ICITKM
Proceedings of the First International Conference on Information Technology and Knowledge Management pp. 131–135
DOI: 10.15439/2018KM20
ISSN 2300-5963 ACSIS, Vol. 14

A Contemplating approach for Hive and Map reduce for efficient Big Data Implementation

Gopinadh Sasubilli[1], Uday Shankar Sekhar[2], Ms.Surbhi Sharma[3], Ms.Swati Sharma[4]

[1]*Sr. Solution ArchitectMc Donald's, IL.*
[2]*Senior Associate,EY*
[3]*Assistant Professor, ACERC Ajmer Rajasthan, India*
[4]*Research Scholar, ACERC Ajmer,Rajasthan, India*
[1]*kumarattangudiperich@kpmg.com,* [2]*uday.sekhar@ey.com,*[3]*surbhi2690@gmail.com,*
[4]*swati199410@gmail.com*

***Abstract*—In the reference current scenario, data is incremented exponentially and speed of data accruing at the rate of petabytes. Big data defines the available amount of data over the different media or wide communication media internet. Big Data term refers to the explosion in the quantity (and quality) of available and potentially relevant data. On the basis of quantity amount of data are very huge and this quantity has been handled by conventional database systems and data warehouses because the amount of data increases similarly complexity with it also increases. Multiple areas are involved in the production, generation, and implementation of Big Data such as news media, social networking sites, business applications, industrial community, and much more. Some parameters concern with the handling of Big Data like Efficient management, proper storage, availability, scalability, and processing. Thus to handle this big data, new techniques, tools, and architecture are required. In the present paper, we have discussed different technology available in the implementation and management of Big Data. This paper contemplates an approach formal tools and techniques used to solve the major difficulties with Big Data, This evaluate different industries data stock exchange to covariance factor and it tells the significance of data through covariance positive result using hive approach and also how much hive approach is efficient for that in the term of HDFS and hive query. and also evaluates the covariance factors after applying hive and map reduce approaches with stock exchange dataset of around 3500.After process data with the hive approach we have conclude that hive approach is better than map reduce and big table in terms of storage and processing of Big Data.**

***Index Terms*—Big Data, Hadoop; MapReduce; HDFS; Grid Computing; Big Table; Hive.**

I. Introduction

BIG data is comparable to tiny knowledge however it's larger in terms of volume, selection and rate. Massive knowledge may be the next big factor within the IT space. Massive knowledge generates price from the storage and process of terribly giant quantities of digital data that can't be processed by standard info systems. The larger a part of the knowledge is delivered, put away, listed and handled over the net, prompting the enlargement in size of data systematically. This substantial live of data introduce over the net is alluded to as "Big Data". Massive knowledge characterized by the info quantity (volume), data speed (velocity) and differing types of knowledge (variety).

Volume: Volume denotes the dimensions of knowledge over the web. Presently it's in petabytes and is predicted to be raised to zettabytes. Knowledge from the good phones, sensors embedded into everyday objects can presently lead to billions of recent knowledge.

Fig. 1. Characteristics of Big Data

Velocity: Velocity inputs cover the speed of input generation and data managing. Online gaming systems support millions of concurrent users, each producing multiple inputs per second. [2].

Variety: Variety covers the type of input. Input can be constructed (text), unstructured (data generated from social networking sites and sensors) or semi-structured (data from web pages, web logs-mail etc).

Two more characteristics have also been included- Veracity and Value.

Veracity- It means how much the data is related to truth or facts.

Value- It covers the processing input and how the data can be combined with other data to extract meaningful information from it.

II. Proposed Work

In the present paper, we have proposed distinctive apparatuses and strategies which are utilized to beat the regular is-

sues identified with huge information. Term Big Data examination includes devices, calculations, and design that break down and change substantial and monstrous volumes of information [10].Big information investigation is an innovation empowered technique for empowering an association to have an aggressive edge over others by dissecting business sector and client patterns. Investigation on on-going information, online value-based information gives further experiences of the patterns to settle on opportune and exact choices. For the Computation reason for wide range volume of information [5]Big Data Computing is worried about the preparing, changing, dealing with and capacity of data. Frameworks, for example, Map Reduce, Hadoop, Grid Computing, and Big Table[8] have made composition and executing specially appointed huge information investigation and calculation simple. As web indexes have changed data get to, different types of enormous information registering can and will change the exercises like restorative and logical research, protection undertaking and so on. This paper focus on the following technologies:

A. Hadoop

Hadoop is actually a large scale batch data processing system. Hadoop developed as an establishment for huge information handling undertakings, for example, logical examination, business and deals arranging, and preparing huge volumes of sensor information, including from web of things sensors. Hadoop is supportable for distributed cluster system, parallel data processing system and worked as a platform for massively scalable applications. Facebook, Apple, Google, IBM, Twitter and hp are the famous hadoop users. Hadoop provide access to the file system called HDFS (hadoop Distributed File system). Basic capabilities of the hadoop include some packages like Apache Flume, Apache HBase, Apache Hive, Apache Pig, Apache Oozie- and many more. Hadoop is beneficial in terms of cost efficient and reliable and scalable data processing. Different components of Hadoop system are explained below[10]:

B. HDFS Architecture

HDFS stands for Hadoop Distributed File System. It is an essential component of Hadoop which is used to store huge datasets. The main task of HDFS is to distribute the data to Various clusters of computers (machines) and then processing of this data is done. The advantage of using HDFS is that it coordinates the work among machines and if any one of them fails, Hadoop continues to operate by shifting the work from one machine to another without losing data or interrupting work [11].

C. MapReduce

MapReduce is a parallel programming framework that allows operations to be applied over large datasets. The main task of MapReduce is to divide the problem into smaller parts and then run those subparts in a parallel fashion. MapReduce consists of two functions: Map and Reduce.

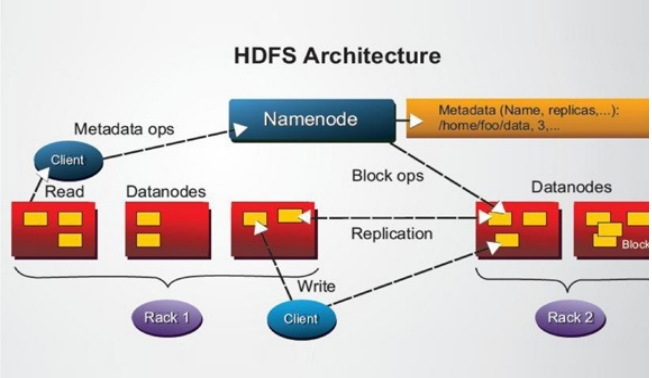
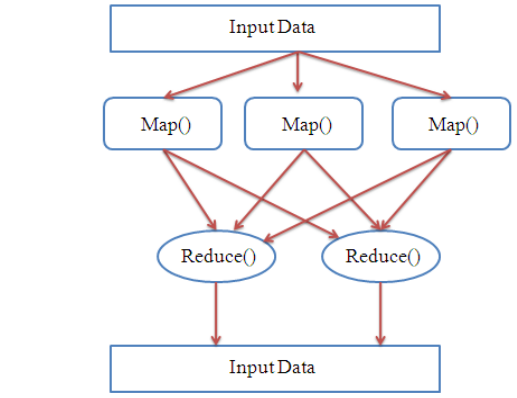


Fig .2. Architecture of HDFS

Map: This function generates a key/value pair and performs sorting and filtering of data.

Reduce: This function combines all the intermediate values and gives the output.



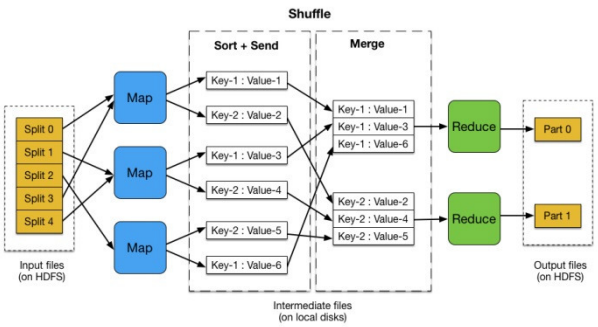


Fig.3.(a) Architecture of MapReduce (b) Internal structure of MapReduce

D. Grid Computing

A grid is a system in which a number of servers are connected to each other through a high speed internet. It is a distributed computing model in which the servers are geographically apart from each other and the users can access the data transparently from any location[12]. Although Grid

is beneficial as it provides hardware for storage of data but it has a drawback that current Grid infrastructure is not capable enough to handle Big Data. Thus research is still going on to find a solution to this problem so that it can deal with large volume of data.

Challenges with grid computing [12]:

- Data movement
- Data replication
- Resource management
- Job submission

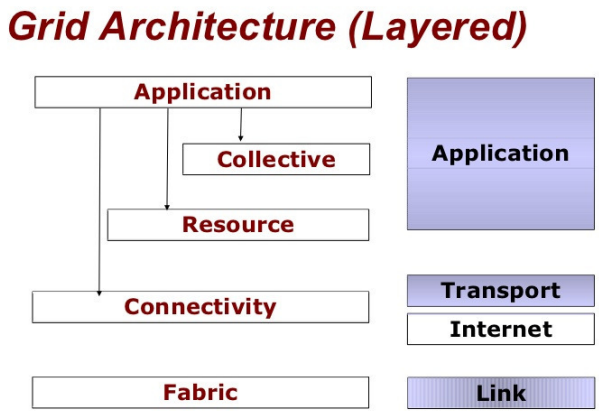


Fig. 4 Architecture of Grid Computing

a. **Bigtable**

Bigtable is a component to implement Big data which is similar to distributed storage system developed by Google. Bigtable is used for handling huge volume of data. It can-handle data up to petabytes. It is a distributed and sparse map which is matricide by a row key element, column key element, and a timestamp value; each value in the map is anon disturbed array of bytes.

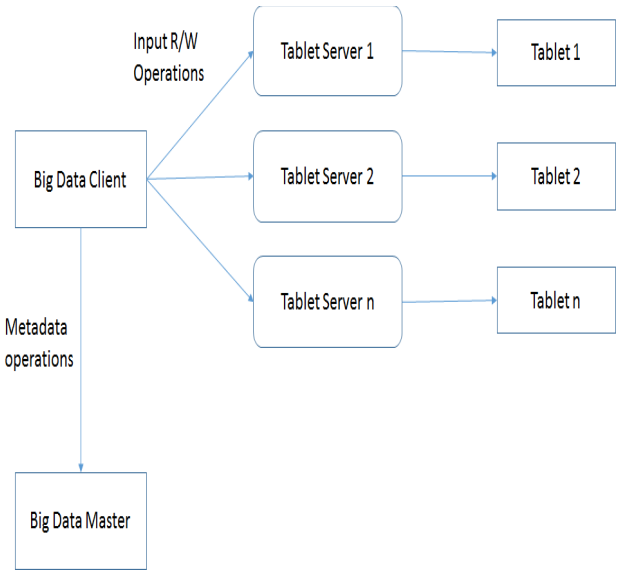


Fig.5. Architecture of Bigtable

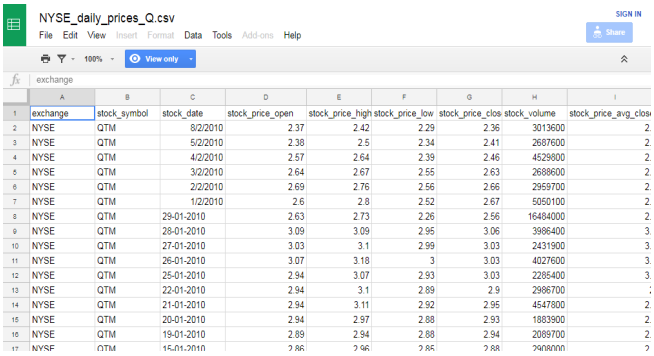
Challenges with Bigtable implementation are:

- Reliability storage: durability and replication
- Sequential storage
- Structured storage

III. IMPLEMENTATION OF BIG DATA USING HIVE

Hive is techniques supported the SQL, it in addition uses a lot of customary and in some cases program secret writing that we might have to be compelled to implement Map Reduce programming. We have got to use Hive to interrupt down the stock and large knowledge set info, at that time we might have the advanced and entity relative calculus primarily based question to use the SQL skills of Hive-QL and connected info is overseen during a specific map and scale back mapping. it'll depreciated the advancement time and may administrate joins between the dataset (Eg. Stock info, Industrial data).Hive in addition has its main servers, by that we will gift our Hive queries from anywhere to the Hive server, that is employed to executes them. Hive SQL queries area unit being modified over into define employments by Hive compiler, and software system engineers have to be compelled to solve this advanced programming and solved the problems connected with massive knowledge and organization knowledge. For applying this methodology we have a tendency to could have to be compelled to use a dataset happiness to exchange and Dataset contains following properties:

- Data is being organized above all arrangement.
- It would judge joins to cipher Stock variance.
- It may well be sorted out into composition of various forms of be a part of.
- In neutral condition, info size would be extreme high.


NYSE_daily_prices_Q.csv

exchange	stock_symbol	stock_date	stock_price_open	stock_price_high	stock_price_low	stock_price_clos	stock_volume	stock_price_avg_clos
NYSE	QTM	8/2/2010	2.37	2.42	2.29	2.36	3013600	2.
NYSE	QTM	5/2/2010	2.38	2.5	2.34	2.41	2687600	2.
NYSE	QTM	4/2/2010	2.57	2.64	2.39	2.46	4529800	2.
NYSE	QTM	3/2/2010	2.64	2.67	2.55	2.63	2688600	2.
NYSE	QTM	2/2/2010	2.69	2.76	2.56	2.66	2959700	2.
NYSE	QTM	1/2/2010	2.6	2.8	2.52	2.67	5050100	2.
NYSE	QTM	29-01-2010	2.63	2.73	2.26	2.56	16484000	2.
NYSE	QTM	28-01-2010	3.09	3.09	2.95	3.06	3986400	3.
NYSE	QTM	27-01-2010	3.03	3.1	2.99	3.03	2431900	3.
NYSE	QTM	26-01-2010	3.07	3.18	3	3.03	4027600	3.
NYSE	QTM	25-01-2010	2.94	3.07	2.93	3.03	2285400	3.
NYSE	QTM	22-01-2010	2.94	3.1	2.89	2.9	2986700	2.
NYSE	QTM	21-01-2010	2.94	3.11	2.92	2.95	4547800	2.
NYSE	QTM	20-01-2010	2.94	2.97	2.88	2.93	1883900	2.
NYSE	QTM	19-01-2010	2.89	2.94	2.88	2.94	2089700	2.

Fig.6 Stock Exchange Dataset(.csv) file

- Issues related with map reduce are solved with Hive:
- Used Hive setup on Cloudera.
- Create Hive Table:
- Use 'make table' Hive command to create the Hivetable for our conideredcsv format dataset

- hive > create table STOCK (trademark String, stock_symbols String, stock_datestart String , stock_price_opens twofold, stock_price_uptwofold, stock_price_bottom twofold, stock_price_closed twofold, stock_quantity twofold, stock_price_adj_close twofold)

- Load .csv info into Hive Table: hive> stack info neighbourhood inpath '/home/cloudera/STOCK.csv' into table STOCK;

V. This can stack the dataset from the required space to the Hive table 'STACK' as created on top of but this dataset are place away into the Hive-controlled record framework namespace on HDFS, with the goal that it can be bunch ready more by MapReduce employments or Hive queries.

VI. Calculate the Covariance factor.

VII. We can figure the Covariance for the gave stock dataset to the inputted year as beneath utilizing the Hive select inquiry:

VIII. Select a.STOCK_SYMBOLS,b.STOCK_SYMBOLS, month(a.STOCK_DATESTART), (AVG(a.STOCK_PRICE_UP*b.STOCK_PRICE_UP) (AVG(a.STOCK_PRICE_UP)*AVG(b.STOCK_PRICE_UP)))from STOCK a join STOCK b on a.STOCK_DATESTART=b.STOCK_DATESTARTwherea.STOCK_SYMBOLS<b.STOCK_SYMBOLS and year(a.STOCK_DATESTART)=2008 Group by a.STOCK_SYMBOLS, b. STOCK_SYMBOLS, month(a.STOCK_DATESTART);

- This Hive select query will trigger up the MapReduce work as below:

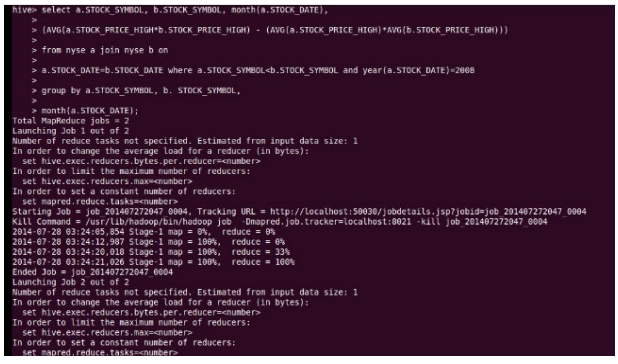
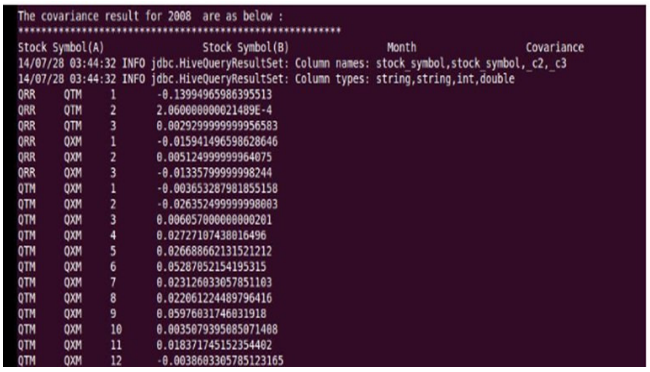


Fig 7. Hive Approach Implementation



From the variance issue, stock dataset recommend the subsequent conclusions: For Stocks QRR and QTM, these are having additional positive variance than negative variance, therefore having high chance that stocks can move along same means.

1. For Stocks QRR and QXM, these are for the foremost half having negative variance. Therefore there exists an additional distinguished chance of stock prices acquiring a reverse course.

2. For Stocks QTM and QXM, these are typically having positive variance for particularly else months, therefore these tend to maneuver an analogous means the bulk of the circumstances. So this discourse analysis comprehends the attendant 2 crucial objectives of giant data advances:

(a) Storage: it's the deepest connected issue for huge stock data into HDFS, the arrangement provides considerably additional strong, strength, scalable, and elastic.

(b) Processing: In several Hive composition it relies on a typical SQL information, we tend to could get the advantage of running SQL queries on the large dataset likewise and may method the massive quantity of GBs or TBs of data with basic SQL queries.

IV. Conclusion And Future Scope

We have conclude that map reduce approach is limited for small level data set and required a larger amount of storage to hold the map level and reduced data set recursively but we have used Hive approach to evaluate covariance among our considered data set and it shows the result that the covariance between QTM and QXM parameter is positive. Another factor is that the amount of storage over HDFS is limited under hive approach and processing is programmed with hive SQL Query which is used to take a shortest time for execution for petabytes amount of datasets. Legitimate and powerful examination of in-depth volumes of data can prompt speedier advances in varied logical teaches and enhance the profit and accomplishment of various enterprises.

The difficulties incorporate the difficulty of in-depth volume, however additionally no uniformity, unclear structure, blunder coping with, protection, favorableness, security cradle, combination, and illustration. These specialized difficulties area unit found an immense assortment of use areas and consequently force an immense value. Besides, these difficulties would require would force transformative arrangements and can require an intensive

type of apparatuses, systems, and applications to manage. With a particular finish goal to accomplish the bonded benefits of massive information, this stuff should be taken underneath thusly thought so most capability will be determined to select up an associate aggressive edge.

To take out the simplest have the benefit of Hadoop, the in-depth analysis must be applied and revolutionary tools and techniques must be developed to rigorously comprehend and properly reply to numerous challenges.

REFERENCES:

[1] Lawal Muhammad Aminu, "Implementing Big Data Management on Grid Computing Environment", International Journal of Engineering and Computer Science ISSN: 2319-7242, Volume 3, Issue 9, September 2014, Page No. 8455-8459

[2] Agrawal et al., 2011; Baer et al., 2011 Agrawal, D., Das, S., &Abbadi, A. (2011), Big Data and Cloud Computing: Current State and Future Opportunities. ACM EDBT Conference, March 22–24, 2011, Uppsala Sweden. http://dx.doi.org/10.1145/1951365.1951432

[3] Baer, T. (2011). 2012 Trends to Watch: Big Data. Ovum Report, OI00140-041.Baer, T., Sheina, M., and Mukherjee, S. (2011). What is big data? The big architecture.Ovum Report, OI00140-033.

[4] S. Vikram Phaneendra & E. Madhusudhan Reddy "Big Data-solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).

[5] Kiran kumara Reddi & Dnvsl Indira "Different Technique to Transfer Big Data: survey" IEEE Transactions on 52(8) (Aug.2013) 2348 {2355}

[6] Jimmy Lin "MapReduce Is Good Enough?" The control project.IEEE Computer 32 (2013)

[7] Umasri M. L, Shyamalagowri. D, Suresh Kumar. S "Mining Big Data:- Current status and forecast to the future" Volume 4, Issue 1, January 2014 ISSN: 2277 128X.

[8] Albert Bifet "Mining Big Data in Real Time" Informatica 37 (2013) 15–20 DEC 2012

[9] Zan Mo, Yanfei Li Research of Big Data Based on the Views of Technology and Application American Journal of Industrial and Business Management, 2015, 5, 192-197 Published Online April 2015 in SciRes.

[10] Harshawardhan S. Bhosale1, Prof. Devendra P. Gadekar2 "A Review Paper on Big Data and Hadoop" International Journal ofScientific and Research Publications, Volume 4, Issue 10, October 2014 1 ISSN 2250-3153

[11] https://www.quora.com/What-are-the-main-features-of-Hadoop

[12] https://www.slideshare.net/sandpoonia/1-grid-computing

[13] http://stackoverflow.com/questions/782913/googles-bigtable-vs-a-relational-database