JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 10, ISSUE 03, 2023

Data Behaviour Analysis using Intelligent Big Data Analytics

P. Venkateshwara Rao[1], Adnan Faheem[2], Afreena Begum[2], Varshith Aadi[2]

[1,2]Department of Information Technology

[1,2] CMR Engineering College, Kandlakoya, Medchal, Hyderabad.

ABSTARCT

Intelligent big data analysis is an evolving pattern in the age of big data science and artificial intelligence (AI). Analysis of organized data has been very successful but analysing human behaviour using social media data becomes challenging. The social media data comprises a vast and unstructured format of data sources that can include likes, comments, tweets, shares, and views. Data analytics of social media data became a challenging task for companies, such as Dailymotion, that have billions of daily users and vast numbers of comments, likes, and views. Social media data is created in a significant amount and at a tremendous pace. There is a very high volume to store, sort, process, and carefully study the data for making possible decisions. This project proposes an architecture using a big data analytics mechanism to process the huge social media datasets efficiently and logically. In addition, this work employing parallel processing techniques called spark, which will create multiple threads and then distribute work between those thread to perform task parallelly and then send result back to spark. All existing algorithms works on single thread but spark will distribute works in multiple threads so its paralleling processing will be faster and suitable for big data applications.

This proposed work uses hive, spark, and Hadoop where first two will be used to store the data and spark will be used to read and process that data. Here, the dataset of reviews is gathered from Dailymotion website as .csv file and then extracting useful information such as most talk countries with many likes and then extracting likes, view, and comments from so many categories called fashion, entertainment, news etc. Finally, this project compares the execution time of processing with and without spark algorithm.

Keywords: Data behaviour analysis, Big data analysis, Artificial intelligence.

1. INTRODUCTION

Intelligent big data analysis is an evolving pattern in the age of data science, big data, and artificial intelligence (AI). [1] Data has been the backbone of any enterprise and will do so moving forward. Storing, extracting, and utilizing data has been key to any operations of a company. When there were no interconnected systems, data would stay and be consumed in one place. With the onset of Internet technology, the ability and requirement to share and transform data have been exploited. With the spread of social media, the nature of data has changed. Social media can consist of billions of users who continuously provide their digital traces with incredible velocity. [2] As the data comes from many sources and in an unstructured format, it is not easy to handle in traditional relational databases. The need for handling unstructured data gives birth to another type of data called big data, which is unstructured, semi-structured, and unpredictable. This data is created real-time, and the amount of data is increasing daily. [3] The data generated from these social media sites can take the form of text, images, videos, and documents. Only structured data can be processed and stored using an RDBMS. Big data is used to process data with a huge volume that is not possible to process using old database techniques and traditional relational databases, within an acceptable processing time.

Big data is characterized by a large volume of data with a large variety and higher velocity Data generated moves through cables, either TV or internet, and data on local TV cables broadcast with

large volume, variety, and velocity. [4] The amount of data generated every day in the world is increasing exponentially. The rate of data growth is surprising, and this data comes at a speed, with variety (not necessarily structured), and contains a wealth of information that can be key for gaining an edge in competing businesses. The ability to analyse this massive amount of data brings a new era of innovation, productivity growth, and consumer surplus. "Big data is the term for a collection of data sets so large and complex that it becomes difficult to process it using traditional database management tools or data processing applications".[5] The challenges include capturing, curating, storing, searching, sharing, transferring, analysing, and visualizing this data. This section discusses the related literature.

[6] Big data is described with 5V's instead of 3V (volume, velocity, and variety) and included veracity and value. The widely known big data examples are social networking sites, such as Facebook, YouTube, Dailymotion, Google, and Twitter. These sites receive a tremendous amount of data regularly with different variety, velocity, and veracity. The data include value as well. [7] As the number of users increases, the amount of data also increases day by day. Users and data both keep growing on these sites, and this amount of data is a big challenge for owners and companies. This data contains all useful information that needs to be processed in a concise period. To generate more revenue and increase sales, the companies need the processed and analyzed data. The analysis of this data is not possible through relational or traditional database systems within a given time frame as the resources of this traditional system are not sufficient to accomplish processing and storing this huge amount of data; hence, Hadoop comes into the existence for fulfilling this need. In recent years, a large amount of unstructured data is generated from social media sites, such as Facebook, Twitter, Google, and some Dailymotion forums in the form of images, text, videos, and documents, to access and analyse this type of data, this work is best for practicing in the entire field. [8] Twitter and Facebook are some of the most famous social media platforms, and the companies find that it is very crucial for obtaining customer feedback and maintaining goodwill.

Dailymotion is one of the best video-sharing social media websites. [9] It is a viral platform that publishes community feedback through its videos and comments, likes, dislikes, published videos, and subscriber information for a particular channel. The analysis of this type of data is important for acquiring knowledge about users, categories, and interests of users. Most of the production companies have their channels to share daily their movie trailers for getting user feedback before releasing them to the general public. Furthermore, individual users upload their videos to get more subscribers and views. These data points are critical for owners to analyse data to understand the views and feelings of customers about their video and service. Dailymotion has billions of users, who watch hours of videos on their site and generate a massive amount of views. It is estimated that more than a hundred hours of videos are watched per minute, and this amount is increasing day by day. To analyse such a huge amount of data, relational databases are not applicable. Users can use this data to understand how much their marketing program is effective. They can check their view counts and subscribers based on the date range that will show them the peak and downtime of views in a particular time. This will also help to check social trends and behaviour of people over time. For example, users can check how many views their videos have received and how much people have liked their video or product. They can also analyse likes and dislikes from the diverse nature of people around the world.

In this research, [10] we utilized Apache Spark to process datasets of social media. Apache Spark is a parallel and distributed platform that overcomes the challenges faced by the traditional processing mechanisms. The main objective of the project is to demonstrate the use of Apache Spark parallel and distributed framework technologies with other storage and processing mechanisms. The social media data generated from Dailymotion is taken under consideration in this article.

2. LITERATURE SURVEY

A. P., Chiplunkar N. N, et al, (2018) [11] authors analysed tweets streamed in real time. They have used Apache Flume to capture real-time tweets. As an analysis, they have proposed a method for finding recent trends in tweets and performed sentiment analysis on real-time tweets. The analysis is done using Hadoop ecosystem tools such as Apache Hive and Apache Pig. Performance in terms of execution time is compared for analysis of real-time tweets using Pig and Hive. From the experimental results, conclusion can be drawn that Pig is more efficient than Hive as Pig takes less time for execution than Hive.

Rodrigues A. P., Rao A., Chiplunkar N. N, et al, (2017) [12] In this work Authors have considered a real-time streaming data on political issue which are loaded in the JSON (JavaScript Object Notation) format. In JSON format, every data is represented in key/value pairs and separated by comma. This paper is organized as follows. describes various methods used for sentiment analysis. The proposed methodology is discussed. explains and analyses the results obtained from their proposed method. Finally, the conclusion and future work are drawn.

Blomberg J. et al, (2012) [13] In this paper, they will outline the concept and execution of two social media analytics applications that use SAS to address law enforcement issues. The applications incorporate social media in very different ways. The first is as an investigative tool to find social media related to specific people. Using an adaptation of our Social Network Analysis (SNA), they present Facebook and Twitter searches of multiple suspects in an easily digestible form for the analyst. The second application focuses on monitoring social media across a much broader spectrum, looking for the proverbial "needle in a haystack". In this example, they show how to collect and analyse historical Twitter data to try to understand precursors to dangerous activity at events, such as riots at concerts or flash mobs.

Mahalakshmi R., Suseela S.et al (2015) [14]. In this research, they utilized Apache Spark to process datasets of social media. Apache Spark is a parallel and distributed platform that overcomes the challenges faced by the traditional processing mechanisms. The main objective of the project is to demonstrate the use of Apache Spark parallel and distributed framework technologies with other storage and processing mechanisms. The social media data generated from Dailymotion is taken under consideration in this article.

Barros, C. P., and Couto, E.et al (2013) [15] This paper considers productivity changes in European airlines between 2000 and 2011 with a particular focus on the impacts of the events of September 11th, 2001, and subsequent shocks to the system, including fuel price fluctuations. The period has seen significant changes in the structure of the European industry both as reflection of these shocks but also as the result of on-going market forces. For example, there has been consolidation of some of the largest carriers, such as Iberia and British airways in 2011, the restructuring of several after bankruptcy including Alitalia – Compagnia Aerea Italiana SPA taking over Alitalia – Linee Aeree Italiana after it went bankrupted in 2008, and the growth and demise of a number of low-cost carriers.

Xia Q., Yin X., He J., Chen F. et al (2018).[16] Real-time recognition of human daily motion with smartphone sensor. *Int. J. Performability Eng.* 14 593–602. This present paper proposes a method for real-time identification of typical movements in people's daily life, which is based on seven lightweight feature vectors from the time-domain of smartphone sensor. The method can provide monitoring and tips for people's physical health. Motions of six kinds are chosen – stillness, walking,

running, walking upstairs, walking downstairs and cycling – as recognition objects, following SM Camhi's [5] research on the relationship between the health of human cardiac metabolism and the exercise intensity. By using lightweight feature vectors from the time domain, the method can effectively reduce the computing load of the smartphone and thereby realize the real-time recognition of daily movements.

Lee N. R., Kotler P. et al (2011). [17] *Social Marketing: Influencing Behaviors for Good.* Thousand Oaks, CA: Sage Publications. In this paper, they will outline the concept and execution of two social media analytics applications that use SAS to address law enforcement issues. The applications incorporate social media in very different ways. The first is as an investigative tool to find social media related to specific people. Using an adaptation of our Social Network Analysis (SNA), they present Facebook and Twitter searches of multiple suspects in an easily digestible form for the analyst.

Wang J., Yang Y., Wang T., Sherratt R. S., Zhang J.et al [18] (2020). Big data service architecture: a survey. *J. Internet Technol.* 21 393–405. This paper is devoted to analyzing the current big data service architecture, which is composed of three main layers. In the data collecting and storage layer, data sources in big data services are needed to be collected by corresponding equipment, and then the data in "pre-processed" state will be stored and processed in a distributed file system or database system. In the data processing layer, different processing frameworks are adopted according to different forms of data. The in-depth analysis of big data is currently mainly based on large-scale machine learning technologies, which can deeply mine the potential value of data. Finally, visualization tools are used to present results to data service consumers.

Cui Y., Kara S., Chan K. C. et al [19] (2020). Manufacturing big data ecosystem: a systematic literature review. *Robotics In* this paper presents a systematic literature review 21 of the state-of-the-art of big data in manufacturing. Six key drivers of big data 22 applications in manufacturing have been identified. The key drivers are system 23 integration, data, prediction, sustainability, resource sharing and hardware. 24 Based on the requirements of manufacturing, nine essential components of big 25 data ecosystem are captured. They are data ingestion, storage, computing, 26 analytics, visualization, management, workflow, infrastructure and security. 27 Several research domains are identified that are driven by available capabilities 28 of big data ecosystem.

Grover V., Lindberg A., Benbasat I., Lyytinen K.et al [20] (2020). The perils and promises of big data research in information systems. J. Assoc. Inf. Syst. 21:9. This paper addresses the key factors that cause social marketing programs (typically consisting of discrete programs or interventions, but also including broader-scale initiatives) to fail. It argues that understanding these failures offers greater insight to researchers and practitioners than publications solely focused on successes. Focus: this paper discusses the causes of the failure of social marketing programs, an area that has largely been ignored in extant research. Research Question: What causes social marketing programs to fail? Importance: As the majority of practitioner-oriented social marketing research focuses on how to develop a successful program, they identified a tendency to ignore failed programs. they suggested that both researchers and practitioners can arguably learn more useful lessons from failures rather than successes. Thus, this paper contributes to social marketing literature by exploring the key causes of social marketing failures. Methods: they conducted ten semi-structured interviews with social marketing practitioners recruited using a purposive sampling technique.

3. PROPOSED SYSTEM

3.1 pyspark:

PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, large-scale data processing. If you're already familiar with Python and libraries such as Pandas, then PySpark is a good language to learn to create more scalable analyses and pipelines.

Apache Spark is basically a computational engine that works with huge sets of data by processing them in parallel and batch systems. Spark is written in Scala, and PySpark was released to support the collaboration of Spark and Python. In addition to providing an API for Spark, PySpark helps you interface with Resilient Distributed Datasets (RDDs) by leveraging the Py4j library.

The key data type used in PySpark is the Spark data frame. This object can be thought of as a table distributed across a cluster, and has functionality that is similar to data frames in R and Pandas. If you want to do distributed computation using PySpark, then you'll need to perform operations on Spark data frames and no other Python data types.

Py4J is a popular library which is integrated within PySpark and allows Python to dynamically interface with JVM (Java Virtual Machine) objects. PySpark features quite a few libraries for writing efficient programs. Furthermore, there are various external libraries that are also compatible, including:

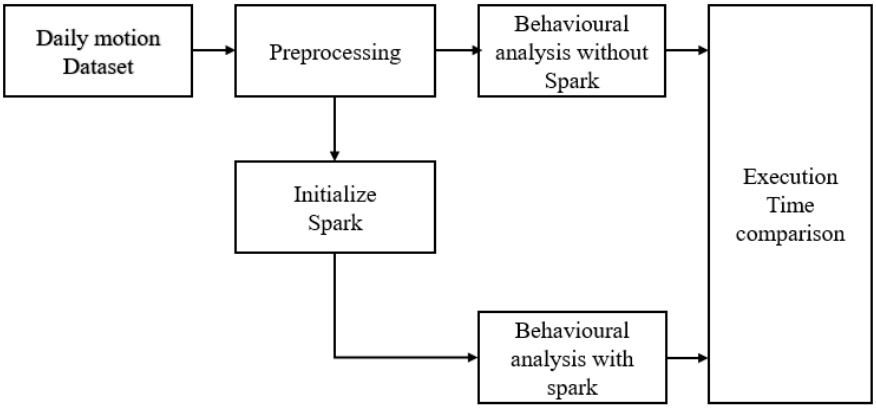


Fig. 1: Block diagram of proposed system.

PySparkSQL - A PySpark library to apply SQL-like analysis on a huge amount of structured or semi-structured data. You can also use SQL queries with PySparkSQL.

MLlib - A wrapper over PySpark and Spark's machine learning (ML) library. MLlib supports many machine learning algorithms for classification, regression, clustering, collaborative filtering, dimensionality reduction, and underlying optimization primitives.

GraphFrames - A graph processing library that provides a set of APIs for performing graph analysis efficiently, using the PySpark core and PySparkSQL. It is optimized for fast distributed computing.

3.2 Apache Spark

Apache Spark is a data processing framework that can quickly perform processing tasks on very large data sets, and can also distribute data processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools. These two qualities are key to the worlds of big data and machine learning, which require the marshalling of massive computing power to crunch through large data stores. Spark also takes some of the programming burdens of these tasks off the shoulders of developers with an easy-to-use API that abstracts away much of the grunt work of distributed computing and big data processing.

ISSN- 2394-5125 VOL 10, ISSUE 03, 2023

From its humble beginnings in the AMP Lab at U.C. Berkeley in 2009, Apache Spark has become one of the key big data distributed processing frameworks in the world. Spark can be deployed in a variety of ways, provides native bindings for the Java, Scala, Python, and R programming languages, and supports SQL, streaming data, machine learning, and graph processing. You'll find it used by banks, telecommunications companies, games companies, governments, and all of the major tech giants such as Apple, IBM, Meta, and Microsoft.is a data processing framework that can quickly perform processing tasks on very large data sets, and can also distribute data processing tasks across multiple computers, either on its own or in tandem with other distributed computing tools. These two qualities are key to the worlds of big data and machine learning, which require the marshalling of massive computing power to crunch through large data stores. Spark also takes some of the programming burdens of these tasks off the shoulders of developers with an easy-to-use API that abstracts away much of the grunt work of distributed computing and big data processing.

From its humble beginnings in the AMP Lab at U.C. Berkeley in 2009, Apache Spark has become one of the key big data distributed processing frameworks in the world. Spark can be deployed in a variety of ways, provides native bindings for the Java, Scala, Python, and R programming languages, and supports SQL, streaming data, machine learning, and graph processing. You'll find it used by banks, telecommunications companies, games companies, governments, and all of the major tech giants such as Apple, IBM, Meta, and Microsoft.

Assumptions of Apache Spark:

Spark makes certain assumptions about the underlying system and data it operates on. Some of

the key assumptions of Spark are:

Data Parallelism: Spark assumes that the data is partitioned across multiple nodes in a cluster, and computations can be performed on each partition in parallel.

Memory-Based Computation: Spark assumes that the data is stored in memory or in a distributed file system, and computations can be performed in-memory for faster processing.

Resilience: Spark assumes that the underlying system may have faults or failures, and provides mechanisms to handle such failures, such as storing data redundantly across nodes and recomputing lost data.

DAG-Based Computation: Spark assumes that the computation can be represented as a Directed Acyclic Graph (DAG) of stages, where each stage contains multiple tasks that can be executed in parallel.

Immutable Data: Spark assumes that the data is immutable and cannot be modified, and instead, new transformations create new RDDs.

Lazy Evaluation: Spark assumes that transformations are evaluated lazily, i.e., they are not executed immediately, but rather when an action is triggered.

Functional Programming: Spark assumes a functional programming paradigm, where operations on RDDs are expressed as transformations and actions on the data.

These assumptions help Spark optimize its execution plan and provide high performance and fault tolerance for large-scale data processing.

4. RESULTS AND DISCUSSION

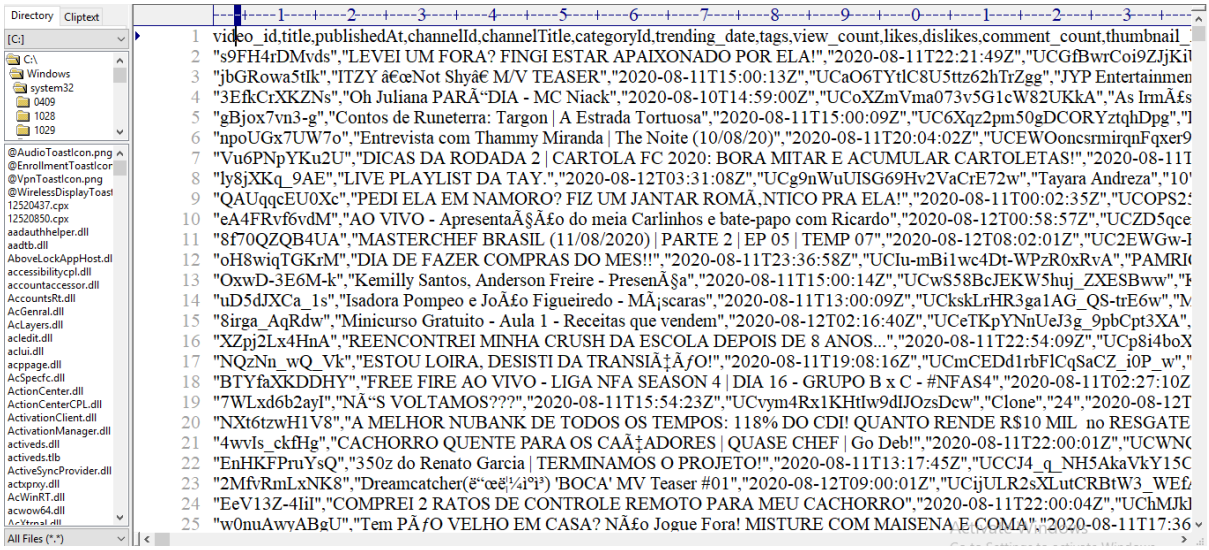
Modules

- Initialize Spark Context: using this module we will initialize SPARK CONTEXT for parallel processing.
- Upload Daily Motion Reviews Dataset: using this module we will upload dataset file path to application.
- Behaviour Analysis without SPARK: using this module we will analyse human behaviour such as their LIKES, DISLIKES from their reviews without using SPARK technology and then capture its execution time.
- Behaviour Analysis with SPARK: using this module we will perform same task of behaviour analysis by using SPARK technology and then capture its execution time.
- Execution Time Comparison Graph: using this module we will plot execution time comparison between without and with SPARK processing.

Now-a-days almost all peoples are using social media to express their views and by analysing this view we can predict person behaviour as their view often describe their personality but this social media contains reviews as TWEETS, POSTS in unstructured format and everyday this unstructured data gather in terabytes and if we want to extract meaningful information such as famous brand, powerful leader, most trending entertainment then this terabytes data processing may take huge time with traditional algorithms so author of this paper employing parallel processing techniques called SPARK.

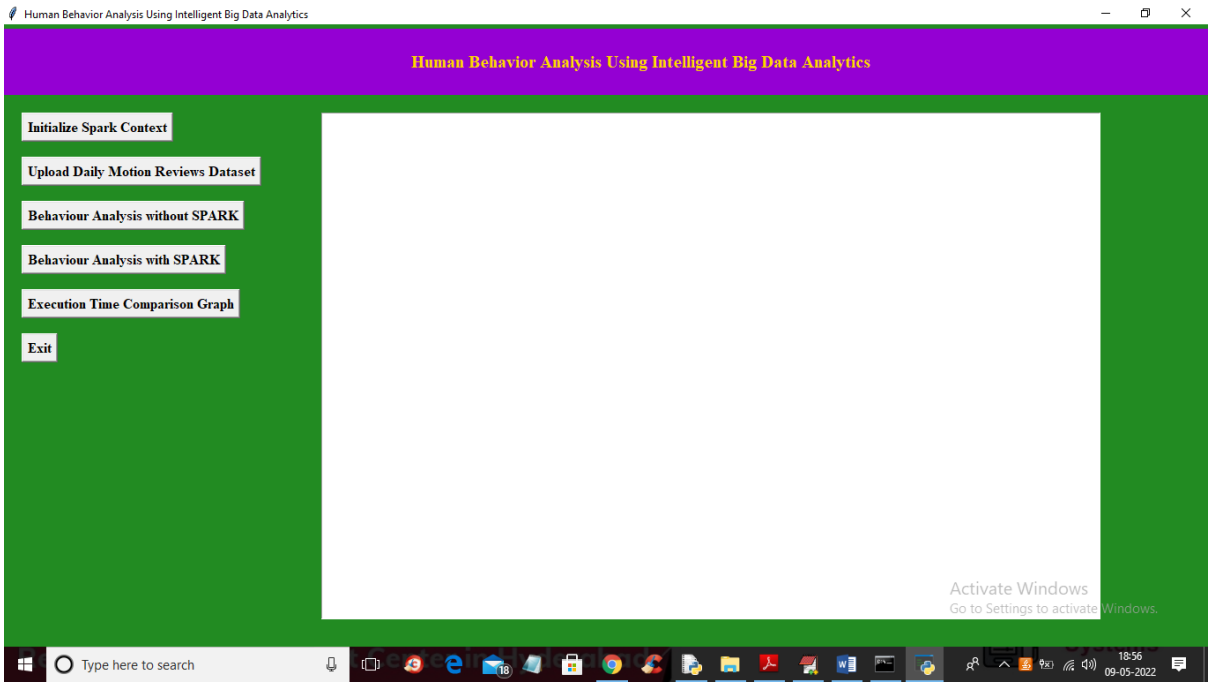
SPARK will create multiple threads and then distribute work between those thread to perform task parallelly and then send result back to SPARK. All existing algorithms works on single thread but SPARK will distribute works in multiple threads so its paralleling processing will be faster and suitable for BIG DATA applications.

So, this project used HIVE, SPARK and HADOOP where HIVE and HADOOP will store data and SPARK will read and process that data. In proposed work, we have gathered reviews from DAILYMOTION website as CSV file and then extracting useful information such as MOST TALK COUNTRIES with many LIKES and then extracting LIKES, VIEW and COMMENTS from so many categories called FASHION, ENTERTAINMENT, NEWS etc. We have compared the execution time of SPARK processing and without spark processing and this experiment proves that SPARK is faster than traditional single thread processing. We are using below CSV dataset of DAILYMOTION website to extract useful information

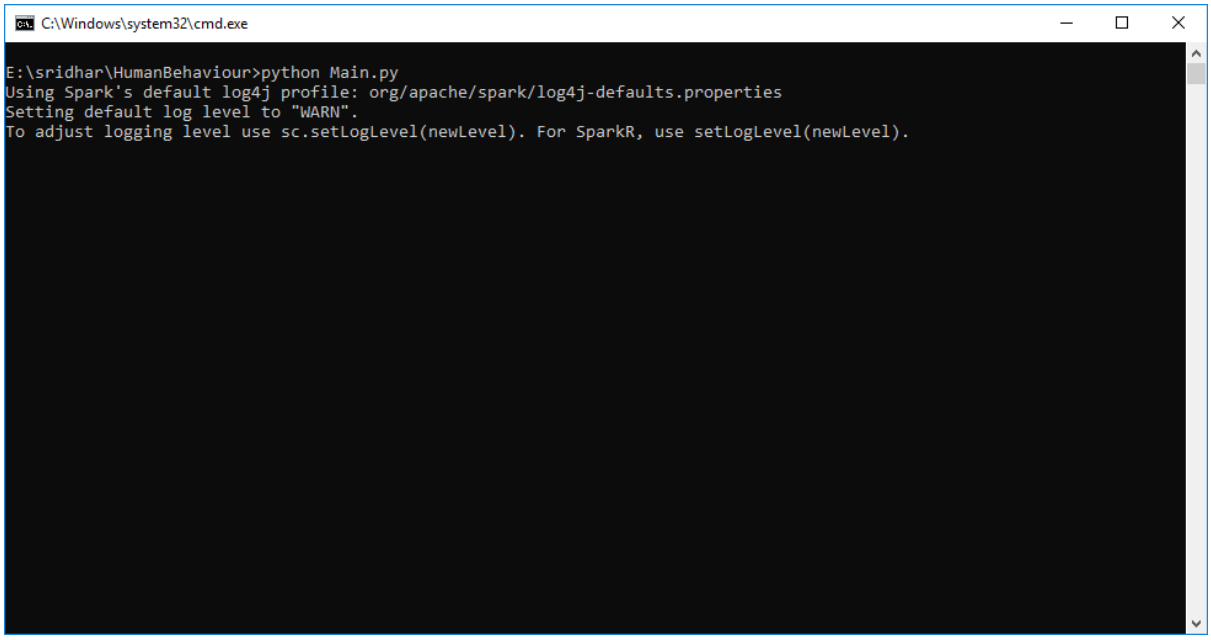


In above dataset screen first row contains dataset column names and remaining are the dataset values.

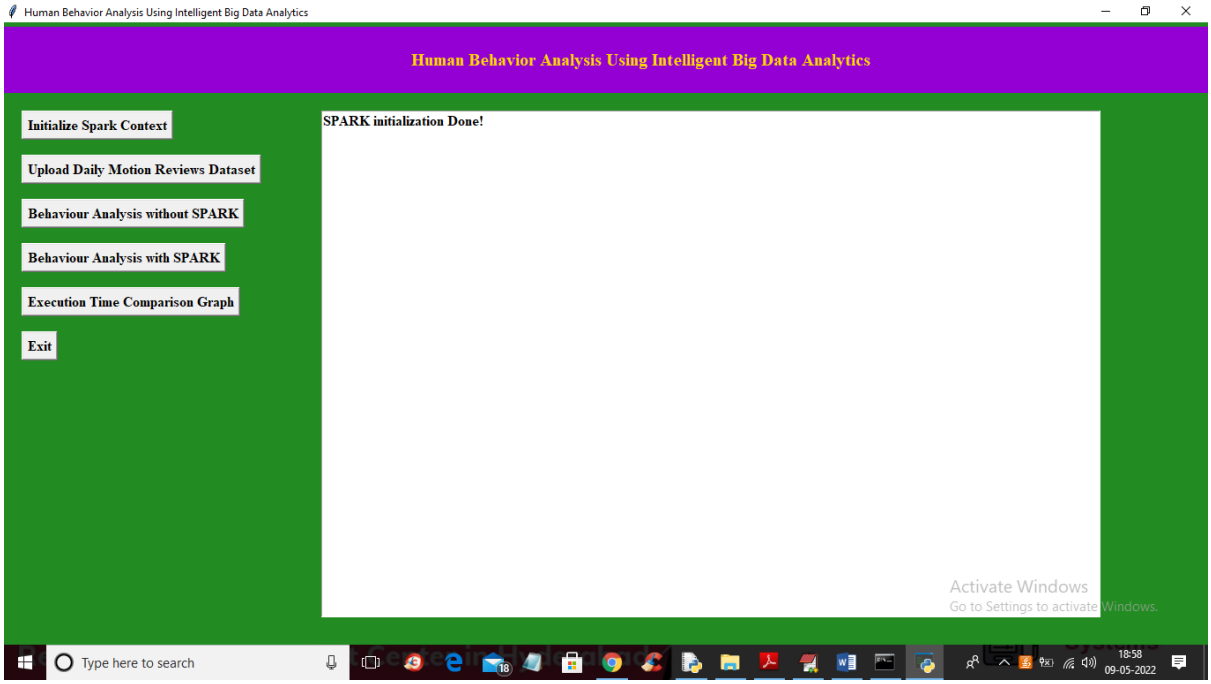
SCREEN SHOTS



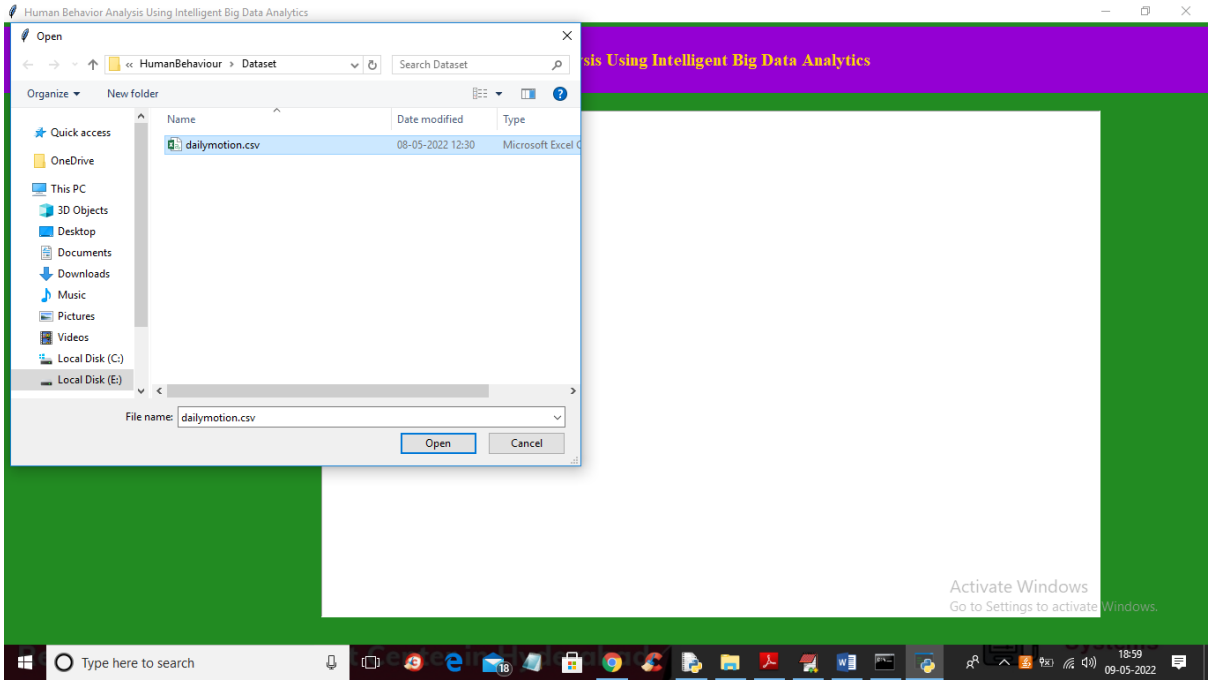
In above screen click on 'Initialize Spark Context' button to setup spark context and get below output after initialization



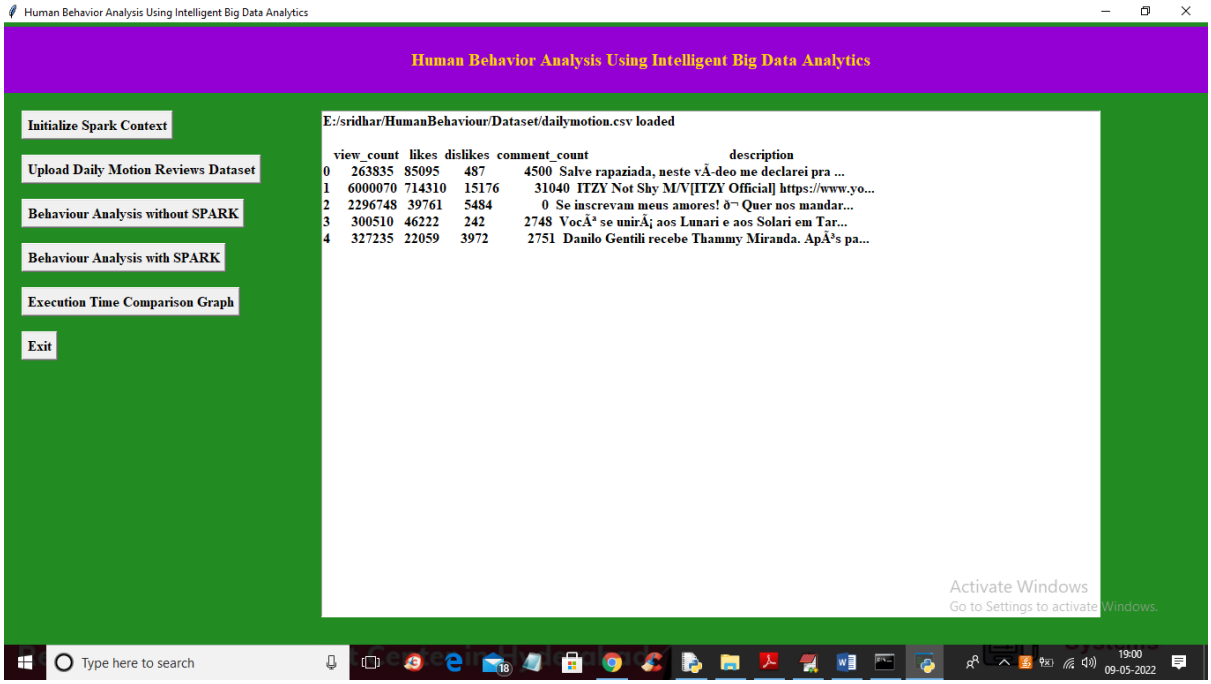
In above screen we can see SPARK object is getting initialized and after initialization will get below output



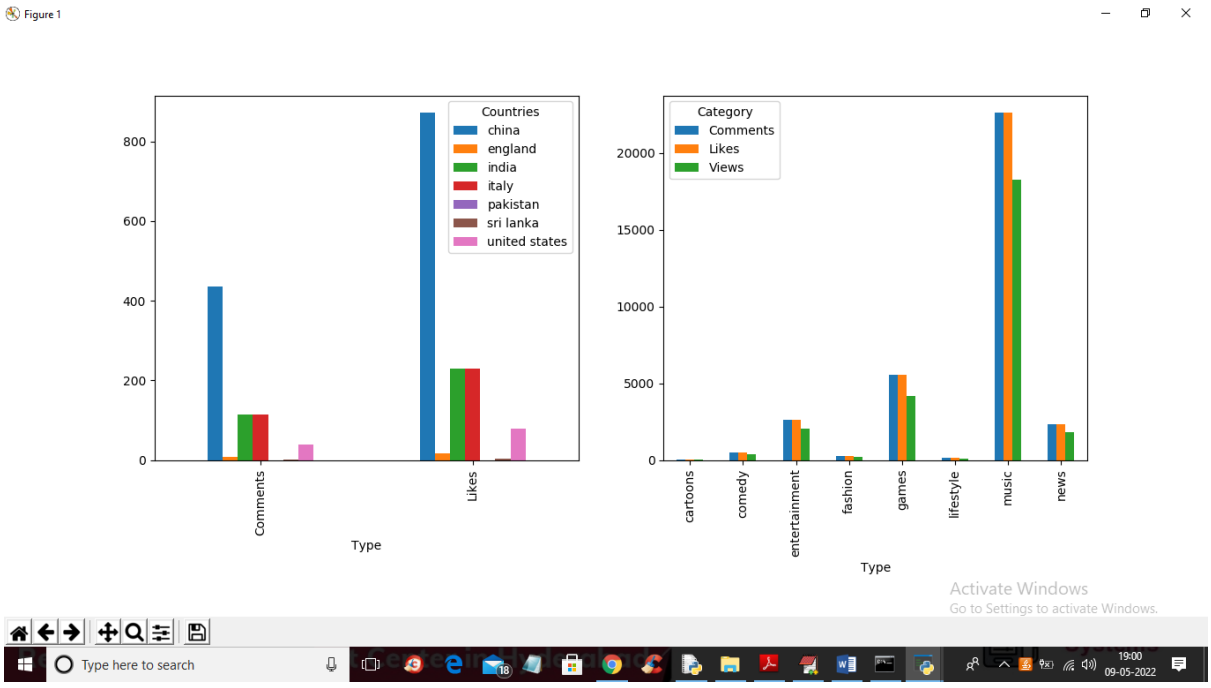
In above screen SPARK context is initialized and now click on 'Upload Daily Motion Reviews Dataset' button to upload dataset and get below output



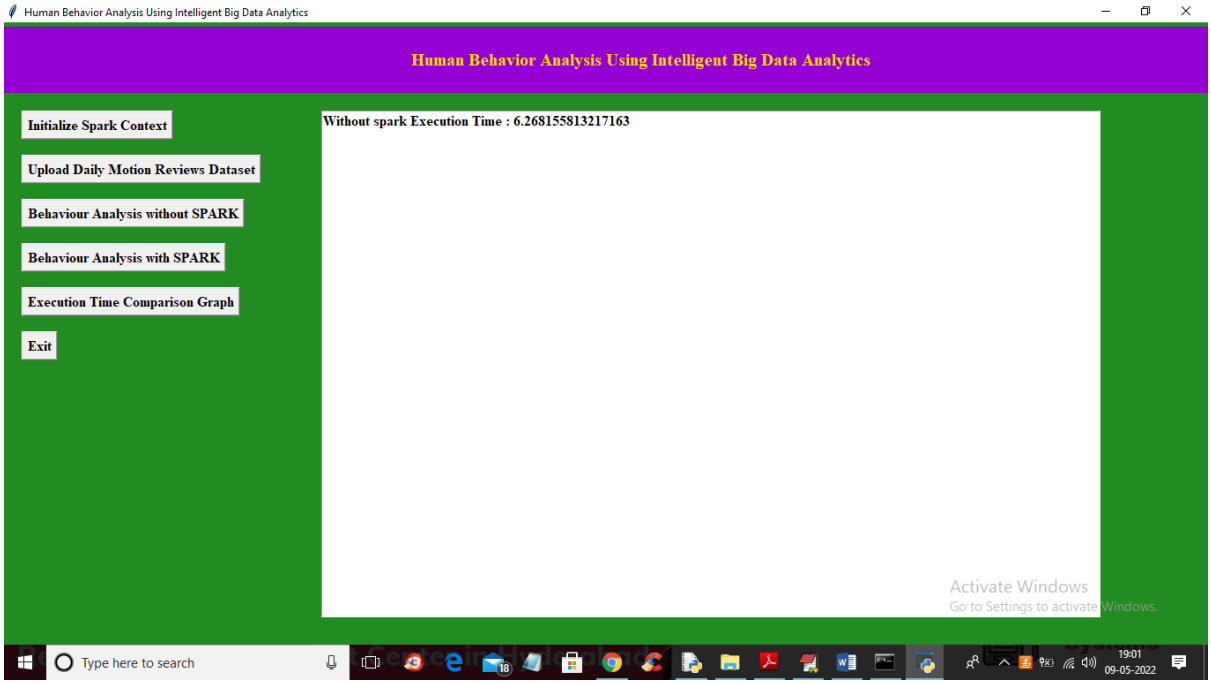
In above screen selecting and uploading dataset file and then click on 'Open' button to load dataset and get below output



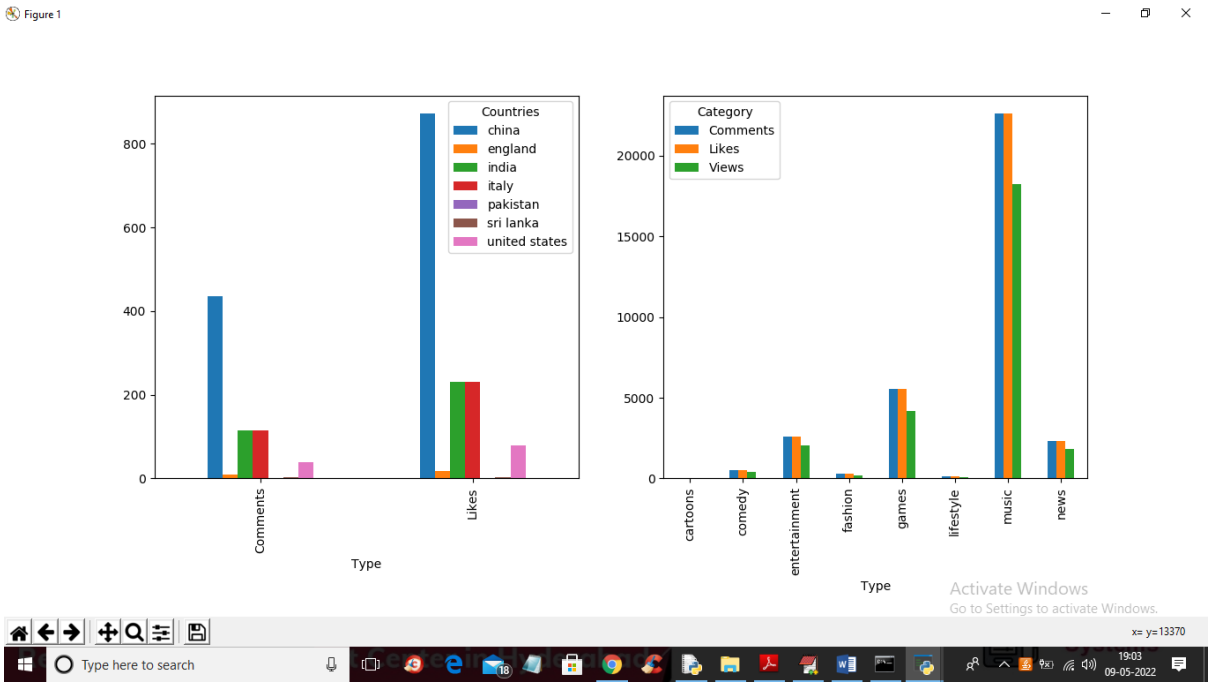
In above screen dataset loaded and now click on 'Behaviour Analysis without Spark' button to get below output



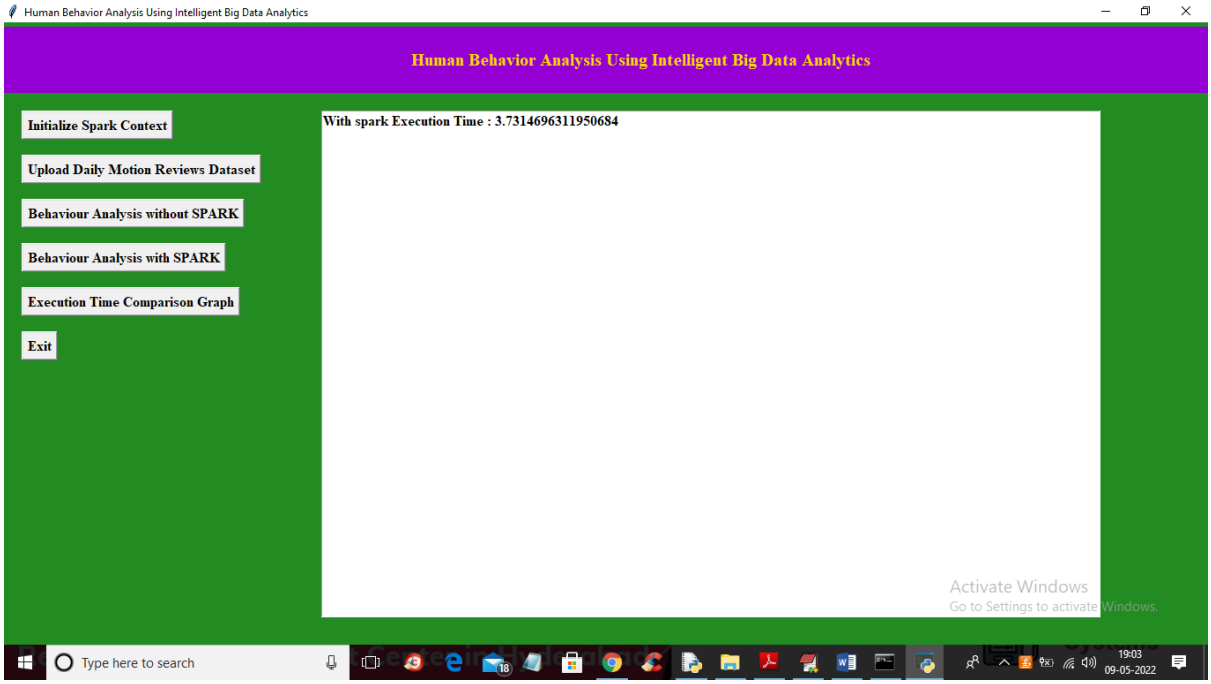
In above graph we have identify behaviour of persons like on which fashion or country they talk most with more LIKES and below screen showing execution time of WITHOUT spark processing



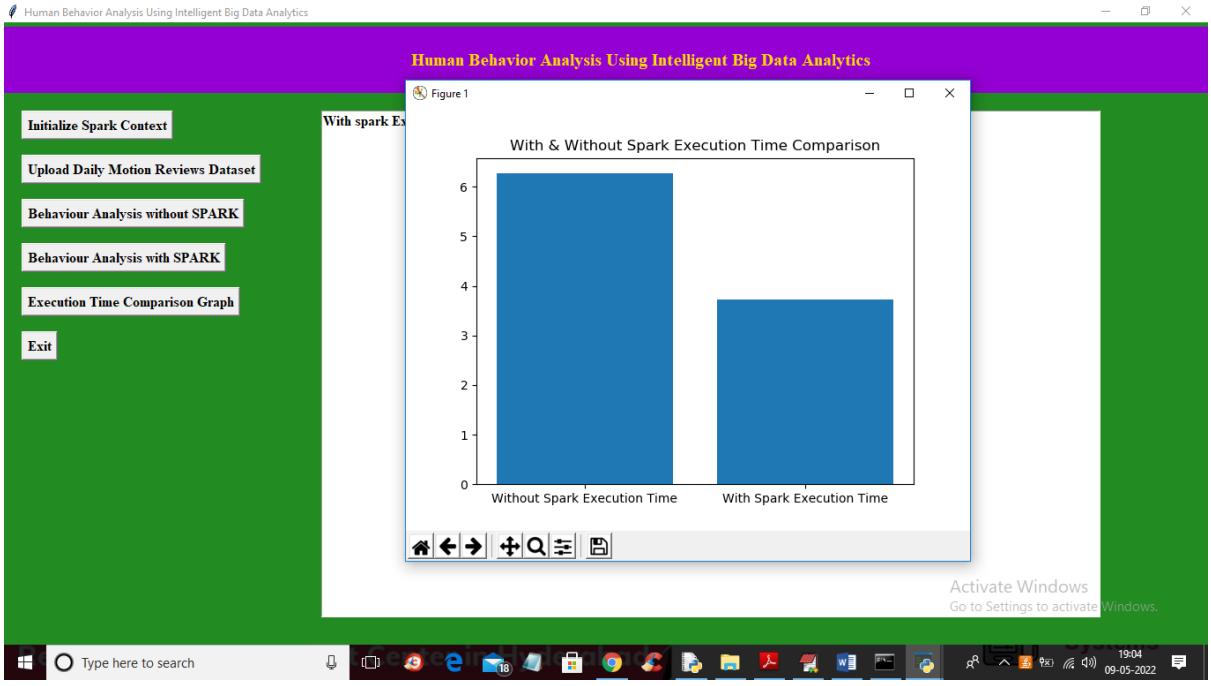
In above screen without spark processing it took 6.26 seconds and now click on 'Behaviour Analysis with SPARK' button to process same data using SPARK and get same output



In above screen with SPARK also we got same output but the difference is execution time and in below screen I am showing SPARK execution to process same data



In above screen SPARK took 3.73 seconds which is lesser than existing single thread processing and now click on 'Execution Time Comparison Graph' button to get below output



In above graph x-axis represents technique names and y-axis represents execution time and we can see SPARK processing took less execution time so it's faster than traditional processing so BIG DATA processing with SPARK can be efficient

5. CONCLUSION

This project proposed an architecture using a big data analytics mechanism to process the huge social media datasets efficiently and logically. In addition, this work employed parallel processing techniques called spark, which will create multiple threads and then distribute work between those thread to perform task parallelly and then send result back to spark. All existing algorithms worked on

single thread but spark will distributed works in multiple threads so its paralleling processing will be faster and suitable for big data applications. This proposed work used hive, spark, and Hadoop where first two will be used to store the data and spark will be used to read and process that data. Here, the dataset of reviews is gathered from Dailymotion website as .csv file and then extracting useful information such as most talk countries with many likes and then extracting likes, view, and comments from so many categories called fashion, entertainment, news etc. Finally, this project compared the execution time of processing with and without spark algorithm.

REFERENCES

[1] P., Chiplunkar N. N. (2018). Real-time twitter data analysis using Hadoop ecosystem. Cogent Eng. 5:1534519. 10.1080/23311916.2018.1534519.

[2] Rodrigues A. P., Rao A., Chiplunkar N. N. (2017). "Sentiment analysis of real time Twitter data using big data approach," in *Proceedings of the 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution*

[3] Blomberg J. (2012). *Twitter and Facebook Analysis: It's Not Just for Marketing Anymore*, Vol. 309. Denver, CO: SAS Global Forum.

[4] Mahalakshmi R., Suseela S. (2015). Big-SoSA: social sentiment analysis and data visualization on big data. *Int. J. Adv. Res. Comp. common. Eng*

[5] Xia Q., Yin X., He J., Chen F. (2018). Real-time recognition of human daily motion with smartphone sensor. *Int. J. Performability Eng.* 14 593–602.

[6] Barros, C. P., and Couto, E. (2013) This paper considers productivity changes in European airlines between 2000 and 2011

[7] Lee N. R., Kotler P. (2011). *Social Marketing: Influencing Behaviours for Good.* Thousand Oaks, CA: Sage Publications.

[8] Cui Y., Kara S., Chan K. C. (2020). Manufacturing big data ecosystem: a systematic literature review. *Robotics Compute. Integr. Manuf.* 62:101861. 10.1016/j.rcim.2019.101861

[9] Grover V., Lindberg A., Benbasat I., Lyytinen K.et al (2020). The perils and promises of big data research in information systems. *J. Assoc. Inf. Syst.* 21:9.

[10] Wang J., Yang Y., Wang T., Sherratt R. S., Zhang J.et al [(2020). Big data service architecture: a survey. *J. Internet Technol.* 21 393–405.