CHAPTER – 1

AN OVERVIEW OF BIG DATA ANALYSIS

1.1 INTRODUCTION

In computation, Data is altered into a form that is competent for processing and analyzing. In real world, data is immense, has a complex structure and represents complex structures which are beyond the capability of existing architecture. Data gets generated in continuous stream from innumerable sources in today's digital world and is freely available. Data that emerges from infinite sources is classified in various forms such as structured, unstructured and semi-structured. Hence, the rapid growth of digitalized data provides vast opportunities for data analytics.

Data Analytics is the scientific and statistical tool for analyzing raw data to renovate information for acquiring knowledge. Data analytics collaborates with data to formulate complex decisions from diverse perspectives for facing the real world challenges. The role of analytics is to assemble, store, process and analyze data to address empirical methods in real world for decision making. It is broadly classified into *descriptive*, *inferential*, *predictive* and *prescriptive analytics* [1].

Data analytics extends as a process of analyzing massive real time streaming data, which varies in data structure called as *Big Data Analytics*. Big data acts as a frontier for innovation, competition, productivity and business forecasting since the data is exponentially growing [2]. The analytics on such huge data reveals hidden patterns, unfound correlations, market trends, consumer requirements and future recommendations, which assist in critical decision-making process [3].

Big Data Analytics facilitates various organizations as equipment for handling, managing, analyzing and evaluating the data to identify new opportunities. Efficient operations, cost reduction, smart decision making, service systems and advanced product development are some features of Big Data Analytics. Advanced storage structures are practiced for representing Big Data. Traditional database systems are untrained to handle and analyze such massive Big Data Structures. A novel approach is appropriate for managing huge data. Hence an expert system is incorporated for analyzing Big Data as a pre-processing step using data analytics.

The key characteristics that focus Big Data assist the organizations to maximize the benefits. The various challenges of Big Data are data capturing, cleansing, association, storage retrieval, processing, indexing, exploring, distributing, relocating, mining, analyzing and visualization on huge datasets, which is rapidly streaming in real time environment.

The analytical tools applicable on Big Data facilitate both technical companies and academic users for processing an expected span of time and for acquiring knowledge from the data [4]. Big Data analyses transform the information by extracting several features and performs forecast on the data, which relates to the organizations' perspectives.

Figure – 1.1: Characteristics of Big Data

The features of Big Data are illustrated as ten V's as described in figure 1.1, which represents the properties [2]:

- ***Volume:*** Massive Size of Real time streaming data
- ***Variety:*** Data in various forms of text, images, music, videos, graphs, plots etc.
- ***Velocity:*** Speed of the data flow emerging from various sources
- ***Veracity:*** Authentic data arriving from numerous sites
- ***Validity:*** Data storage in terms of legality and periods of time
- ***Value:*** Significance of data stored and computing the worthiness
- ***Variability:*** Changing the data along with the flow of data

- ***Venue:*** Storage and retrieval of data from a specific location in data centers
- ***Vocabulary:*** Language readability and understanding of grammatical notations.
- ***Vague:*** Data is immense and does not make sense unless it is analyzed.

1.2 OVERVIEW OF SPARK USING MACHINE LEARNING

Big Data is a large data processing system developed to solve the issues of various traditional methods. As it has large voluminous data, privacy turns to be a major challenging issue. To provide acute security the critical modeling techniques requires analytical processing, effective storage and successful retrieval techniques. Therefore, a massive parallel programming model termed as MapReduce is employed on a distributed framework that provides high computational environment for Big Data Analytics.

Ideally, MapReduce model has two significant functions the Map and Reduce to process the huge data structures. In this model, the data is initially allocated to parallel distributed map tasks. Each map task processes the data input and produce key-value pairs. The output produced by each map task is further taken as an input for reduce task. Finally, passing through many phases of reduce task such as shuffle, merge and sort function the MapReduce model emerges with a final summation output. The process of constructing such fusion of information/models has tremendous impact on the quality of the final system. In concise MapReduce models are basically designed for guiding the experimental study in analytics field of research.

In spite of the significant features the MapReduce has many limitations. MapReduce model suffers from high disk reads and writes, low throughput, which results in low performance of a cluster, low latency and poor reading structure. To overcome the drawbacks of MapReduce framework, proficient models have emerged for a distributed environment like *Kafka*, *Spark*, *Flume* and many more. These advanced technologies have replaced MapReduce model by suppressing the above stated limitations.

Spark is a well-developed tool, which replaces MapReduce for managing the execution of tasks across cluster of computers. A cluster or group of machines utilized by Spark framework pools the resources of machines by allowing usage of cumulative resources at a single point of time. Spark supports applications with working sets while retaining scalability and fault tolerance features intact of MapReduce. To accomplish these goals Spark uses RDDs

(Resilient Distribute Datasets) for partitioning the datasets on cluster computers by avoiding data loss.

Spark is the earliest system that allows an efficient general-purpose programming language for an interactive process of large datasets on a cluster. Spark provides a unified engine provisioned with ease usage and faster attitude than Hadoop for large scale data processing. It implements iterative Machine Learning workloads by interacting and scanning datasets parallel to sub-second latency [35].

Today's organizations create dissimilar data from diverse sources, which are user centric. Developing personalization, recommendation and predictive insights are major goals for advancement in organizations. The existing model MapReduce proved to be unsuccessful in processing such huge data generated from organizations. Spark is designed with advanced features that incorporates various features like simplicity, scalability, easy integration, compatibility and speed. These are employed to solve challenges faced by organizations. Diverse use cases are applied that adapt quickly with the iterative models to meet the requirements of organizations.

The recent popularity of Spark models invokes significant interest in implementing scalable versions of Machine Learning algorithms. Machine Learning is the existing and current automated technology involved in handling and evaluating the massive data structures. The process of developing an amalgam of information/models contributes to the quality of the final system. These models are designed for guiding the experimental study in analytics field of research.

Machine Learning techniques are broadly classified into *supervised* and *unsupervised learning*. Supervised learning techniques emphasize on accurate predictions whereas unsupervised learning work on compact descriptions of the data. Data classification, complex pattern recognition, predictions/intelligent decisions and clustering are some of the major features of Machine Learning techniques.

Machine Learning techniques learn to understand the complex datasets for critical decisions and tune the features for extraction of high performance. The expected performance depends on the trained features underlying in the parallel programming framework. Test features are obtained from the learning models which are exercised on training datasets.

Spark has a unique library for Machine Learning called MLlib. It solves wide range of data problems attributed to streaming, graph computation and real-time interactive query processing. MLlib assist by leveraging the scale and speed that builds specialized use cases for varied analytical models. This thesis focuses on implementing Machine Learning techniques on Apache Spark framework that focus on measuring Time and Space complexities on varied datasets.

1.3 ROLE OF MACHINE LEARNING IN BIG DATA ANALYTICS

Machine Learning is a technical tool of data science that creates logic from data by transforming data into knowledge. Many powerful algorithms from the field of Machine Learning are developed to learn patterns, acquire insights and do forecasting from previous events.

In this modern era, a large assortment of data is immersed that starves for knowledge. This abundant data is classified into structured and unstructured form. Structured data is the data which is arranged in tables and unstructured is the data which are in irregular form such as images, documents, text, audio, video etc., Human intervention is reduced by automated machines to build models for processing huge amounts of unstructured data.

Machine Learning provides efficient analysis models for capturing knowledge by improving prediction for data driven decisions. It plays an eminent role in the field of computer science that paves its way in analyzing robust emails, spam filters, convenient text, voice recognition, web search engines, game developments and self – driving cars.

Learning is an activity in which a model is tuned to solve various problems by understanding the characteristics of parallel distributed data. Learning models are developed in accordance with the input data feed into the system. Since the data is complex and massive it is essential to perform computations on a distributed parallel environment using Machine Learning techniques. Due to this exponential growth of data Machine Learning techniques are evolved to adapt new complexities.

Novel intricacies prevailing in organizations require multi-core processors because a single core system fails to handle massive data structures. Hadoop Distributed File System of Big Data stores massive amount of information that scale up exponentially and survive the failure of storage infrastructure without data loss. To interpret analytics clusters and multi core processors are built with inexpensive computers that shift computations to the other machines

in the system. The amalgamation of big data with Machine Learning channelizes business units to handle more complex data.

1.3.1 APPLICATIONS OF MACHINE LEARNING

The following are the applications of machine learning:

- ***Marketing:*** Understanding the site behavior by predicting the possible number of users visiting the business page for product analysis.
- ***Advertising Optimization:*** The optimization techniques are applicable for advertising to promote the business.
- ***Recommendation Engine:*** Techniques of machine learning, forecast the preference of the customers by understanding the previous analytical survey.
- ***Market Basket Analysis:*** Interpreting the customer behavior by providing personalized discounts offered by company to increase the cart size.
- ***Customer Churn Prediction:*** Specific customers receive additional consideration from service teams to secure their loyalty.
- ***Operational Optimization:*** Systems are likely to fail therefore regular checks are operated as preventive measures.
- ***Preventive Maintenance:*** Safe guarding the operations and transactions at regular intervals for avoiding data loss.
- ***Security Monitoring:*** Examining the anomalous user behavior and tracing the malicious one among various users.
- ***Risk Assessment:*** Assessing the risk of certain transactions from past experience, continuous supervision and analysis.
- ***Fraud Detection:*** These methods scrutinize the abnormal characteristics of a typical user for detecting the fraud and cybercrimes.
- ***Network Monitoring:*** Observing the network failure tracks the strength and weakness of a network path by increasing screening.

1.4 FEATURES OF HADOOP IN BIG DATA ANALYTICS

In today's scenario, every organization faces lots of challenges especially in analyzing the stream of data. But large amount of data when read and write through multiple disks in parallel mode consumes maximum time in data analytics. These challenges are effectively addressed by Big data.

In a nutshell, Hadoop is an effectual analysis and reliable data storage system. The proficient storage is provided by Hadoop distributed file system and data analysis is processed by MapReduce model. To handle the massive data High Performance Computing (HPC) systems are used. Application Program interfaces with High Performance Computing Systems are used for transmitting information are connectively referred as Message Passing Interface.

The file systems in HPC are hosted on Storage Area Networks (SAN) for a proper access on cluster machines. HPC file systems are capable of computing intensive jobs. The problem raised during the usage of these systems is the network band width. In certain cases, when the data is too large for transmission the employed bandwidth and entire system collapses resulting in performance issues because the computing nodes need to be idle for a long time.

The first problem is to solve node failure where hardware pieces are prone to fail one or other node at a certain point of time. Hadoop Distributed File System (HDFS) system creates triplet instances of same data scattering them on different servers. A general way of avoiding data loss is through replication. Any failure or data loss can be handled by utilizing the redundant copies. This mechanism works by RAID concept in the HDFS.

The second problem is the major analysis task that requires merging of data from various disks. The distributed and extracted data from many sources needs to be combined and this is currently the most formidable challenge. The Spark framework that replaces MapReduce technique is one among the tangible solution to address the problem of repeatable disk reads and disk writes.

1.5 MOTIVATION

Data emerging from diverse sources should be used in an innovative way. Generally existing data is used for analytics. Through this analytics, prosperity of any organization can be evaluated. This evaluation leads to the development of prediction. Advancement in technology, enthusiastic behavior of organizations, smartness in thinking and critical decision-making are some of the features, which lead to the development of forecasting. Predictive analytics is the solution for better understanding from the existing data and assist in forecasting.

Decision makers seek insights from predictive solution for developmental activities and to participate proactively in challenges for future capitalization. It aids the industrial experts by revealing the association between unstructured and structured data by drawing relevant

inferences and analyze consequently. It supports many companies to face the harness of big data by finding pricing products, trends, maintain inventory, market analysis etc.

Predictive analytics reduces risk and loss for an organization by identifying the fraudulent transactions. It even improves efficiency in increasing profit margins by data analytics and systematic reasoning. Customer preferences, response and feedback are analyzed by predictive analytics for business development. Leveraging Big Data in collecting information from social streams helps in initiating marketing campaigns that target interests and preferences of customers.

Predictions play a vital and significant role, for instance, organizations for their marketing and sales, government agencies to satisfy the requirements of people, travelers for planning a trip, meteorological department for weather forecast and farmers for their crops etc. Ideal trends, constant analysis, passionate to forecast and novel challenges provide insights to analyze and predict for future. Due to the peculiar challenges prevailing in the present world, this research work mainly targets in addressing the accurate prediction using Machine Learning techniques.

1.6 SCOPE OF THE RESEARCH

Predictive analytics solutions extend to understand behaviour, attitudes, services and preferences of consumer by giving some offers. The critical factors in the behavioural study of consumer such as price-driven, brand-driven, quality-driven, and quantity-driven and characteristic-driven are identified by predictive analysis. Based on this, the analytics are categorized as *functional* and *industrial*.

1.6.1 FUNCTIONAL ANALYTICS

The functional analytics breaks down the component functions such as marketing, accounting, HR, etc. These functions are sub-divided into smaller functions and so on, till it reaches to a suitable problem-solving functional level. Marketing mix, cross sell, patterns, deficiencies, deviations, issues and trends are various functional areas of prediction analytics.

1.6.2 INDUSTRIAL ANALYTICS

Predictive analytics has varied applications in other sectors like banking, retail, healthcare, insurance, telecommunications, pharmaceuticals and manufacturing. Features which advance in applications of predictive analytics are discussed as below:

- ***Anti-theft:*** Biometric sensors read the identity of a person and make it impossible to start the engine of a car if any unknown person attempts to do so.
- ***Traffic:*** Navigates to alternative paths at peak hours by predicting traffic.
- ***Collision Avoidance:*** A stronger vibration and a sound alert is alarmed to track imminent collisions.
- ***Food:*** An enrouting method drives suggesting daily food preferences using a recommendation system in restaurants.
- ***Reliability:*** Predicting the failure of internal parts of a machine and intimating the service request for the machine.

1.7 PROBLEM DESCRIPTION

The supervised learning techniques on real time data sets are implemented using MapReduce programming model. Machine Learning techniques are tuned for processing the huge real time datasets. These techniques from Machine Learning are applicable for prediction, pattern matching, deep learning, recommendation systems and many more.

MapReduce uses Machine Learning techniques as a single predictor. It demonstrates parallel map and reduce the running of jobs using iterative algorithm. The MapReduce model results in high disk rates and low throughput. Using the existing MapReduce model, deploying analytics is a forthcoming challenge for the organizations because the end users gather information from existing data for further data analytics.

To build predictive intelligent system there is a need to tune existing models by exercising Machine Learning techniques. Machine Learning methods are useful to predict from the existing features of data. As data is distributed on cluster machines analytics is a major confront to Hadoop systems to channelize processing efficiently. Optimization can be achieved by minimizing the total execution time bounding by best possible prediction from existing techniques. Thus this proposed model depicts the best prediction from supervised learning techniques for handling the real time data with limited time bounds.

The existing MapReduce framework faces several shortcomings resulting in an inefficient cluster performance due to high rate of disk I/O requests and low throughput for poor resource utilization. To overcome this issue there is a necessity to build automated system where learning algorithms is identified by computing prediction on data sets using Spark framework.

The proposed model uses Apache Spark framework by tuning the Machine Learning techniques for prediction. A comparative analysis examines various techniques considering time and space complexities of various cluster jobs. Orange is an open source tool that favours comparative analysis. Random Mean Square Error (RMSE) is computed as a measure of relevance for each algorithm.

1.8 OBJECTIVES

The primary objective of this research is to tune Big Data methodologies using supervised approach. This research is a fair attempt to study the complex learning techniques. Prediction outcome is extracted from the existing data sets by utilizing Machine Learning methods. The captivating objectives are discussed below:

- **To develop a predictive model that forecast from the test data**
  The supervised learning methods are used as automated tool for prediction on data set. It incorporates methods to transform, evaluate and predict from the data by computing RMSE.
- **To identify the framework for distributed platform for analytical modeling**
  Data is parallely distributed on cluster environment, which appropriately adapts the dataset for analytical processing.
- **To compare Machine Learning Algorithms for accurate prediction**
  RMSE is calculated on each learning method and collective evaluation is examined by considering time and space complexities.
- **To evaluate new model with MapReduce and Machine Learning techniques**
  This model is evaluated on Apache Spark framework by replacing the existing MapReduce for prediction through supervised algorithms.
- **To predict the data from existing feature selection**
  Apart from the considered dataset there is an elite chance of possibility to append further features to increase the accuracy of prediction.

1.9 LIMITATIONS AND DELIMITATIONS

The following are the various limitations and delimitations of this research:

i. ***Time and Space as Complexity Features:***The data collected for analysis is massive and multifaceted. Time and Space are considered for evaluating the complexities of read and write operations while implementing on Spark project.

ii. ***Data Size:*** Enormous data is collected as a training set to learn Machine Learning methods. A collection of hundred and fifteen years of temperature data, which consist of both annual and seasonal, is used for predictive analysis.

iii. ***Predictive Modeling:*** Different models of analytics can be applied on a dataset. But this research work focuses on predictive modeling. According to this model, data is partitioned to train test sets for predictive analytics.

iv. ***Data Analytics:***Data Analytics mainly concentrate on evaluating the features in the temperature dataset. The datasets are analyzed using Machine Learning algorithms in comparison with MapReduce and Spark model.

v. ***Supervised Learning Differences:***Supervised learning for prediction is categorized as classification and regression models. Inferring the similarities and differences of various regression models are considered on the test dataset.

vi. ***Other Learning Methods:***The proposed model is not applicable for *Unsupervised* and *Reinforcement* learning techniques.

1.10 SOCIETAL CONTRIBUTION

This thesis portrays the usage of Machine Learning techniques for Big Data analytics. The proposed model incorporates the parallel processing on distributed environment like Spark to predict various features from datasets. The central core of the thesis is the comparative study on various Machine Learning algorithms on a proposed model for predictive analytics.

The following are the unique features of research study:

1. Proposing a novel model to obtain best prediction from various Machine Learning algorithms.
2. Time and space are the prime parameters for assessing performance metrics of this model.
3. MapReduce and Spark models are interpreted for computational operations on clusters after job allocation.
4. The data drawn from past experience are trained as test data using supervised learning.
5. At times applying certain irrelevant and redundant features on the model may affect the performance negatively due to outliers. For instance, considering the temperature dataset, the redundant and irrelevant features like amount of rainfall, salinity and salt concentration in the ocean water may impact negatively.

1.11 ORGANIZATION OF THESIS

The thesis is divided into six chapters.

Chapter 1 presents an introduction on Big Data Analytics followed by the detailed description on MapReduce technique stretched upon the Machine Learning approaches. In particular, data analytics focus on the analytical tools for incorporating these techniques on Spark environment. Extensively Big Data analytics deal with these learning methods by analyzing the performance using Spark framework.

A comprehensive review on various methods that handle Big Data using Machine Learning methods is categorized in chapter 2. It gives a glimpse of all possible data structures used for managing the big data and discourses on various state of work. Apparently, an analytical approach on MapReduce implemented on Machine Learning methods is articulated in this chapter.

Chapter 3 is an assortment of theoretical facts that apprehends significant tools and method. These methods connect data analytics approaches with Machine Learning for Big Data computation in the field of research. This knowledge assists in understanding the concept in an appropriate way. The tools used in implementing the proposed system are focused in this chapter.

The fine analysis across the embarked boundaries of each module incorporated in the developed system is demonstrated in chapter 4. It depicts the flow of actions carried out to fetch entire functionality of the developed system. In addition, the implemental chores are elucidated in a distinct way.

Chapter 5 shows the experimental results of the developed system. It highlights the accuracy of the integrated components of developed system. The discussions about the developed model display the functionality in a positive venture.

Finally, chapter 6 has the concluding part with the summarized portions of work carried out. It is followed by future directions which pave the way for further enhancement or extensions of the work. It also addresses open issues for future work.

1.12 SUMMARY

An elaborate view on Big Data Analytics stating the important features and causes of technical data evolution has been presented in this chapter. Since, exponential growth of data

demands for parallel automation which is carried out successfully by Machine Learning approaches. This chapter provided detailed description and a viable solution using Machine Learning techniques under various structural data backgrounds. In addition, it also stated the role and the importance of Spark MLlib library in construction of Machine Learning applications. Lastly, clear depiction on supervised learning techniques, features of Big Data analytics and advantages of Spark framework upon Hadoop structures have been emphasized.