CHAPTER - 2

REVIEW OF LITERATURE

2.1 INTRODUCTION

This chapter gives a comprehensive review on MapReduce programming model and the analytical processing constructed upon various platforms. It provides a detailed description on various data structures edificed for Spark and Machine Learning framework in a distributed environment for parallel processing. The entire chores of learning methods applicable for Spark model and the demonstrative works prevalent among these fields are discussed in this chapter.

2.2 MAPREDUCE AND HADOOP DISTRIBUTED FILE SYSTEM

MapReduce plays an involuntarily action that parallelizes and executes the program on a large cluster of commodity hardware. Being a data processing tool, large amount of data is processed on several machines using distributed environment. Hadoop an open source distributed file system implement MapReduce for efficient storage on parallel databases. Though MapReduce programming model has mentionable benefits the following segment describes few limitations that affect its performance on clusters.

- MapReduce is mainly used in Google for sorting, data mining and Machine Learning. This programming model deal with computational fault-tolerance in parallel distributed background. Locality optimization facilitated with read/write data operations to local disk saves network bandwidth. Redundant execution reduces the impact of slow machines, machine failures and data loss [1].
- A Comparative study on MapReduce with other database structures was discussed in this paper [2]. Startup latency, Data shuffling, Textual formats, Natural indices and unmerged output are significant advantages. MapReduce provides a heterogeneous storage system for data loading with fine grain fault tolerance and execution of more complicated function through simple SQL queries [2].
- The MapReduce implementation relies on an in-house cluster management system for distribution and execution of user tasks on shared machines. This cluster management resembles the other systems such as Condor [3]. Redundant execution that recover data loss caused by failures and locality aware scheduling to reduce the amount of data sent

across congested network links are fundamental similarities targeted in MapReduce programming model [4].

- Locality optimizations are motivated from techniques such as active disks where the computation is channelized into processing elements which are adjacent to local disks. This technique condenses the amount of data sent across I/O subsystem or the network. This system support commodity processor in which the small number of disks is directly connected instead of running directly on disk controller processor [5].
- This paper compared the performance of MapReduce and parallel databases. The comparative results addressed the shortcomings of MapReduce model that targets mainly on engineering considerations, reading unnecessary data, merging results and data loading [6].
- MapReduce comprises of row-oriented fashion to scan the input data that apparently slow the execution of analytical tasks when compared to other advanced knowledge databases in interactive searches [7]. In Trojan layout data is organized inside data blocks according to the incoming workloads implemented on top of HDFS. This layout replaces the MapReduce framework due to specific features such as co-locating attributes according to query workloads, different data block replicas and mimic fractured mirrors [8].
- Mastiff is optimized system for time based big data analytics. It exploits a systematic combination of a column group store and a light weight helper structure for diverse workloads to boast query performance. The MapReduce framework is replaced by this system for its high data loading scheme and high query performance [9].
- Hadoop software stack consists of MapReduce execution engine, pluggable distributed storage engine and a range of procedural to declarative interfaces for Big Data analytics. Negative aspect of Hadoop performance is explicitly proved by Starfish introduced by Herodotos. These self-tuning system lifecycle analytics automatically that parallely runs knob for performance tuning to optimize resources, time and money [10].
- Job optimization and physical data structures like data layouts and indexing are implemented using Hadoop MapReduce. This research study discussed the state of approach for performance improvement [11].
- ETL iterative tools are implemented to transfer data from warehouses to marts for analytical processing. The novel challenges like data mining, data quality validation,

cleansing, profiling, statistical algorithms, hierarchical and drill down functionality are substituted with ETL process designed by Intel [12] for better performance.

- Google designed a scalable distributed file system for large data intensive applications. The aspect of fault tolerance on execution of commodity hardware delivers high aggregate performance to a large number of clients. Thus MapReduce facilitate high aggregate throughput to parallel readers and writers in a DFS [13].
- Apache Hadoop provides a versatile ecosystem with all data processing tools featuring automation, scaling and built-in error torrents. Apache Hadoop ecosystem plays a significant role in day-to-day practices of enterprises and research groups. In this article author emphasized the importance of Hadoop system by comparing the relational databases because of the increasingly large volumes of digital knowledge [14] [15].
- Hive QL is a declarative SQL language implemented on MapReduce using Hadoop, which enables to plug MapReduce scripts into queries for solving the problems dealing with massive data sets [16].
- 'Hadoop, MapReduce and HDFS: A Developers Perspective' article discussed the complexity behind the programming model MapReduce on a cluster of commodity server. MapReduce is modeled as independent platform service layer in enabling data processing and analyzing [17].
- Cloud Technology is progressed by handling massive and novel datasets by increasing value to businesses and processing power. Cloud vendors upload the data in their clouds for effectiveness and easy-of use of data analytical algorithms. HDFS systems are designed to hold data by providing high-throughput and access to this information on the cloud systems [18].

2.3 MAPREDUCE USING MACHINE LEARNING

A diverse body of existing works has been focused on MapReduce using Machine Learning approaches. The following is an assortment of various analytical algorithms for processing large data sets using Hadoop Distributed File System.

- A statistical query model uses the least square regression in two-step process. Initial step computes applicable statistics on the data for processing. The subsequent step concentrates on computation by using data mining approaches on summations [19]. This model implements diverse Machine Learning algorithms on top of MapReduce framework.