International Journal of Future Generation Communication and Networking
Vol. 13, No. 1, (2020), pp.549 - 558

Spark Framework for Streaming and Generating Predictive Business Intelligence

Madadi Vijayakamal[1], D. Vasumati[2]

[1]*Research Scholar, Dept. of CSE, JNTUniveristy Hyderabad, kamalmvcse@gmail.com*
[2]*Professor Dept of CSE, JNTUCEH, JNTU Hyderabad, roshan44@gmail.com*

Abstract

Apache Spark is one of the stream processing frameworks that can be associated with cloud computing. Real time streaming data is processed with machine learning and natural language processing. Apache Spark is used to explore process mining as well. Process mining is for discovering business processes and diagnose the difference between real processes and processes discovered from event logs. This kind of business intelligence can help improve business processes in real world applications. In presence of very huge number of business processes, usage Spark provides scalability and performance in real time. The data given as input is divided into batches with a time window for processing. In the process, the framework discovers business intelligence as per the algorithms defined. An empirical study made with Spark and its performance is compared with that of Apache Flink. The empirical results revealed that the performance of Apache Spark is better than that of Flink.

Keywords: *Business process, event logs, Apache Spark, process mining, business intelligence*

1. Introduction

As cloud and its distributed programming frameworks like Apache Spark emerged, large volumes of data are stored and cloud and processed. With respect to business process event longs also this is true. Due to this, enterprises are not willing to lose data anymore. They wanted to save every detail of data and get it processes for comprehensive business intelligence. Hadoop is one of the frameworks which is used for processing large volumes of data. It supports MapReduce programming approach to handle large volumes of data with its features parallel process as explored in [8]. When compared with Spark, Hadoop takes more time for execution of algorithms used to discover business processes from event logs in the real time applications. Therefore, Spark is better candidate framework for processing streaming data. In fact, it works on live streaming data.

The input data is divided into number of batches and scheduled for processing. After processing, the result is sent to a distributed file system. As the data is processed in a time window, the result is appended to the output generated. With its ability to processing streaming data, Spark performs better than its counterparts like Hadoop. With its abilities for queries, it can analyze data live and provide business intelligence as needed.

ISSN: 2005-4238 IJFGCN

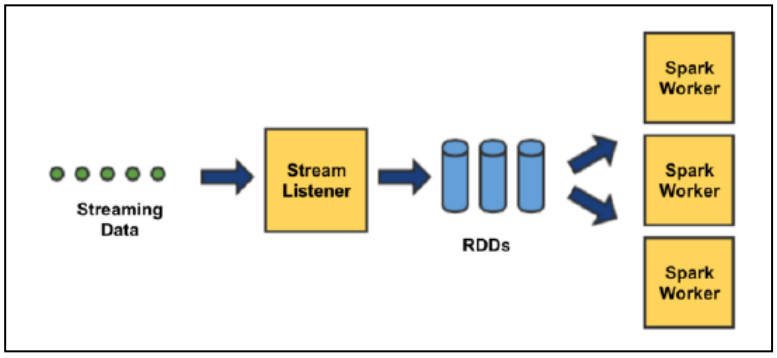

Figure 1: Architectural overview of Spark

Spark exhibits resiliency in distributed data processing or streaming processing. It takes streaming data input. There is stream listener implemented that works as per the availability of streaming data. Then the RDDs come into picture before allocating streaming processing to different worker nodes. Spark is well known for its processing of huge amount of data in batches in order to acquire business intelligence. In distributed environments, it is also used to have business intelligence from the processes that are discovered from event logs. In other words, it can be used for process mining as well.

2. Literature Survey

This section provides review of literature on mining of huge amount of data in the real world by using distributed programming frameworks like Spark. Gutfreund [1] proposed a methodology for test mining and NLLP using Apache Spark. The framework is used to work with real time data that is unstructured in nature. The processing results in patterns that can be used to recognise information related to customer relationships management (CRM). Aslst [2] on the other hand concentrated on process mining that is used analyse event logs and discover processes from it. Conformance checking is technique used to discover processes with certain validations and make a model. According to Ramkrushna *et al*. [3], Spark is widely used open source framework to work with large volumes of data. The data of enterprises in the real time is processed and patterns are recognized. This will result into output that can be used to understand the business intelligence and make well informed decisions.

Caya *et al*. [4] proposed a conceptual framework that exploits business intelligence that arrives from data analytics. The results are obtained from sports data and the patterns are identified by stream processing using Spark. Wani *et al* [5] focused on acquiring business intelligence from the big data. It is made by processing large volumes of data using open source frameworks like Hadoop. They also used Apache Spark and concluded that Apache Spark can process streaming data and provide required business intelligence more efficiently. There are many contributions found in [7]- [21] from which the importance of mining processes is found. However, there is need for empirical study to gain business intelligence from such data using Spark.

3. Existing System

Process mining is the phenomenon in which event logs are analyzed to discover business processes. The data is taken as input from a client with web based interface. There is interaction between client and server for analyzing the patterns in the data. Then there is extraction of process models from the underlying data. The outcomes are also analyzed with different metrics. There are changes in the process models and there are important observations as provided below.

1. There is maximization of the web site reaching to final state.

ISSN: 2005-4238 IJFGCN

2. There is increase in the performance of the process models with each visit.

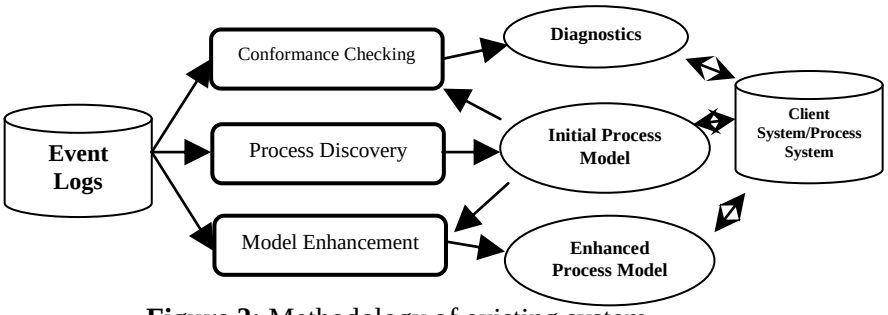

Figure 2: Methodology of existing system

As presented in Figure 2, the process mining with event logs is divided into main operations known as process discovery, model enhancement and conformance checking. There are other operations involved known as diagnosis, initial process model and improved process model.

3.1 Process discovery

This activity is started with the input in the form of event log. There is automated learning from the entries of event logs. The events that are related to particular process are ordered. The process models extracted from event logs are used for identifying problems in the business processes and provide business intelligence to stakeholders. When actual processes are analyzed, the problems in them are found and understood. The process models discovered also provide some sort of template that can be used later for further processing.

3.2 Conformance Checking

Conformance checking the process of finding whether discovered processes conform to actual business processed. In other words, it finds the exact difference between actual business processed in the real world and the business processes discovered from event logs. It also checks quality of processes involved. It finds cases that are not in conformance with expected behavior of business processes. It also finds some part of the processes that exhibit higher deviations. It finds the quality of process models that are discovered. In [21], genetic algorithm is used to find models with conformance checking. It will also lead to another activity known as model enhancement. There are many reasons for this activity as it can improve quality of processes.

3.3 Model Enhancement

After discovering a model and conformance checking of it, it is then possible to enhance process model. Log entries can be used in order to improve processes discovered. It is used to enhance models that can be used to diagnose any further issues. Every event log entropy has a case number, resources associated with it and timestamps. It maintains elapsed time between events in the process. Calculation of bottlenecks is made with the time difference that occurs pertaining to event logs as explored in [11]. The model enhancement is made with different algorithms that are used to determine changes to be made in order to enhance. Process mining in the existing system has limitations in predictive business intelligence. To overcome this drawback, the proposed system is used as discussed in Section 4.

ISSN: 2005-4238 IJFGCN

4. Proposed System

In the proposed system, Spark is used for process mining improvement. Spark is one of the famous open source distributed computing frameworks that are available in the real world. It can get streaming data from different sources and the data is kept in RAM. The process engine is used in order to evaluate it [9]. Data is maintained in the form of in-memory RDDs that are resilient structures associated with dataset. Hadoop is also used to storage and process large volumes of data. However, it takes more time process execution of jobs. Spark on other hand is much faster than its counterpart known as Hadoop. There are many abstractions in the Spark. It has libraries for different real world activities. Spark supports machine learning approaches and its SQL and ETL operations. Spark is good candidate for processing data in distributed environment. In fact, Spark can access large volumes of data from distributed environments. Its operations are done on RDD and other abstractions possible by considering RDD's. The operations are performed in the form of SQL kind of commands.

```
>>>rows = sqlContext.sql('select * from t_table').take(2)
>>> for x in rows:
print x
Row(Message=u'onsider the ethical plagiarism [Reference: 141031-003226] We are
escalating …
Row(Message=u'onsider the ethical plagiarism [Reference: 141031-003226] Incident
No: 141031-003226 Mail forwarded from …
Command took 5.03s
```

Listing 1: Spark allocating data multiple machines

Spark is used to deal with large volumes of data in the real world applications. There are datasets known as Spark resilient distributed dataset that is used for processing. It does not allow an RDD to be added to the one which is already there. When data arrives afresh in the form of streams, Spark has mechanism to handle efficiently. Both the static and dynamic portions of data it can handle and it can work with both continuous and discretized data as explored in [10].

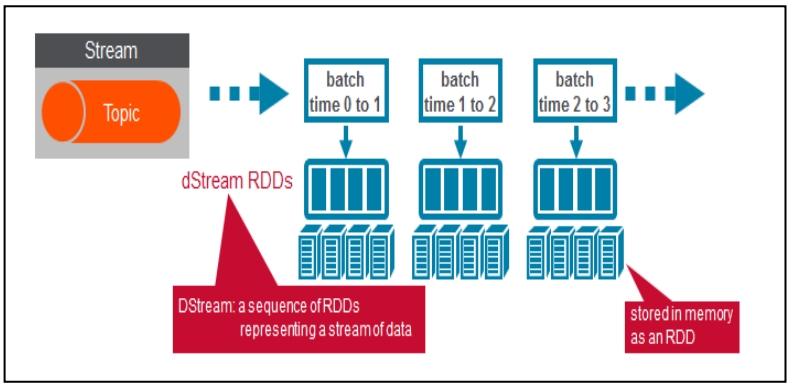

Figure 3: Processing Stream which is sequence of RDDs

As presented in Figure 3, Spark can process data in batches. The data streams may be in the form of Dstream RDDs. Many in-memory RDDs are used in order to get processed.

ISSN: 2005-4238 IJFGCN

```
sc = StreamingContext(sc, 60)
rec= sc.socketTextStream(host,port)
fRDD = rec.map(lambda x: x.split('\t'))
textRDD = fRDD.map(lambda x: x[10])
wRDD=tRDD.flatMap(lambda x: x.split(' '))
pair = wRDD.map(lambda x: (x,1))
wordcount = pair.reduceByKey(lambda x,y: x+y)
wordcount.print()
```

Listing 2: An expert from the code

As presented in Listing 2, an excerpt of Spark streaming processing code is obtained. It is used to obtain streaming context before actually establishing the streaming session and processing large volumes of data.

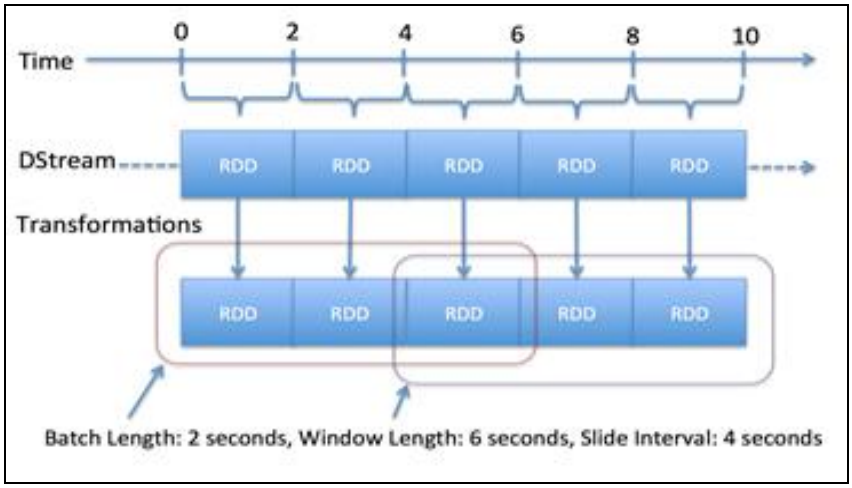

Figure 4: Transformations of Stream

As can be seen in Figure 4, it is clear the CRM data is being processed in large scale. The data is in the form of delimited records. The data is taken and the operations are made on the data in the time window. The processing done faster. Streaming data is processed with as much speed as possible. It can discover business intelligence from process mining that is used to make well informed decisions.

4.1 Natural Language Processing

There is involvement of NLP based on the algorithms employed with Spark distributed programing framework. The language processing is carried out by using certain operations. There are different processes involved in the text based analysis. There are operations like translations, concordance used to extract phrases and sentences and so on. There are key phrases that may precede or follow. There is possible usage of unigram and bigram as needed. It can also use trigram as explored in [11]. There is the usage of n-grams in a multiword context as well.

ISSN: 2005-4238 IJFGCN

```
import ree
def makeNGrams(n,text):
splitwords = text.split(" ")
filteredwords=[]
for i in splitwords:
if ree.match("[a-zA-Z]+", i):
filteredwords.append(i)
splitwords = filteredwords
ngrams = []
for j in xrange(len(wrd) -(n-1)):
gram = ''
for k in xrange(n):
gram += words[j+k] + ' '
ngrams.append(gram.strip())
return ngrams
```

Listing 3: NLP processing code

As shown in Listing 3, there is code for NLP activities. It has procedures like splitting documents into sentences, phrases and words. There is concept of unigram, bigram and trigram in order to have efficient processing of streaming data in Spark.

4.2 Output of Concordancer

With empirical study in Spark, it is understood that there is processing speed in Spark when compared with Hadoop. The output of the concordancer is provided as follows.

ID	Output
141002-001480	that you are having trouble with your purchase. In order to
140929-007053	scussed I am having trouble with advanced author searches
140929-003433	hear you are having trouble with access to Science Direct.

Figure 5: The table containing ID and output

As presented in Figure 5, the table contains ID and output. The textual data is broken down into n-grams. The non letters are got rid of by splitting word. Then the n-grams are transformed into words that are space separated. The application returns the list of n-grams.

5. Experimental Results

Apache Spark and Apache Flink are the two framework for streaming data processing. The CRM case study related process log events are taken live for processing that includes process discovery and identification of inconsistencies. The experimental results are provided in this section.

ISSN: 2005-4238 IJFGCN

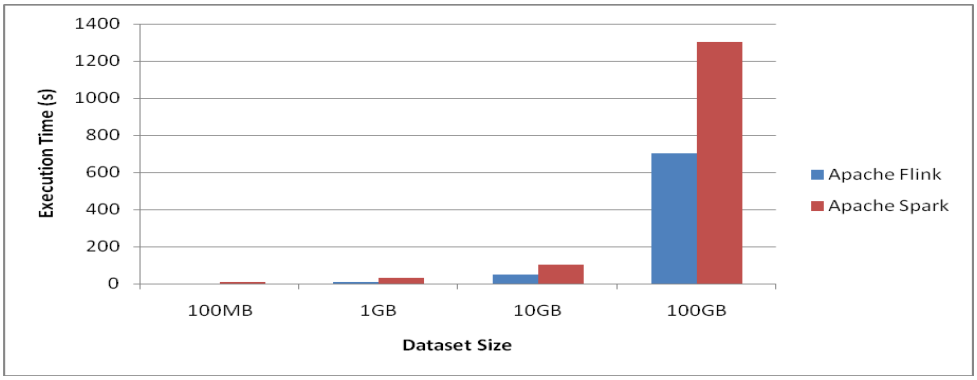

Figure 6: The time taken for processing data of different size

The execution time is compared as shown in Figure 6. Different size of input data is considered in horizontal axis while the time taken is shown in the vertical axis. The results showed that the influence of the size of data is apparent and at the same time the Apache Spark has shown better performance over Apache Flink.

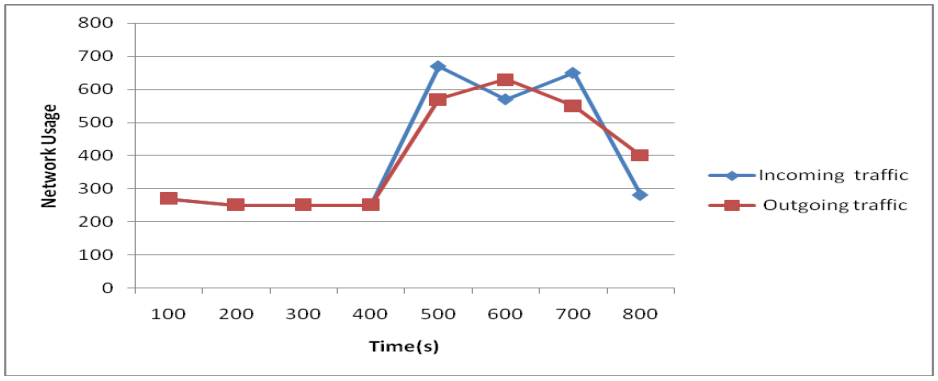

Figure 7: Network usage dynamics of Apache Flink

As presented in Figure 7, the network usage dynamics is provided. The elapsed time is taken in the horizontal axis while the network usage in terms of incoming and outgoing traffic is shown in vertical axis. As time elapsed is increased, the network usage is also increased and the later on decreased. The results showed that there is difference between the incoming traffic and outgoing traffic.

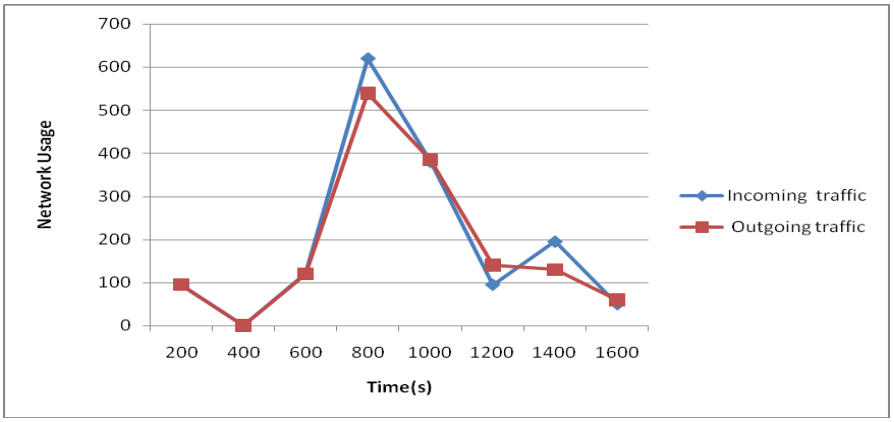

Figure 8: Network usage dynamics of Apache Spark

As presented in Figure 8, the network usage dynamics is provided for Apache Spark. The elapsed time is taken in the horizontal axis while the network usage in terms of incoming

ISSN: 2005-4238 IJFGCN

and outgoing traffic is shown in vertical axis. As time elapsed is increased, the network usage is also increased and the later on decreased. The results showed that there is difference between the incoming traffic and outgoing traffic. There is difference between the results of both the frameworks as well.

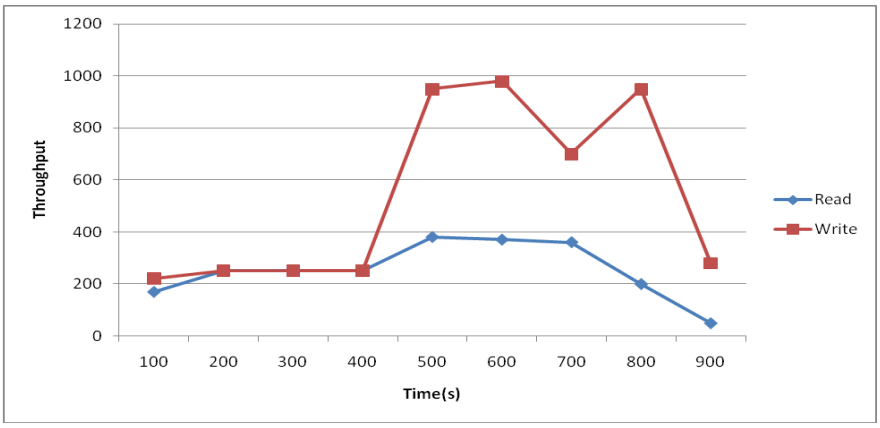

Figure 9: Throughput exhibited by the Apache Flink in terms of Read and Write operations

As shown in Figure 9, the elapsed time it taken in horizontal axis and vertical axis shows the throughput value. The elapsed time has its influence on the throughput. It is understood that Read traffic is less than the Write traffic over a period of time.

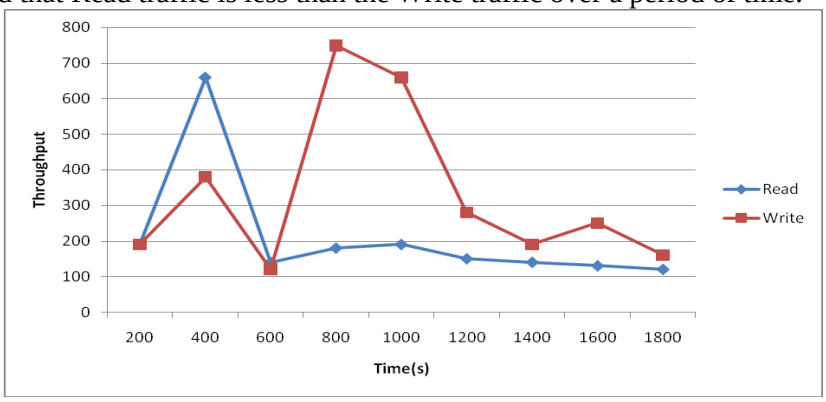

Figure 10: Throughput exhibited by the Apache Spark in terms of Read and Write operations

As shown in Figure 9, the elapsed time it taken in horizontal axis and vertical axis shows the throughput value. The elapsed time has its influence on the throughput. It is understood that Read traffic is less than the Write traffic over a period of time. It is related to Apache Spark. However, there is difference in the results of both the frameworks.

6. Conclusion

In this paper Apache Spark is used to process event logs related business processed that are in the CRM application. The Spark in the cloud eco-system was capable of finding business intelligence from event logs much faster than predecessor. There are queries used to process streaming data. There is analysis of unstructured data. Especially the data related to events that are part of processes is used. Business processes are discovered and then the difference between actual business processes and the business processes discovered from event logs are compared. From the results, it is understood that business intelligence that is used to make right decisions is made possible by screaming data processing with Apache Spark. In future, we intent to have a full-fledged framework for process discovery, identification anomalies and process quality enhancement.

ISSN: 2005-4238 IJFGCN

References

[1] Keith Gutfreund, Big Data Techniques for Predictive Business Intelligence, Journal of Advanced Management Science Vol. 5, No. 2, March 2017

[2] Wil van der Aalst, Using Process Mining to Bridge the Gap between BI and BPM, Eindhoven University of Technology, The Netherlands.

[3] Ramkrushna C. Maheshwar ; D. Haritha D. Haritha , Survey on high performance analytics of bigdata with apache spark , Advanced Communication Control and Computing Technologies (ICACCCT), 2016.

[4] Olivier Caya ; Adrien Bourdon, A Framework of Value Creation from Business Intelligence and Analytics in Competitive Sports , System Sciences (HICSS), 2016 49th Hawaii International Conference.

[5] Mudasir Mudasir Ahmad Wani1, Suraiya Jabin1 , Big Data: Issues, Challenges and Techniques in Business Intelligence

[6] T. White, Hadoop: The Definitive Guide, 3rd ed. Sebastopol, CA: O'Reilly, 2012, pp. 11-12.

[7] M. Bhandarkar, "MapReduce programming with Apache Hadoop," in Proc. 2010 IEEE International Symposium on Parallel & Distributed Processing, Atlanta, GA, Apr. 2010.

[8] M. Grover, T. Malaska, J. Seidman, and G. Shapira, Hadoop Application Architectures,1st ed. Sebastopol, CA: O'Reilly, 2015.

[9] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in Proc. HotCloud 2010, 2nd USENIX Workshop on Hot Topics in Cloud Computing, Boston, MA, 2010.

[10] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, Learning Spark, Lightning-Fast Big Data Analysis, 1st ed. Sebastopol, CA: O'Reilly, 2015.

[11] C. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, 2nd ed. Cambridge, MA: MIT Press, 2000.

[12] A. M. Turing, "Computing machinery and intelligence," Mind, vol. 59, no. 236, pp. 433-460, Oct. 1950.

[13] (Jan. 8, 1954). IBM Press Release. 701 Translator. [Online]. Available: http://www-03.ibm.com/ibm/history/exhibits/701/701_translator.html

[14] C. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, 2nd ed. Cambridge, MA: MIT Press, 2000, ch. 1.4.5, pp. 31-34.

[15] D. Jurafsky and J. H. Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 2008.

[16]I. Karthika, K. P. Porkodi ,"Fraud Claim Detection Using Spark" International Journal Of Innovations In Engineering Research And Technology ISSN: 2394-3696 Volume 4, Issue 2, Feb.-2017

[17]I.Karthika, K.P.Porkodi," Automatic Monitoring And Controlling Of Weather Condition Using Big Data Analytics , International Journal of Advanced Research in Computer and Communication Engineering Vol. 6, Issue 1, January 2017

[18]I. Karthika, P. Gokulraj, S. Saravanan" Prediction of sales using Big data analytics" Journal Of Advances In Chemistry Vol 12, No 20

ISSN: 2005-4238 IJFGCN

[19]I. Karthika, S. Priyadharshini " Survey on Location based sentiment analysis of Twitter data" © 2017 IJEDR | Volume 5, Issue 1 | ISSN: 2321-9939

[20]S Saravanan, V Venkatachalam,"Advance Map Reduce Task Scheduling algorithm using mobile cloud multimedia services architecture" IEEE Digital Explore,pp21-25,2014.

[21]S Saravanan, V Venkatachalam,"Enhanced bosa for implementing map reduce task scheduling algorithm" International Journal of Applied Engineering Research,Vol 10(85),pp60-65,2015.

Authors

M. V Kamal, who could complete B.E in CSE from Gulbarga University, and M. Tech in Software Engineering from JNTU Hyderabad has been pursuing Ph.D in Data Mining from the JNT University, Hyderabad. He is having 18 years of experience in academia. He has published several papers in both National and International Papers and attended several National and International Conferences and organized Seminars etc. His area of interest includes Data Mining and Information Security

D Vasumathi, completed her B.Tech, and M.Tech from Jawaharlal Nehru Technological University Hyderabad. She did her Ph.D (Research) in the area of Data Mining from JNT University, Hyderabad. Presently she is working as Professor in Dept. of CSE, JNTUCEH and having more than 25 years of experience in teaching. She is a member for several professional bodies like CSI, IEEE and ISTE. She had presented and published several papers in National and International Conferences and also in IEEE Explorer. She was chair for several conferences. She did for Editorial board member for several papers of National and International events. Her area of interest includes, Data warehousing and Data Mining, Information Security and Information Retrieval Systems.

ISSN: 2005-4238 IJFGCN