A Narrative Approach for Developing and Configuring BIG DATA Cloud Using SPARK

Avishek Kumar Singh[1], Prajit Paul[2], and Deepak Kumar Singh [3], Jeet Mukherjee[4], Subhadeep Banerjee[5], Shibram Pramanik[6]
[1,2,3,4,5,6]*Asansol Engineering College, Asansol, W.B, India*

***Abstract*—In this paper 'Developing and configuring Big Data cloud using spark' is an attempt to build a cloud cluster for the analysis of big data using spark framework. The aim of this work is to develop and configure a cloud using a number of physical computers to analyze and process big data. For developing the cloud we use 4 physical computers and connect them using LAN. And to analyze and processing of data spark framework will be used. We used Windows Operating System in this work. As the result of this project we will get a system which helps to process and analyze huge amount of data and we will learn how Apache Spark framework helps with big data processing and analytics with its standard API.**

Index Terms: Spark, Hadoop

I. INTRODUCTION

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size. The best examples of bigdata can be found both in the public and private sector. From targeted advertising (Behavior analysis, Profile's segmentation), education, and already mentioned massive industries (healthcare, insurance, manufacturing or banking), to real-life scenarios, in guest service or entertainment (Netflix, Amazon Prime). Big data is processed using some frameworks. Spark, Hadoop, Flink, Storm and Samza are the examples of Big data frameworks. Spark is a framework - in the same way that Hadoop is - which provides a number of inter-connected platforms, systems and standards for Big Data projects . Spark enables applications in Hadoop clusters to run up to 100 times faster in memory and 10 times faster even when running on disk. Spark lets you quickly write applications in Java, Scala, or Python. In our project we will use Spark framework to create a Big data cloud to analyze a huge amount of data. [1] Data which are very large in size is called Big Data. Normally we work on data of size MB like WordDoc , pdf , Excel or maximum GB like Movies, Codes, but data in Peta bytes e is called Big Data. The term not only refers to the data, but also to the various frameworks, tools, and techniques involved.

Apache Spark is an open-source unified analytics engine for large-scale data processing. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.

Originally developed at the University of California, Berkeley's AMPLab in 2009, the Spark codebase was later donated to the Apache Software Foundation, which has maintained it since.

It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing. You'll find it used by organizations from any industry, including at FINRA, Yelp, Zillow, DataXu, Urban Institute, and CrowdStrike. Apache Spark has become one of the most popular big data distributed processing framework with 365,000 meetup members in 2017

Latest versions

i. Spark 3.1.1 (Mar 02 2021)
ii. Spark 3.0.2 (Feb 19 2021)
iii. Spark 2.4.7 (Sep 12 2020)

With increasing popularity of spark there is a comparison arises between spark and Hadoop . Hadoop is also a big data framework as Spark but Spark included in most Hadoop distributions these days. Due to two big advantages, Spark has become the framework of choice when processing big data,

overtaking the old MapReduce paradigm that brought Hadoop to prominence.

A cluster manager is used to acquire cluster resources for executing jobs. Spark core runs over diverse cluster managers including Hadoop YARN, Apache Mesos, Amazon EC2 and Spark's built-in cluster manager (i.e., standalone). The cluster manager handles resource sharing between Spark applications. On the other hand, Spark can access data in HDFS, Cassandra.20 HBase, Hive, Alluxio and any Hadoop data source.

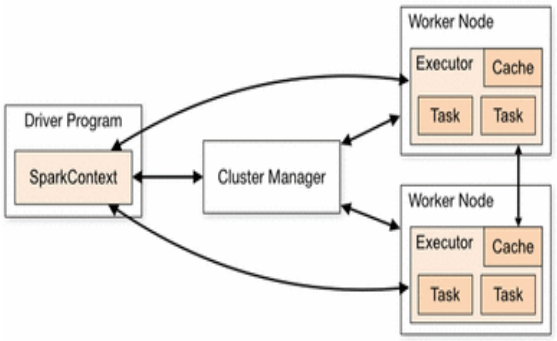

Fig-1 Spark Architecture

II. HELPFUL HINTS

A. Figures

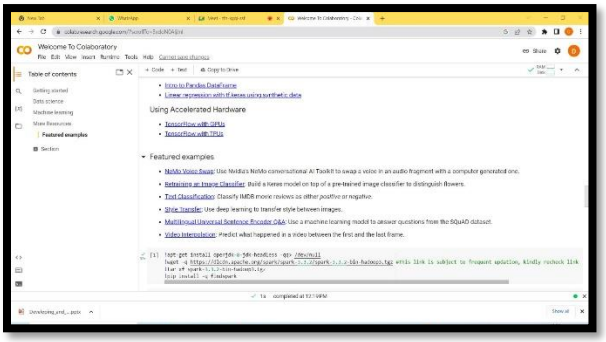

Fig-2 Installing of Java and downloading Of SPARK

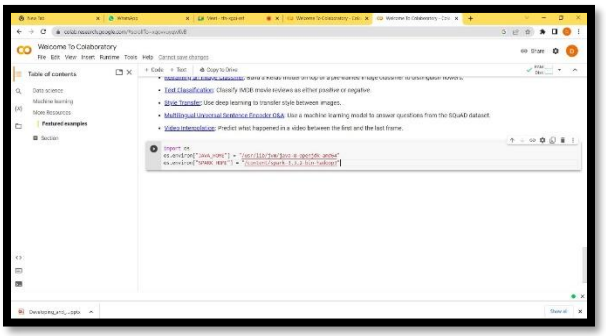

Fig-3 Importing and installing OS and setting up environment for SPARK and Java

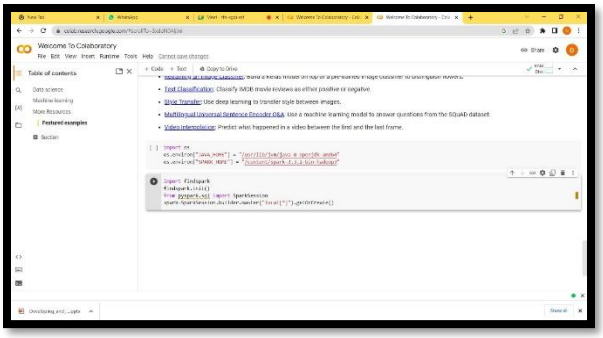

Fig-4 Setting up Py-Spark and Importing Spark Session

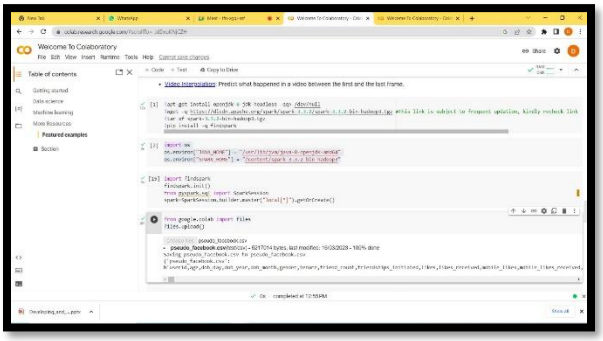

Fig-5 Uploading of Big Data file (pseudo_facebook.csv)

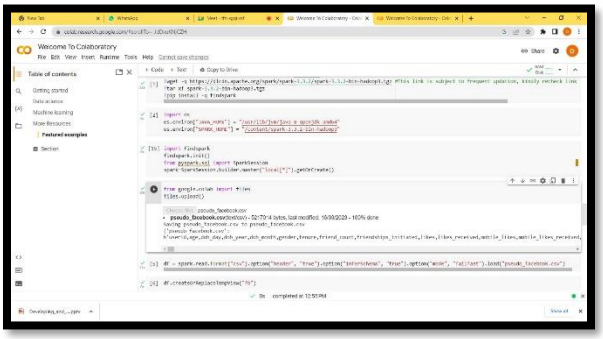

Fig-6 Creating of data frame

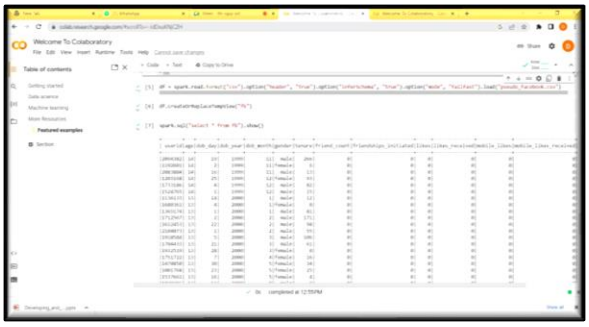

Fig-7 Displaying of Data frame

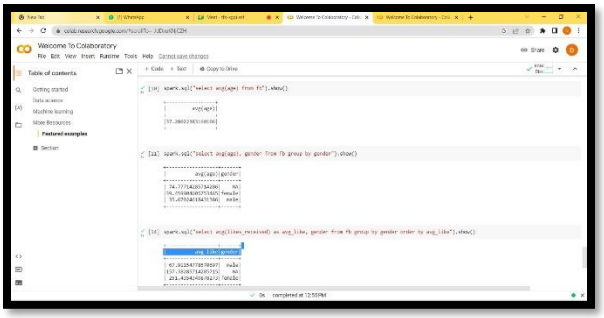

Fig-8 Observation
1.Average age group using face book
2.Average age group using face book as per
3.Average age group receiving likes as per gender.

IV. CONCLUSION

The Spark ecosystem is growing day by day with new features being added. Also the industry adoption is also increasing at a very fast pace as Spark has a lot to offer to the Big Data and Data Science world. In this project we will try to develop a Big data cluster using a number of physical computers and as the result we will be able to analyze huge amount of data using Apache Spark framework. By working on this project we able to learn more about Big Data and also how Spark works. We require number of physical computers for this project but we have only limited resource, as of now we simulates the project and we will try to implement it properly when we get proper resources

III.ACKNOWLEDGMENT

It is my great privilege to express my profound and sincere gratitude to our Supervisor, Prajit Paul for providing me a very cooperative and precious guidance at every stage of the present project work being carried out under his/her supervision. His valuable advice and instructions in carrying out the present study has been a very rewarding and pleasurable experience that has greatly benefited me throughout the course of work.

We would like to convey my sincere gratitude towards Dr Kuntal Ghosh, Head of the Department of Electronics & Communication Engineering, Asansol Engineering College for providing us the requisite support for time completion of our work. We would also like pay my heartiest thanks and gratitude to all the teachers of the Department of Electronics & Communication Engineering, Asansol Engineering College for various suggestions being provided in attaining success in our work.

I would like to express my earnest thanks to my other colleagues along with all technical staffs of the Department of Electronics & Communication Engineering, Asansol Engineering College for their valuable assistance being provided during my project work.

Finally, I would like to express my deep sense of gratitude to my parents for their constant motivation and support throughout my work.

V. FUTURE SCOPE

The future use of the project can vary depending on the context and the specific goals of the project. Here are some potential future uses:

1) Production deployment: Once the project is completed, it can be deployed in a production environment to process and analyze large volumes of data in real-time. The Spark cluster can be scaled up or down based on the data processing needs, and the application can be fine-tuned to optimize performance.

2) Optimization and enhancement: The Spark application can be continuously optimized and enhanced to improve its performance and scalability. This can involve fine-tuning the Spark configuration, optimizing the data processing algorithms, and adding new features to the application.

3) Data-driven decision-making: The insights gained from processing and analyzing large volumes of data can be used to inform data-driven decision-making. The Spark application can be integrated with other tools and systems to provide real-time insights and automate decision-making processes.

4) Research and development: The Spark application can be used as a basis for research and development in various fields, such as machine learning, natural language processing, and image processing. The application can be extended to support new data processing tasks and algorithms, and new data sources can be added to the Spark cluster.

Overall, the future use of the project depends on the specific goals and context of the project. The project can be continuously optimized and enhanced to support a

wide range of data processing tasks and decision-making processes

REFERENCE

[1] https://www.guru99.com/what-is-big-data.html Date-19 March 2023

[2] https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/ Date-8 January 2012

[3] https://www.rfwireless world.com/Terminology/ Advantages-and-Disadvantages-of-Big-Data. html

[4] https://www.edureka.co/blog/Spark architecture/ Date- 21 July 2015

[5] https://www.knowledgehut.com/blog/big-data/ apache-spark- advantages-disadvantages Date- 20 June 2020

[6] https://www.upgrad.com/blog/apache-spark-applications-usecases/ Date- 19 December 2016

[7] Getting started with SQL-by Thomas Nield Date-16 September 2012

[8] Introduction to Hadoop, Spark by Raj Kamal | Preeti Saxena Date- 17 August 2014

[9] Fundamental of Data Engineering by Joe Reis & Matt Housley Date-15 August 2016

[10] https://spark.apache.org/docs/latest/ Date- July 28 2013

[11] https://www.inderscience.com/jhome.php?jcode =ijbdi/ Date-June 12 2017

[12] Agarwal, S., Mozafari, B., Panda, A., Milner, H., Madden, S., Stoica, I.: Blinkdb: Queries with bounded errors and bounded response times on very large data, Proceedings of the 8th ACM European Conference on Computer Systems. ACM, New York, pp 29–42 (2013). doi: 10.1145/2465351.2465355 Date-May 30 2019

[13] Journal of Big Data: https://journal of big data. springeropen.com/ [Accessed January 23 2019]

[14] International Journal of Big Data Intelligence: https://www.inderscience.com/jhome.php?jcode =ijbdi[Accessed September 15, 2017].

[15] https://www.computer.org/csdl/journal/tb[Accessed August 19,2016].