will learn that how big data is tackling these complex situations:

Most of the Big Data tools and framework architecture are built keeping in mind about the following characteristics:

Data distribution: The large data set is split into chunks or smaller blocks and distributed over N number of nodes or machines. Hence the data gets distributed on several nodes and becomes ready for parallel processing. In Big data world this kind of data distribution is done with the help of Distributed File System or DFS.

Parallel processing: The distributed data gets the power of N number of servers and machines in which data is residing and works in parallel for the processing and analysis. After processing, the data gets merged for the final required result. The process is known as Map Reduce which is adopted from Google's Map Reduce research work.

Fault tolerance: Generally we keep the replica of a single block (or chunk) of data more than once.
Hence even if one of the servers or machine is completely down, we can get our data from a different machine or data center. Again we might think that replicating of data might cost lots of space.

Technologies used in Big Data Analytics:

Before starting technological reviews , how Big Data Analysis are made , how the data is being transformed into Structured from Unstructured Repositories it is needed to have basic knowledge of few common technologies used in this domain as :

1. **NoSQL Database.**
2. **Map Reduce.**

Now we will study these technologies in details as:

1. NoSQL

NoSQL is the technology that is used to efficiently extract unstructured data from extremely large data repositories. Hive, Cassandra is two examples of open source NoSQL frameworks. In old, structured data repositories like relational databases such as Oracle, SQL server we used SQL but for unstructured data as JSON, we use NoSQL.

Why NoSQL?

Relational and NoSQL data models are very different. The relational model takes data and separates it into many interrelated tables that contain rows and columns. Tables reference each other through foreign keys that are stored in columns as well. When looking up data, the desired information needs to be collected from many tables (often hundreds in today's enterprise applications) and combined before it can be provided to the application. Similarly, when writing data, the write needs to be coordinated and performed on many tables.

NoSQL databases have a very different model. For example, a document-oriented NoSQL database takes the data you want to store and aggregates it into documents using the JSON format. Each JSON document can be thought of as an object to be used by your application. A JSON document might, for example, take all the data stored in a row that spans 20 tables of a relational database and aggregate it into a single document/object. Aggregating this information may lead to duplication of information, but since storage is no longer cost prohibitive, the resulting data model flexibility, ease of efficiently distributing the resulting documents and read and

write performance improvements make it an easy trade-off for web-based applications.

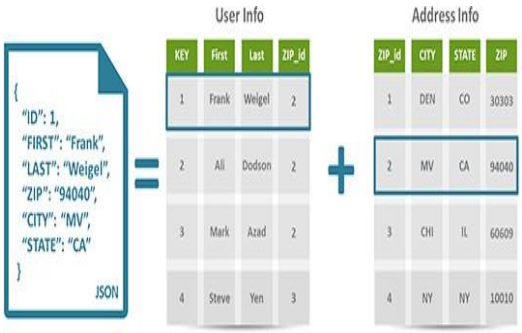


Figure 1

Another major difference is that relational technologies have rigid schemas while NoSQL models are schema less.

Auto-shading
A NoSQL database automatically spreads data across servers, without requiring applications to participate. Servers can be added or removed from the data layer without application downtime, with data (and I/O) automatically spread across the servers. Most NoSQL databases also support data replication, storing multiple copies of data across the cluster, and even across data centers, to ensure high availability and support disaster recovery.

ROLE OF HIVE AND CASSANDRA FRAMEWORKS IN NO SQL:-

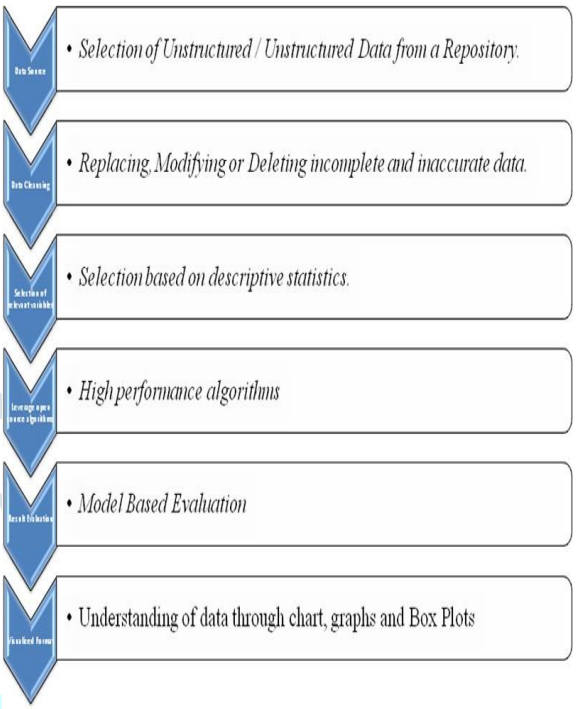
Hive:
The Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL.

Cassandra:
Apache Cassandra is an open-source distributed database management system designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Cassandra offers robust support for clusters spanning multiple datacenters, with asynchronous master less replication allowing low latency operations for all clients.

Proposed Solution:

Implementation of 6 Phases Big Data Analytics Framework Mechanism:

So, Now I am going to elaborate my work with following phases as follows:



1. **Data Source**

(How we retrieve data from any repository of unstructured or structured Source).

2. **Data Cleansing**

(Replacing, Modifying or deleting incomplete and inaccurate data)

3. **Selection Of relevant Variables**

(Selection based on descriptive statistics).
High-performance

4. **Leverage Open - Source Algorithms**

(High performance algorithms).

5. **Result Evaluation** (Model Based Evaluation)

6. **Visualized Format**

(Understanding of data through chart, graphs and Box Plots).

CONCLUSION:

So, after studying how Big Data Analytics work, on which tools/Frameworks some of which we have read. Now, comes the final conclusion implementation process of whole Analytics Workflow into 6 different phases which have been made simpler to understand the basic mechanism made for beginners in this field.

In my paper I have tried to provide basic knowledge of Big Data, Big Data Analytics, Some of the tools and technologies used to implement the workflow structureOf Big Data Analytics mechanism- which is the basic theme of my paper.

REFERENCES

1. Improving Decision Making in the World of Big Data http://www.forbes.com/ sites/christopherfrank/2012/03/25/improving-decision-making-in- the-world-of-big- data/
2. World's data will grow by 50X in next decade, IDCstudy predicts http://www.computerworld.com/s/article/9217988/World-s-data- will-grow-by-50X-in- next-decade-IDC-study-predicts
3. The 2011 Digital Universe Study: Extracting Value from Chaos http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index. htm
4. Kognitio WX2 overviewhttp://www.dbms2.com/2008/01/26/kognitio-wx2/
5. IDC Releases First Worldwide Big Data Technology and Services Market Forecast, Shows Big Data as the Next Essential Capability and a Foundation for the Intelligent Economy http://outsourcing.ulitzer.com/node/2195534
6. Gartner Hype Cycle 2012 for Emerging Technologies http://sembassy.com/wp- content/uploads/2011/10/gartner-hype- cycle-2012.gif
7. Gartner Hype Cycle 2012 http://www.gartner.com/id=2065716
8. Big data: The next frontier for innovation, competitionand productivity http://www.mckinsey.com/Insights/MGI/Research/Technology- and-Innovation/ Big-data-The-next-frontier-for-innovation
9. Big data: The next frontier for competitionhttp://www.mckinsey.com/ features/big-data

★★★