A Review On Spark VS Hadoop which is better for Big Data Analytics

Introduction

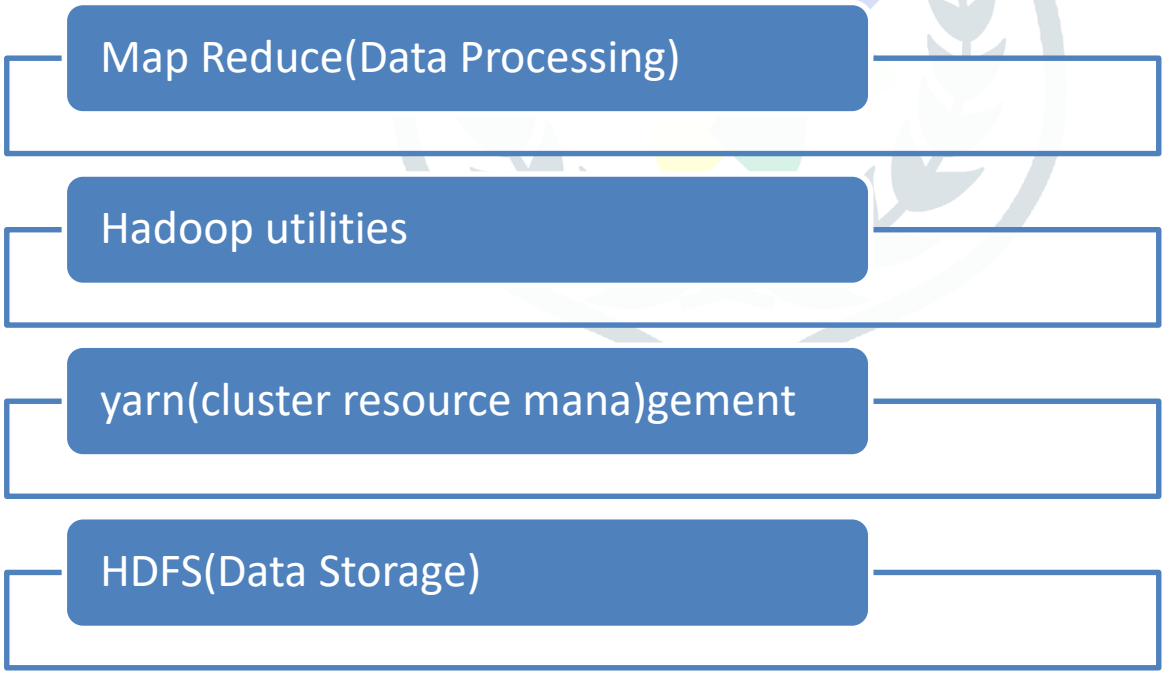
The main focus of this paper is to compare the performance between Hadoop and Spark on some applications, such as iterative computation and real-time data processing. The runtime architectures of both Spark and Hadoop will be compared to illustrate their differences, and the components of their ecosystems will be tabled to show their respective characteristics. In this paper, we will highlight the performance comparison between Spark and Hadoop as the growth of data size and iteration counts, and also show how to tune in Hadoop and Spark in order to achieve higher performance. and how to verify the correctness of the running results. Inthis chapter, an overview of Hadoop and Spark is introduced to get a basic understanding of their frameworks, **including** their key component sand how data flows in MapReduce and Spark respectively. Next, their runtime architectures are dissected to better comprehend how an application works in Hadoop and Spark separately.In addition, their ecosystems are illustrated to show the characteristic and functionality of each element.

The Overview of Hadoop and Spark

Hadoop

Hadoop is a framework for big data analytics.and use distributed processing of big data.across clusters of computers.Apache Hadoop mainly consist of four core components:--

Map reduce,common utilities,yarn(Yet Another Resource Negotiator) and HDFS(Hadoop Distributed File System)



- **MapReduce:--** is a programming model which provides support for parallel computing,locality-aware scheduling, fault-tolerance, and scalability on commodity clusters [2].MapReduce separates the data processing into two stages: the Map stage and the Reduce stage.MapReduce is a programming model suitable for processing of huge data. Hadoop is capable of running MapReduce programs written in various languages: Java, Ruby, Python, and C++
  - Map phase
  - Reduce phase.

the whole process goes through four phases of execution namely, splitting, mapping, shuffling, and reducing.

- Map tasks (Splits & Mapping)

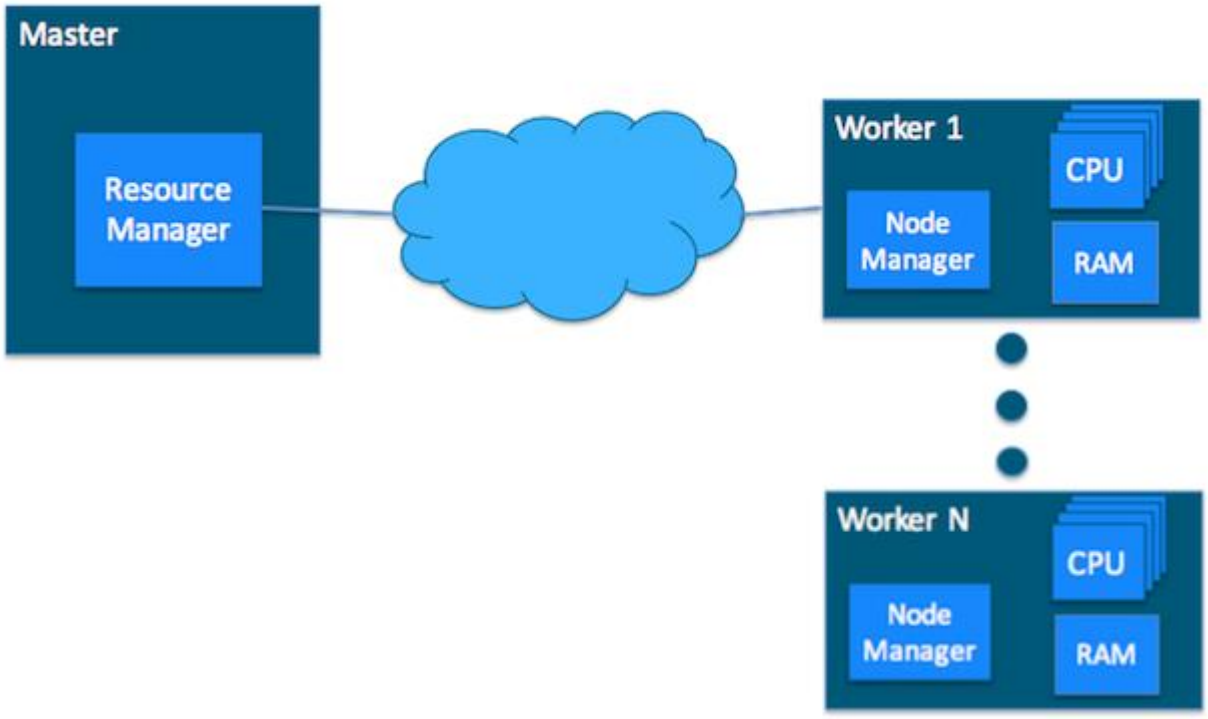
- Reduce tasks (Shuffling, Reducing)

The complete execution process (execution of Map and Reduce tasks, both) is controlled by two types of entities called a Jobtracker: Acts like a master (responsible for complete execution of submitted job)Multiple Task Trackers: Acts like slaves, each of them performing the job For every job submitted for execution in the system, there is one Jobtracker that resides on Namenode and there are multiple tasktrackers which reside on Datanode.

❖ **YARN (Yet Another Resource Negotiator):--** is the resource management layer for the Apache Hadoop ecosystem. YARN is a cluster resource management framework in Hadoop.

In a YARN cluster, there are two types of hosts:

1. The *ResourceManager* is the master daemon that communicates with the client, tracks resources on the cluster, and orchestrates work by assigning tasks to *NodeManagers*.
2. A *NodeManager* is a worker daemon that launches and tracks processes spawned on worker hosts.

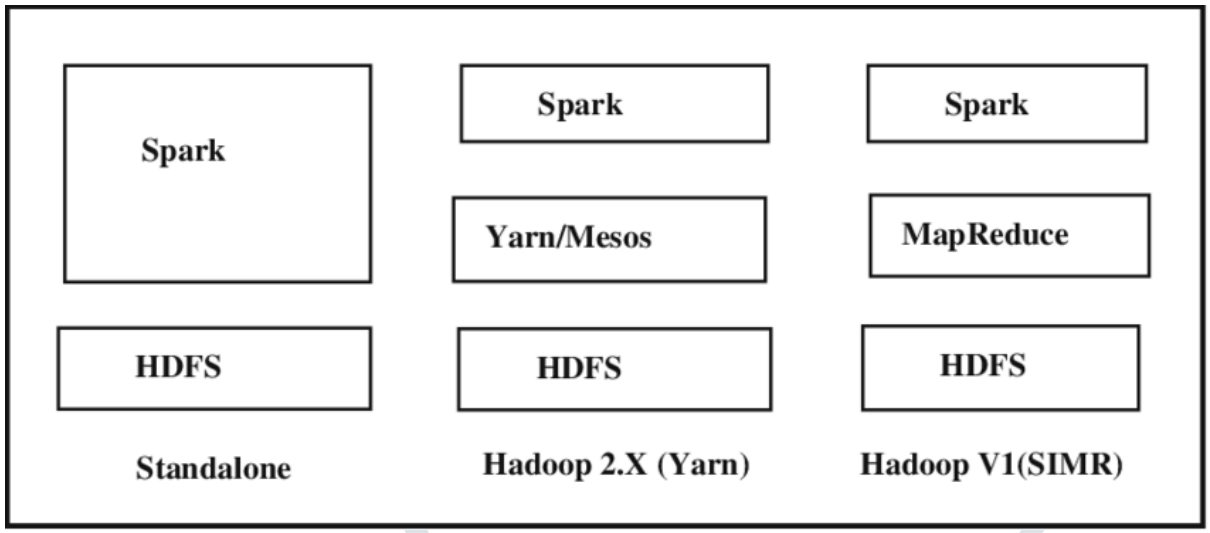


❖ **HDFS (Data Storage)** is a file system which stores big data, links data blocks logically, and streamsdata at high bandwidth to applications in a distributed system [6]. It separates file systemmetadata from application data. The former are presented on NameNode, and the later arestored on DataNode. Also, HDFS replicates data across clusters to achiseve reliability in case of failure of nodes. Hadoop is believed to be reliable, scalable, and fault-tolerant. It is well known that MapReduce is a good fit for applications of processing big data, but it is a poor fit for iteration

❖ **Apache Spark:--**Apache Spark is a lightning-fast cluster computing designed for fast computation. It was built on top of Hadoop MapReduce and it extends the MapReduce model to efficiently use more types of computations which includes Interactive Queries and Stream Processing. This is a brief tutorial that explains the basics of Spark Core programming. park was introduced by Apache Software Foundation for speeding up the Hadoop computational computing software process.Spark is not a modified version of Hadoop and is not, really, dependent on Hadoop because it has its own cluster management. Hadoop is just one of the ways to implement Spark.Spark uses Hadoop in two ways – one is storage and second is processing. Since Spark has its own cluster management computation, it uses Hadoop for storage purpose only Spark is one of Hadoop's sub project developed in 2009 in UC Berkeley's AMPLab by Matei Zaharia. It was Open Sourced in 2010 under a BSD license. It was donated to Apache software foundation in 2013, and now Apache Spark has become a top level Apache project from Feb-2014.

❖ Spark Built on Hadoop

There are three ways of Spark deployment as explained below



- **Standalone** − Spark Standalone deployment means Spark occupies the place on top of HDFS(Hadoop Distributed File System) and space is allocated for HDFS, explicitly. Here, Spark and MapReduce will run side by side to cover all spark jobs on cluster.
- **Hadoop Yarn** − Hadoop Yarn deployment means, simply, spark runs on Yarn without any pre-installation or root access required. It helps to integrate Spark into Hadoop ecosystem or Hadoop stack. It allows other components to run on top of stack.
- **Spark in MapReduce (SIMR)** − Spark in MapReduce is used to launch spark job in addition to standalone deployment. With SIMR, user can start Spark and uses its shell without any administrative access.

❖ sPerformance Measure and Matrices of Spark And Hadoop

Key Points	Hadoop	Spark
Data Processing	Basic data processing engine	Data analytics engine
Usage	Batch processing with huge volume of data	Process real time data from real time events like twitter and facebook
Latency	High latency computing	Low latency computing
Data	Process data in batch mode	Can process interactively
Ease of Use	Hadoop map reduce model is complex,need to handle low	Easier to use,abstraction enables a user to process data

	level apis	using high level operators.
Scheduler	External Job Scheduler Is required	In memory computation,no external scheduler is required
Security	Highly Secured	Less Secure As Compare to Hadoop
Cost	Less Costly	Costlier than Mapreduce since it has in memory solution

The Ecosystem of Hadoop And Spark

Ecosystem / Elements	HADOOP	SPARK
Distributed File System	HDFS,FTP file system,windows,Amazon s3	HDFS, Cassandra, Amazon-S3, Kudu
Distributed Resource Management	YARN framework	It is adaptable. YARN, Mesos, or Built-in Standalone Manager which provides the easiest way to run applications on a cluster
SQL Query	HIVE: A data warehouse component	SPARK SQL
Machine Learning	Mahout: A Machine learning component	MLib
Stream Processing	Storm: real-time computational Engine	Spark Stream
Graph Processing	Giraph: A framework for large-scale graph processing	GraphX
Management Interface	ZooKeeper: A management tool for Hadoop cluster.	No support
Stream tool	Flume: a service for efficiently transferring streaming data into the Hadoop Distributed File System	No support

Data Flow Processing	Pig: a high level scripting data flow language which expresses data flows by applying a series of transformations to loaded data [12].	No support
NoSQL database	HBase: based on BigTable, and column-oriented	No support

Conclusion

if your project is based on structured data such as customer's names and addresses, Hadoop would suffice. Your job would be done at reduced price at without any external need of installing Spark over Hadoop which just means more cost and time. Besides, Spark's security and support needs more betterment.The conclusion can be it would be the best if you can use both Spark and Hadoop at a time to gain from the faster processing speed of Spark, its advanced analytics and excellent integration support, with the cost effectiveness of Hadoop. Hadoop and Spark make an umbrella of components which are complementary to each other. Spark brings speed and Hadoop brings one of the most scalable and cheap storage systems which makes them work together. They have a lot of components under their umbrella which has no well-known counterpart. Spark has a popular machine learning library while Hadoop has ETL oriented tools. However, Hadoop MapReduce can be replaced in the future by Spark but since it is less costly, it might not get obsolete. **Spark** has overtaken **Hadoop** as the most active open source **Big Data** project. While they are not directly comparable products, they both have many of the same uses. ... Although **Spark** is reported to work up to 100 times faster than **Hadoop** in certain circumstances, it does not provide its own distributed storage system.

References :

1..Varsha B.Bobade,"Survey Paper on Big Data and Hadoop",International Research JournalofEngineering and Technology (IRJET) ,Volume:03 Issue: 01 | Jan-2016, e-ISSN: 2395-0056 p-ISSN: 2395-0072.

2.S. Justin Samuel, Koundinya RVP, Kotha Sashidhar and C.R. Bharathi, "A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES",VOL. 10,NO.8, MAY 2015 ISSN 1819-6608,ARPN Journal of Engineering and Applied Sciences.

3.Ms. Vibhavari Chavan, Prof. Rajesh. N. Pursue"Survey Paper On Big Data",Vibhavari Chavan et al, / (IJCSIT) International Journal ofComputer Science and Information Technologies, Vol. 5 (6) ,2014, 7932-7939.

4.Ankush Verma ,AshikHussain Mansuri ,Dr.Neelesh Jain "Big Data Management Processing with Hadoop MapReduce and Spark Technology: A Comparison" 2016 Symposium on Colossal Data Analysis and Networking (CDAN) ,978-1-5090-0669-4/16/$31.00 © 2016 IEEE.

5.Wei Huang, Lingkui Meng, Dongying Zhang, and Wen Zhang, "In-Memory Parallel Processing of Massive Remotely Sensed Data Using an Apache Spark on Hadoop YARN Model" ,IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONSAND REMOTE SENSING, VOL. 10, NO. 1, DECEMBER 2016

.6.Katarina Grolinger, Michael Hayes, Wilson A. Higashino, Alexandra L'Heureux1David S.Allison ,MiriamA.M. Capretz, "Challenges for MapReduce",978-1-4799-5069-0/14 $31.00©2014IEEEDOI10.1109/SERVICES.2014.4.

7.Xiuqin LIN, Peng WANG, Bin WU, "LOG ANALYSIS IN CLOUD COMPUTING ENVIRONMENT WITH HADOOP AND SPARK",978-1-4799-0094-7/13/$31.00©2013

8.K..Naga Maha Lakshmi et al., International Journal of Computer Engineering In Research Trends,Volume3, Issue 3, March-2016, pp. 134-142

.9. Sunil B. Mane et.al, "Product Rating using Opinion Mining", International Journal of Computer Engineering In Research Trends, 4(5):161-168,may 2017