Advanced Big Data Analytics For Ecommerce

Snehal Nrupnarayan[1], A. R. Itkikar[2]

**[1,2]Sipna College of Engineering and
Technology, Amrawati, Maharashtra-444607,
India**

Abstract

The assistance of big data analytics is essential for optimizing e-commerce sites. The amount data being generated is of huge scale and performing analytics over such a huge data is a complex task. Further deriving the business logic through such analysisis the actual motive of the process. Conventionally, Hadoop is the most widely used Big data framework. But it has some limitations which came across recently. The process focuses on using different analytics framework “Apache Spark” which will be able overcome to Hadoop’s limitations and provide efficient optimization outcome through it’s analytics.

Keywords: *Big Data, Hadoop, HDFS, Spark, Data Analytics.*

1. Introduction

Big Data analytics is the process of collecting, organizing and analyzing large sets of data (called Big Data) to discover patterns and other useful information.Big Data analytics can help ecommerce sites to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysis typically is worked out towards derving the knowledge.

To analyze such a large volume of data, Big Data is typically performed using specialized software tools and applications for predictive analytics, data mining,text mining, forecasting and data optimization. Collectively these processes are separate but highly integrated functions of high performance analytics. Using Big Data tools enables to determine which data is relevant and can be analyzed to drive better business decisions.

The focus of Data Analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows. Driven by specialized analytics systems and software, big data analytics can point the way to various business benefits, including new revenue opportunities, more effective marketing, better customer service, improved operational efficiency and competitive advantages over rivals.

Hadoop is Big data framework widely used in variety of Big data analytics projects. It finds a lot of implementations in the ecommerce optimization strategies. It has been the most popular choice of the analysts when it comes to the Big Data Analytics.

2. Data Processing Framework : Hadoop

Hadoop is an open source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

It’s large computing clusters are prone to failure of individual nodes in the cluster. Hadoop is fundamentally resilient – when a node fails processing is redirected to the remaining nodes in the cluster and data is automatically re-replicated in preparation of future node failures.

In Hadoop, with a parallel and distributed algorithm, MapReduce process large data sets. There are tasks that need to be performed: Map and Reduce and,

3. Limitations of Hadoop Framework

1)Slow Processing Speed:

MapReduce requires a lot of time to perform these tasks thereby increasing latency. Data is distributed and processed over the cluster in MapReduce which increases the time and reduces processing speed.

2)Support for batch processing only:

Hadoop supports batch processing only, it does not process streamed data, and hence overall performance is slower. MapReduce framework of Hadoop does not leverage the memory of the Hadoop cluster to the maximum.

3)No Real-Time Data Processing:

Apache Hadoop is designed for batch processing, that means it take a huge amount of data in input, process it and produce the result. Although batch processing is very efficient for processing a high volume of data, but depending on the size of the data being processed and computational power of the system, an output can be delayed significantly. Hadoop is not suitable for Real-time data processing.

4)Latency:

In Hadoop, MapReduce framework is comparatively slower, since it is designed to support different format, structure and huge volume of data. In MapReduce, Map takes a set of data and converts it into another set of data, where individual element are broken down into key value pair and Reduce takes the output from the map as input and process further and MapReduce requires a lot of time to perform these tasks thereby increasing latency.

To overcome these limitations, we need to have framework which lies in the category of in-memory data processing.A framework which will stage pipelines to use the processed data over and over again. Using such framework for Big data analytics in ecommerce field which result into a considerable increase in speed of deriving business logic and will be best suited for it's optimization. One such framework is "Apache Spark".

4. Proposed Approach

Applying Big data Analytics other than conventional optimization techniques will result into a much better efficiency and accuracy of optimization. The conventional tool used for big data analytics is Hadoop. The proposed approach is to apply analytics using Apache spark, a new data analytics framework which has overtaken most of the open Source Big data Projects. It's getting preference over Hadoop nowadays and is much faster than Hadoop in certain circumstances.

The main issue with Apache Spark is that it does not provide it's own distributed file system. The file system is necessary for a systematic organization of files or data. While dealing with "Big data", an efficient file system is must. Thus, proposed approach will have a combination of Hadoop and Spark. Hadoop framework will provide a distributed file system HDFS (Hadoop Distributed File System) for systematic organization on top of which will reside the "Analytics Layer" of Apache Spark.

Architecture:

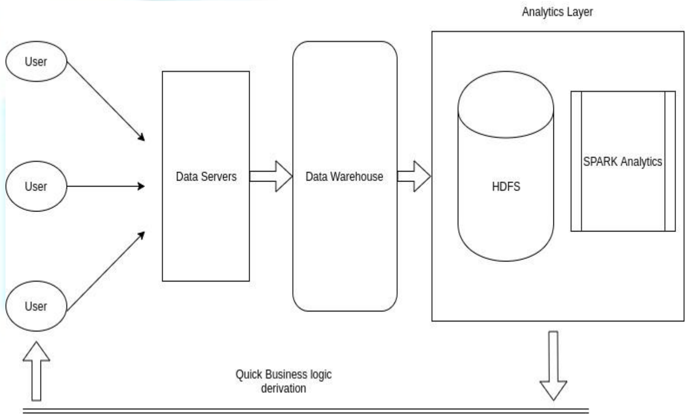

Fig 1:Analytics using Spark

5. Conclusion

Big data has been on the rise in the recent years. Various advancements in the field of social media, consumer electronics and personal computing has led to accumulation of huge data which has a relative connection to users. This huge scale of data is beneficial for a core analysis of customers and enhance the business logic to reap more profits. Big data analytics is advantageous and rewarding, but also it a very time consuming process. Conventionally used Hadoop framework has been the most extensively used Big data tool but it has it's cons which have came across in the recent years with ever increasing size of the data.Thus, this paper provides an improved analytics process with a new category of framework "Apache Spark" which could overcome the disadvantages of Hadoop and provide a new way of implementation for Big Data Analytics in the field of e-commerce.

References

[1] BIG DATA DRIVEN E-COMMERCE ARCHITECTURE,Ahmad Ghandour ,Department of MIS, College of Business Administration

[2] A Survey on Big Data Market: Pricing, Trading and Protection,Fan Liang, Wei Yu, Dou Any, Qingyu Yangz, Xinwen Fux and Wei.

[3] Big data analytics in E-commerce: a systematic review and agenda for future research Shahriar Akter1 & Samuel Fosso Wamba2

[4] The Hadoop Distributed File System: Architecture and Design writeen by Dhruba Borthakur

[5] STRATEGIC ANALYSIS WITH BIG DATA ,Pooja G. Joshi 1, Om P. Bankar 2, Ajay K. Ghode 3,Neeta Patil 4

[6] KP-S: A Spark-Based Design of the K-Prototypes Clustering for Big Data Mohamed Aymen Ben HajKacem ; Chiheb Eddine Ben N'Cir ; Nadia EssoussiFLEXChip Signal Processor *(MC68175/D)*, Motorola, 1996.

[7] Advanced Analytics with Spark, by Juliet Hougland, Uri Laserson, Sean Owen, Sandy Ryza and Josh Wills (O'Reilly Media)

[8] Fast Data Processing with Spark, Krishna Sankar and Holden Karau (Packet Publishing)