Journal of Advanced Research in Technology and Management Sciences
Volume: 01 Issue: 03 ISSN: 2582-3078 November 2019

Available online at: http://www.jartms.org

AN EFFICIENT APACHE SPARK FRAMEWORK ANALYSIS FOR ENGINEERING STUDENT'S INFORMATION BEHAVIOUR

RAJESH KUMAR.P, B. VIJAYA, V.SAMBA SIVA, A N SREEDHAR

Assistant Professor, Computer science Engineering, Chadalawad Ramanamma Engineering College
Assistant Professor, Computer science Engineering, Chadalawad Ramanamma Engineering College
Assistant Professor, Computer science Engineering, Chadalawad Ramanamma Engineering College
Assistant Professor, Computer science Engineering, Chadalawad Ramanamma Engineering College
rajeshkumar0540@gmail.com,vijayabathini@gmail.com,
sambasivareddycse@gmail.com,ansreedhar01@gmail.com

ABSTRACTION

The past few years have Seen a major changes in huge data volumes in big data world. Spark as in memory cluster computing to analyse the large amount of data in a frame work. Spark offers a rich set of big data application through the programming interface. Spark programming framework provide an effective open source solution for managing and analyze the big data. Now a day's engineering education plays a key role in Andhra Pradesh, India.Engineerin education growing importance of day to day usage of learning management system. The LMS and technical education system has made significant contributions to produce one of the largest technical usages of engineering education students in Andhra Pradesh. This paper discussing the Spark working model procedure as well as Learning management's activities by the students and their information behaviour.

KEYWORDS: Human information interaction, Big data, Spark, RDD,LMS

INTRODUCTION

Human information interaction[1] is a field of study focus on how and why humans use information for decision, learn, plan, make sense, discover and more other activities.HII is a broad area of research, and researches are interested in many different aspects like information behaviour, information search and information retrieval.HII research is the investigation of relationship between humans and information rather than technology. Here we are focusing on engineering student's information behaviour through the spark analsis.human information behaviour models are action models that describe the current activities of day to day activities by the engineering students. We can say these activities called learning management system.HIB is the human behaviour in relation to sources and channels of information include both passive information and information use like browsing, face to face communication in the Educational data. engineering student can interact with Learning management system(LMS) as the learning activities, syllabus, course materials, curriculum, examination patterns, attendance, results, timetables, data related to administrations, infrastructure facilities provided, information about students, staffs and teachers, data from social media, campus sensor data, data from log in details to the online course page etc.Human have been generating data for thousands of years then the Big data is becoming one of the most talked technology trends in now a day. Big data contain large collection complex data sets it is very difficult to process the database with the traditional applications.

Spark[2] is a lightning and fast cluster computing technology to design a fast computation. it is based on HadoopMapReduce and extends the MapReduce model which includes interactive queries and stream processing . The main feature of Spark is its in-memory cluster computing that increases the processing speed of an application. Spark is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries and streaming. Apart from supporting all these workload in a respective system, it reduces the management burden of maintaining separate tools. the following will illustrate the components of spark

Apache Spark core:

The spark core contains the basic functionality including task scheduling, memory management, Fault recovery, interacting with storage system and more. Spark Core is also home to the API that defines resilient distributed datasets (RDDs), which are Spark's main programming abstraction. RDDs represent a collection of items

distributed across many compute nodes that can be manipulated in parallel. Spark Core provides many APIs for building and manipulating these collections

Spark SQL

Structured data can be analyzed with spark SQL through spark package. it support many more data bases and quires like Hive tables, parquet and JSON.SQL queries with RDD manipulate the programming languages like java,scala,python and many more with the single application with combining SQL with complex analytics.

Spark Streaming

The input data can be enables the process of live stream of data. it manipulate the RDD API and application data can be stored in memory on disk. It streams was designed to provide the same degree of fault tolerance, throughput and scalability as spark core why because it's also a core component

MLlib

MLlib provides multiple types of machine learning algorithms including data mining patterns like classification, regression, clustering, and collaborative filtering, as well as supporting functionality such as model evaluation and data import. It also provides some lower-level ML primitives, including a generic gradient descent optimization algorithm. All of these methods are designed to scale out across a cluster.

GraphX

GraphX is a library manipulating graphs and performing graph parallel computation.sprak RDD API allowing to create a directed graph with arbitrary properties attached to each vertex and edge.GraphX also provides subgraph and mapVertics with graph algorithm like PageRank and triangle counting.

Apache spark has a real time analytic power to manage massive amount of data in real time. We can do the fast computation and processing the data.

Apache Spark is used to manage massive amounts of data and to provide real-time analytic power. In this model, we used Apache Spark for fast computations and for managing and processing big data. The second section of this paper provides a literature review. The third section presents the proposed system, including data collection, factors affecting student performance and dropout, NLP solutions and proposed Data Points for Decision Support Systems. The fourth section presents the experiment and the fifth section is the conclusion

LITERATURE REVIEW

The general motivation for this research is assisting engineering students to achieve success. To learn about student behaviour first we need to identify the information needs and seeking behaviour of today's prospective.
[3]This Study provides a big data analytics in higher education and integrated learning solutions. The technology is capable to deal with academic authority and advisors at educational institutions in making decisions concerning individual students.[4]This study focus on the design and implementation of modern and hybrid real time data pipeline architecture by using Apache spark. the analysis about student activities as well as behavior of students in the academics' can be stored and process the data by using Hive reports. [5]In this paper we can identify the advance machine learning architecture of multilayer perception and empirical analysis .the results are encouraging and corroborate our proposed framework over traditional big data analysis methods that use either spark or deep learning as individual elements [9] this paper identified the huge amount of heterogynous data is generated in educational syatem.the data is rich but information is poor, the Hadoop plays an efficient role in performing meaningful real-time analysis on the huge volume of data and able to predict the emergency situations before it happens [10] it derives about the usage of big data analytics in higher education. It help plot out an effective and sustainable management model. Big data can improve enrolment, student performance and optimization of resources. [11] it explored Big Data Analytics and its relevance in Educational systems with a view of helping educational institutions to adopt Big Data Analytics.[12] this study provides evidence on the actual information

seeking behaviour of students in a digital scholarly environment, not what they thought they did. It also compares student information seeking behaviour with that of other academic communities.

OBJECTIVIES OF THE STUDY

The study examined the information seeking behaviour of engineering students in Andhra Pradesh, India.the focus was on obtaining information on the nature of academic, administrative and social media needed by the students. The objectives of the study were

- ✓ To study the information needs and information seeking behaviour of the engineering colleges students in Andhra Pradesh region
- ✓ To identify the educational ecosystem are making it possible to collect, manage and maintain massive amounts of data in engineering education
- ✓ To investigate the big data analytics can be implements in the engineering university/college in Andhra Pradesh
- ✓ To design the engineering students predictive analytics to avoid the failures in there course

RESEARCH METHODOLOGY

The primary data was collected through online questionnaires and interviews of engineering students in Andhra Pradesh state. For this we used the email, facebook, whatsapp.A qualitative and qualitative research design was used in depth of the investigation for information behavior. Totally 1000 questionnaires were distributed and 850 were received after duly filled and the response rate is 85.0%.

We used different methods like online questionnaire, interviews and face-to-face interviews in depth where we used to collect the data from their information seeking behavior in the engineering education. The tools designed to elicit research information in order to identify the factors including socio-economic factors which are liable to affect the information seeking behavior. Such as nature of the information need the purpose it would serve or the existence of external barriers posed by the environment of engineering Students in the Andhra Pradesh. The questionnaire method was adopted for collecting research data keeping in view of the objectives of the study. It was also felt necessary to adopt interview method to overcome the limitation of the questionnaire especially among some students, who could not respond to questionnaire owing to shortage of time. Under such circumstances, interview schedule was employed in obtaining research data from the engineering Students of Andhra Pradesh as a supplement to questionnaire method

Big Engineering Data Collection and Apache Spark streaming frame work

Big data[13] is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage, search, sharing, transfer, analysis and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger

a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

Human Generated Data is emails, documents, photos and tweets. We are generating this data faster than ever. Just imagine the number of videos uploaded to You Tube and tweets swirling around. This data can be Big Data too. Machine Generated Data is a new breed of data. This category consists of sensor data, and logs generated by 'machines' such as email logs, click stream logs, etc. Machine generated data is orders of magnitude larger than Human Generated Data.

In this aspects engineering student[15] can interact with learning management system and platforms to learning activities ,courses information consisting a curriculum such learning objects assyllabuses, learning material and activities, examination results and courses' evaluation, to other kind of data related to administrative, educational and quality improvement processes and procedures. The limited exploitation of big educational data and the size and type of these data within the context of higher education signifies the need for special techniques to be applied in order to discover new beneficial knowledge that currently is hidden within data. Big data can be successfully used to analyses the Educational data. Recently, big data and Analytics together have shown promise in promoting different actions in higher education. These actions concern "administrative decision-making and organizational resource allocation", prevention of students at risk to fail by early identify them, development of effective instructional techniques and transform the traditional view of the curriculum to reconsider it as a network of relations and connections between the different entities of data gathered and regularly produced from LMSs, social networks, learning activities and the curriculum . More specifically, one of the identified areas in which big data and Analytics are appropriately applicable for investigation and improvement in higher education is the curriculum and its contents, as a major part of big educational data

Apache spark streaming

Spark is an Apache project advertised as "lightning fast cluster computing". It has a thriving open-source community and is the most active Apache project at the moment. Spark provides a faster and more general data processing platform. Spark lets you run programs up to 100x faster in memory, or 10x faster on disk, than Hadoop. Last year, Spark took over Hadoop by completing the 100 TB Daytona GraySort contest 3x faster on one tenth the number of machines and it also became the fastest open source engine for sorting a petabyte.Spark also makes it possible to write code more quickly as you have over 80 high-level operators at your disposal. The Spark core is complemented by a set of powerful, higher-level libraries which can be seamlessly used in the same application. These libraries currently include SparkSQL, Spark Streaming, MLlib (for machine learning), and GraphX, each of which is further detailed in this article. Additional Spark libraries and extensions are currently under development as well.

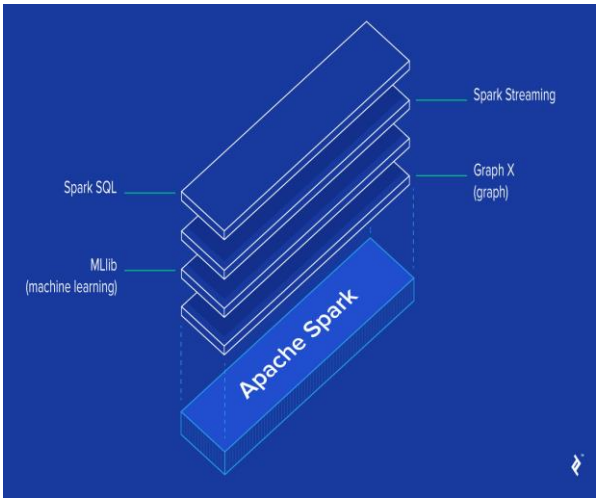

Figure1:apache spark streaming

Figure 1 shows at a high level, every application consists of a driver program that runs the users main function and executes various parallel operations on a cluster. Spark is essentially a computing framework for processing large datasets using a cluster of nodes. Unlike a database, it does not provide a storage system, but it works in conjunction with external storage systems. Spark supports variety of data sources because it is used within the distributed storage system like HDFS, HBase, Cassandra and AmazonS3.

Information behaviour of engineering students

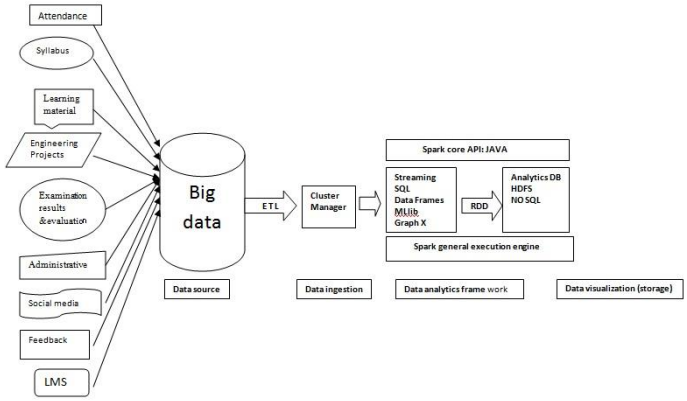

Figure2: ecosystem of engineering data analysis in spark

In the Educational data , engineering student can interact with Learning management system(LMS)[20] as the learning activities, syllabus, course materials, curriculum, examination patterns, attendance, results, timetables, data related to administrations, infrastructure facilities provided, information about students, staffs and teachers, data from social media, campus sensor data, data from log in details to the online course page etc. . Most of the data generated in this sector is of video and audio type followed by text and image. The number of students who could qualify for admission and those who acknowledged their acceptance. In the state of Andhra Pradesh fees reimbursement and scholarship are implements through the student web portal. Alumni data is about those who have already completed their education from the institute. It includes information about their current contact details, job positions, workplace etc**.** Course data includes data about each course undertaken by the students and the students enrolled in each course. Facilities data includes data about infrastructure facilities like the number of classrooms and labs, equipment, and software provided in the labs etc. The students' digital interaction with the University generates flow data. This includes login details to the online course page, notes downloaded, biometric given at library, the time they stayed online for, the number of books issued in his name, dates of issuing and returning the books etc. Today, faster internet speeds and cloud computing supports similarly interactive virtual learning environments. Learners select among modular videos, practice problem sets, and explore supplemental reference material at their convenience.

The Figure2 have been selected for the nature of data, tools employed in data collection and information behaviour. The available data can be access by the process engine such as the engineering legacy database system [22] e,g, students logins,acamic. firstly the data is dump into spark processing engine. Then the performance of action as It stores the data in its raw format,Provides real time accesss and Its convenient for batch processing.

For the data ingestion Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. thus, all the engineering data can be connected to sqoop, which can be scheduled to ingest data at regular intervals.API connector enables the direct storing the structured and unstructured format. The approximation for each student can requires 2MB the we can analyse the maximum of 4lakhs students in Andhra Pradesh in case of the engineering data as LMS, records, acadamics and social media,

requires minimum storage of 1PB.The collected data is now stored in a single location as Parquet files (columnar), in order to save space and facilitate distributed/random access. This distributed storage enables access to any data variable, as it comes from one large table, and further enables the processing engine to run iterative machine learning queries.

Spark[21] is to create new RDDs, transformation and calling operations on RDDs to compute the results.RDD in spark is to distribute collection of objects in cluster. Each object is split into multiple partitions. This may be computed on different nodes of the cluster. This experiment shows the information behavior analysis is more efficient than traditional process. Here we are analyzing the various types of data like structured, semi-structured and unstructured. The structured data is a relational data model is the most used data model Relation, a table with rows and columns Every relation has a schema defining each columns' type The programmer must statically specify the schema The structured data as engineering student syllabus, course materials, curriculum, examination patterns, attendance, results, timetables, projects.Semi-Structured Tabular Data One of the most common data formats A table is a collection of rows and columns Each column has a name Each cell may or may not have a value. Unstructured data (text data) the students were asked to write about five sentences stating their opinion of the engineering course. The students' opinions were collected using a Google form. Out of the 500 participants, only 126 students completed the questionnaire. The questionnaire contained one field, the student ID, and three open questions to motivate students to write more sentences. This data can be run in the programming scala as [22]

In your machine you need to install scala, sbt(build tool), java 8 and spark cluster(aws / use cloudera vm).

- Open command prompt and type 'sbt new scala/hello-world.g8'.
- Write your logic in src/main/scala/**Main.scala** file.
- To compile/ run / package the scala project open the root path of the project in command prompt
  To compile- sbt compile
  To create a package - sbt package
  To run that project - sbt run

Operating or Deploying a Spark Cluster Manually .The most common way to launch spark applications on the cluster is to use the shell command spark-submit. When using spark-submit shell command the spark application need not be configured particularly for each cluster as the spark-submit shell script uses the cluster managers through a single interface. Spark-submit script has several flags that help control the resources used by your Apache Spark application. Spark-submit flags dynamically supply configurations to the Spark Context object.
Below is the source code for the Word Count program in Apache Spark -

```
import org.apache.spark.SparkContext
import org.apache.spark.SparkContext.
import org.apache.spark.
object SparkWordCount {
  def main(args: Array[String]) {

   val sc = new SparkContext( "local", "Word Count", "/usr/local/spark", Nil, Map(), Map())
    val input = sc.textFile("input.txt")
    Val count = input.flatMap(line ⇒ line.split(" "))
    .map(word ⇒ (word, 1))
    .reduceByKey(_ + _)
    count.saveAsTextFile("outfile")
    System.out.println("OK");
  }
}
```

Linking with Apache Spark
The first step is to explicitly import the required spark classes into your Spark program which is done by adding the following lines -

```
import org.apache.spark.SparkContext
import org.apache.spark.SparkContext
import org.apache.spark
```

Creating a Spark Context Object
The next step is to create a Spark context object with the desired spark configuration that tells Apache Spark on how to access a cluster. The below line of code in the word count example does this

val sc = new SparkContext("local", "Word Count", "/usr/local/spark", Nil, Map(), Map())

"Word Count": This is the name of the application that you want to run.
"local": This parameter denotes the master URL to connect the spark application to. /usr/local/spark- This parameter denotes the home directory of Apache Spark.
Map() : The first map specifies the environment whilst the second one specifies the variables to work nodes.

Creating a Spark RDD
The next step in the Spark Word count example creates an input Spark RDD that reads the text file input.txt using the Spark Context created in the previous step-
val input = sc.textFile("input.txt")

Spark RDD Transformations in Wordcount Example
The below lines of spark application code transform the input RDD to count RDD

Val count = input.flatMap (line ⇒ line. Split ("")).map (word ⇒ (word, 1)).reduceByKey (+ _)

In the above piece of code, flatMap () is used to tokenize the lines from input text file into words. Map () method counts the frequency of each word. reduceByKey () method counts the repetitions of word in the text file.

In that spark cluster, please check the spark version is same as the build.sbt spark Let's create a Spark RDD using the input file that we want to run our first Spark program on. You should specify the absolute path of the input file-

scala> val inputfile = sc.textFile ("input.txt")

Now is the step to count the number of words -

a. Each line is split into words using **flatMap** RDD transformation. **flatMap** works applying a function that returns a sequence for each element in the list, and flattens the results into the original list.
b. Each word is read and key-value pairs are created for each one of them using **map** transformation. This will assign the value '1' to each of the work-keys.
c. Finally the values of similar keys are added to get the final word count using **reduceByKey** function.

scala> val counts = inputfile. flatMap (line => line. Split (" ")).map (word => (word, 1)).reduceByKey (_+_)
You will get the following output:

```
scala> val counts = inputfile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_+_)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:29
```

The next step is to store the output in a text file and exit the spark shell.

scala>counts.saveAsTextFile ("output")

```
scala> counts.saveAsTextFile("output")

scala> exit
```

Go to the output directory (location where you have created the file named output). Use 'ls' command to list the files present in the directory. On successful execution of the word count program, the file ls will be created as shown below -

Using the cat command, print the contents of the output file to find the occurrence of each word in the input.txt file -

```
himanshu@himanshu-VirtualBox:~$ cd output
himanshu@himanshu-VirtualBox:~/output$ ls
part-00000  _SUCCESS
himanshu@himanshu-VirtualBox:~/output$ cat part-00000
(fail,1)
(Led,1)
(next,1)
(nor,1)
(rose,1)
(smallness,1)
(countryman,1)
(are.,1)
(Had,1)
(here,1)
(woody,1)
(put.,1)
```

EXPERIMENTS AND RESULTS

In this practical experience of engineering students behaviour could be analysed for answering our research survey questions and the result will focus the seeking behaviour of engineering students in Andhra Pradesh. Learning activate environment in the education curriculum and admistration works can massively turn into huge amount of data. Big data analytics will collect the large amount of data for objective evolution. From this point the student can succussive their learning process in their academic and it gives the institutions a better understanding of what is happening and what can be done because they are able to read about other issues. We need to implements the big data analytics in the engineering university/college in Andhra Pradesh. Big data is used to understanding and gathering information on strengths and weakness of an education system in order to improve the education system

CONCLUSIONS

The research will explore the predictability of student success from learning analytics on big data sets. We seek to analyse a rich set of data of student activities as gathered via their interaction with a Learning management system. Educational institutions are generating huge volumes of data through the admission process, evaluation and teaching learning. From this analysis a model of academic performance among other measurable for success. The Big Data paradigms are needed in current world to add value to the processes of educational institutions. The benefits and potential to use them for quality improvement purposes in healthy education.

. **Reference**

1. Raya Fidel, Human Information Interaction: An Ecological Approach to Information behaviour, Publisher: The mit press, august 2013
2. Mohammed Guller Big Data Analytics with Spark A Practitioner's Guide to Using Spark for Large Scale Data Analysis, Aprèss ,2015
3. Amal S. Alblawi , Ahmad A. Alhamed "BIG DATA AND LEARNING ANALYTICS IN HIGHER EDUCATION" ed:IEEE,2017,pp.21-25
4. Abdelmajid Chaffaik, Larbi Hassouni, Houda Anoun" Real-Time Analysis of Students' Activities on an E-Learning Platform based on Apache Spark, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 7, 2017,pp.101-109

5. Anand Gupta, Hardeo Kumar Thakur "A Big Data Analysis Framework UsingApache Spark and Deep Learning. arXiv:1711.09279v1 [cs.DB] 25 Nov 2017
6. B. Daniel and R. Butson, "Foundations of big data and analytics in higher education," 2014.-
7. B. K. Daniel, "Overview of Big Data and Analytics in Higher Education," in Big Data and Learning Analytics in Higher Education, ed: Springer, 2017, pp. 1-4.
8. O. Aljohani, "A review of the contemporary international literature on student retention in higher education," International Journal of Education & Literacy Studies, vol. 4, 2016.
9. Deepa. A, Dr.E.Chandra Blessie (2017**)** Big data analytics for accreditation in the higher education, IJCSIT, Vol. 8 (3) , 2017, 357-360,ISSN:1975-9646
10. Vatsala, Rutuja Jadhav and Sathyaraj R (2017) A Review of Big Data Analytics in Sector of Higher Education ISSN: 2248 -9622, Vol. 7, Issue 6, (Part -2) June 2017, pp.25-32
11. Julius Murumba Elyjoy M. Micheni (2017) Big Data Analytics in Higher Education: A Review ISSN (e):2319 – 1813 ISSN (p): 2319 – 1805 || Volume || 6 || Issue || 6 || Pages || PP 14-21 || 2017
12. David Nicholas and Eti Herman. Assessing information needs in the age of the digital consumer Routledge. 2009 (ISBN 978-1-85743-487-3)
13. Bernard Marr ,Big data and Big data in practices , 2nd edition ,willy publication, ISBN-13: 9781118965825,2015
14. S. L. Thomas, "Ties that bind: A social network approach to understanding student integration and persistence," The journal of higher education, vol. 71, pp. 591-615, 2000
15. B. Daniel and R. Butson, "Foundations of big data and analytics in higher education," 2014.
16. S. J. Jones, "Technology review: the possibilities of learning analytics to improve learner-centered decision-making," The Community College Enterprise, vol. 18, p. 89, 2012.
17. B. K. Daniel, "Overview of Big Data and Analytics in Higher Education," in Big Data and Learning Analytics in Higher Education, ed: Springer, 2017, pp. 1-4.
18. O. Aljohani, "A review of the contemporary international literature on student retention in higher education," International Journal of Education & Literacy Studies, vol. 4, 2016.
19. S. L. Thomas, "Ties that bind: A social network approach to understanding student integration and persistence," The journal of higher education, vol. 71, pp. 591-615, 2000.
20. Dr. J Meenakumari, Jayashree M. Kudari. Learning Analytics and its challenges in Education Sector a Survey. International Journal of Computer Applications, 2015, 0975 – 8887.
21. Surveylink: https://docs.google.com/forms/d/e/1FAIpQLSemqLiQewRnPe5dC4EVAR_Bdm_-N9jD-1KHJuS0gHBff1HVPQ/viewform
22. Spark documentation, available at *http://spark.apache.org/docs/latest/*