RESEARCH ARTICLE OPEN ACCESS

Review: Apache Spark and Big Data Analytics for Solving Real World Problems

Mrs. Dipali Dayanand Ghatge
Department of Computer Science and Engineering
Karmaveer Bhaurao Patil College of Engineering, Satara- 415001
Maharashtra - India

ABSTRACT

Big Data analysis is having an impact on every industry today. Industry leaders are capitalizing on these new business insights to drive competitive advantage. Apache Hadoop is the most common Big Data Framework, but the technology in evolving rapidly and to cope with that the latest innovation is Apache Spark. This paper discusses the basics of Apache spark and some real world use cases and applications for Big Data analytics with Apache Spark.

Keywords:- Big Data Analytics, Apache Spark, In Memory Computation

I. INTRODUCTION

In today's world of computers human life is very much dependent upon the technology. Our personal, professional and social life is fully surrounded by technology. By some or the other means we are dealing with some kind of data. This data is originated through mobile phones, computers, laptops, cameras and many other electronic gadgets. Due to immense growth of data the challenges of data management arise. Data Management deals with not only storing the data but it also involves the accessing, analyzing and securing it. Big Data analysis is involves the collection of the data from different sources, organizing it so that the accession and the analysis will become easier. This analysis helps us to dig out the hidden facts and information from the huge data collection. Analysis is found useful for categorizing and ranking the data as per its importance with respect to the application.

This paper focuses on open source tool Apache Spark. It is the best alternative for faster big data analysis. Spark supports in-memory computing, which is faster than disk based engine like Hadoop. This paper is organized as follows: section II explains concept of big data analytics and focuses on the key barriers to the Big Data Analytics. Section III explores the basics of Apache Spark, section IV will discuss some case studies of Big Data analytics using Apache Spark and finally the conclusion is stated.

II. BIG DATA ANALYTICS

Data is our most valuable resource. Organizations use this data for enhancing situational awareness among people forecasting market dynamics in financial services, for early detection analysis in health care. There is some desired value is obtained from the vast amount of data called as Big Data by the government organizations and private firms[1].

<u>Some Big Data Facts-</u>

1. 2.5 Quintillion bytes data is created every day.
2. 90% of the world data is created in last two years
3. 80% of the world’s data is unstructured.
4. Facebook processes 500 TB per day.
5. 72 hours of videos are uploaded to youtube every minute.

This data must be analyzed to gain the insights and to act on complex issues this is what big data analytics is. Big Data analytics is the process of collecting, organizing and analyzing the large sets of data i.e Big Data to discover the patterns and other useful information.

The big data analytics can be categorized into following categories:

1) Descriptive analytics: what happened?
2) Diagnostics analytics: What did it happen?
3) Predictive analytics: What is likely to happen?
4) Prescriptive analytics: What should I do with it?

What makes the Big Data analytics critical?

1) Data volume is very large and it is steadily growing.
2) Data volume has variety of data ie structured, unstructured etc.
3) Volume and speed of the data creates challenge for architectural management and analytics services.

Organizations are using new big data technologies and solutions such as Hadoop, MapReduce, Hadoop Hive, Spark, Presto, Yarn, Pig, NoSQL databases and more to support their big data requirements.

III. APACHE SPARK

Apache Spark is a data analytics, cluster computing framework. It is an open source technology originally developed in AMPLab UC Berkely. Sprks fits into Hadoop open source community. It builds on the top of the hadoop distributed file system(HDFS). Spark is not tied to the two stage map reduce paradigm. It promises the performance 100 times faster than Hadoop Mapreduce. Apache Spark provides the primitives for in memory cluster computations. In memory cluster computing allows user programs to load data into a cluster's memory so that they could be queried repeatedly. This makes Spark well suited for machine learning algorithms. Spark became the Apache top level project in February 2014. The companies like Yahoo, Intel contributed in development of Apache Spark. Apache Spark is written in Scala, Java and Python languages. It is supported by Linux, MAC OS and Windows [5].

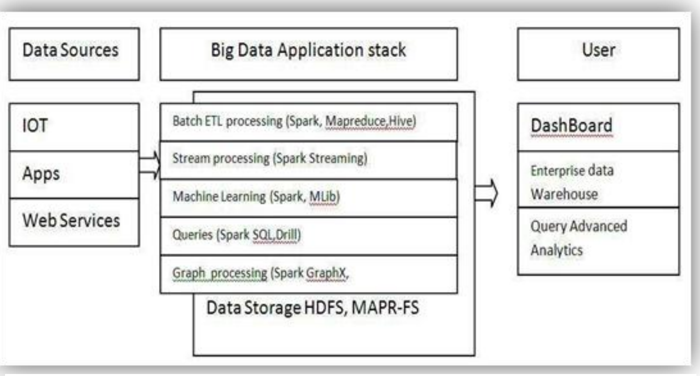
Data Sources
Big Data Application stack
User
IOT
Apps
Web Services
Batch ETL processing (Spark, Mapreduce,Hive)
Stream processing (Spark Streaming)
Machine Learning (Spark, MLib)
Queries (Spark SQL,Drill)
Graph processing (Spark GraphX,
Data Storage HDFS, MAPR-FS
DashBoard
Enterprise data Warehouse
Query Advanced Analytics

Figure 1 Apache Spark in Big Data Application Stack

Besides In memory computation Apache Spark extends Map Reduce with support for more components like streaming and interactive analysis. You can continue to use your existing big data investments.Spark is fully compatible with Hadoop.It can run in YARN, and access data from sources including HDFS, MapR-FS, HBase and HIVE. In addition, spark can also use the more general resource manager Mesos.

How does spark fit into big data application stack?

On the left we see different data sources. There are multiple ways of getting data in using different industry standards such as NFS or existing Hadoop tools. The stack in the middle represents various Big Data processing workflows and tools that are commonly used. You may have just one of these workflows in your application, or a combination of many. Any of these workflows could read/write to or from the storage layer. As you can see here, with Spark, you have a unified stack. You can use Spark for any of the workflows. The output can then be used to create real-time dashboards and alerting systems for querying and advanced analytics; and loading into an enterprise data warehouse.

IV. CASE STUDIES OF THE BIG DATA ANALYTICS USING APACHE SPARK.

a) Mapping brain activity at scale with cluster computing.

Large network of neurons are to be monitored during behavior for understanding brain function requires Due to advances in recording technology the size and complexity of neural data has been tremendously increased. Analyzing such a huge data is very critical. Scientists from HHMI Janelia research Campus present a library of analytical tools called thunder built on the open-source Apache spark platform for large-scale distributed computing. The library implements a variety of univariate and multivariate analyses with a modular, extendable structure well-suited to interactive exploration and analysis development. We demonstrate how these analyses find structure in large-scale neural data, including whole-brain light-sheet imaging data from fictively behaving larval zebrafish, and two-photon imaging data from behaving mouse. The analyses relate neuronal responses to sensory input and behavior, run in minutes or less and can be used on a private cluster or in the cloud. our open-source framework thus holds promise for turning brain activity mapping efforts into biological insights. Neural data pose unique challenges for analytics. The data are complex, and the 'right' analysis is rarely obvious. Every analysis provides a lens through which to see the data, and it is often necessary to try different analyses interactively, whether by varying parameter choices or developing entirely new algorithms [2].

The need for flexible analytics is especially crucial for large data sets; the more complex and heterogeneous the response properties and dynamics, the wider the variety of analyses needed to reveal their structure. Prototyping analyses for small data sets are straightforward on a workstation using existing tools, but for large data sets, especially those that exceed the memory of one machine, this becomes intractable. Large-scale neuroscience thus demands a flexible platform for creating analyses and inspecting results.

b) Real-time shopper engagement at the shelf:

Today's consumer is highly connected at all points of the shopper journey. This connectivity continues in the store, and getting that real-time dialogue with your customer when they are most eager to find relevant product information and recommendations bespoke to their needs, is a mobile moment that cannot be missed. Trax Smart Shopper is a mobile application that provides consumers with real-time information on products in the grocer's aisle. With Trax, manufacturers and retailers can add value, choice and differentiation to increase sales potential by empowering their consumers to shape their own shopping experience. Making the most of their time in the store, consumers can quickly and easily discover, filter, review and locate products on the shelf. And when favourite products are not available, shoppers are provided with alternative product recommendations instantly. Trax Smart Shopper is powered by Trax's proprietary image recognition technology, a platform developed from breakthrough computer-vision algorithms that were specifically designed for retail. Retailers and brands leverage Trax to deliver a superior level of customer service through unprecedented real-time engagement and understanding of their customers' path to purchase in-store [4].

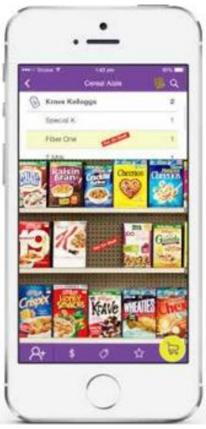

Figure 2 Trxa's Shelf Navigation Option

The shopper starts with their shopping list, moving between the aisles using Trax's shelf navigation option to locate their sought after items.

When the shopper comes across a particular item on their list and discovers that it's missing on shelf, the next best recommendation pops up as a suggestion. This recommendation is presented as a targeted offer, responding to their specific preferences and requirements, whilst maximizing the mobile moment with this customer.

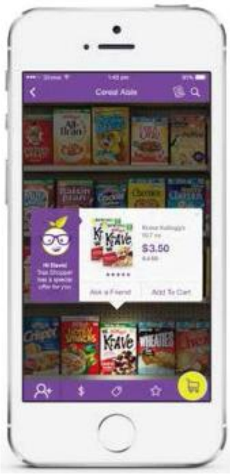

Figure 3 Recommandation pop up

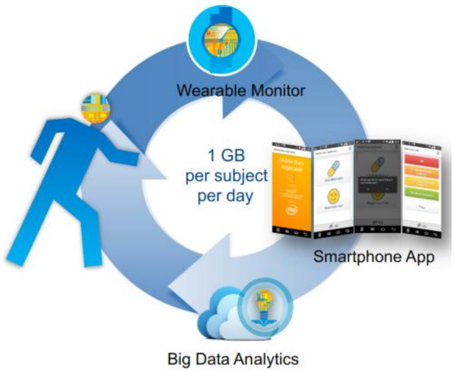

Figure 5 Wearable Monitor

c) **A wearable monitor**: Michel J Fox Foundation for Parinson's research in association with Intel has invented a wearable monitor to monitor the activities of different parts of the body. 1GB of data per patient per day is collected utilizing Spark and Hadoop combination. This data is analyzed to monitor the real time information regarding what is patient is doing? And what is the specific movement of different parts?. This wearable can monitor sleep patterns, way of walking, balance, vibration etc. On the basis of the observations done the anomaly detection becomes easier. And patient is 24 X 7 under observation. The tremendous data is collected which can be used for doctors to take medication action and the same can be utilized by the researchers for doing research and study.

d) **Real-Time Retail Shelf Occupancy Analysis**: This real time shelf occupancy analysis is an analytical system for malls and big stores, where the shelves are analyzed to know missing items, sold items and misplaced items.

1) Image of a shelf is uploaded.

Figure 6 Image Upload

2) Current image is sent to the server for Planogram comparison

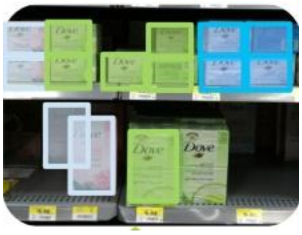

Figure 4 Planogram Comparision

3) Planogram comparison analyzes the self images and highlights the misplaced items, missing items etc.

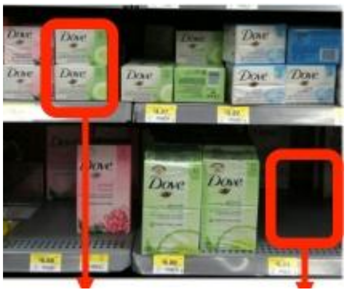

Figure 7 Analysis

4) The analysis is presented in the form of graphs for presenting clear insights regarding the shelf
5) The analysis also creates the alerts regarding missing items, out of stock items etc. using this information the stock verification is done in few seconds.
6) The analysis also provides the information regarding the sales of the particular item, which gives gives the idea to the shop owner about the customer's choice

V. CONCLUSION

Apache Spark framework helps with big data processing and analytics with its standard API to find the solutions to many real world problems. Spark makes it easy and inexpensive to combine different processing types, which is often necessary in production data analysis pipelines. In addition to that, it reduces the management burden of maintaining separate tools. Vast data and instant analysis can be achieved by using Apache Spark for big data analysis.

REFERENCES

[1] Abdul Ghaffar Shoro & Tariq Rahim Soomro, "Big Data Analysis: Ap Spark Perspective", Global Journal of Computer Science and Technology: Software & Data Engineering, Volume 15 Issue 1, 2015.

[2] Jeremy Freeman et.al, "mapping brain activity at scale with cluster computing", nature methods ADVANCE ONLINE PUBLICATION, doi:10.1038/nmeth.3041,7 july 2014.

[3] Sujee Maniyam, "Apache Spark fast and Easy data processing",Elephant Scale LLC,SNIA Analytics and Big Data Summit,2015.

[4] Trax Image Recognition,"Trax Smart Shopper",2016.

[5] Spark mlib, Apache Spark performance,https: //spark.apache.org/mllib/ ,Retrived Jan 2016.

[6] Big Data: what I is and why it mater, 2014,http://www.sas.com/en_us/insights/big-data/whatis-big-data.html.

[7] www.brighttalk.com