International Journal of Engineering & Technology, *7 (4) (2018) 4839-4843*

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet. v7i4.25997
Research paper

Study With Comparing Big-Data Handling Techniques using Apache Hadoop Map Reduce VS Apache Spark

N. Deshai[1]*, B.V.D.S.Sekhar[1] ,S.Venkataramana[1], V.V.S.S.S.Chakravarthy[2], P.S.R.Chowdary[2]

[1]Department of Information Technology, S.R.K.R Engineering College, Bhimavaram, Andhra Pradesh, India-534201
[2]Department of Electronics & Communication Engineering, Raghu Institute of Technology, Visakhapatnam, AP, India
**Corresponding author E-mail:desaij4@gmail.com*

Abstract

Current digital world face trouble with massive information, again it made a demand for latest and advanced software frameworks for efficiently processing present world large data. Because digital world information is double rapidly, generally but existing and traditional tools for Big Data (BD) are becoming insufficient since enormous data processing towards to distributed, parallel, and group (Batch). Main essential thing is to evaluate tools and technologies, one important thing must follow the understanding of what to evaluate for. Even growing multiple options the intention of choosing Big Data functions for the digital world will be difficult. In the existing tools had merits, disadvantages and lack of many limitations but many had an overlapping custom. This survey looks at the major attention on BD the basic area is associated with analytics tools. In the current digital world (DW), exactly every computation perform on online as interactive processing also introduce apache free access tool to overcome restrictions and issues in Hadoop by Apache open Spark.

Keywords:*Big Data; Hadoop; Map Reduce; Spark .*

1. Introduction

In Digital World (DW) things will depend on much utilize the internet, also all regions of computers generate gigantic size of datasets as multi structure [1]. From the vast size, multiple kinds of multiplicity with more speed on DW data emerges the BD issues. Big data(BD) become latest and present most popular technology particularly in digital world data growing in datasets are the large volume in size, the variety of information, finally more speedily generate the informationpeople more demanded to evaluate the new objection and challenge. Because follows traditional tools and services are not adequate suitable for processing the DW outsized data the support of highly distributed and very huge scale processing, BD tools offer advanced result to BD issues. Now world BD issues which would work out by latest developing BD techniques. various organizations, industries and companies, applications working on BD regarding processing DW information in smooth, faster, highly accuracy, efficiency, good efficiency and highly reliable manner [2]. In DW each click, play, pause, request, message on n//w, error, bug, service on time could be recorded. A Petta to Zetta size of data certainly gathered each second in the current digital planet. But the process, store and analysis the DW is tremendously tiny level. Every post on face book has different formats of data, finally each second its scale like GB-TB. On YouTube each second nearly 4000 hours videos upload happened, it consequently million to billions of views [3]. In current DW there is endless generating information like structured, unstructured and semi-structured from various organizations, Social n/w and industry more ever this information has useful results for taking decisions and makes prediction and in another angle those organization don't discover regarding how efficiently utilize world BD become most excellent service their curiosity finally Apache tools especially Spark would assist us to determine knowledge from mixture of applications [4]. The big DW has vast information involve 5 dimensions as shown in Fig.1.

1.1. Volume

Means more industries and applications are overwhelmed with frequently produce information in different formats and types naturally generate peta-bytes to zetta-bytes of information [15].

1.2. Velocity

Information is growing at unequalled speed severe demanding challenge for coming days.

1.3. Variety

As designing multiple technologies more ever the dw information which comes nowadays in several syntaxes. Also, those formats handling becomes a critical task for many industries and organizations.

1.4. Veracity

This is very difficult to make an intellectual decision on DW information in various organizations because many times this information is not authentic also Spam.

1.5. Value

As value attribute, many analysts attempt to estimate needs and usages of large information in BD for strengthening and living standards.

The Large size of biometric information focuses these five attributes regarding solving BD issues are store, manage and retrieve. The methods in BD analysis follow steps:

1) Gain and gather data.
2) Clean and discover DW data
3) Aggregate and combine DW data
4) Model and study DW datasets
5) Interpretation.

This section presents a brief discussion of big data and their problems Rest of the paper covers in section 2 describes big data technologies. Section 3 covers Apache Spark Features. Section 4 discusses Conclusion on Hadoop limitation.

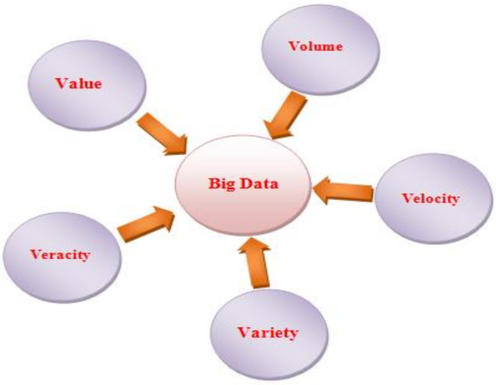

Fig. 1: Big Data Characteristics.

2. Big data techniques

2.1. Hadoop

This is a world popular technology for processing Big Data, helps to resolve the scalability issue with the help of distributed storage, parallel processes. While Apache open Hadoop is simply a popular programming framework to efficiently processing digital world data, it offers the more extensible platform, which allows any organization, industries applications [5, 16, 17].

Following Hadoop modules:

Common: Its utility libraries which help to join other components.

HDFS: an excellent distributed system on files which supports to access data about each application with good throughput, fault-tolerant and efficient processing manner.

YARN: It helps to assign resources smoothly to every required job because YARN intermission can perform efficient job arrangement and resource organization during process execution.

Map-Reduce: it efficiently processes the large information as parallel.

2.2. HDFS

It is capable to powerfully store huge information storage device used by apache open resource tool hadoop computations [6,18]. It's working based on Name node and Data Node to design a distributed File system. It offers and supports efficient performance, easy to use data on large clusters. Latest source especially Apache frameworks are Map Reduce is the present world best programming framework for to handle DW huge size of information known as BD. These are free access because the open free source technology to manage DW data as paralleling, distributing and fault-tolerant. Regarding MR model the incoming information divide into additional memory chunks and allocate to various map tasks and feed as I/P to reduce tasks.

2.3. Map reduce (MR)

MR like more famous function to for handling BD aspects with parallel, distributed and high scalable strategies [7]. The Year of 2006 innovative, latest java programming framework called map reduce presented by hadoop version 1, but discomfort from serious limitation and restriction is accepted nodes maximum of four-thousand only. To determine concurrency depended on aspects on the system with working performance regarding dividing jobs into the number of tasks of equal size for this convenient hadoop MR component is used, while it supports to handle granularity with controlling total map tasks.

Finally, hadoop and Map Reduce have some limitations than Apache Spark. Because of lack of speed, latency, stream processing, low security, no caching, no abstraction that's why to turn to Apache Spark to overcome the map reduces limitations.

2.4. Apache spark

Open Hadoop source and latest spark are two advanced BD tools. Hadoop process DW datasets on disks and spark powerfully process DW datasets on memory. Spark can run application 100 times quicker than Hadoop [8]. This much speed has a main important an essential task in several applications because DW computation needs short response time. In addition, apache latest tool spark becomessignificant technology perform BD analysis regarding steaming processing both Hadoop and latest spark working and handing on distributing manner efficient processing popular structure name as MR. Spark working on top of the HDFS as shown in Fig.2.

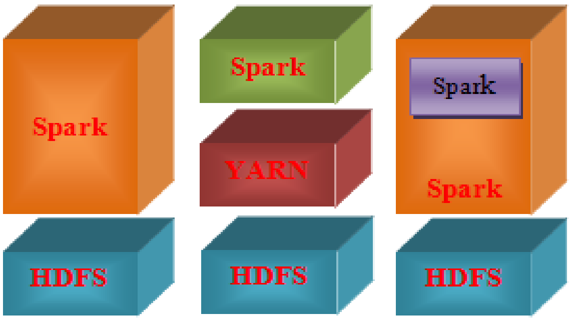

Fig. 2: Standalone, Over Yarn, Spark in MR (SIMR).

3. Apache spark

Today's we seriously and urgently require a latest advanced efficient software frameworks for easily, smoothly, and efficientlystore and processing such huge data [9].

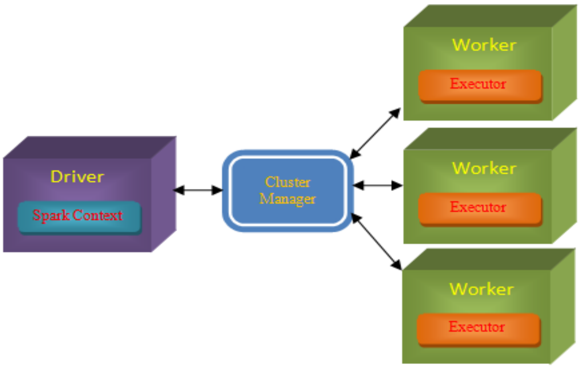

Fig. 3: Spark Architecture.

Because in current digital world information is generated at too fast from various organizations and applications also size of the database is always growing. Another latest free open access tool for BD analysis is Apache Sparkwill working as parallel and distributed mode regarding data blockage also analysis Spark establish RDD(Resilient Distributed-Dataset) which could support like application programming interface(API) fixed on Data structure.

Spark Architecture working: over the master/slave style during two major daemons and an efficient cluster administrator.

1) Master based Daemon-[Driver oriented Process]
2) Worker based Daemon-[Slave oriented process]

Every Spark clustering had only one Master but much amount of workforce.

In typically Spark apache driving which execute by executors individually and users execute them through horizontal cluster also a different System.

Working of Driver in Spark: Driver like essential and main entry way in spark shell. Driver running each submission of major functions then put in spark context. Driver encompasses several components.

RDD like unchangeable fault-tolerant dispersed gathering objects which would work as parallel manner[10]. With Map-Reduce compute models handle every map-reduce tasks but read and write on disk causes more delay in execution. However, RDD helps preserve the information in RAM for successfully handle that information on write programs in a distributed manner because it supports and dealing with the distributed shared memory. Again Spark can be the latest tool for efficiently utilize in-memory LRU cache with exclusion in-memory full position and which place total datasets upon local file systems in shuffle performance.

3.1. Spark resilient distributed datasets (RDD)

Features from spark are most (RDD), fault-tolerant multiples elements which will work based on parallel. Spark as an efficient s/w model that characterizes an unchallengeable number objects which will divide above a computing cluster as shown in Fig.4. Process an RDD will split over the computing clusters and implement as a parallel batch-wise process, the main significant thing is fast and extra scalable parallel performing RDD would make frame common text files SQL, DB, NOSQL [9].

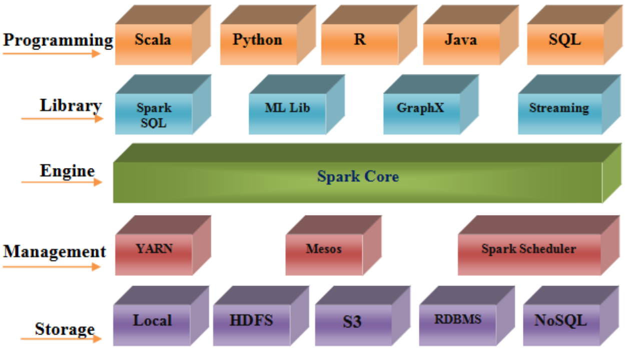

Fig. 4: Spark Framework.

Spark core: from apache many projects were created to finish the jobs through spark core.

Spark SQL (SSQL): SSQL running through data frames (Table) like latest data arrangement method to handle ordered and semi-ordered data. Tables provide SQL helps by commands connection with ODBC/JDBC controller.

Spark Streaming: To get fast process by spark API streaming. Spark streaming could job with many databases like HDFS, Flume, Kafla

Machine Learning Library: MLib is the latest features from Spark it has two packages MLib will place above on RDD and ML which places above on table to designing pipeline.

Spark Graphx: Users easily can see, convert and add changing upon graph also collections. it follows computation on pergel abstraction.

3.2. Features on spark

Many features on spark as shown in Table.1compare with hadoop Map Reduce.

Speed: ever and ever significant task for efficient and smooth process large data. Because today's organizations in the digital world to smoothly process high voluminous dimension of information as easy as feasible. Spark has Lighting fast processing tool build it much speeder to deal with difficult processing very easily and smoothly (Resilient Distributed Dataset) it supports to accumulate information clearly in memory, which supports in decrease lengthy read/write to disc [11].

Usability: It allows much program language made it dynamic. It accepts you quickly designing applications in object-oriented java, Scala, Python and R.

Computing on memory: hold datasets in RAM though it supports quickly admission the datasets in memory. Analytics velocity up iterative ML it protects data Read/write around tour from to disc.

Table 1: Features of Hadoop and Spark

Properties	Hadoop MR	Spark
Develop	Java	Scala
Support	C,C++, Ruby, Groony, Perl, Python	Scala, Python, R,SQL, java Because rich API
Real-time Analysis	Poor (i.eBatch)	Good
Speed	Slow	100X Fast than MR
Mode	No Interactive	High
Latency	High	Low
Locality	Low	High
Throughput	Low	High
API	Low	High
Storage	Disk	Memory
Libraries	More	None
Elasticity	Yes	No
Difficulty	Not Easy	Easy because of RDD
Streaming	Only in Batch mode	Real-time
Manage	Difficult	Easy
Usage	Complex	Simple
Fault-tolerant	High	High
Scheduler	Needs External Scheduler	Doesn't need because it has own Flow scheduler
Recovery	Good because Fault tolerant	Good because DAG
Cost	Low cost	Cost
Security	More	Low
OS support	Basic processing Engine	Analytics Engine
Community	Turn to Spark	Strong Community
Scalability	High	High
SQL Support	Hive	SQL
Machine Learning	Mahout	MLib
Line of code	1,20,000	20,000
Caching	Poor because can't cache	Good performance because it caches
Hardware requirement	On commodity Hardware	High Hardware

Popular to complicated Analytics: Sparks has an efficient apparatus for interact queries Streaming information ML it depends on map reduced function this finally user tin simply join mutually in a single workflow.

Instantaneous stream processing: spark holds real time-stream handing beside with mixture of supplementary frame works stream procedure is easy, fault-tolerant integrated Spark as open, simple to accessible, extra influential and accomplished big data tools to deal with digital world bid data issues open access spark now the latest component becoming major stream also more in-demand big data tools & framework across all organizations and industries.

Resilient Distributed Datasets (RDD): RDD has numerous information essentials that are dividing into a number of the division then store in-memory in spark workers node of the cluster. Regarding datasets, Spark permit two fundamentals of RDD.

1) Hadoop datasets which are designed by files store on HDFs
2) Collections regarding parallel which depends on existing scale collection. RDD's follows two kinds of function.
3) Transformations

4) Actions

DAG: conversion is like accomplishment which transfer from A to B in Data Partition.

Acyclic: Transformation unable reverse to the earlier partition DAG execute a series of calculation on huge information where each one node is executed by the border is execute on top of information.

The abstraction of DAG supports to remove the multiple times of MR implementation functions also offers efficiently processing strategy developments across Hadoop [12].

DAG scheduler, Task scheduler, Backend scheduler and Block manager are foremost responsibility is to exchange user code to original spark jobs execute on the cluster. Working of Executor in Spark: like distributed agent goal is efficiently run several tasks. Applications had an individual executor which runs the whole application in a static way .Since for executors can join/delete executor dynamically suitable through every workload.

Batch processing (BP): It executes a group of jobs in the program which allows various files as input then process it finally give output. In previous projects follows various BP techniques those are MR, Hadoop Spark and Pregal [12]. These methods would process again analyze huge information in batch, dispersed also parallel manner. Especially spark was fast and general purpose engine on gigantic scale processing also helps good scalability and fault-tolerant of MR spark presents a dispersed memory model called RDD the purpose is in- memory performance over several nodes in fault-tolerance fashion.

Stream process: now digital planet needs latest online processing tool which process continuously on system input data .because online tasks demands fast interaction response i.e. the processing rate must not be slow compared with speed on system incoming data. Few powerful platforms to streaming include storm tool, Apache open access spark and s4.

Spark: a spark stream is a valuable tool in inside the spark which helps with getting scalability, more output, fault-tolerant working on live information streaming. the information will be gathered from several fields based on kafka, flume, kinesis and TCP based sockets which could be efficiently processed by large algorithm articulated like map(),reduce(),join(),window().

Query processing: an efficient structure can help convert client queries into strong information retrieval and handling functions then run those functions on the large quantity of nodes. Many platforms efficiently perform distributed computation are apache Hive, Pig Latin and SSql called query processing. The processing designers follow declarative mode to identify their jobs, and convert into suitable to optimized operations. Ssql place on the peak of apache spark also joins both API and Mlib to handle formatted and semi formatted information by sq lot API.

Interactive processing: It supports to users for efficiently perform analysis, review and compare high interactive way directly in structure or graphical model. Google's dremel, drill, spark are distributed tools to gain high interact analysis (ISS) on data-intensive computations. Spark has most powerful feature is ISS by the help of such feature the designers can integrate spark libraries like spark streaming (assume streaming), Ml algorithm (repeating things) and graphx (showing analysis) [13].

Limitations of MR over Spark: MR allows processing on course grained operations but its processing is too vast especially on iterative MR unable to collect/cache middle information from memory but very quickly write intermediate data to disk per each step. variety of RDD: RDD working by using maps (), Flat Map () called map partitions RDD, also using coalesce An RDD offers efficient repartition called coalesced RDD Data after that sorted in HDFS called Hadoop RDD Other useful RDD types: series of File RDD, piped RDD, Co Graped RDD, and Shuffled RDD [14].

Finally, MR has same drawbacks velocity is very slow also poor performance because MR persists reverse to disk. Also, MR is writing in java, so tricky to program carrier to Spark. MR can do a group (batch) process only more tolerant. Hadoop Features: Open source, information is enormously available, highly scalable, fault-tolerant information is trustworthy stored, not extremely costly simple

In this survey Article, we performed as efficient and analytical comparative review on Hadoop MR Software form in favor of easily processes and store immense data [15]. Another one related to spark becomes as reference framework regarding instantaneous streaming. Apache processing types: this survey Article we also cover the diversity of processing elements.

4. Conclusion

Hadoop secondary popular component as Map Reduce is the world first distributed software framework to handle and manage two digital world problems as efficiently store and process massive information. But MR faces trouble from many limitations and drawbacks regarding speed, real-time, latency, and streaming. Latest Spark has more features inApache open source projectand MR are highly efficient and effective programming frameworks also so popular ,distribute, parallel computing tools to processing huge DW data for producing best results but today's MR frameworks has many limitations compared to spark especially speed, streaming, interactive nature, latency and caching because spark follows RAM in- memory. In this survey and study article, we focus and compare the working and lack of features of both MR and Spark programming frameworks. Because Spark brings latest features also remove the many MR overheads.

References

[1] BingbingRao, liqiang Wang„ *"A Survey of Semantics-Aware Performance Optimization for Data-Intensive Computing",*3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, (IEEE), (2017), pp.81-88.

[2] S Agarwal, S Kandula, N Bruno, M C Wu, I Stoica, J Zhou," Re-optimizing data-parallel computing", In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation. USENIX Association, (2012),

[3] J Ahrens, B Hendrickson, G Long, S Miller, R Ross, D Williams," *Data-intensive science in the us doe: case studies and future challenges*", Computing in Science & Engineering,(2011).https://doi.org/10.1109/MCSE.2011.77.

[4] A Alexandrov, R Bergmann, S Ewen, J C Freytag, F Hueske, A Heise, O Kao, M Leich, U Leser, V Markl," *The stratosphere platform for big data analytics*", The VLDB Journal, (2014).https://doi.org/10.1007/s00778-014-0357-y.

[5] M Armbrust, R S Xin, C Lian, Y Huai, D Liu, J K Bradley, X Meng, T Kaftan, M J Franklin, A Ghodsi," *Spark sql: Relational data processing in spark*", In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, ACM, (2015), pp.1383-1394. https://doi.org/10.1145/2723372.2742797.

[6] Sara Landset, Taghi M Khoshgoftaar, Aaron N Richter, TawfiqHasanin, "*A survey of open source tools for machine learning with big data in the Hadoop ecosystem*", Journal of Big Data, (2015),pp.1-36.

[7] Jai PrakashVerma, Bankim Patel, Atul Patel, "*Big Data Analysis: Recommendation System with Hadoop Framework*", IEEE International Conference on Computational Intelligence & Communication Technology, (2015), pp.92-96.https://doi.org/10.1109/CICT.2015.86.

[8] YavuzCanbay, serefsagiroglu," Big data anonymization with spark", Diego García Gil, Sergio RamírezGallego, Salvador García, Francisco Herrera," *A comparison on scalability for batch big data processing on Apache Spark and Apache Flink*", Big Data Analytics, (2017), pp.1-12.

[9] Amir Bahmani, Alexander B Sibley, Mahmoud Parsian, KourosOwzar, Frank Mueller,"*SparkScore: Leveraging Apache Spark for Distributed Genomic Inference",* International Parallel and Distributed Processing Symposium Workshops (IPDPSW)Chicago, IL, USA, IEEE, (2016), pp.435-442.

[10] Jian Fu, Junwei Sun, Kaiyuan Wang,"*SPARK—A Big Data Processing Platform for Machine Learning*",International Conference on Industrial Informatics - Computing Technology, Intelligent Technology, Industrial Information Integration, IEEE, (2016), pp.48-51.

[11] AsmelashTekaHadgu, Aastha Nigam, Ernesto Diaz Aviles," *Large-scale learning with AdaGrad on Spark"*, 2015 IEEE International Conference on, Santa Clara CA, IEEE, (2015), pp. 2828-2830.

[12] Hang Tao, Bin Wu, Xiuqin Lin, *Budgeted mini-batch parallel gradient descent for support vector machines on Spark*, In 2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS), Hsinchu, (2014), pp. 945-950.

[13] SauptikDhar, Congrui Yi, Naveen Ramakrishnan, Mohak Shah,*ADMM based Scalable Machine Learning on Spark*, in Big Data(Big Data), 2015 IEEE International Conference on, Santa Clara CA, (2015), pp. 1174-1182.

[14] Zhijie Han, Yujie Zhang, "*A Big Data Processing Platform Based on Memory Computing*, in Parallel Architectures, Algorithms and Programming (PAAP)", 2015 Seventh International Symposium on,Nanjing, (2015), pp. 172-176.

[15] E.Dede, B.Sendir, P.Kuzlu, J Weachock, M Govindaraju, L Ramakrishnan, "*Processing Cassandra Datasets with Hadoop - Streaming Based Approaches"*, IEEE Transactions on Services Computing, (2015), pp. 46-58.

[16] N.Deshai, G.P.S.Varma, S.V.Ramana, "A study on analytical framework to breakdown conditions among data quality measurements" in International Journal of Engineering & Technology, Vol 7(1.1), pp: 167-172, 2018.

[17] N.Deshai, S.Venkataramana, I.Hemalatha, G.P.S.Varma, "A Study on Big Data Hadoop Map Reduce Job Scheduling", International Journal of Engineering & Technology, Vol 7(3.31), pp: 59-65, 2017.

[18] N.Deshai, P. Swamy, G.P.S.Varma, "Big Data Challenges and Analytics Processing Over health Prescriptions", Jouonal of Advance Research in Dynamical & Control Systems, 15-Special Issue Vol 7(3.31), pp: 650-657, Oct'2017.