RESEARCH ARTICLE OPEN ACCESS

A Review Document on Apache Spark for Big Data Analytics with Case Studies

Vivek Francis Pinto [1], Sampath Kini [2], Igneta Mcluren Dsouza [3]
Department of Computer Science and Engineering [1] & [2]
NMAM Institute of Technology, Nitte
Department of Computer Network and Engineering [3]
RVCE, Bengaluru
India

ABSTRACT
Evolution in technology has given rise to usage of new methods for collecting data. On the other side the size of the data that is getting collected is of huge size which is categorized as Big Data. Big Data has three main characteristics namely Volume, Velocity and Variety. Organization are looking to process this huge data that would help them in getting additional value to their business. The review document gives an insight about Big data, Hadoop Map Reduce and Apache Spark and review of few case studies in which Apache Spark has been used.
Keywords: — Big Data, Apache Spark, Map Reduce, Big Data Analytics

I. INTRODUCTION

Over the past years, many organizations have collected data as part of their business activities, proposals, feedbacks, product reviews, customer entries and through many other mean and it was only collected rather than processing for a value out of. Before a decade, organizations started to use these data for marketing and to target only few selected customers. Currently the same organizations are beating the data using improved data processing techniques to get the additional information that adds revenue to their business and increases customer satisfaction.

The word Big Data has become a buzz word from then in the field of information technology. A processing methodology called as Big Data Analytics started its play by giving processing capability to process the data using Hadoop' s Map Reduce. However there was need of even more speeder processing technology which was addressed by Apache Spark which runs program 100 times faster than the Hadoop Map Reduce in memory and ten times faster when on disk [1].

Further, on the business front, the Data that has been generated has given raised to improved business tactics to be implemented upon understanding different factors using data processing. Data analytics is used in variety of industries to understand customer behaviour [2], to understand a patient's health related aspect or patterns [3], to understand loopholes of system [4], to understand the trend, to decide on number of units of the products to be needed, to detect unusual activities in the transaction [5] and by the government of India in rural development planning [6].

II. BIG DATA ANALYTICS HADOOP MAPREDUCE

Hadoop MapReducer is a technology for implementing application in order to process huge amount of data in size of multi tera-byte using thousands of node or clusters in a reliable manner[7].

A. Mapper

Given a data file, based on the implantation logic written, Mapper of Hadoop maps the input key/value pairs to a set of intermediate key/value pair [7].These tasks will transform input into intermediate records. Further a concept of Combiner can be used by the user in order to do local aggregation of the intermediate output.

B. Reducer

Once the mapping is done and intermediate output in generated, Hadoop processing will perform the Reduce operation, the operation which reduces a set of intermediate values which share a key to a smaller set of values [7].Further reducer works in three phases. Shuffle phase which fetches the relevant partitions in the sorted output. Sort which groups the input keys and finally the Reduce phase.

Hadoop's Mapper main goal is to process the large amount of input data and break down into tuples or key/value pair. Reducer takes the input as the output of Mapper phase and still reduces the tuples of smaller size.

C. Common Terminologies of MapReduce Framework

1. HDFS : Hadoop Distributed File System distributed file system designed to run on commodity hardware[8]
2. Master Node: Node where the job tracker executes and takes the job request from the client.
3. Slave Node: Node where the Map and Reduce tasks are executed.

4. Data node: Before the computation or processing is done, the data resides in Data Node.
5. Name Node: Node that is associated with the Hadoop Distributed File System.
6. Task: The execution of the Mapper and Reducer is called as Task
7. Job Tracker: Entity responsible to schedule the jobs and track the status of entire job.
8. Task Tracker: Entity responsible for a single task to be tracker and this reports to Job tracker.

On a business and technology view ,Big Data analytics is used to get the insight on the data that can increase the revenue of the organization that uses it and it can be broadly be classified as Diagnostic analytics ,Descriptive analytics ,Predictive analysis and perspective analysis. The Volume, Variety and Velocity factors make the Big Data analytics as very critical in today's environment of business.

III. APACHE SPARK

As stated earlier, even though MapReduce helped in getting good insight and increased revenue generation, the processing speed as the size of data increased was a challenge. To address this challenge, Matei Zaharia initially started implementation of open-source cluster computing framework. Developed at Berkeley's University of California and later gave to Apache Software Foundation [9].Spark provides implementation using Java, Scala and Python and R.

Fig 1 shows that the time taken by Hadoop for processing data is 110 times more than that it is taken by Spark. Due to the usage of Directed Acyclic Graph execution engine which creates a DAG of task stages to be performed on the cluster [10].

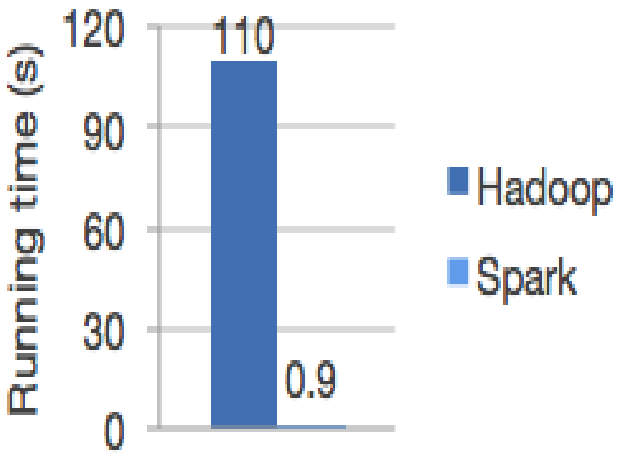

Fig. 1 Processing speed of Hadoop and Spark

Fig 2 shows the built in libraries of Apache Spark, namely Spark SQL, Spark Streaming, MLib Libraries and GraphX which can be used in any combination for implementing an application [1].

Spark SQL is a module for working with structured data which helps to seamlessly write the structured queries in the spark programs. Provides a uniform data access that can be used to access variety of data from sources like Hive, Avro, Parquet, JavaScript Object Notation and Java Database Connections. Further Spark SQL library provides libraries to access existing Hive warehouses. [11].Apache Spark SQL modules are shown in Fig 3.

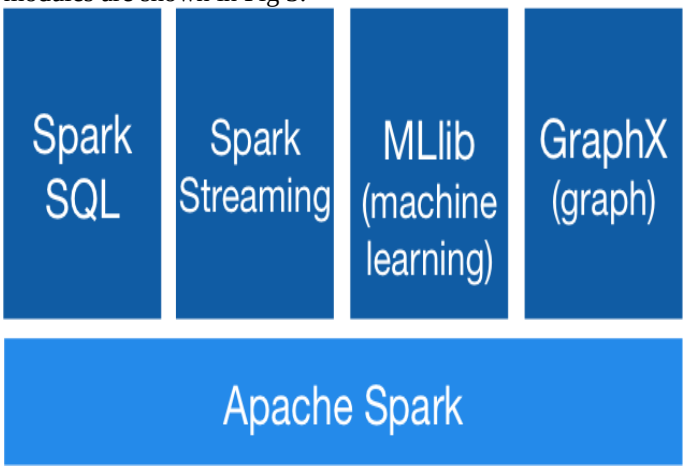

Fig. 2 Apache Spark Inbuilt Libraries

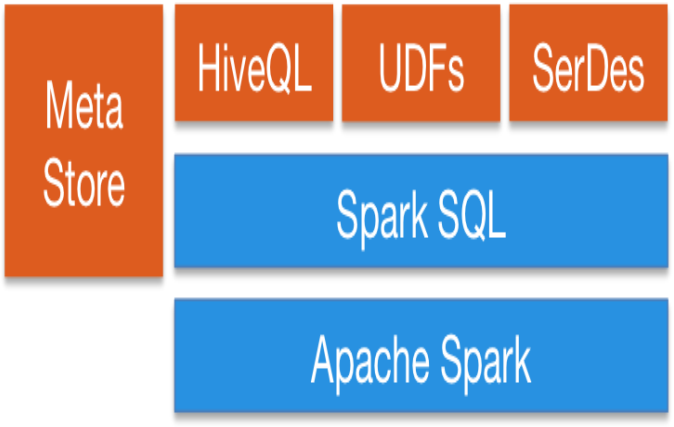

Fig. 3 Apache Spark SQL Library

Spark Streaming is a module which consists of language integrated application programming interfaces for processing of streams which helps to write the streaming job using Java, Python and Scala. Spark streaming has functionalities which helps to recover from a lost work and from an operator state. It can be used to read data from HDFS, Kafka, Twitter, Flume and ZeroMQ [12]

Spark Machine Learning Library called as MLib provides machine learning capabilities. This library interoperates with Python and R libraries and it can be integrated to any sources like HBase and Local files and can be plugged to Hadoop workflow [13].

GraphX Library of Spark provides functionalities that help to do graph parallel computation, to view the data on graph. The data structure behind the GraphX is Resilient Distributed Dataset which is used to represent subject or resource objects. Further GraphX has rich graph algorithm libraries which can be used based on the need.

Some of the companies which have already used Apache Spark are

1. Pintrest uses the spark streaming API to get deep insight into user engagement details.
2. Uber, in order to build the continuous ETL pipeline uses Spark Streaming, Kafka and HDFS.
3. Conviva which is a video company uses Spark for managing live traffic and for optimizing the videos.
4. Netflix uses the real-time stream processing for providing recommendations to its customers.

The survey performed by authors of Apache Spark brings out the following facts about "Why companies should use in-memory computing like Apache Spark?" which are as follows [14].

1. 91 percent use of Spark was to utilize its performance gain.
2. 77 percent use of Spark was contributed with the fact that it is easy to use.
3. 75 percent use of Spark was contributed because it is easy to deploy.
4. 64% percent use of Spark was to utilize its capabilities of advanced analytics.
5. 52% percent use of Spark was to utilize its real-time streaming capabilities.

IV. CASE STUDIES OF USING APACHE SPARK

This section gives a review of usage of Apache Spark in E-commerce and Travel and Media and Entertainment Industry. Spark is being used to transform the business and to attain higher revenue generation with the optimized service. Apache Spark has been used in Media, Finance, Retail, Healthcare and Travel.

A. Apache Spark in E-Commerce Industry

As many E-commerce companies started collecting huge amount of data, they also started to use Spark for getting value out it. Companies wanted to study on the kind of products its customer is buying so that recommendation and new products can be showed to customers. Using Spark, the purchase transaction details can be passed to streaming clustering algorithms. The results obtained by these algorithms can be combined with data from social media, reviews, feedback, and customer comment to give better recommendation to customer.

Shopify is stores partnership company which wanted to understand what its customers are selling so that it can plan to do partnership with such companies of which the customer selling more product. However without using Spark they were unable to achieve it. Upon using the Spark capabilities they were able to process 67 Million records in a minutes and it has created a list of companies it can probably do partnership. [14]

EBay one of the giant e-commerce company has used Apache Spark to provide better and matching offers to targeted customers and also to improve the customer experience. With the help of Hadoop YARN, generic tasks have been run by EBay which used clusters in the range from 2000 to 20,000 nodes with help of 1000TB RAM. [14]

Alibaba being one of Chinese e-commerce Company having mission to make it easy to do the business anywhere using its platform. Using their web portal, Alibaba provides business to consumer and consumer to consumer and business to business sales service. Alibaba collects hundreds of peta bytes of data and processing takes lot of time. That is when Alibaba started using Apache Spark library, especially GraphX features of Spark. The everyday interaction of customers is being expressed as complex, large scale graphs which needed distributed data processing engine and interactive algorithms. Apache provides its GraphX mining algorithms which are also reusable module. These have support for PageRank, Triangle Counting, and Shortest Path. Usage of these helped Alibaba to establish many relationship between user and their items in the stores. Based on these relationship, Alibaba is taking good and effective decision on their business [15].

B. Apache Spark in Media & Entertainment Industry

As the invention of new hand held gadgets has led to change is generation using these handheld devices for their day to day needs, not back seat at media as well. As younger generation is pond of playing games using these devices, gaming industries have regularly increased their target sectors. So there was need to do analytics in order to give the best related service to the users. In gaming industry Apache Spark helped in coming up with business opportunities like targeted advertisement, different complexity levels in games and player retention. In media, the companies could refresh their news sections based on user's activities and last history of videos.

Yahoo being one among the search engines has main usage by users in News section.to cater different types of users Yahoo used capabilities of Spark for personalizing the webpage sections and advertisement section based on user. Upon using Machine learning capabilities of Spark, Yahoo was able to find out quickly what categories of news are being read by different category of users. According to their development code line statistics, initially they had used C++ coding which used 15000 lines and with Spark it used just 120 lines of Scala code [14].

Conviva, a private company that offers solutions for online video analytics and video optimization. Conviva ecosystem provides the analytics that helps in providing the best quality of experience to the consumer who is interested in video [16]. Using Apache Spark, Conviva was able to remove the screen buffering upon learning the real-time network conditions and speed, to deliver quality of service to its customer. The analysed information is stored in the video player to manage the live video traffic resulting from 4 billion video feeds every month in order to give the maximum pay-through there by providing its customer with a continuous video viewing experience.

Netflix, world's largest streaming service spread across 250 countries having 80+ million user base. Netflix used Apache Spark as distributed framework and implemented different algorithms that helped them to give good service to consumers.

Apache Spark was used as the engine of their Recommendation pipeline. Spark enabled Netflix to use a single unified framework for feature generation, model training and validation process. Spark Machine libraries are being used by Netflix extensively [17].

V. CONCLUSIONS

The review study on Apache Spark has enabled to conclude that, using Apache Spark, processing of the data can be done much faster than the Hadoop MapReduce framework. Further, many of organizations have already been benefited in their business. Usage of Apache spark has enabled companies to come up with faster and reliable quick solutions that help them in processing their vast amount of data to get good value addition insight from the data. Very technology rich libraries of the Apache Spark has provided companies with capabilities to implement faster solutions to address different business improvement needs across Retail, Healthcare, Media, Gaming Industry, Defense sector, E-commerce industries.

REFERENCES

[1] Apache Spark website. [Online]. Available :https://spark.apache.org/

[2] Matt Hawkins(2015), "Customer Analytics".[Online].Available: https://www.datameer.com/product/industries-use-cases/customer-analytics/

[3] Doug Adamson (2016), "Big Data in Healthcare Made Simple: Where It Stands Today and Where It's Going". [Online]. Available : https://www.healthcatalyst.com/big-data-in-healthcare-made-simple

[4] Andrea Eichhorn and Kim Minor(July 2013)," Augmenting Your Claims Fraud System with Big Data Capabilities".[Online].Available : http://www.ibmbigdatahub.com/blog/augmenting-your-claims-fraud-system-big-data-capabilities

[5] Palshikar G.K. (2014) Detecting Frauds and Money Laundering: A Tutorial. In: Srinivasa S., Mehta S. (eds) Big Data Analytics. BDA 2014. Lecture Notes in Computer Science, vol 8883. Springer, Cham

[6] Rikita Patil and Nevil Shah ,"Bigdata analytics applications in government sector of India", International Journal of Scientific Research and Engineering(IJSRE),Vol 1,Jan 2017,Page(s):50-54

[7] MapReduce Tutorial(2017,Sep 17):Apache Hadoop 3.0.0 [Online].Avaialble:https://hadoop.apache.org/docs/current/hadoop-map reduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

[8] HDFS Architecture Guide (2017, April 8): Hadoop 1.2.1 Documentation.[Online].Available: https://hadoop.apache.Org/docs/r 1.2.1 /hdfs_design.html

[9] https://en.wikipedia.org/wiki/Apache_Spark

[10] Directed Acyclic Graph in Apache Spark(2017 Apr 8) [Online]:Avaialbel:http://data-flair.training/blogs/dag-in-apache-spark/

[11] Apache Spark Website. [Online]. Available: https://spark.apache. org/ sql/

[12] Apache Spark Website. [Online]. Available: https: //spark.apache . org/ streaming/

[13] Apache Spark Applications [Online].Available: https://intellipaat.com /tutorial/spark-tutorial/apache-spark-applications/

[14] Top 5 Apache Spark Use Cases(2016,Jun 16).[Online].Available: https://www.dezyre.com/article/top-5-apache-spark-use-cases/271

[15] Andy Huang and Wei Wu(2014,Aug 14)[Online].Available: https://databricks.com/blog/2014/08/14/mining-graph-data-with-spark-at-alibaba-taobao.html

[16] About Conviva on Website[Online].Available:https://www .conviva. com/about/

[17] DB Tsai (2017,Feb 8) ,"Netflix's recommendation ml pipeline using apache spark".[Online].Available: https://spark-summit.org/east-2017/events/netflixs-recommendation-ml-pipeline-using-apache-spark/