National Conference on Engineering Innovations and Solutions (NCEIS – 2018)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 6 | ISSN : 2456-3307

Study of Machine Learning Techniques using Apache Spark

Soumya Manjunath Hegde , Shilpa .M, Soujanya .C .S, Urvashi Grover
Eighth semester, Department of ISE, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India

ABSTRACT

The challenges in the field of big data analysis is growing due to the huge volume of data collected on daily basis by social media, weather forecast, mobile data etc. In this survey paper, there is a look on different aspects of usage of Apache spark, be it, the framework, the libraries, the spark technologies etc. The spark platform provides various algorithms to analyse machine learning techniques and implement them on other virtualization platforms such as VMware vSphere. Further, Spark is used on different platforms to achieve high performance, overcome latency and achieve efficiency. The papers, studied here, have drawn parallelism between the Hadoop and the Spark and the latter has proved to be the best platform as it is hundred times faster and more efficient.

Keywords: weather forecast, virtualization, Hadoop, Spark, latency

I. INTRODUCTION

Big data analytics is the biggest challenge from past few decades because of the huge amount of data that is generated every day. There are many open-source technologies which are used to handle massive data volumes. One such technology is Apache Spark. Apache Spark is a lightning-fast cluster computing technology, designed for fast computation. It is based on Hadoop Map Reduce and it extends the Map Reduce model to efficiently use it for more types of computations which include interactive queries and stream processing. Some of the most popular companies that are using Apache Spark are Uber, Pinterest, conviva,data etc. The basic components used in Apache Spark are Spark Streaming, Spark SQL,Spark MLlib,GraphX,prediction with logistic regression. Spark is used for some of the prominent applications such as machine learning, fog computing,interactive analysis,event detection etc. This section provides a brief introduction of big data and spark platform.

II. BASICS OF SPARK

Spark is referred to as distributed framework which is based on Hadoop MapReduce algorithms.In (1) Spark's features include Memory Computing which helps in storing the intermediate and the output results of spark jobs in the memory which is an advantage over Hadoop MapReduce. Memory Computing improves the efficiency of computing. So, the Spark can be used for iterative applications such as machine learning and data mining. Spark provides Resilient distributed dataset which provides rich set of operations to manipulate the data. The API in Spark is available in JAVA, Python, R and Scala languages. The processing speed of the spark is said to be 100x faster than Hadoop MapReduce.

A. Framework

The framework method for the management and analysis of qualitative data has been used since the 1980s.The method originated in large-scale social

policy research but is becoming an increasingly popular approach in medical and health research.In this paper(2) we are using the popular concept called cascade learning for the advanced machine learning architecture of multilayer perception(MLP) and distributive computational abilities of apache sparks combined together in a framework. Spark is mainly used to handle the large magnitude of data efficiently. Framework and artificial intelligence are used to solve the real world problems usingBig data analytics. The problem in real world data is time and space constraints. To overcome these challenges, themachine learning, cascading, big data analysis and deep learning combined ideas have been used.To solve traditional machine learning the novel framework is used. Using the novel framework, traditional machine learning tools will improve the accuracy and speed of the system. The main reasons choosing the novel frame work is to enhance future set, compute time, continue learning and improve.

B. Spark MLlib

The changing and improving trends in big data analysis is by far a big concern in the field of machine learning which is why the big data machine learning platforms such as Apache Spark MLlib have been developed. In paper(3), there is some light cast on the libraries of Apache Spark.One of the major libraries of Apache Spark, Spark MLlib is the most prominent platform available for big data analysis to carry out various techniques. It consists of more than 55 algorithms that support data and process parallelization. It also provides APIs in different languages to evaluate machine learning methods. It surpassed the performance of Hadoop in terms of running time when the same algorithm was run on Weka library components used in Hadoop and spark MLlib on Apache Spark. Spark MLlib offers fast, flexible and scalable implementation of a variety of machine learning components. It offers options for distributed processing by parallel processing. It decreases the processing time required and, at the same time, increases time to interpret analytic results.

III. SPARK FRAMEWORK USED IN VARIOUS FIELDS.

Intrusion detection systems monitor network or systems for policy violation or malicious activity.In(4)the advantage of iterative algorithms on Spark is mentioned. For example,the intrusion detection algorithm is highly time consuming and occupies large amount of memory. In order to solve this problem, the usage of a parallel Principal Component Analysis (PCA) combined with support vector machine (SVM) algorithm based on Spark platform is proposed (SP-PCA-SVM). Principal component analysis is used for training and predicting the data and fusion of bagging integration strategy and SVM algorithm is used on the spark distributed framework. Spark platform is considered because it reduces the training time and improves model learning efficiency. In paper [2] Parallel SVM algorithm is used to effectively deal with large scale datasets. Parallel SVM is based on the iterative map reduce provided by the Spark environment. The applications like improving the efficiency of iterative algorithms using Spark platform highlights the parallel computing feature of Spark.

A. Spark Streaming

Tocope with streaming data, various stream-processing-based frameworks have been proposed, such as Storm, Flink and Spark Streaming. In (5) ,one of the spark machine learning libraries Spark Streaming is discussed to process online flow of data. The major task is to handle the explosive growth of internet traffic. The need for spark streaming came into picture when traditional network analysis methods were no longer suitable for processing huge traffic of data on single machine due to poor processing ability. The application, Spark streaming is used in internet traffic monitoring system. The system involves 3 components, the collector, managing systems and stream processor. The collector collects and stores the data packets, the stream processor processes the data collected in the

collector and the managing system behaves as abridge between the former two. The in- memory computing feature of spark uses RDD and spark streaming processes and analyses the data.

B. Enterprise Big Data

Big data is data that is too large to process using traditional methods. Enterprises have large amounts of data and this data has to be safe and secured.In this paper,(6)Enterprises give strong controls and strategies to prevent cyber-attacks and the data is not leaked. It is confidential. Employees and data scientists have access to analyse and derive insights from the data but there are insufficient controls and employees are usually permitted access to all information about the customers of the enterprise including sensitive and private information.

In this paper, author speaks about Shade. Shade is a system that allows a spark cluster which contains sensitive data which can be accessed in different manners.The framework Shade includes two mechanisms Spark LAP and Spark SAM.The Enterprises analyse the data to understand their users and know their behaviour and requirements so that better customization can be provided. Spark can be used for a wide variety of data analysis tasks such as statistical querying or machine learning.

C. Spark BDD

Spark-BDD is a pioneer platform which provides and allows programmers to exercise on BDD interactive debugging capabilities to set break points or trace through a program for execution.In this paper(7),Spark BDD commences as a support to debug analytic programs. Debugging toolkits on Spark provides an interactive query interface which focuses on bringing interactive capabilities to the spark platform. It supports all the features through a 3 key mechanism; 1)data lineage information 2)incremental dataflow computation 3)runtime level profiling. It also enables function hot swapping and replay. It is an ultimate application of using debugging in use case applications and features supported by different distributed debugger.

D. Spark-SIFT

Apache Spark is an open-source cluster-computing framework. In this paper, (8) author speaks about SIFT (Scale, Invariant, Feature, and Transform) image feature extraction algorithm is implemented in Spark- SIFT framework. Image processing has an important phase that is, feature extraction.

Spark is a memory based data processing framework with faster speed.The framework contains three part, the base interface of image processing,the sift algorithm in the spark, and the sequence of images. Many problems arise .One of them is load unbalance. This happens when size of images to deal have wide difference. In this paper, the solution to this problem is the segmentation of image feature extraction algorithm in spark. Feature extraction takes a long time in processing, especially in large-scale image retrieved systems. The feature of spark are running faster, spark owns DAG execution engine, which support the iterative calculation of data in the memory, scala is the program language supported in the spark. Scala is effective, extensible and can deal with a job in simple code.Good generality, spark BSDA includes spark core, spark SQL, Spark Streaming, MLlib and GraphX components.

E. Extreme Learning Machine Algorithms Based on Spark

The non-iterative ELM algorithm helps in generating weights of hidden layers and determines the output layer weights by analysing. In paper(9), the method discussed brings in convenience to many time sensitive applications by reducing learning time. The VMware vSphere virtualisation platform to analyse and manage architecture based services, application services, complex data centre etc. it brings in more flexibility, serviceability and effectiveness through virtualised and distributed basic architecture services; monitors the availability and accessibility of resources. The Feed Forward neural network parallel

algorithm is based on spark platform and the establishment of VMware vSphere ,here, helps to perform experiments as an experiment platform. Again, here, the spark highlights its advantages on using in memory processing and distributed processing based on spark. Spark serves as a cluster computing platform and supports task scheduling process at every phase by processing RDD objects in generating non-acyclic graphs. The new ELM,neural network algorithm has fats training speed, less artificial interference and strong data generalisation ability.

IV. REAL WORLD APPLICATION OF APACHE SPARK

A. Weather Data Analysis

Weather data is used to predict the atmospheric changes. The real time data is analysed(10)In this paper, weather is of most concern. Weather forecasting is a challenge in human civilization. There are many methods and algorithms that have been developed to predict weather forecasting. Big data is the key concept used to manage large amounts of data. Hadoop is a platform designed to run in situations where analytics are used that are deep, extensive like clustering. Real time analytics with spark streaming is designed to analyse the real time data. There is a robust and an efficient technique for analysing the weather data set using spark. The weather data is collected from sensors and power stations. Spark overcomes the drawbacks of Hadoop in terms of processing speed.

B. Agricultural Information System

Agricultural information is vast and the data collected here is in big amounts. Big data technology plays an important role in spatial analytics by which decision making is enhanced. In this paper(11),Agricultural domain consumes huge volume of data. In this paper, the author proposes a spark based information management system for agriculture. Big data analytics supports the development and delivery of agricultural information and services to make farming economical and sustainable. Spatial data is very important in agricultural domain. This data is important in agricultural domain. This data is important to develop flexible and includes all types of function. This paper deals with spark based agricultural information system on big data by developing analytical and visualization services.

C. Target Prediction in Drug Discovery

Initially,the machine learning predictors programs was written in C and C++. These programs would take too long to run in parallel because we do not use the multiple nodes.In(12)it is prediction of drug discovery we uses apache spark to enable existing program single node into the multiple node cluster pipeline , using apache spark we can speed up to 8 nodes in a system.Here spark is mainly uses to evaluate the intermediate storage into various forms. Apache spark has two categories one is runtime system to schedule work units on a cluster pipeline in the form of graph. Second is creating dependency graph using programming model.In programming model we mainly concentrated on resilient distributed data (RDD). Spark has control on programming over intermediate RRDs storage, using different combination of RDD. Spark programs will express more algorithms. Because of all these RDD is split up and created one task per partition.

D. Study on forecast of shared Bicycle

Now-a-days, shared bicycle projects have developed continuously, the problem is that of storing the vast amount of information about the usage of bicycles. In(13),they are using spark MLlib for shared bicycle to form three different prediction models. The three different prediction models are multiple linear regression, decision tree, random forests. In shared bicycle proposal we have enormous data, to segregate them using spark machine learning framework. Apache Spark is fastest and used to handle huge data processing. Data processing and MySQL databases are used to store the information in the form of data tables. Data processing will retrieve the data set and make website, each action of data is stored in a CSV

file using available information and then information is stored in some format. SQL processing we consider CSV file, if CSV file generated more than 36 million GB data then we uses spark SQL to manage data processing. To read the original CSV file uses spark streaming.

E. Road Traffic Event Detection

In real time,twitter has become a very famous and a trending social network. Twitter is a powerful source of information used to detect the traffic in a particular area because it has the information about real time event happening in the surrounding. People tweet on whatever they see or feel like in their day to day life. In(14) ,they have considered the real world application to Detection of the road traffic using spark based on the twitter datasets. In twitter per year 200billion tweets means 6000 tweets are generated per seconds using classification techniques to assigning class labels to the systems. Based on the tweet data the system will fetch the information tweets related to the traffic. We invoke the logistic regression and support vector machine (SVM) classifier for classification of dataset in the tweets.Other than above techniques we undergo some more techniques to extract the useful information from the tweet dataset. Techniques are: statistics, natural language processing and machine learning. Spark is mainly used for scalability of data in tweet. Using spark we can execute many analysis and pattern classification techniques. But SVM supports only for binary classification where logistic supports both binary and multiclass classification.

F. Mobile Big Data

Mobile big data is a concept that describes a massive amount of mobile data that cannot be processed using a single machine. In this paper (15),MBD analytics is currently a high focus topic aimed at extracting meaningful information and patterns from mobile data. Deep learning is a solid tool in MBD analytics. The framework used in this is Apache Spark,which provides an open source cluster computing platform . This enables distributed learning using many computing cores on a cluster where continuously accessed data is cached to running memory, thus speeding up the learning of deep models several fold. The learning time of deep models is decreased as a result of the parallel Spark based implementation. Paper contains the challenges of MBD such as 1) Large scale and high speed mobile networks 2) portability 3) crowdsourcing. The definition of deep learning is mentioned like this, Deep Learning is a new branch of machine learning that can solve broad set of complex problems in MBD analytics. The advantages of DeepLearning in MBD analytics are mentioned. 1) Deep learning scores highly accurate results which are a top priority for growing mobile systems. 2) Deep learning generates intrinsic feature that are required in MBD analytics. 3) Deep learning can learn from unlabelled mobile data, which minimizes the data labelling effort. The authors then mentioned the importance of spark in deep learning models of MBD analytics. The parallelization using spark of deep model is performed by slicing the MBDinto many partitions. Each partition is contained in a resilient distributed dataset that provides an abstraction for data distribution from the spark engine. The author mainly concentrates on Spark platform because it tackles the problem of volume, velocity, and volatility aspects of MBD. Volume aspect by parallelizing the learning task into many tasks, Velocity by its streaming extensions, Volatility aspect is addressed by significantly speeding up the training of deep models. Author implements a deep learning model by considering the mobile activity dataset using spark environment.

V. SURVEY PAPER

Survey on high performance analytics of big data with Apache Spark

The main components of the spark which is also called ecosystem of the spark are presented in detail

in (16). To work with structured data spark contains Spark SQL package. This enables users to query using SQL. Spark does not provide normal SQL interface, instead of that Spark SQL allows programmers to merge different SQL queries with programmatic manipulations that are supported by RDDs in Scala, Python, R and Java. Machine learning functions like regression,collaborative filtering ,classification and clustering are provided by spark with the help of MLlib package. Data Analysis or Machine learning techniques on data can be applied effectively using this package. Spark Streaming component of the spark permits the processing of live streams of data. To perform parallel computations and manipulate graphs Spark provides library called GraphX. Spark core includes different components for memory management, interacting with storage system, fault recovery, task scheduling and many more. Apart from these Spark also includes HDFS, web interface, parallel library etc.

VI. CONCLUSION

Spark being one of the best open-source platforms from data cleansing to any data mining technique. It is a cluster computing framework which works upon fault tolerance and data parallelism. It uses in-memory processing by which it overtakes Hadoop and other memory management issues such as serialization etc. RDD is a fundamental data structure of spark which is a distributed collection of objects which process data in parallel. It results in faster and efficient processing of data. It uses various libraries to handle to different data sets and techniques by implementation of suitable algorithms such as SIFT algorithms, intrusion detection algorithm etc.Hence, spark is a widely used platform for big data analysis for continuously growing and changing trends in technology

VII. REFERENCES

[1] SPARK—A Big Data Processing Platform for Machine Learning. Jian Fu, Junwei Sun,Kaiyuan Wang. Wuhan, Hubei, China : IEEE, 2016.

[2] A Big Data Analysis Framework Using Apache Spark and Deep Learnig. Anand Gupta, Hardeo Kumar Thakur. Delhi, India : s.n., 2017.

[3] Big Data Machine Learning using Apache Spark MLlib. Mehdi Assefi, Ehsun Behravesh, Guangchi Liu, Ahmad P. Tafti. USA : s.n., 2017.

[4] Research of Intrusion Detection Algorithm Based on Parallel SVM on Spark. Hongbing Wang, Youan Xiao and Yihong Long. Wuhan, Hubei Province, China : s.n.

[5] Online Internet Traffic Monitoring System Using Spark Streaming. 2018.

[6] Shade: A Differentially-Private Wrapper For Enterprise Big Data. Alexander Heifetz, Vaikkunth Mugunthan and Lalana Kagal. Cambridge, USA : s.n., 2017.

[7] Spark-BDD: Debugging Big Data Applications. Tyson Condie, Muhammad Ali Gulzar, Matteo Interlandi, Miryung Kim,Todd Millstein Sai, Deep Tetali, Seunghyun Yoo. California, Los Angeles : s.n.

[8] Spark-SIFT: A Spark-Based Large-Scale Image Feature Extract System. xinming Zhan, YaoHua Yang, Li Shen. china : s.n., 2017.

[9] Parallelization of a Series of Extreme Learning Machine Algorithms Based on Spark. Tiantian Liu, Zhiyi Fang, Chen Zhao, Yingmin Zhou. china : s.n.

[10] Weather data analysis using Spark – An In-memory Computing framework. Ms.D.Jayanthi, Dr.G.Sumathi. INDIA : s.n., 2017.

[11] Towards Development of Spark Based Agricultural Information System including Geo-Spatial Data. Purnima Shah, Deepak Hiremath,Sanjay Chaudhary. Ahmedabad, India : s.n., 2017.

[12] Scaling Machine Learning for Target Prediction in Drug Discovery using Apache Spark. Dries Harnie, Alexander E Vapirev, Jorg Kurt Wegner, Andery Gedich, Marvin

Steijaert,Roel Wuyts and Wolfgang De Meuter. Belgium : s.n., 2015.

[13] Research on the forecast of Shared Bicycle rental demand based on spark machine learning framework. Zilu Kng, Yuting Zuo, Zhivin Huang, Feng Zhou, Penghui Chen. china : s.n., 2017.

[14] Real Time Road Traffic Event Detection using Twitter and Spark. Ketan R. Pandhare, Medha A Shah. India : s.n., 2017.

[15] Mobile Big Data Analytics Using Deep Learning and Apache Spark. Mohammad Abu Alsheikh, Dusit Niyato, Shaowei Lin, Hwee-Pink Tan, and Zhu Han. 2016.

[16] Survey on High Performance Analytics of Bigdata with Apache Spark. Ramkrushna C. Maheshwar, D. Haritha. India : s.n., 2016.