International Journal of Advance and Innovative Research
Volume 4, Issue 4 (II): October - December, 2017
ISSN 2394 - 7780

HADOOP MAPREDUCE : FUTURE OF STOCK PRICE ANALYSIS

Anshuman Sharma
Assistant Professor, Department of Computer Science, Hindu College, Amritsar

ABSTRACT

Stock Market trade a lot many shares on daily basis. The prices of a stock are fluctuates each minute. Lot of data is being generated on daily basis. Beside many technical aspects like PE, EPS, dividend etc many indirect factors may influence the market price of shares corresponding to a particular company or particular sector of companies. These may include weather forecasting for rainfall, past record of stock performance data, markets of neighboring countries, list of competitors, volume of shares traded, feedback from various surveys etc. So this huge data due to various indirect factors also be need to analyze and predict the price of a particular share in the future. In addition to the huge volume of data that is generated on daily basis, data fetched from various sources may be structured or unstructured like messages regarding shares from a from websites dealing with shares, news articles etc. It is important to analyze this data to identify patterns, trends and predict the future prices. This can be done using, big data analysis using Hadoop Mapreduce ecosystem tools. This paper discusses the Hadoop tools that can be used in analysis of prices of stocks and thus predict which shares to buy and at what time.

Keywords: Big data, Hadoop, Stock market analysis, Pig

INTRODUCTION

A large number of people invest in shares either directly or indirectly using their fund managers, SIP etc. Many investors need various kinds of reports, charts and answer to various queries to predict the future stock prices. The stock prices data and various other data is usually available from stock exchanges like NSE web sites is structured and can be analyzed using existing tools like SPSS etc. But quite often we need to analyze data over a period of 25 years and more. Analysts may also want to consider data such as stock prices in neighboring countries of similar sectors, weather forecasting to predict effect on agriculture, and many other factors. In addition in future due to Internet of Things, many devices will also be used to provide data. Such huge volume of data being structured may also be unstructured as it may be fetched from various other sources. This data can be termed as big data due to its 5V's – Volume as huge volume of data is to be analyzed, Variety as data may be of different types, Velocity as data is generated on daily basis, Value as it is important to find correct meaning of data, Veracity as in some cases there may be uncertainty and inconsistency in data.[3,4] Big data extracts and organizes the valued information from the rapidly growing data sets collected from multiple and autonomous sources in the minimal possible time, using several statistical and machine learning techniques. [5] The big data real time analysis can be done using the Hadoop Mapreduce ecosystem in a quick and efficient way. Moreover, as most of the tools are open source so cost is not an issue for analyzing data as compared to other proprietary tools available.

To practically analyze big data, various tools that are part of Apache Hadoop ecosystem can be used. Hadoop is a framework that enables applications to work on large amounts of data on clusters in parallel and distributed fashion. Some tools are (a) Flume used for ingesting unstructured/semi structured data from social media sites into HDFS. (b) Scoop used for ingesting structured data from social media sites into HDFS. (c) HDFS which is a distributed file system that stores the data on various computers called nodes, enabling a high bandwidth across the cluster. To implement a parallel computational algorithm, MapReduce, is used. It divides the main task into small chunks. These small chunks are mapped by processing parallel thereby increasing efficiency. The results obtained are combined into a final output, the reduce stage.[1] (d) YARN (Yet another resource Negotiator) used for processing big data. (e) Hive which is developed by Facebook and used for analytics. It uses hive query language (Hive QL) which is similar to SQL. (f) Apache PIG, programming language is configured to assimilate all types of data (structured/unstructured, etc.). It is comprised of two key modules: the language itself, called PigLatin, and the runtime version in which the Pig Latin code is executed.[8] (g) Apache Spark which is an open source real time processing engine used in big data analytics. (h) Apache Mahout and Spark MLib used for machine learning . (i) Zookeeper allows a centralized infrastructure with various services, providing synchronization across a cluster of servers. Big data analytics applications utilize these services to coordinate parallel processing across big clusters.[8] (j) Apache Ambari for management and coordination of Apache Hadoop Cluster. (k) HBase is a column-oriented database management system that sits on top of HDFS. It uses a non-SQL approach. (l) Cassandra is also a distributed database system. It is designated as a top-level project modeled to handle big data distributed across many utility servers. It also provides r