International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 04 Issue: 10 | Oct -2017 www.irjet.net p-ISSN: 2395-0072

Review on apache spark technology

Kalyani K. Pathrikar[1], Prof. Arundhati A. Dudhgaonkar[2]

[1]PG Student, Department of Master of computer Applications. MGM's JNEC, Aurangabad, Maharashtra.
[2.] Assistant professor, Department of Master of computer Applications. MGM's JNEC, Aurangabad, Maharashtra.

Abstract - *Apache spark is general purpose cluster computing system. It is faster than hadoop because of main memory processing .Apache spark is an open source big data processing framework constructed around the speed ,accessibility and sophisticated use. Apache spark is lightning-fast cluster computing designed for fast computation.*

***Key Words*: Apache spark, Big Data, MapReduce, RDD, Hadoop.**

1.INTRODUCTION

Apache spark is developed in 2009 in UC Berkeley's AMPLab, and open sourced in 2010 as an Apache project. Spark engine is established in memory processing as well as disk base processing. [4]

It run on top of the existent hadoop cluster and access hadoop data store [HDFC] can also development structured data in hive and streaming data from HDFC, twitter.

1.1 Enlargement of apache spark

As apposite a common belief spark is not a alter version of hadoop & is not actually dependent on hadoop because it has its own cluster management.
Spark uses hadoop in 2 ways.

1] Storage.
2] Processing.

The main feature of apache spark is in main memory cluster calculated that to grow processing speed of application. [3]

1.2 Why we use apache spark

Following are important reason to use apache spark.

1. Apache Spark is a quick and general purpose engine for large-scale data processing.
2. It is a full, top level Apache project.
3. It is easy to install.
4. It is execute in Scala, which is power full object oriented languages.
5. It is able than old hadoop map reduce.
6. Popular feature of apache spark is capable to join dataset across many contrasting data source.[2]

1.3 Feature of apache spark:-

There are following feature

1] Speed: - It is 100 times faster in memory, and 10 times faster when executing on disk. It is use idea of a RDD [Resilient Distributed Dataset].

2] Easy to use:-Developer writes applications in Java, Scala, or Python. Developer can formulate & run their application on their conventional programming language.

3] Map reduces: - It support SQL quires.

4] Provide multiple languages: - Spark support built in API in java, python, Scala.

5] It run everywhere:-Apache spark run on hadoop, cloud. [1]

1.4 Apache spark architecture:-

There are three main components

1] Data storage:-

Spark uses hadoop distributed file system [HDFC] for data storage purpose It work with any hadoop appropriate data source including HDFC.

2] Compute interface:-

API support application developer to create Spark based application using standard API interface .Spark support API for Scala, java & python programming language.

3] Resource management:-

Spark can be arranging as a standalone server or it can be on a distributed computing framework like YARN. [4]

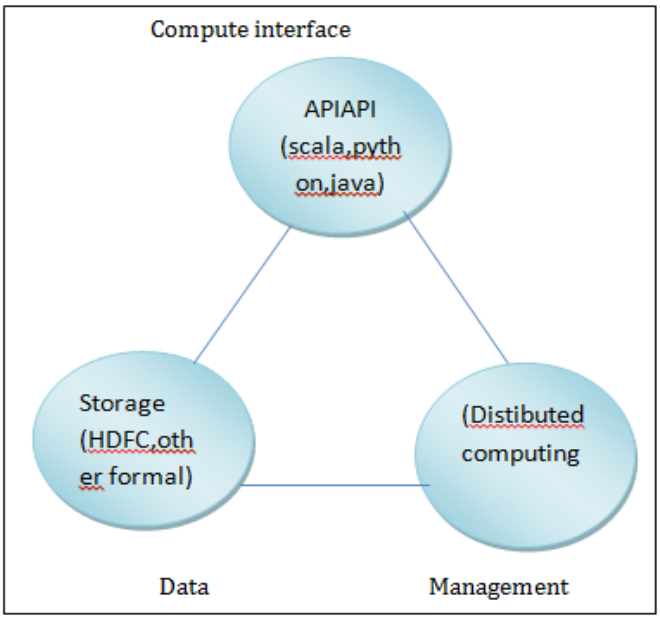

Fig: Apache spark architecture

1.5 Data processing in apache spark

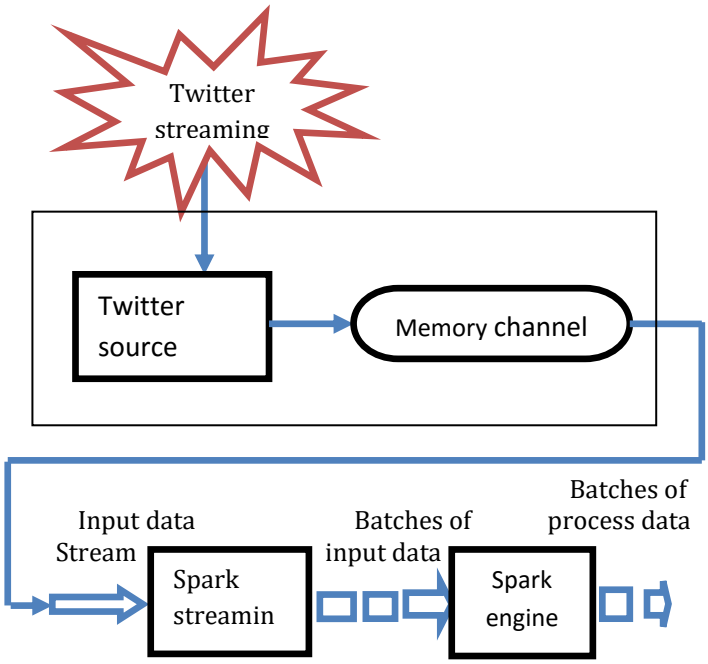

Fig -2: Apache spark data processing

In this huge amount of data and process analyze huge volume of data to concentrate some meaningful information. Following fig. shows data processing in happening using twitter streaming API and apache spark.

Twitter streaming is used to access twitter big data using apache spark .The amount of data proceed of each scenario processing time & result is represent in tabular format or graphical format.[2]

1.6 Component of apache spark

Apache spark is general purpose cluster calculating engine which is very fast & reliable. There are following five components.

1] Spark SQL:-

It provides structure & semi structure data. It is top of the spark cores that propose new data abstraction called schema RDD. [1]

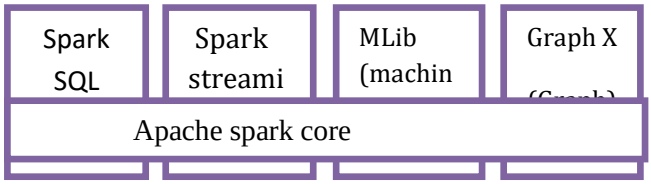

Fig: Apache spark component

2] spark streaming:-

Spark streaming can be used for processing the real time streaming data. It fast scheduling capabilities to preform streaming analytics. It divide data into mini batches and it perform RDD transaction on those data. [1]

3] MLib [Machine learning libraries]:-

Spark Mlib is 9 times faster as a hadoop disk-based version of apache mahout. MLib is scalable machine learning library consisting of common learning algorithm & utilities ,including classification regression, clustering. [1]

4] Graph X:-

It is distributed graph-processing framework on top of spark.[1]

5] Apache spark core:-

Spark core is the general execution engine for spark platform that all other functionality is built is built upon. [1]

1.6 Spark major use cases over hadoop

1. Iterative algorithm in machine learning.
2. Iterative data mining and data processing.
3. Spark is fully apache hive compatible data warehousing system that can run 100x faster than hive.

4. Stream processing long processing and fraud detection in live streams for alert aggregation & analysis. [2]

2. CONCLUSIONS

In this paper we studied feature apache spark, apache spark architecture & component of spark technology. In additional we have processing how to process data in apache spark. Apache spark is an effective open source processing engine built around speed, easy to use, elaborate analytics. [1] In this paper we have cover the brief description of spark & working spark major use case over hadoop. In spark technology is use in twitter.

REFERENCES

1. A Review Study of Apache Spark in Big Data Processing.
2. Big Data Analysis: Ap Spark Perspective.
3. http://spark.apache.org/
4. Big Data Processing with Apache Spark