



## A Multidisciplinary Approaches of Big Data and Data Science Using Hadoop and Map Reduce

<sup>1</sup>Dr. Kusum Yadav, <sup>2</sup>Dr. Rajender Bathla

<sup>1</sup>Asst Prof, College of Computer Science and Engineering, University of Hail Kingdom of Soudi, Saudi Arabia

<sup>2</sup>Head, Department of Computer Science & Applications, HCTM affiliated to Kurukshetra University,  
Kurukshetra, Haryana, India

DOI: [10.23956/ijarcsse/V7I1/01102](https://doi.org/10.23956/ijarcsse/V7I1/01102)

---

**Abstract**— *Data Science and Big Data have an important role among the modern science and emerging technology. Infact, we can say that these are the two side of the same coin. Data Science is establishing basic foundations to become a profession. The data science community can follow a road map for how data science can be more useful for various social and commercial sectors. Today we are surrounding by data like oxygen. The growth of data first presented challenges to cutting edge businesses such as Google, yahoo, Amazon, Microsoft, Face book, twitter etc.*

**Keywords**— *Big Data, Data Science, Hadoop, Map Reduce, Data Security, Privacy.*

---

### I. INTRODUCTION

Big Data and Data Science is the hottest and powerful trends in the business and IT world right now. We are living in the age of Big Data where due to the rapid growth development and computational power are used. There are much confusion and debate about the definition of data Science and a new role bread of sexy bird called the data scientist. Data science and Big data Analytics are exciting new areas that combine scientific inquiry, substantive expertise, coding and statistical knowledge. One of the main challenges for business and policy makers when using big data is to find people with the appropriate skills. Data Science is no longer only the domain of computer scientists and engineers. Good and best data science requires experts that combine substantive knowledge with data analytical skill, which makes it a prime area for social scientists with an interest in quantitative methods.

### II. ROLE OF BIG DATA

Big Data is an analytical process of optimizing, collecting, organizing and analysing a large set of data to discover the pattern, logic and other useful information. Big data is the process of changing data in to information, which then changes in to knowledge.

Big data Analytics the use of advance analytics technique against very large, diverse data set that include types such a s structured, unstructured and streaming, batch and different sizes from Terabyte to Zeta bytes. Big Data has one or more of the following characteristics, high volume, high velocity or high variety. Bid data comes from sensor, devices video, audio, Network log files, and transactional applications, web and social media much of it generated in real time and in a very large scale.

“Big Data is a collection of data sets so large and complex that it becomes difficult to process using on hand data base management tools. Scientist breaks down Big data in to many dimension: Volume, Velocity, Variety, Varsity. Big Data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools [3].Scientists break down Big Data into many dimensions: Volume, Velocity, Variety, Veracity and Value [4, 5]

**2.1 Volume** – The amount of data is at very large scale. The amount of information being collected is so huge that modern database management tools are unable to handle it and therefore become obsolete.

**2.2 Velocity** – We are producing data at an exponential rate. It is growing continuously in terabytes and peta bytes.

**2.3 Variety** – We are creating data in all forms -unstructured, semi structured and structured data. This data is heterogeneous is nature. Most of our existing tools work over homogenous data, now we require new tools and techniques which can handle such a large scale heterogeneous data.

**2.4 Veracity**-The data we are generating is uncertain in nature. It is hard to know which information is accurate and which is out of date.

**2.5 Value**-The data we are working with is valuable for society or not. IBM estimates that every day 2.5 quintillion bytes of data are created ,out of which 90% of the data in the world today has been created in the last two years .This data comes from sensors used to gather climate information, posts to social media sites, digital pictures and videos uploaded on internet, purchase transaction records, and cell phone conversation.

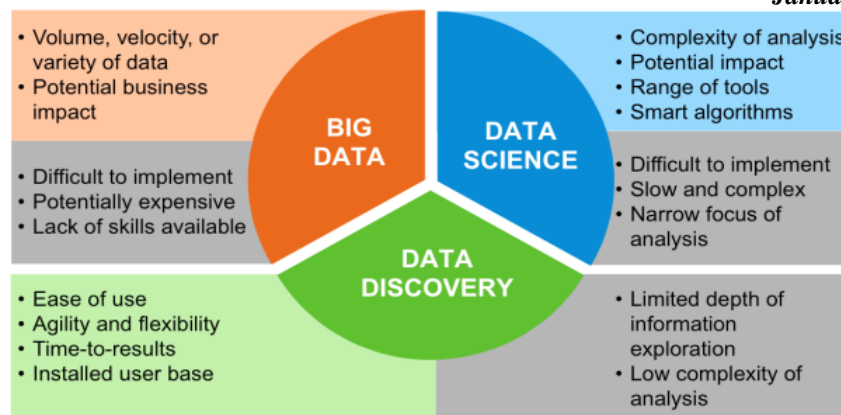


Fig 1 Big Data Discovery is the combination of Big Data, Data Science, and Data Discovery.

### III. DATA SCIENCE

Data Science is a field of systematic interdisciplinary survey & study to elucidate relationships across and with I formal, social natural and special sciences strategies through the application of scientific methods. Interdisciplinary area include analytical process, mathematics probability and statistics, logic modelling, machine learning , algorithms communications, traditional sciences, business, public policy and philosophy.

The following are the defined Data Science specialization and its types.

- 3.1 Blue sky data Science
- 3.2 Basic Data Science
- 3.3 Applied data Science
- 3.4 Data Science

### IV. ROLE OF DATA SCIENTIST

A person who studies or has expert knowledge of the interdisciplinary field of data science. Data Science requires skill ranging from traditional computer science to mathematical to art. A Data Scientist represent an evolution from the business or data analyst role has been described as “part Analyst, Part Artist. The Data Scientist will be responsible for designing and implementing process and layout for complex, large scale data set used for modelling, data mining and research purposes.

### V. MULTIDISCIPLINARY APPROACH OF DATA SCIENCE

**5.1 Basic Tools:** No matter what type of company you’re interviewing for, you’re likely going to be expected to know how to use the tools of the trade. This means a statistical programming language, like R or Python, and a database querying language like SQL.

**5.2 Basic Statistics:** At least a basic understanding of statistics is vital as a data scientist. An interviewer once told me that many of the people he interviewed couldn’t even provide the correct definition of a p-value. You should be familiar with statistical tests, distributions, maximum likelihood estimators, etc. Think back to your basic stats class! This will also be the case for machine learning, but one of the more important aspects of your statistics knowledge will be understanding when different techniques are (or aren’t) a valid approach. Statistics is important at all company types, but especially data-driven companies where the product is not data-focused and product stakeholders will depend on your help to make decisions and design / evaluate experiments.

**5.3 Machine Learning:** If you’re at a large company with huge amounts of data, or working at a company where the product itself is especially data-driven, it may be the case that you’ll want to be familiar with machine learning methods. This can mean things like k-nearest neighbours, random forests, and ensemble methods – all of the machine learning buzzwords. It’s true that a lot of these techniques can be implemented using R or Python libraries – because of this, it’s not necessarily a deal breaker if you’re not the world’s leading expert on how the algorithms work. More important is to understand the broad strokes and really understand when it is appropriate to use different techniques.

**5.4 Multivariable Calculus and Linear Algebra:** You may in fact be asked to derive some of the machine learning or statistics results you employ elsewhere in your interview. Even if you’re not, your interviewer may ask you some basic multivariable calculus or linear algebra questions, since they form the basis of a lot of these techniques. You may wonder why a data scientist would need to understand this stuff if there are a bunch of out of the box implementations in learn or R. The answer is that at a certain point, it can become worth it for a data science team to build out their own implementations in house. Understanding these concepts is most important at companies where the product is defined by the data and small improvements in predictive performance or algorithm optimization can lead to huge wins for the company.

**5.5 Data Mining:** Often times, the data you’re analyzing is going to be messy and difficult to work with. Because of this, it’s really important to know how to deal with imperfections in data. Some examples of data imperfections include missing values, inconsistent string formatting (e.g., ‘New York’ versus ‘new York’ versus ‘ny’), and date with a breakdown of the 8 skills need to get the job. Formatting (‘2014-01-01’ vs. ‘01/01/2014’, UNIX time vs. timestamps, etc.). This will be most important at small companies where you’re an early data hire, or data-driven

companies where the product is not data-related (particularly because the latter has often grown quickly with not much attention to data cleanliness), but this skill is important for everyone to have.

**5.6 Data Visualization & Communication:** Visualizing and communicating data is incredibly important, especially at young companies who are making data-driven decisions for the first time or companies where data scientists are viewed as people who help others make data-driven decisions. When it comes to communicating, this means describing your findings or the way techniques work to audiences, both technical and non-technical. Visualization wise, it can be immensely helpful to be familiar with data visualization tools like GG plot and d3.js. It is important to not just be familiar with the tools necessary to visualize data, but also the principles behind visually encoding data and communicating information.

**5.7 Software Engineering:** If you're interviewing at a smaller company and are one of the first data science hires, it can be important to have a strong software engineering background. You'll be responsible for handling a lot of data logging, and potentially the development of data-driven products.

**5.8 Thinking Like A Data Scientist:** Companies want to see that you're a (data-driven) problem solver. That is, at some point during your interview process, you'll probably be asked about some high level problem – for example, about a test the company may want to run or a data-driven product it may want to develop. It's important to think about what things are important, and what things aren't. How should you, as the data scientist, interact with the engineers and product managers? What methods should you use? When do approximations make sense. Data science is still nascent and ill-defined as a field. Getting a job is as much about finding a company whose needs match your skills as it is developing those skills. This writing is based on my own firsthand experiences – I'd love to hear if you've had similar (or contrasting) experiences during your own process.

## VI. BIG DATA REVOLUTION

From consumer to companies, people have an unquenchable appetite for data all that can be done with it. Not only are we relying on data for movies suggestions and gift recommendations but are depending on data for multidisciplinary climate and energy research, building adaptable roads and buildings, better foresighted health care, new ways to identify fraud and keeping a check on consumer behaviour and sentiment. Data has become a factor of production, according to the economist's report almost on par with labour and capital. IDC has predicate that the digital world will be 44 times in 2020 of what it was in 2009. Totalling a whopping 35 zeta bytes has reported the member of customers who are storing a peta bytes or more of data to grow from 1,000 to 100, 00 with in the next 10 years.[9]

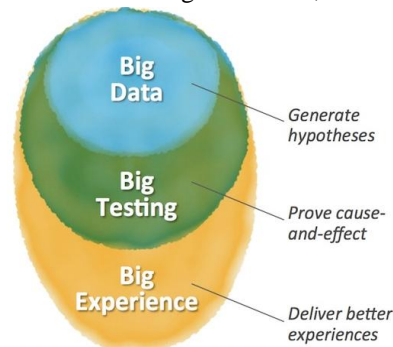


Fig 2. Big Data behaviour

## VII. BIG DATA ANALYTICAL TOOLS

There are varieties of tools and techniques are developed by various organization to process and analyse Big Data. Big Data captured today is unstructured, from sensors used to gather climate information, posts on social media site, digital pictures and videos, purchase transaction records.[4] All unstructured data can be analyze through applications support parallelism with the help of computing clusters. The following are the tools in area of Big Data Analytics.

**7.1 Hadoop:** Hadoop cluster is a special type of computational cluster designed specifically for storing and analyzing huge amount of structured and unstructured data in a distributed computing environment.[4]Hadoop is a large scale, open-source software framework that supports a large data sets in distributed computing environment. Distributed analytic frameworks, such as MapReduce, are evolving into distributed resource managers that are gradually turning Hadoop into a general-purpose data operating system, says Hopkins.[5] With these systems, he says, "you can perform many different data manipulations and analytics operations by plugging them into Hadoop as the distributed file storage system." [3]Hadoop is dedicated to scalable, distributed, data-intensive computing. Hadoop used to handle thousands of petabytes of data in different clusters. Hadoop is the core platform for structuring Big Data, and solves the problem of formatting it for subsequent analytics purposes.[6]Hadoop uses a distributed computing architecture consisting of multiple servers using commodity hardware, making it relatively inexpensive to scale and support extremely large data stores. Apache Hadoop, a nine-year-old open- source data-processing platform first used by internet giants including Yahoo and Facebook, leads the big –data revolution. Cloudera introduced a commercial support for expertise in 2008, and MapR and Hortonworks piled on in 2009 and 2011, respectively.[7] Clusters runsHadoop's Open source distributed processing software on low-cost commodity computers. Typically one machine in the cluster is designated as the NameNode and another machine the Job Tracker; these are the masters. The rest of the machines in the cluster act as both

DataNodes and TaskTracker; these are the slaves. Hadoop clusters are often referred to as “shared nothing”. Hadoop clusters are known for boosting the speed of data analysis applications. If a cluster’s processing power is overwhelmed by growing volumes of data, additional cluster nodes can be added to increase throughput.[8]Hadoop clusters also are highly resistance to failure because each piece of data is copied onto other cluster nodes, which ensures that the data is not lost if one node fails. As of early 2013, Facebook was recognized as having the largest Hadoop cluster in the world. Other prominent user includes Google, Yahoo and IBM.[4][9]

**7.2 Hadoop Architecture:** At a high-level, Hadoop operates on the philosophy of pushing analysis code close to the data it is intended to analyze rather than requiring code to read data across a network. As such, Hadoop provides its own file system, named as *Hadoop File System* or *HDFS*. When you upload your data to the HDFS, Hadoop will partition your data across the cluster (keeping multiple copies of it in case your hardware fails), and then it can deploy your code to the machine that contains the data upon which it is intended to operate.[10]

HDFS organizes data by keys and values. Each piece of data has a unique key and a value associated with that key.[8] Relationships between keys, if they exist, are defined in the application, not by HDFS. And in practice, we analyze a problem domain in order realize the full power of Hadoop (see the next section on MapReduce).[11]

The components that comprise Hadoop are:

**7.3 HDFS:** The Hadoop file system is a distributed file system designed to hold huge amounts of data across multiple nodes in a cluster. Hadoop provides both an API and a command-line interface to interacting with HDFS.

## VIII. MAP REDUCE APPLICATION

The next details of MapReduce, but in short, MapReduce is a functional programming paradigm for analyzing a single record in your HDFS. It then assembles the results into a consumable solution. The Mapper is responsible for the data processing step, while the Reducer receives the output from the Mappers and sorts the data that applies to the same key.

**8.1 Partitioner:** The partitioner is responsible for dividing a particular analysis problem into workable chunks of data for use by the various Mappers. The Hash Partitioner is a partitioner that divides work up by “rows” of data in the HDFS, but you are free to create your own custom partitioner if you need to divide your data up differently.

**8.2 Combiner:** If, for some reason, you want to perform a local reduce that combines data before sending it back to Hadoop, then you’ll need to create a combiner. A combiner performs the reduce step, which groups values together with their keys, but on a single node before returning the key/value pairs to Hadoop for proper reduction.

**8.3 Input Format:** Most of the time the default readers will work fine, but if your data is not formatted in a standard way, such as “key, value” or “key [tab] value”, then you will need to create a custom Input Format implementation.

**8.4 Output Format:** Map Reduce applications will read data in some Input Format and then write data out through an Output Format. Standard formats, such as “key [tab] value”, are supported out of the box, but if you want to do something else, then you need to create your own Output Format implementation.

Additionally, Hadoop applications are deployed to an infrastructure that supports its high level of scalability and resilience.[12]These components include.

**8.5 NameNode:** The NameNode is the master of the HDFS that controls slave DataNode daemons; it understands where all of your data is stored, how the data is broken into blocks, what nodes those blocks are deployed to, and the overall health of the distributed filesystem.

**8.6 Secondary Name Node:** The Secondary Name Node monitors the state of the HDFS cluster and takes “snapshots” of the data contained in the NameNode. If the NameNode fails, then the Secondary NameNode can be used in place of the Name Node. This does require human intervention, however, so there is no automatic failover from the Name Node to the Secondary Name Node, but having the Secondary Name Node will help ensure that data loss is minimal. Like the Name Node, each cluster has a single Secondary Name Node.

**8.7 Job Tracker:** The Job Tracker daemon is your liaison between your application and Hadoop itself. There is one Job Tracker configured per Hadoop cluster and, when you submit your code to be executed on the Hadoop cluster, it is the Job Tracker’s responsibility to build an execution plan. This execution plan includes determining the nodes that contain data to operate on, arranging nodes to correspond with data, monitoring running tasks, and relaunching tasks if they fail.

**8.8 Task Tracker:** Similar to how data storage follows the master/slave architecture, code execution also follows the master/slave architecture. Each slave node will have a Task Tracker daemon that is responsible for executing the tasks sent to it by the Job Tracker and communicating the status of the job with the Job Tracker.

Figure 2 shows the relationships between the master node and the slave nodes. The master node contains two important components: the Name Node, which manages the cluster and is in charge of all data, and the Job Tracker, which manages the code to be executed and all of the Task Tracker daemons. Each slave node has both a Task Tracker daemon as well as a Data Node: the Task Tracker receives its instructions from the Job Tracker and executes map and reduce processes, while the Data Node receives its data from the Name Node and manages the data contained on the slave node. And of course there is a Secondary Name Node listening to updates from the Name Node.[5][13] Hadoop recently became very popular with several different vendors offering distributions with a set of optimizations and features. Data meer is committed to supporting all of the Hadoop distributions and allows easy migration from one to other. Data meer isolates the end user from the lower level technical details and provides an simple though powerful web bases application on top that abstracts all interactions with Hadoop. Some of distributors preferring Hadoop platforms are: Apache, Amazon, Cloud era, EMC, Horton works, IBM, MapR, Microsoft etc.

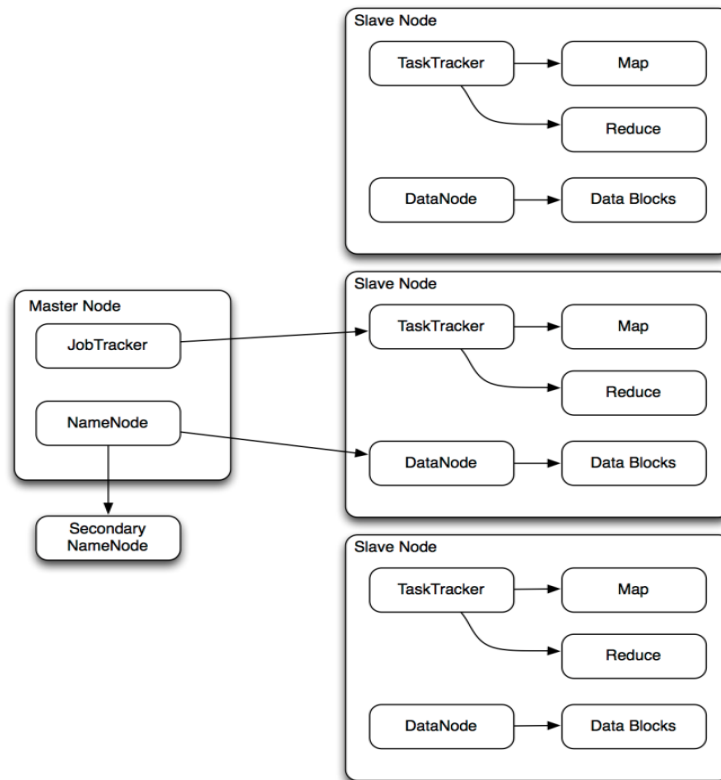


Fig 2: Hadoop application and Infrastructure interactions.

### IX. MAP REDUCE

Map Reduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.[14] A Map Reduce program is composed of a Map () procedure that performs filtering and sorting and a Reduce () procedure that performs a summary operation. The "Map Reduce System" manages the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance. The model is inspired by the map and reduce functions commonly used in functional programming, although their purpose in the Map Reduce framework is not the same as in their original forms. The key contributions of the Map Reduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved for a variety of applications by optimizing the execution engine once.[15] As such, a single-threaded implementation of Map Reduce (such as MongoDB) will usually not be faster than a traditional (non-Map Reduce) implementation, any gains are usually only seen with multi-threaded implementations. The use of this model is beneficial only when the optimized distributed shuffle operation (which reduces network communication cost) and fault tolerance play. Optimizing the communication cost is essential to a good Map Reduce algorithm.[16] Google invented Map Reduce to deal the issues of how to parallelize the computation, distribute the data, and handle failures.

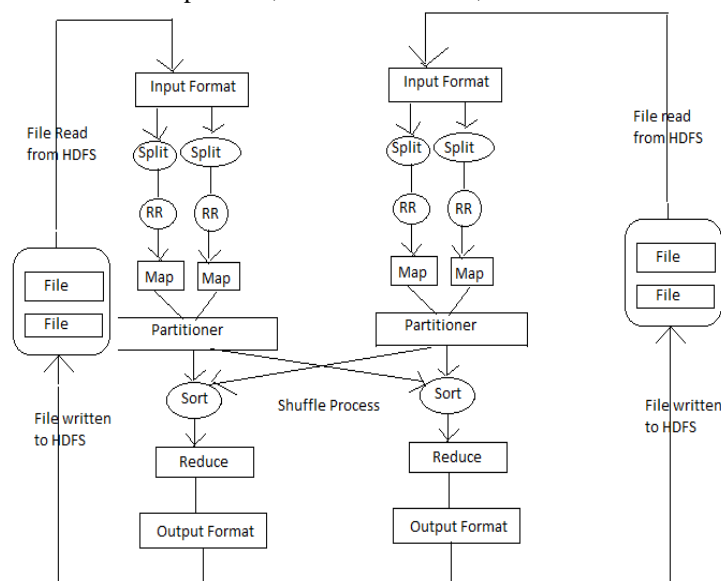


Fig 3: Flow Chart for Map Reduce

**9.1 Prepare the Map() input:** The system splits the input files into M pieces and then starts up M Map workers on a cluster of machines.

**9.2 Run the user-defined Map() code:** The Map worker parses key-value pairs out of the assigned split and passes each pair to the user- defined Map function. The intermediate key-value pairs produced by the Map function are buffered in memory. Periodically, the buffered pairs are written to local disk, partitioned into R regions for shading purposes by the partitioning function that is given the key and the number of reducers R and returns the index of the desired reducer.

**9.2.1 Shuffle the Map output to the Reduce processors:** When ready, a reduce worker reads remotely the buffered data from the local disks of the map workers. When a reduce worker has read all intermediate data, it sorts the data by the intermediate keys so that all occurrences of the same key are grouped together. Typically many different keys map to the same reduce task.

**9.2.2 Run the user-defined Reduce () code:** The reduce worker iterates over the sorted intermediate data and for each unique intermediate key encountered, it passes the key and the corresponding set of intermediate values to the user's Reduce function.

**9.2.3 Produce the final output:** The final output is available in the R output files (one per reduce task). Optionally, a combiner can be used between map and reduce as an optimization. The combiner function runs on the output of the map phase and is used as a filtering or an aggregating step to lessen the data that are being passed to the reducer. In most of the cases the reducer class is set to be the combiner class so that we can save network time.[17] Map Reduce provides programmers a simple parallel computing paradigm. Inspired by functional programming, the computing paradigm of Map Reduce is really simple and easy to understand. Because of automatic parallelization, no explicit handling of data transfer and synchronization in programs, and no deadlock. This model is very attractive. Map Reduce is also designed to process very large data that is too big to fit into the memory.[18] To achieve that, Map Reduce employs a data flow model, which also provides a simple I/O interface to access large amount of data in distributed file system. It also exploits data locality for efficiency. In most cases, we don't need to worry about I/O at all. Now let's look into several examples, which will also help us understand the limitations of Map Reduce.

### **Example 1: Sort**

The essential part of the Map Reduce framework is a large distributed sort. So we just let the framework do the job in this case while the map is as simple as emitting the sort key and original input. The reduce operator is an identity function.

1. Public class Sorting Thread
2. Sorting(Array List al, int from Index, int to Index, intfno)
3. If(loc>begin+1)
4. Sorting stL=new Sorting (arl,begin,loc-1,fno)
5. Sorting left start
6. If (loc<end-1)
7. SortingThread stR=new Sorting Thread (arl,loc+1,end,fno)
8. Sorting right start
9. Public void run()
10. Set left=begin
11. Set right=end
12. While(left<right)
13. Record left=array list(left)
14. Record right= array list(right)
15. Value left= Record[fno]
16. Value right= Record[fno]
17. If (valR>0)
18. Right—
19. Else If(left==right)
20. Set Loc=left
21. Split(left)
22. End If
23. Set temp1=arl(left)
24. Set temp2=arl(right)
25. Swap(left,temp2)
26. Swap(right,temp1)
27. While(right>left)
28. Rec1=arl.get(left)
29. Rec2=arl.get(right)

```
30. valueL=recL[fno]
31. valueR=recR[fno]
32. If(valL>0)
33. Break
34. Else left++
35. End while
36. If(left==right)
37. Loc=left
38. Temp1=arl(left)
39. Temp2=arl(right)
40. Swap arl(left,temp2)
41. Swap arl(right,temp1)
42. End run
43. End class
```

### **Example 2: Join**

A join combines records from two or more data sets by a common key. There are several ways to implement join in Map Reduce. A straightforward approach is the reduce-side joins that take advantage of the identical keys to the same reducer. In practice, join, aggregation, and sort are frequently used together, e.g. finding the student scores maximum during the period. In Map Reduce, this has to be done in multiple phases. The first phase filters the data base on the click timestamp and joins the client and click log datasets. The second phase does the aggregation on the output of join and the third one finishes the task by sorting the output of aggregation.

```
1. Public class JoinThread
2. JoinThread(ArrayList r1,ArrayList r2,ArrayList r3,int r)
3. Public void run()
4. String A[]=rsl1.get(rno)
5. String B[]=rsl2.get(rno)
6. String C[]=rsl3.get(rno)
7. For(i=1;i<=4;i++)
8. A=parse integer(A[i])
9. B=parse integer(B[i])
10. C=parse integer(C[i])
11. Total[i-1]=a+b+c
12. End for
13. End run
14. End class
```

### **Example 3: Aggregation**

Aggregation is a simple analytic calculation such as counting the number of access or users from different colleges. Word count, the hello world program in the world of Map Reduce, is an example of aggregation

```
1. Public class Frequency Thread
2. FrequencyThread(String nm ,Array List al1,ArrayList al2)
3. Public void run()
4. If (alA.contains (name))
5. Int i= alA.index Of(name)
6. Int f= Integer parse int((String)alB.get(i))
7. f++
8. set alB(I,f+""")
9. Else
10. alA.add(name)
11. Add 1 to alB
12. End if
13. End class
```

Map Reduce is useful for batch processing on terabytes or peta bytes of data stored in Apache Hadoop.

## **X. SCOPE IN REAL-WORLD FOR BIG DATA ANALYTICS**

With data growing so rapidly and the rise of unstructured data accounting for 90% of the data today, the time has come for enterprises to re-evaluate their approach to data storage, management and analytics.[19] Legacy systems will remain necessary for specific high-value, low-volume workloads, and complement the use of Hadoop -optimizing the data management structure in your organization by putting the right system.

The cost effectiveness, scalability, and streamlined architectures of Hadoop will make the technology more and more attractive.

- \* Consumer product companies and retail organizations are monitoring social media like Face book and Twitter to get an unprecedented View into customer behavior, preferences, and product perception.
- \* Manufacturers are monitoring minute vibration data from their equipment, which changes slightly as it wears down, to predict the optimal time to replace or maintain. Replacing it too soon wastes money; replacing it too late triggers an expensive work stoppage
- \* Manufacturers are also monitoring social networks, but with a different goal than marketers: They are using it to detect aftermarket support issues before a warranty failure becomes publicly detrimental.
- \* The government is making data public at both the national, state, and city level for users to develop new applications that can generate public good.

### **XI. LIMITATION OF HADOOP/MAP REDUCE**

- \* Cannot control the order in which the maps or reductions are run.
- \* For maximum parallelism, you need Maps and Reduces to not depend on data generated in the same Map Reduce job (i.e. stateless)
- \* A database with an index will always be faster than a Map Reduce job on un indexed data
- \* Reduce operations do not take place until all Maps are complete (or have failed then been skipped)
- \* General assumption that the output of Reduce is smaller than the input to Map; large data source used to generate smaller final values.

### **XII. CONCLUSION**

This paper is a systematic study of various tools for Big data analytics. Big data is very challenging research area. Data is too big to process using conventional tool of data processing. Academia and industry has work together to design and develop new tools and technologies which effectively handle to processing of Big Data. Big Data is an emerging trend and there is immediate need for new machine learning and data mining techniques to analyse massive amount of data in future. After years of practice, the community has realized these problems and try to address them in different ways. In this paper we conclude Hadoop& Map Reduce tools. These tools are used for storing of structured/unstructured data in large amount this can be done through HDFS in which data is stored in clusters. Map Reduce is used to process such a large amount of data in systematic manner. By using Hadoop, Map Reduce algorithms we can access such a large amount of data in very efficient time. We try to simulate a university system by using the basic algorithms of Hadoop and Map Reduce. We can also process this data with the queries or through the

### **REFERENCE**

- [1] A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices" Noids:2013,pp. 404-409,8-10 Aug. 2013.-
- [2] Jimmy Lin, Chris Dyer," Data-Intensive Text Processing with Map Reduce", Manuscript prepared April 11, 2010.
- [3] Tom White," Hadoop: The Definitive Guide", O'Reilly Media, Inc., 2009.
- [4] Lu, Huang, Ting-tin Hu, and Hai-shanChen,"Research on Hadoop cloud computing Model and its applications" Hangzhou, china: 2012, pp.59-63, 21-24 Oct.2012.
- [5] Geczy, P., Izumi,N.,&Hasida, K. (2012). Cloud sourcing: Managing cloud adoption. Global Journal of Business Research, 6(2), 57-70
- [6] Cole, B. (2012). Looking at business size, budget when choosing between SaaS and hosted ERP. E-guide: Evaluating SaaS vs. on premise for ERP -systems. Retrieved from [http://docs.media.bitpipe.com/io\\_10x/io\\_104515/item\\_548729/SAP\\_sManERP\\_IO%23104515\\_EGuide\\_061212.pdf](http://docs.media.bitpipe.com/io_10x/io_104515/item_548729/SAP_sManERP_IO%23104515_EGuide_061212.pdf)
- [7] Cloud era MapReduce Algorithms, 2009 Cloud era, inc.
- [8] <http://searchcloudcomputing.techtarget.com/definition/Infrastructure-as-a-Service-IaaS>
- [9] <http://www-01.ibm.com/software/data/infosphere/hadoop/what-is-big-data-analytics.html>
- [10] [https://en.wikipedia.org/wiki/Big\\_data](https://en.wikipedia.org/wiki/Big_data)
- [11] <http://www.computerworld.com/article/2690856/big-data/8-big-trends-in-big-data-analytics.html>
- [12] <http://searchbusinessanalytics.techtarget.com/definition/Hadoop-cluster>
- [13] <http://www.informit.com/articles/article.aspx?p=2008905>
- [14] [https://en.wikipedia.org/wiki/Big\\_data](https://en.wikipedia.org/wiki/Big_data)
- [15] [http://www.technologytransfer.eu/article/98/2012/1/What\\_Is\\_Big\\_Data\\_and\\_Why\\_Do\\_We\\_Need\\_It\\_.html](http://www.technologytransfer.eu/article/98/2012/1/What_Is_Big_Data_and_Why_Do_We_Need_It_.html)
- [16] <http://asmarterplanet.com/blog/2014/10/big-data-analytics-caring.htm>
- [17] <http://arunxjacob.blogspot.in/2011/04/hdfs-file-size-vs-allocation-other.htm>
- [18] Apache Giraph Project, <http://giraph.apache.org/>
- [19] [http://www.webopedia.com/big\\_data\\_analytics.html](http://www.webopedia.com/big_data_analytics.html)
- [20] <http://searchbusinessanalytics.techtarget.com/essentialguide/Guide-to-big-data-analytics-tools-trends-and-best-practices>



**ABOUT AUTHOR**



**Dr. Kusum Yadav** has obtained her master's degree in Computer Application from Indira Gandhi National Open University, New Delhi (India) and Ph.D., degree in Computer Science from JJJ University, Rajasthan (India). She is currently working as an Assistant Professor, College of Computer Engineering & Science, University of Hail, Kingdom of Saudi Arabia - a well known University of the Kingdom. She has 12 years' teaching experience at both UG and PG Level. She guides PG/M.Phil., and Ph.D., Research Scholar in the field of Computer Science. She is serving as a Member of Editorial Boards, Technical Committees and Reviewer of many National and International Journals including Thomson Reuters and Elsevier. Her research interests include Cloud Computing, Image Processing, Network Security and Data Mining. She has published a number of research papers in National and International Journals and attended and presented research papers in several National and International conferences.



**Dr. Rajender Kumar Bathla**, author of the present paper is working as an Assistant Professor in Computer Science & Engineering Department at HCTM Technical Campus, Kaithal. He started his professional career from Haryana College of Technology & Management, Kaithal in 2004 just after graduating his MCA Degree from MDU Rohtak and later he accomplished M.Tech. Degree with specialization in Information Technology from M.M. University Mullana, Haryana in 2009 and earned Ph.D. Degree also in 2013 in the Discipline of Computer Science & Engineering, Faculty of Technology from Private State University, (India) under supervision and guidance of his Senior Colleague & Advisor Prof. (Dr.) V. N. Maurya, Former Founder Director, Vision Institute of Technology Aligarh, U.P. Technical University, India. His research areas are mainly Software Testing and Data Structures and published several research papers in Indian and Foreign journals and Conferences. **Dr. Rajender Kumar Bathla** is an associated with several research and profession bodies, **IEEE, ISTE, CSTA, UACEE, IACSIT, IIST, IAEST, IAENG and ISA**. Apart from this, he is also on role of Editor and Reviewer of several Foreign leading International journals such as of **IJSEA USA', 'IJOAIEST-INDIA', 'IJECSE- TAIWAN', 'IJOAR&T –USA**.