# THE INTEGRATING E-GOVERNANCE WITH BIG DATA ANALYTICS USING APACHE SPARK

**Dr. C K GOMATHY[1], A. SURYA TEJA[2], CH. MANI KANTA[3], CH. VENKATA MAHESH[4]**

---

**ABSTRACT:-** The constant innovations and rapid developments in the IT industry have revolutionized the thinking andmindset of the people throughout the world.

Government departments have also been computerized to provide transparent, efficient and responsible government through e-governance. The government have been providing access to various websites or portal for filing complaints, uploading or downloading forms, pictures, data or PDFs to avail the government services. Enlightened citizens are frequently using the portal to access governmentservices. Thus, the size and volume of data that need to be managed by government departments have been increasing drastically under e-governance. The traditional database managementsystem is not designed to deal with such mix type of data. Moreover, the speed at which the overnance generated data need to be processed isanother big challenge being faced by traditional database system.

All the abovesaid concerns can be solved by using the emerging technology - Big Data Analytics techniques. Big data analytic techniques can make the government more efficient and transparent by processing structured, unstructured or mixed types data at a great speed. In this paper, we shall understand the scenario for the need or the emergence of big data analytics in e-governance and knowhow of Apache Spark. This paper proposes a practical approach to integrate big data analytics with e-governance using Apache Spark. This paper also reflects how major issues of traditional database management system (mixed type datasets, speed and accuracy) can be resolved through the integration of big data analytics and e-governance.

**Keywords:** E-governance; Big data analytics; Apache Spark

## I. INTRODUCTION

Due to the advancement in technology, various industries or domains like transport,tourism, hotel, banks and so on have been digitized and generating large amount of data. People are using the Internet to generate forms, reports, graphs, periodic or to do online shopping on discounted rates. Social media (Facebook, Instagram, blogs, twitter etc.) or entertainment industries are using computers toshare pictures, audios and videos. According to a genral survey posted on wikipedia till April 2019, 56.1% of population has been accessing Internet services [10]. Government websites have also been generating massive amount of data by uploading or downloading pictures or credentials like finger prints, retina scan, forms, reports of the citizens. The big data analytic techniques have been designed to store, process and analyze such a mixed mode data.

Revised Manuscript Received on February 24, 2020.

* Correspondence Author

Poonam Salwan*, Research Scholar, I.K. Gujral Punjab Technical University, Jalandhar,Punjab, India.

E-mail: poonam12_sharma@yahoo .com Dr. Veerpaul Kaur Maan, A.P., Giani Zail

Singh Punjab Technical University, Bathinda, Punjab, India.

E-mail: veerpalkaur1@rediffmail.com Thus, integrating e-governance with big data is the need of the hour. The main objectiveof this research paper is to provide an insight of emerging needs to deal with huge data under e-governance, introduction to Apache Spark framework to store, process and deal with big data at greater speed and accuracy. This paper progresses by expanding the literature review on e-governance, Big data, Big data analytics, followed by Apache Spark and proposed a newsystem framework.

## II. EXISTING SYSTEM

As both the terms – e-governance and Big dataanalytics are very vast, let us try to understand them one by one. The first section will deal with e-governance and second section deal with big data.

### A. E-governance

E-governance refers to the process of providing government services online. It makes the government system more efficient, transparent and accountable. Citizens can access government services by using the web portals that have been created to provide all services at one click. They can easily upload or download forms, photos, data and so on. The biggest example of the success of e-governance is Aadhar (UDAI) portal run by Indian government. This portal stores variety of data like text, images, PDF, retina scan, name, age, address and other related information of their stakeholders [2]. As of 2018, India has a population of over 1.355 billion people, and its growth is expected to continue through at least 2050[11]. Managing such

mix type of data being generated at such a huge level is the biggest challenge.

B. Big data

As the name is indicating, a huge amount of data that is very difficult to store, analyze and execute is called Big data. The definition of big data varies from company to company. One company's big datacan be small for others. But when the data does not fit in memory, nor on hard disk and if there is continuous demand for processing, then it is considered as big data [5]. For some companies, data upto 10 TB is considered as big data. While, for some other companies,

1 PB of data is considered as big data.

C. Big data analytic

The term Big data analytics refers to the process of analyzing raw datasets to understand their hidden patterns and behaviours using qualitative and quantitative techniques. The analyic techniques are basically used in B2C (Business to Customer) applications to collcet, categorize, store, process and analyze the trends and future expectations. Thus, Big data also helps in decision making. The various techniques of Big data analytics are as follows in Fig. 2:

1. Descriptive analytics: It uses historic or traditional datasets and apply predictive or trends analysis on them.

2. Predictive analysis: It indicates what will happen in future. It also indicates what will be the situation, trends or outcome in that particular time span.

3. Prescriptive analysis: It helps to take best possible solution from multiple options. Predective analytics become much mature and stable with age and experience.

## III. PROPOSED SYSTEM

The e-governance can be integrated with various components provided by Hadoop to face the challenges occurred due to big data. Here, a new system is proposed to design and integrate Apache Spark with Hadoop to provide a framework to deal e-governance big data at a greater speed and accuracy.

A. Experimental environment set up

The basic requisites of the proposed system are:

(1) Download and install Java: In order to install Apache Spark on the system, you need to download and install Java (version 1.8). Use a secured link from the Internet to download Java.

(2) Download and install Scala: Apache Spark is made up of Scala language. Use the secured link to download Scala. To install Scala [18], run the following commands (see Table I) for extracting the tar files:

**Table I: Installing Scala**

| Sr. No | Requirement | Command(s) |
|---|---|---|
| 1 | Extracting Scala tar file | $ tar xvf scala-<version>.tgz<br>Example: $ tar xvf scala-2.11.6.tgz |
| 2 | Moving Scala to a specific directory | # cd /home/file/downloads/<br># mv scala-2.11.6/usr/local/scala<br># exit |
| 3 | Setting path for Scala | $ export PATH =$PATH:/usr/local/scala/bin |
| 4 | Verifying Scala installation | $ scala -version |

(3) Download and install Apache Spark: Use the secured link to download the latest version of Apache Spark. To install Apache Spark, run these commands (Table II).

**Table II: Installing Apache Spark**

| Sr. No | Requirement | Command |
|---|---|---|
| 1 | Extracting Apache Spark tar file | $ tar xvf spark-<version>.tgz<br>Example: $ tar xvf spark-1.3.1-bin-hadoop2.6.tgz |
| 2 | Moving Spark to a specific directory | # cd /home/hadoop/downloads/<br># mv spark-1.3.1-bin-hadoop2.6/usr/local/spark<br># exit |
| 3 | Setting path for Spark | $ export PATH =$PATH:/usr/local/spark/bin |
| 4 | Verifying Spark installation | $ spark -shell |

B. Data source

The data source, used in this paper, has been taken from open-source data repository (data.world) for research and analysis. This government collected dataset has been published by National University of Educational Planning and Administration, on behalf of department under Ministry of Human Resource Development Department of School Education and Literacy, Government of India.

The current dataset is regarding the status of Elementary Education in India, published in the year 2014-2015 and 2015-

2016. The School Report Cards are available at www.schoolreportcards.in(https://data.world/ind erz/india-district-level-school-report-card). Use Table III to know the detail of the dataset files, their sizes and number of records.

**Table III: Dataset description**

| File Name | Size of File | Number of Columns | Number of Records |
|---|---|---|---|
| Dist. Report Card 2014-15 | 2 MB | 256 | 680 |
| Dist. Report Card 2015-16 | 2 MB | 256 | 680 |

Some of the fields from the selected dataset, for both the years, are shown in Table IV.

**Table IV: Fields from School Report Card**

| AC_YEAR | Data Reported From | Total Schools by Category | Total Schools by Category - Government & Aided | Schools by Category: Boys Only |
|---|---|---|---|---|
| Schools by Category: Girls Only | Enrolment by School Category | Teachers by School Category | Single-Classroom Schools by School Category | Schools Approachable by All Weather Road |
| Schools with Computer | Single Teacher Schools | Enrolment by Grade | Teachers by School Category: Male | Girls Enrolment By School Category |
| Teachers by School Category: Female | Number of Classrooms by School Category | Committee (Government & Aided Schools) | Schools with Enrolment <= 50 | Schools Constituted School Management |

**Table V: Records from dataset**

| AC_YEAR | Data Reported From | | | Schools with Computer (2014-15) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Primary Only | Primary with | Primary with | Upper Primary | Upper Primary | Primary with | Upper Primary | Total |
| | DISTCD | State Name | DISTNAME | SCOMP1 | SCOMP2 | SCOMP3 | SCOMP4 | SCOMP5 | SCOMP6 | SCOMP7 | SCOMPTOT |
| 2014-15 | 0101 | JAMMU & KASHMIR | KUPWARA | 21 | 91 | 1 | 1 | 1 | 38 | 14 | 167 |
| 2014-15 | 0102 | JAMMU & KASHMIR | BARAMULA | 16 | 116 | 6 | 0 | 3 | 76 | 24 | 241 |
| 2014-15 | 0103 | JAMMU & KASHMIR | SRINAGAR | 26 | 193 | 45 | 0 | 2 | 204 | 2 | 472 |
| 2014-15 | 0104 | JAMMU & KASHMIR | BADGAM | 15 | 94 | 8 | 3 | 1 | 67 | 15 | 203 |
| 2014-15 | 0105 | JAMMU & KASHMIR | PULWAMA | 16 | 89 | 4 | 1 | 2 | 64 | 13 | 189 |
| 2014-15 | 0106 | JAMMU & KASHMIR | ANANTNAG | 52 | 160 | 11 | 1 | 3 | 85 | 13 | 325 |
| 2014-15 | 0107 | JAMMU & KASHMIR | LEH (LADAKH) | 19 | 84 | 2 | 2 | 2 | 35 | 1 | 145 |
| 2014-15 | 0108 | JAMMU & KASHMIR | KARGIL | 18 | 65 | 6 | 2 | 1 | 19 | 15 | 126 |
| 2014-15 | 0109 | JAMMU & KASHMIR | DODA | 13 | 35 | 9 | 0 | 4 | 38 | 2 | 101 |
| 2014-15 | 0110 | JAMMU & KASHMIR | UDHAMPUR | 23 | 87 | 30 | 0 | 4 | 63 | 0 | 207 |

Analysis is done to find out which State of India has highest percentage increase in Computers in schools. The fields that are required to be studied for this analysis is AC_YEAR, Data Reported From and Schools with Computer as shown in Table V.

## C. Experiment details

The purpose of this study is to understand the increase or decrease of number of computers in each and every state of India for the last 2 years. Then try to find the state whose percentage of using computer has been increased.

### i. Algorithm

On the basis of available datasets following generic algorithm is designed to read, load and analysis the datasets.

Step 1 Convert the downloaded Excel files into the flat files (.csv).

Step 2 Load the .csv file in to the system to create RDD; to create load DataFrames.

Step 3 Run the SQL commands on DataFrames toperform manipulations.

Step 4 Rum the SQL command to extract the subset data file containing fields - Year, Data Reported From and Schools with Computer 2014-15.

Step 5 Repeat the Step 1 to Step 4 to extract the similar dataset for Year 2015-16.

Step 6 Load the aggregated subset files for both the years using Spark framework.

Step 7 Iterate the files for all the districts of each state using Select command and find total number of computers available in each state for both years.

Step 8 Iterate state-wise to compare the Total Computers in each state.

Step 9 Calculate percentage increase or decrease in the total number of computers in schools for each state as

Formula for percentage change in computer per state = [(Total Computer in 2015-16 per state - Total Computer in 2014-15 per state) / (Total Computer in 2014-15 per state)] * 100

### ii. Implementation and result

The implementation and the corresponding results of the experimental environment are as follows:

Files are loaded in Spark based experimentalenvironment having 1 Executor, 1 Node system with Windows 10, Intel Core i5, 2.4 GHz, 8 GBRAM.

Project Setup

Here major dependencies are on the following factors:

o Spark core

o Spark sql

o Spark csv

Querying .csv data is very easy using the Spark csv library. For that, we will be using SQLContext object. With SQLContext object, we can query the data like we do in any database language. We can perform all the operations on data like SELECT and also write the data into a new file.

After setting up an SBT project, we will start by adding required dependencies into build.sbt.

Project execution

To execute the project, perform the following tasks:

(1) Set up the Apache Spark configuration.

(2) Initiate the process by making the sparkContext object.

(3) Make the sqlContext object to retrieve the desired dataset from the .csv files.

(4) Read the .csv files using sqlContext.read.format() method. Load the data from .csv file into a Resilient DistributedDataset (RDD).

(5) Create the DataFrames objects using method toDF();

(6) Run the appropriate queries using SQL commands to retrieve the desired dataset.

(7) Place the ResultSet in the temporary table or can save it to use it later.

Execution Time

After implementing the Algorithm steps, the following is the result displayed by the Spark-shell: Result is [state : BIHAR, old_SCOMPTOT = 5296, new SCOMPTOT

=6085, difference 789, % diff =

14.89803625377]

Overall experiment completed in 2-4 seconds with Spark-shell on a single node

Around 6-8 seconds, when program is compiled to make jar and then executed on more than one node

iii. Hypothetical analysis

The current dataset has been analyzed on a Windows based Spark-shell. Considering the amount of data to be processed, single node system has been used. However, if data volume increases in size, the datasets can be distributed to N nodes (multiple nodes) and the overall processing will be as shown in Fig. 6.

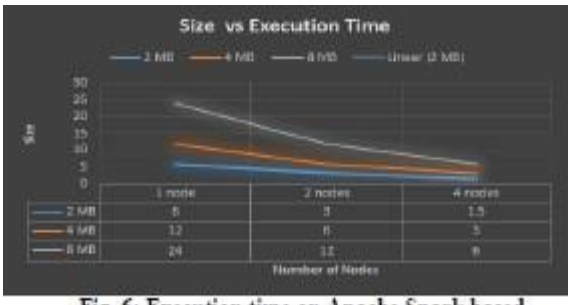


Fig. 6: Execution time on Apache Spark based system(s)

If the same analysis needs to be performed without using any technology stack from Big Data ecosystem, the same process will take much more time for similar records.

It shows that calculation on e-governance data using Apache Spark has taken only few seconds to compute and display the name of state that has highest growth in the use of computers among the schools of all the states in India.

Similar big data framework can be implemented easily to analyze e-governance datasets collected from other fields as well.

## IV. CONCLUSIONS

This paper has given the brief introduction of e-governance, its close relationship with big data and how Apache Spark based big data analytic system can be helpful in analyzing government collected datasets accurately at high speed. This research paper gives an insight about Apache Spark framework, its node-based architecture as well as the implementation ways of setting up and analyzing the government collected datasets.

This paper proposes a new system, through an algorithm, to analyze government collected datasets (Elementary Education in India) by setting up a Spark object. This paper further shows the execution time to perform analysis using Spark-shell on single executor;single nodeWindows based system.

For future references, hypothetical analysis supported by comparison chart has been created to show how execution efficiency of government collected datasets will be increased multiple times using multi-node systems supported by Apache Spark.

## V. REFERENCES

1. Amol Bansod, “Efficient Big Data Analysis with Apache Spark in HDFS, International Journal of Engineering and Advanced Technology”, IJEAT,Volume-4 Issue-6, August 2015.

2. Swapnil Shrivastava, Supriya N Pal, “A Big Data Analytics Framework for Enterprise Service Ecosystems in an e-Governance Scenario”, ICEGOV '17,ACM, New Delhi India, 2017.

3. Sruthika s., N.Tajunisha, “A study on evolution of data analytics

to big data analytics and its research scope",International Conference on Innovations in Information Embedded and Communication Systems, IEEE,2015.

4. Preet Navdeep, Dr. Manish Arora, Neeraj Sharma, "Role of Big Data Analytics in Analyzing e-Governance Projects", International onference on New Trends in Business and Management: An International Perspective, E-journal ISSN2250-348X, 2016.

5. Annu kumari, Shailendra Singh, "A review paper on E-governance: transforming government", International Conference on Cloud System and Big Data Engineering, IEEE, 2016, 689-692.7

6. C.K.Gomathy.(2010),"Cloud Computing: Business Management for Effective Service Oriented Architecture" International Journal of Power Control Signal and Computation (IJPCSC), Volume 1, Issue IV, Oct - Dec 2010, P.No:22-27, ISSN: 0976-268X .

7. Dr.C K Gomathy, Article: An Effective Innovation Technology In Enhancing Teaching And Learning Of Knowledge Using Ict Methods, International Journal Of Contemporary Research In Computer Science And Technology (Ijcrcst) *E*-Issn: 2395-5325 Volume3, Issue 4,P.No-10-13, April '2017

8. Dr.C K Gomathy, Article: A Semantic Quality of Web Service Information Retrieval Techniques Using Bin Rank, International Journal of Scientific Research in Computer Science Engineering and Information Technology ( IJSRCSEIT ) Volume 3 | Issue 1 | ISSN : 2456-3307, P.No:1563-1578, February-2018

9. Dr.C K Gomathy, Article: A Web Based Platform Comparison by an Exploratory Experiment Searching For Emergent Platform Properties, IAETSD Journal For Advanced Research In Applied Sciences, Volume 5, Issue 3, P.No-213-220, ISSN NO: 2394-8442,Mar/2018

10. Dr.C K Gomathy, Article: A Study on the Effect of Digital Literacy and information Management, IAETSD Journal For Advanced Research In Applied Sciences, Volume 7 Issue 3, P.No-51-57, ISSN NO: 2279-543X, Mar/2018

## Author's Profile:-

**1. Mr.A.SURYATEJA**, Student,
B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. ISr Area of Big data analytics

**2.Mr. CH.MANI KANTA**, Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. His Area of Big data analytics

**3.Mr. CH. VENKATA**
MAHESH,Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. Her Area of Big data analytics.

**4. Dr. C.K. Gomathy** is Assistant Professor in Computer Science and Engineering at Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. Her area of interest is Software Engineering, Web Services,Big data Analytics, Knowledge Management and IOT.