AN ADVANCED MAP REDUCE FRAMEWORK IMPLEMENTATION FOR BIG DATA APPLICATIONS

Mr. Prasanth Settivari[1]
3rd Year Student,
Department of Computer Science,
SV U CM & CS, Tirupati.

Prof. G.Anjan Babu[2],
Professor,
Department of Computer Science,
SV U CM & CS,, Tirupati.

Abstract: Big data takes many forms, including messages in social networks, data collected from various sensors, captured videos, and so on. Big data applications aim to collect and analyze large amounts of data, and efficiently extract valuable information from the data. A recent report shows that the amount of data on the Internet is about 500 billion GB. With the fast increase of mobile devices that can perform sensing and access the Internet, large amounts of data are generated daily.

In general, big data has three features: large volume, high velocity and large variety [1]. The International Data Corporation (IDC) predicted that the total amount of data generated in 2020 globally will be about 35 ZB. Face book needs to process about 1.3 million TB of data each month. Many new data are generated at high velocity. For example, more than 2 million emails are sent over the Internet every second.

Mobility services such as Google Maps and Navigation Service provide benefits and convenience to people. These applications are big data applications because the data set size is big and the data update rate is fast [2]. Large amounts of fresh mobility-related data are generated every day, for instance, video surveillance data collected by high-definition cameras at roadsides and junctions.

Typically, the rapidly generated big data are not uploaded to a data center at once. Instead, the fresh big data is quickly stored in local servers temporarily. Previous research works on big data mainly study efficient processing techniques and analytical methods for big data in a clustered environment, and do not consider a geo dispersed big data scenario.

Big data takes many forms, including messages in social networks, data collected from various sensors, captured videos, and so on. Big data applications aim to collect and analyze large amounts of data, and efficiently extract valuable information from the data. For example, analysis of video content is a complex operation. An advance map reduce framework was proposed to support such complex operation on big data applications. This project deals on how the proposed advance map reduce frame work is implemented and how it shall work on a real time big data application.

INTRODUCTION:

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy. There is no hard and fast rule about exactly what size a database needs to be in order for the data inside of it to be considered "big." Instead, what typically defines big data is the need for new techniques and tools in order to be able to process it. In order to use big data, you need programs which span multiple physical and/or virtual machines working together in concert in order to process all of the data in a reasonable span of time. The uses of big data are almost as varied as they are large. Prominent examples we are probably already familiar with including social media network analyzing their members' data to learn more about them and connect them with content and advertising relevant to their interests, or search engines looking at the relationship between queries and results to give better answers to users' questions. But the potential uses go much further! Two of the largest sources of data in large quantities are transactional data, including everything from stock prices to bank data to individual merchants' purchase histories; and sensor data, much of it coming from what is commonly referred to as the Internet of Things (IoT).

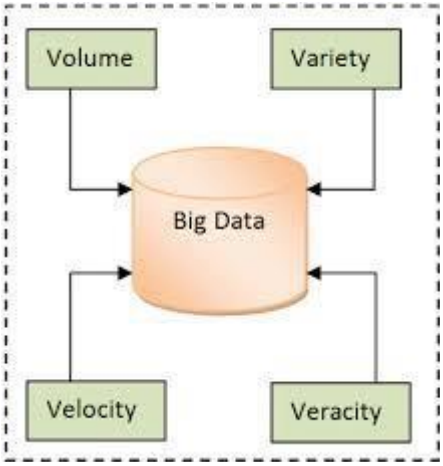

Figure 1: Different forms of Big Data

This sensor data might be anything from measurements taken from robots on the manufacturing line of an automaker, to location data on a cell phone network, to instantaneous electrical usage in homes and businesses, to passenger boarding information taken on a transit system. One of the best-known methods for turning raw data into useful information is by what is known as Map Reduce. Map Reduce is a method for taking a large data set and performing computations on it across multiple computers, in parallel. It serves as a model for how to program, and is often used to refer to the actual implementation of this model. In essence, Map Reduce consists of two parts. The Map function does sorting and filtering, taking data and placing it inside of categories so that it can be analyzed. The Reduce function provides a summary of this data by combining it all together. While largely credited to research which took place at Google, Map Reduce is now a generic term and refers to a general model used by many technologies. Perhaps the most influential and established tool for analyzing big data is known as Apache Hadoop. Apache Hadoop is a framework for storing and processing data in a large scale, and it is completely open source. Hadoop can run on commodity hardware, making it easy to use with an existing data center, or even to conduct analysis in the cloud. Hadoop is broken into four main parts:

The Hadoop Distributed File System (HDFS), which is a distributed file system designed for very high aggregate bandwidth.

YARN, a platform for managing Hadoop's resources and scheduling programs which will run on the Hadoop infrastructure.

Map Reduce, as described above, a model for doing big data processing. And a common set of libraries for other modules to use. Big data takes many forms, including messages in social networks, data collected from various sensors, captured videos, and so on. Big data applications aim to collect and analyze large amounts of data, and efficiently extract valuable information from the data. A recent report shows that the amount of data on the Internet is about 500 billion GB. With the fast increase of mobile devices that can perform sensing and access the Internet, large amounts of data are generated daily. In general, big data has three features: large volume, high velocity and large variety. The International Data Corporation (IDC) predicted that the total amount of data generated in 2020 globally will be about 35 ZB. Face book needs to process about 1.3 million TB of data each month. Many new data are generated at high velocity. For example, more than 2 million emails are sent over the Internet every second.

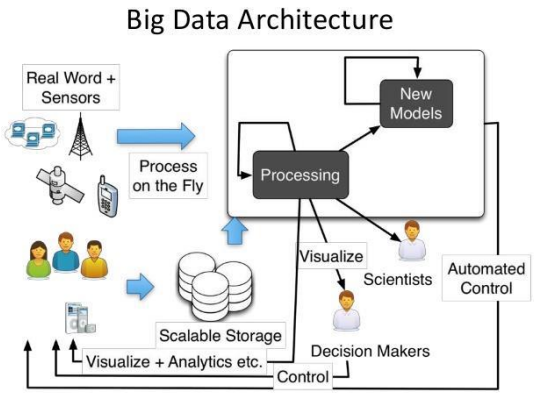

Figure 2: Big Data Architecture

CONTENT ANALYTICS TECHNIQUES

The main objective of this research is as follows:

➢ Hadoop environment setup in a single node.

➢ Implement the Map Reduce program which shall take the video surveillance data as input and extract the frames from the video data.

➢ Implement the Map Reduce program which shall take the video frames as input and extract the required images based on the specific color.

➢ Implement the Map Reduce program which shall the specific color images as input and extract the required images based on the vehicle license.

Running HDFS and Map Reduce on a single machine is great for learning about these systems, but to do useful work they need to run on multiple nodes. There are a few options when it comes to getting a Hadoop cluster, from building your own to running on rented hardware, or using an offering that provides Hadoop as a service in the cloud.

Cluster Specification

Hadoop is designed to run on commodity hardware. That means that you are not tied to expensive, proprietary offerings from a single vendor; rather, you can choose standardized,

commonly available hardware from any of a large range of vendors to build your cluster.

CONCLUSION

This paper focuses mainly on projecting how the advance map reduce frame work can be implemented using hadoop map reduce frame work for a real time big data application and How to extract required data from a video content using the frame work in hadoop environment.

REFERENCES

1. Processing Geo-dispersed big data in an advance map reduce frame work - Hongli Zhang, Qiang Zhang, Zhigang Zhou, Xiaojiang Du, Wei Yu, and Mohsen Guizani IEEE 2015.
2. FFMPEG - https://ffmpeg.org/ffmpeg.html.
3. OpenALPR - http://doc.openalpr.com/
4. Hadoop: The Definitive Guide – Tom White.
5. J. Manyika et al., "Big Data: The Next Frontier for Innovation, Competition, and Productivity," McKinsey Global Inst., May 2011.
6. S. Shekhar et al., "Spatial Big-Data Challenges Intersecting Mobility and Cloud Computing," in Proc.of 11th ACM Int'l. Wksp. Data Engineering for Wireless and Mobile Access, Scottsdale, AZ, 2012, pp. 1–6.
7. D. Huang et al., "Secure Data Processing Framework for Mobile Cloud Computing," Proc. IEEE INFOCOM Wksp. Cloud Computing, Shanghai, China, 2011, pp. 614–18.
8. B. Chun et al., "Clonecloud: Elastic Execution between Mobile Device and Cloud," Proc. 6th Conf. Comp. Sys., New York, NY, 2011, pp. 301–14.
9. R. Yu et al., "Toward Cloud-Based Vehicular Networks with Efficient Resource Management," IEEE Network, vol. 27, no. 5, Sept.–Oct. 2013, pp. 48–55.
10. M. Felemban, S. Basalamah, and A. Ghafoor, "A Distributed Cloud Architecture for Mobile Multimedia Services," IEEE Network, vol. 27, no. 5, Sept.–Oct. 2013, pp. 20–27.
11. D. Huang, T. Xing, and H. Wu, "Mobile Cloud Computing Service Models: A User-Centric Approach," IEEE Network, vol. 27, no. 5, Sept.–Oct. 2013, pp. 6–11.
12. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Commun. ACM, vol. 51, no. 1, Jan. 2008, pp. 107–13.
13. K. H. Lee et al., "Parallel Data Processing with MapReduce: A Survey," ACM SIGMOD Record, vol. 40, no. 4, Dec. 2011, pp. 11–20.

Authors Profile

SETTIVARI PRASANTH, received Bachelor of Computer Science degree from Sri Venkateswara University, Tirupati in the year of 2013-2016. Pursuing Master of Computer Applications from Sri Venkateswara University, Tirupati in the year of 2016-2019. Research interest in the field of Computer Science in the area ofArtificial Intelligence, Machine learning, Big Data, Network Security, Networkingand Software Engineering.