International Journal of Security and Privacy in Pervasive Computing
Volume 13 • Issue 1 • January-March 2021

Analysis of Big Data Using Two Mapper Files in Hadoop

Jyotsna Malhotra, Deaprtment of CSE, Guru Tegh Bahadur Institute of Technology, India
Jasleen Kaur Sethi, USICT, Guru Gobind Singh Indraprastha University, India
Mamta Mittal, Department of CSE, G. B. Pant Government Engineering College, India

ABSTRACT

Nowadays, a large amount of valuable uncertain data is easily available in many real-life applications. Many industries and government organizations can exploit this data to extract valuable information. This information can help the managers to enhance their strategies and optimize their plans in making decisions. In fact, various private companies and governments have launched programs with investments and funds in order to maximize profits and optimize resources. This vast amount of data is called big data. The analysis of big data is important for future growth. This paper depicts big data analytics through experimental results. In this paper, data for New York stock exchange has been analyzed using two mapper files in Hadoop. For each year, the calculation of maximum and minimum price of every stock exchange and the average stock price is done.

KEYWORDS

Big Data, Big Data Analytics, Hadoop, Hadoop Common, Hadoop Distributed File System, Map Reduce, Mapper, Reducer

INTRODUCTION

We are entering in the era of Big Data – a word that refers to the escalation of available information. A massive amount of unstructured data is continuously producing and stored with cheaper cost than earlier. The rate of growth of data is of the order of zeta bytes and is even further growing. This data is diverse and heterogeneous in nature (Gandomi & Haider, 2015). The traditional database management systems are incapable of handling such vast amount of data. To handle this vast amount of data, Big Data came into existence which has new architecture and techniques to manage the data and obtain useful information from it. Hadoop is based on Google's Map Reduce programming environment. In Hadoop, data is stored in Hadoop Distributed File Systems (HDFS) and Map Reduce is used to process data in parallel on different nodes (Mittal et al, 2018).

Hadoop Map Reduce finds applications in many domains like biological analysis, stock prediction, internet traffic measurement etc. CloudBurst is an algorithm proposed by Schatz (2009) used to map genome data for biological analysis like genotyping. It addresses the alignment and ambiguity problems in the gene mapping programs. CloudBurst is based on Hadoop Map Reduce framework and executes parallel execution of multiple nodes to reduce

DOI: 10.4018/IJSPPC.2021010104

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

the execution time and performance significantly over other gene mapping programs. Another application of Map Reduce is the analysis and measurement of internet traffic. When a large amount of traffic data needs to be measured and analyzed, a single server is not sufficient. Data packet sampling and aggregation of data flow can be used to reduce the high volume of monitoring data. Its performance has been improved by Hadoop Map Reduce framework. It has been proved that Map Reduce improves the time for monitoring the internet traffic by 72% when compared to other traditional analysis tools (Lee et al, 2010). A variant of Hadoop Mapreduce has been proposed by Zhao and Wu (2014) called Dache which is data aware cache framework. In Dache, the results are submitted to the cache manager for matching which could reduce the execution time if a match is found. The protocols for the request and reply to the cache manager have also been designed. It has been shown the Dache when used along with Map Reduce reduces the computation time of the jobs submitted and the processor time. One of the popular algorithms used in recommender systems are the Collaborative Filtering. A major drawback of this algorithm is their high computational complexity. Zhao and Shang (2010) proposed a collaborative filtering algorithm based on Map Reduce such that it increases the scalability. It partitions the users so that all the processors complete the task at the same time which enhances the performance.

Although a plethora of research has been carried out for Hadoop Map Reduce, all the work has been focused to reduce the computation time. Not much research has been carried out if the input data to Hadoop is multiple files which may or may not be in different formats or separator. In this paper, the dataset of New York Stock Exchange from two different mapper files has been analyzed using Hadoop Map Reduce.

The remaining paper has been structured in the following manner: Section 2 gives the overview of Big Data. Section 3 presents an overview of the proposed methodology. In Section 4, the results have been presented. A brief conclusion has been provided in Section 5.

OVERVIEW OF BIG DATA

Big Data is found in number of fields like Astrophysics, Mathematics, Energy, Fundamental Science, Medicine, Engineering, and Remote Sensing and Statistical data. In all these fields the majority of challenges are the development of algorithms, visualization techniques, and analytics. To resolve these challenges there are many Big Data analysis tools available like Hadoop that is an open source platform used for parallel execution of Big Data. It processes large datasets across multiple nodes. According to the definition given by McKinsey Global Institute (MGI, 2012):

"Big Data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze (Bhatia et al, 2017).

Big Data is capable of changing the financial value across the following major sectors:

- Public sector department
- Due to widespread use of mobile phones, personal position data of various customers is maintained
- Due to the presence of large Web e-commerce and retail shops such as eBay and Amazon, retail has become the important sector
- Since patient history and information related to health is very important, Health care is a critical area
- Manufacturing (Zicari, 2014)

We can categorize the following challenges in Big Data.

7 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/analysis-of-big-data-using-two-mapper-files-in-hadoop/269506

Related Content

Pervasive Computing and Ambient Intelligence Development: An Educational Perspective
Vladimír Burešand Pavel Cech (2010). *Strategic Pervasive Computing Applications: Emerging Trends (pp. 235-249).*
www.irma-international.org/chapter/pervasive-computing-ambient-intelligence-development/41593

Inscribing Interpretive Flexibility of Context Data in Ubiquitous Computing Environments: An Action Research Study of Vertical Standard Development
Magnus Anderssonand Rikard Lindgren (2009). *International Journal of Advanced Pervasive and Ubiquitous Computing (pp. 1-18).*
www.irma-international.org/article/inscribing-interpretive-flexibility-context-data/3864

Interpreting Health and Wellness Information
Lena Mamykinaand Elizabeth Mynatt (2009). *Auto-Identification and Ubiquitous Computing Applications (pp. 74-90).*
www.irma-international.org/chapter/interpreting-health-wellness-information/5457

A Modified Landweber Iteration Algorithm using Updated Sensitivity Matrix for Electrical Impedance Tomography
Lifeng Zhang (2013). *International Journal of Advanced Pervasive and Ubiquitous Computing (pp. 17-29).*
www.irma-international.org/article/a-modified-landweber-iteration-algorithm-using-updated-sensitivity-matrix-for-electrical-impedance-tomography/92999

Enhancing Location Privacy in WSN: The iHide Case
Leonidas Kazatzopoulos (2013). *Intelligent Technologies and Techniques for Pervasive Computing (pp. 191-203).*
www.irma-international.org/chapter/enhancing-location-privacy-wsn/76788