**Internship Project Report**

Suyog Ahuja

[Eid-ahujasuyog1@gmail.com](mailto:Eid-ahujasuyog1@gmail.com)

Phno. - 9167135062

For 4 weeks — from 1th of June 2019 to 30th of June 2019. I was an intern at JIO. This report summarizes my project and my progress.

Supervisor : Nikunj Bhartia .

**Task Description**

Hadoop framework is becoming a top open source player on field of distributed system offering possibility of storing and processing big sets of data in a scalable manner Nowadays it is also gaining momentum at JIO – there is an increasing interest in solutions based on Hadoop ecosystem in many areas including experiments,

accelerator controls archives etc. It appears as a natural replacement of a traditional relational system whenever ad hoc analytical processing is a dominant part of data work loads.Some of the systems at JIO already have offline replicas of archive data stored on a Hadoop created manually by a system administrator. Such process is very often time consuming and requires applying sequence of actions on the metadata and data itself .The goal of the project is to automatize the process of data loading between Oracle database and Hadoop cluster by creation of a utility that interfaces with both systems via dedicated tools (Apache Spark & Hive) and applies all necessary actions in order to delivered ready-to-read data on Hadoop file system for high-end frameworks (Spark). The tool should support incremental data loading on a time scope bases and should be configurable for each data set separately.

**Current project :-**  **DMP(Device management platform )**

**Description :-** Creation of DashBoard which will represent data in much more systematic and organized manner. For accessing data from hdfs we used HIVE

And created a source (link) between hive table and ZoomData.

**Zoomdata** is the modern business intelligence and data visualization platform for cloud, big data, live streaming data, multisource, and embedded analytics.

**Assignments :**

**Kick-Starter Kit for BigData Developers (1):-**

This starter kit consisted of 5 assignments which included both theory as well as hands on.

Assignment consisted of understanding of **use cases of Big Data.**

**Summary of assignments:-**

The world population as on April 2018 was **7.6 Billion** people out of which**77%**use internet. People constantly browse sites, do shopping, book deals etc. Just **imagine the data** that is created by these millions of people!

Now **from a business point of view**, it is not manually possible to keep a tab on everyone who has viewed or is interested in your content/ campaigns. So to **better understand the customer patterns and trends in market**, this data needs to be analyzed.

The online method of data analysis is **big data analytics**. It is meant for all big or small businesses.

In technical words, it is voluminous information or **relevant statistics** acquired by companies, firms and large organizations. Multiple softwares and data storages (computers, servers, etc..) have been formulated for processing.

Below are the 4V’s of big data:

* ***Volume****–* It is the quantity of data gathered, generated and stored.
* ***Velocity****–*The speed at which all this data is received and also acted upon.
* ***Variety****–*Variety refers to how this continuous inflow, nature, and type of unstructured data.
* ***Value****–*Every bit of information received has value. To dig deeper multiple quantitative techniques for unstructured data is used.

The above 4 V’s of big data are crucial for **gathering, storing, analyzing, managing and consuming** huge sets of information.

**The Big 6 Steps**

**Data Mining**

There are two focus terms: data extraction & data mining. If simply put, data extraction is a process of collecting all data from web pages into your database. Whereas, data mining is a process of identifying valuable insights within that database. Such data is collected by data scientists.

For example, you are an e-commerce grocery site owner. After using various research techniques, you concluded that approximately 70% people wear jeans. This is called data extraction. Now you have to go deeper to understand which age, gender, and type of people use Brand 1 and Brand 2 jeans. This process is known as data mining. Some of the useful data mining tools include RapidMiner, TeraData & Kaggle

**Data Collection**

Big data doesn’t have an “END” button. As the world grows, data will keep on streaming in. Data needs to be extracted constantly. From the above example: there will be people who wear Brand 1 have switched to Brand 2 and so on. The possibilities are endless! Data extraction becomes easier with tools like [Enterprise Web Data Extraction and Analysis - Import.io - Web Scraping, Web Scraper, Data Extraction, Web Extraction, Web Data, Web Harvesting](http://import.io/).

**Data Storing**

Ever imagined how Google must be storing so much of world data? Of course not on traditional systems – files, CDs, DVDs, etc.. Google, Facebook, Apple, etc.run on hyperscale computing environments. Which type of storage you should use depends on the scale of your business. A good data storage system provides an infrastructure which has all the latest data analytics tools and storage space. You can store your data on data storage providers like Cloudera, Hadoop (not for beginners) and Talend. Data storage is one step which here on can be inserted in between any other step.

**Data Cleaning**

Data sets can come in all forms and degrees – some good and some not so good especially if extracted from the web. Therefore, all the data extracted needs to cleaned. In the cleaning process, all the unwanted and inaccurate data is filtered out. After this process, you will only be left with what you actually want to focus on. Cleaning promotes structuring your data well. For example, you know number and type of people wearing jeans all over. While cleaning, you can remove all the duplicate entries, wrong data, unwanted regions or information and more. You can make use of DataCleaner or OpenRefine for this purpose

**Data Analysis**

The biggest part of big data is the analytics! What is big data analytics? While analyzing the data you come across your audience pattern, behavior and so on. Exploratory research method proves to be very helpful in analyzing big data. Analytics is about asking a specific question and finding answers to it. Qubole and Statwing are powerful data analytics tools. For example, you might ask – does my audience like to wear two pocket jeans? Which color is most preferred by them,etc..

**Data Consumption**

Data is consumed in various verticals which include:

* Identifying retail trends in the market using which businesses can highlight their top selling products.
* It is used by Government bodies in order to reach out to the correct demographics, geographies, and ethnicities.
* Marketers find big data extremely useful to figure out which advertisement works for their products.

Big data is consumed at many places depending on the specific goals you want to achieve.

**Hadoop:-**

[**Hadoop**](https://www.edureka.co/blog/hadoop-tutorial/)is an open source Distributed Data Processing framework from Apache. It is used to process Big-Data applications in clustered computer systems.

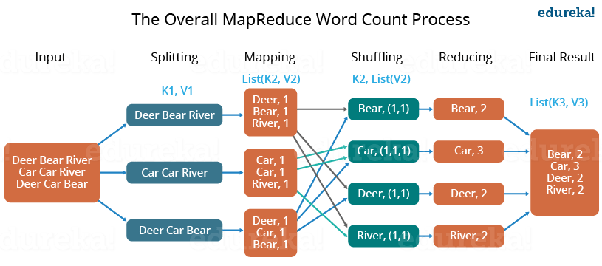
Hadoop can handle various types of data formats. Data formats which could either be Structured (for example Excel files such as Hospital data set) or Unstructured (for example: such as What’s app images, voice messages, videos etc) and Semi-structured (for example .CSV files which do not have any schema about the data.)

At its core, Hadoop has two major layers namely −

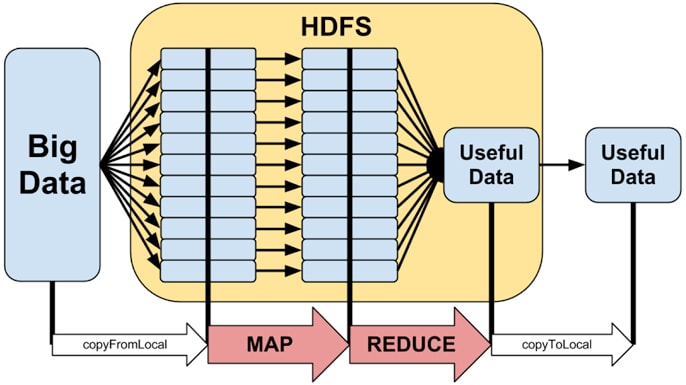
* Processing/Computation layer (MapReduce), and
* Storage layer (Hadoop Distributed File System).

[**MapReduce**](https://www.edureka.co/blog/mapreduce-tutorial/)

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte data-sets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework.



[**Hadoop Distributed File System**](https://www.edureka.co/blog/hdfs-tutorial)



The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low-cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

Apart from the above-mentioned two core components, the Hadoop framework also includes the following two modules −

* **Hadoop Common** − These are Java libraries and utilities required by other Hadoop modules.
* **Hadoop YARN** − This is a framework for job scheduling and cluster resource management.

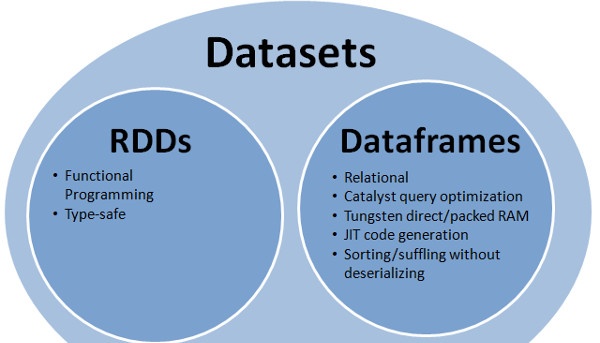
**Spark:-**

Apache Spark is in-memory data processing engine. It provides robust, flexible and user friendly platform for batch processing, stream processing, machine learning and large scale SQL.

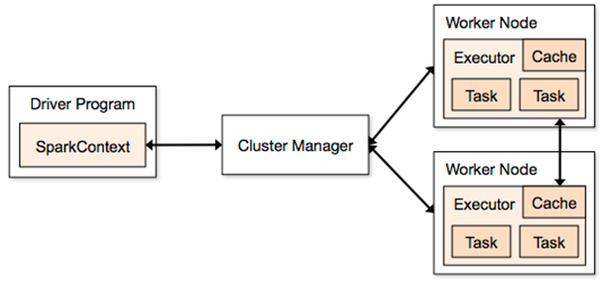
Apache Spark was mainly designed for data scientists. It provides high level API’s is many languages for example Java, Scala, Python and R. For execution of graphs, it has an optimized engine. It has largest open source project in data processing. Key feature of Spark is in-memory processing. It is not a new computing concept. With an underlying design of in-memory processing, there is a long list of database and data processing products. With in-memory processing, processing speed can be increased.



There is no need to fetch data from the disk again and again so obviously time is saved. Apache Spark has DAG computation engine that helps in-memory computations and acyclic data flow that results in high speed. Spark provides API’s for three types of data sets (RDD). Resilient Distributed Data are immutable distributed collection of data that can be manipulated using functional transformations (MapReduce, Filter etc).



The changes applied to data in Spark are through compositional functional transformations. This approach in formulating and resolving data processing problems is favored by many data scientists. Apache Spark is known for its effective use of CPU cores over many server nodes. Along with Standalone Cluster Mode, Spark also supports other clustering managers including Hadoop YARN and Apache Mesos.



Being a distributed computing framework, it is essential for Spark to have robust cluster management functionality. Spark provides real time stream processing. One problem with Hadoop MapReduce was it could only handle and process already present data but this data was not real time. Spark Streaming solves this problem for data scientists. Lazy evaluations in Apache Spark help to increase system’s efficiency. The transformations in Spark’ RDD are lazy in nature and couldn’t provide the desired results right away, rather a new RDD is formed for existing one but this increases the efficiency of the system. Apache Spark helps to ensure that data loss is reduced to zero. That’s why it provides fault tolerance through Spark abstraction- RDD.

**Overview :**

This internship really helped me in understanding how traditional database can’t be used for handling big data and why is it necessary to use Hadoop for big data.

This internship also helped me in learning new languages such as **scala & hive** which are used for aggregation on large data set.

Mentor here were very helpful , I would like to thank Nikunj Bharia and Kaustubh Daware for sharing their knowledge with me also I would like to thank JIO for giving me this opportunity to learn.