

**SIX WEEKS SUMMER TRAINING**

REPORT

On

**Big Data**

Submitted by

**VIJAY SINGH KHOLIYA**

**Registration No : 11603123**

**Programme Name : B.Tech CSE**

Under the Guidance of

**Miss. Gurpreet Kaur**

**School of Computer Science & Engineering**

**Lovely Professional University, Phagwara**

**(June-July,2018)**

**DECLARATION**

I hereby declare that I have completed my six weeks summer training at Allsoft Solutions pvt. Ltd., Mohali from 4-Jun-2018 to 16-Jul-2018 under the guidance of Miss. Gurpreet Kaur. I have declare that I have worked with full dedication during these six weeks of training and my learning outcomes fulfill the requirements of training for the award of degree of Bachelor of Technology in Computer Science and Engineering, Lovely Professional University, Phagwara.

Name of Student : Vijay Singh Kholiya

Registration No. : 11603123

Date : 30 June 2018

**ACKNOWLEDGEMENT**

It is my pleasure to be indebted to various people, who directly or indirectly contributed in the development of this work and who influenced my thinking, behavior, and acts during the course of study.

I am thankful to Ms. Gurpreet Kaur for her support, cooperation, and motivation provided to me during the training for constant inspiration, presence and blessings.

Lastly, I would like to thank the almighty and my parents for their moral support and my friends with whom I shared my day-to-day experience and received lots of suggestions that improved my quality of work.

VIJAY SINGH KHOLIYA



**TABLE OF CONTENTS :**

1. Introduction
2. Technology Learnt - Big Data
3. Core Components of Hadoop
4. History of Hadoop
5. Hadoop Ecosystem
6. Core Components – HDFS
7. Features of HDFS
8. Implementation of Big Data in real life
9. Reason for Choosing the Big Data technology
10. Learning Outcomes
11. Why this Big Data?
12. Future of Big Data
13. Gantt Chart
14. Bibliography

**INTRODUCTION**

The term Big Data refers to all the data that is being generated across the globe at an unprecedented rate. This data could be either structured or unstructured. Big data is a term for data sets that are so large or complex that traditional data processing application softwares are inadequate to deal with them.

The basic requirements for working with big data are the same as the requirements for working with datasets of any size. However, the massive scale, the speed of ingesting and processing, and the characteristics of the data that must be dealt with at each stage of the process present significant new challenges when designing solutions.

The goal of most big data systems is to surface insights and connections from large volumes of heterogeneous data that would not be possible using conventional methods.

Challenges include capture, storage, analysis, search, sharing, transfer, visualization, querying, updating and information privacy.

**Big Data has certain characteristics and hence is defined using 4Vs namely:**

1. **Volume:-** Big data is always large in volume. It actually doesn't have to be a certain number of petabytes to qualify. If your store of old data and new incoming data has gotten so large that you are having difficulty handling it, that's big data. Remember that it's going to keep getting bigger.
2. **Velocity :-**Velocity or speed refers to how fast the data is coming in, but also to how fast we need to be able to analyze and utilize it. If we have one or more business processes that require real-time data analysis, we have a velocity challenge. Solving this issue might mean expanding our private cloud using a hybrid model that allows bursting for additional compute power as-needed for data analysis.
3. **Variety:-** Variety points to the number of sources or incoming vectors leading to databases. That might be embedded sensor data, phone conversations, documents, video uploads or feeds, social media, and much more. Variety in data means variety in databases – we will almost certainly need to add a non-relational database if you haven't already done so.
4. **Veracity :-**Veracity is probably the toughest nut to crack. If we can't trust the data itself, the source of the data, or the processes we are using to identify which data points are important, we have a veracity problem. One of the biggest problems with big data is the tendency for errors to snowball. User entry errors, redundancy and corruption all affect the value of data.  We must clean our existing data and put processes in place to reduce the accumulation of dirty data going forward.

**TECHNOLOGY LEARNT- BIG DATA :**

Big data is an evolving term that describes any voluminous amount of structured, semistructured and unstructured data that has the potential to be mined for information.

Big data is a broad, rapidly evolving topic. While it is not well-suited for all types of computing, many organizations are turning to big data for certain types of work loads and using it to supplement their existing analysis and business tools.

Big data systems are uniquely suited for surfacing difficult-to-detect patterns and providing insight into behaviors that are impossible to find through conventional means.

[Apache Hadoop](https://intellipaat.com/big-data-hadoop-training/#certification) is a Big Data framework that is part of the Apache Software Foundation. Hadoop is an open source software project that is extensively used by some of the biggest organizations in the world for distributed storage and processing of data on a level that is just enormous in terms of volume.

It consists of computer clusters built from commodity hardware. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model.

Hadoop splits files into large blocks and distributes them across nodes in a cluster.

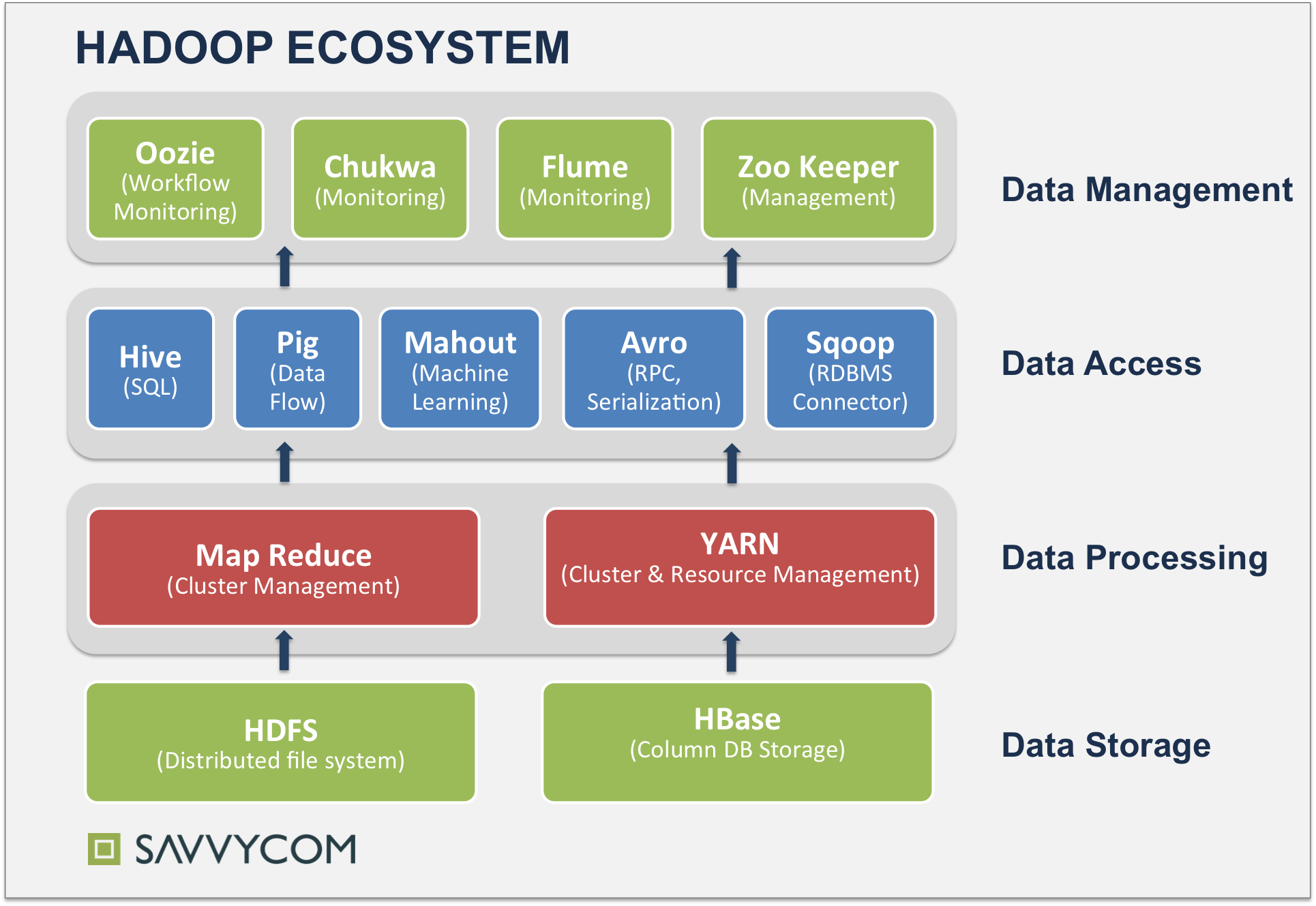
**Core Components of Hadoop:**

* Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
* Hadoop MapReduce – an implementation of the MapReduce programming model for large scale data processing.

**History of Hadoop:**

* The genesis of Hadoop came from the Google File System paper that was published in October 2003.
* This paper spawned another research paper from Google – MapReduce: Simplified Data Processing on Large Clusters.
* Development started on the Apache Nutch project, but was moved to the new Hadoop subproject in January 2006.
* Doug Cutting, who was working at Yahoo! at the time, named it after his son's toy elephant.

**HADOOP ECOSYSTEM:**

****

**CORE COMPONENTS – HDFS :**

1. Hadoop File System was developed using distributed file system design.
2. It is run on commodity hardware.
3. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware.
4. HDFS holds very large amount of data and provides easier access.
5. To store such huge data, the files are stored across multiple machines.
6. HDFS also makes applications available to parallel processing.

**Features of HDFS :**

1. It is suitable for the distributed storage and processing.
2. Hadoop provides a command interface to interact with HDFS.
3. The built-in servers of namenode and datanode help users to easily check the status of cluster.
4. Streaming access to file system data.
5. HDFS provides file permissions and authentication.

HDFS follows the master-slave architecture and it has the following elements :

1. **Name Node :**

* The system having the namenode acts as the master server and it does the following tasks:
* Manages the file system namespace.
* Regulates client’s access to files.
* It also executes file system operations such as renaming, closing, and opening files and directories.

1. **Data Node :**

* The datanode is a commodity hardware having any operating system and datanode software. For every node (Commodity hardware/System) in a cluster, there will be a datanode. These nodes manage the data storage of their system.
* Datanodes perform read-write operations on the file systems, as per client request.
* They also perform operations such as block creation, deletion, and replication according to the instructions of the namenode.

**Implementation of Big Data in real life :**

1. Public Sector Services.
2. Healthcare contributions.
3. Learning Services.
4. Insurance Services.
5. Industrialized and Natural Resources.
6. Transportation Services.
7. Banking Sectors and Fraud Detection

**Big Data Contributions to Healthcare :**

The big data is in extended use in the field of medicine and healthcare. As the technology raises the cost of health care is also increasing more and more. Big data is a great helping hand in this issue.

It is a great help for even physicians to keep track of all the patients’ history. The link to the patient’s history can be accessed only by the patient and his particular physician.Once a patient gets treated his name and his data will be stored in the database safely forever and whenever required, the doctor can have a view of it.

A large number of medical devices are there which are big data oriented. Today data is used to such an extent that doctor prescribes the medicines without even visiting the patient by knowing the heartbeat and temperature through the heart and temperature monitoring watch fitted on the patient’s hand that stays in a remote place.

**Big Data Contributions to Public Sector :**

Big data provides a large range of facilities to the government sectors including the power investigation, deceit recognition, fitness interconnected exploration, economic promotion investigation and ecological fortification.

Big data is even used to examine the food based infections by the FDA. Big data results are fast which outputs to quicker well-being. Also in the investigation of a huge volume of communal complaints uses the big data analytics.

This same analytics are utilized in the course of health check statistics in urgency and resourcefully for quicker pronouncement manufacture and to become aware of mistrustful or falsified declarations.

**Big Data Contributions to Learning :**

Big data has great influence in the education world too. Today almost every course of learning is present online. Along with the online learning, there are many examples of the use of big data in the education industry.

Applications named as the Bubble Score allow teachers to convey multiple-choice assessments through mobile devices and notch up paper tests through the cameras of the mobile phones. Equipment like this usually assists teachers to send out the outputs to rank books and trail development all along distinct characteristics.

**Big Data Contributions to Insurance Services :**

Be deficient in modified services, be short of adapted charging and the need of beleaguered services to fresh fragments and to specific market segments are some of the main challenges.

Big data is the technology tool that is being used in the production to offer purchaser insights for see-through and simpler commodities, by finding out and foreseeing buyer behavior from side to side information obtained from internet websites including the social media as well as CCTV video recording.

Scam discovery has also been improved.

**Big Data Contributions to Industrial and Natural Resources :**

The high demand of the natural sources on this earth is challenging the high volume as well as the velocity of big data. Similarly, a great quantity of data commencing the built-up industry is unexploited. The unused data avoids advanced eminence of merchandise, power competence, dependability, and improved income boundaries.

In the natural wealth industry, big data enables for analytical modelling to sustain judgment creation that is used to consume and incorporate huge amounts of information from geographical information, graphical information, manuscript and chronological statistics.

Big data has as well been worn in finding the solution to the development of confrontations and to grow aggressive improvements in the middle of former settlements.

**Big Data Contributions to Transportations :**

A number of claims of big data by the public sector, private associations and personal use include-

* Private sector uses the big data in traffic management, direction preparation, intellectual transportation arrangements and overcrowding administration
* Private sector uses the big data in income administration, industrial improvements, logistics and for reasonable benefit
* Personal use of the big data comprises direction forecasting to accumulate on petroleum and period, for tour activities in seeing the sights etc.

**Big Data Contributions to Banking Zones and Fraud Detection :**

Big data is hugely used in the fraud detection in the banking sectors. In banking sectors as the big data is implemented, it finds out all the mischief tasks done.

It detects the misuse of credit cards, misuse of debit cards, archival of inspection tracks, venture credit hazard treatment, business clarity, customer statistics alteration, public analytics for business, IT action analytics, and IT strategy fulfillment analytics. The SEC uses this big data in order to keep a track of all the commercial market movements.

In businesses big data helps a lot in knowing the shopping patterns of customers and CRM tactics of the competitors so that they can apply them in their businesses in order to improve the sales.

**Reason for Choosing the Big Data Technology :**

* Big Data is the emerging technology of the 21st century. It has fascinated the engineers a lot in this decade due to various reasons like the exponential growth of this technology in terms of market share, investment by business companies on big data technologies.
* Companies these days generate a gigantic amount of information irrespective of which industry they belong to and there is a need to store these data which are being generated so that they can be processed and not miss out on important information which could lead to a new breakthrough in their respective sector.
* In order to process these data, Big Data analytics is necessary. Another analyst states that “In the next few years, the size of the analytics market will evolve to at least one-third of the global IT market from the currents one-tenths”.
* Hence the need for professionals with experience in the field of analytics is in huge demand as organizations are looking to benefit themselves from the power of Big Data.
* According to an article on Forbes written by Louis Columbus based on a study “2014 IDG Enterprise Big Data Research”, it was found that an average enterprise will spend about $8M on Big Data related initiatives.
* The study also says that 74% predict Big Data will be in mainstream use in at least one business unit or department.
* A Forbes article on Big Data jobs in 2015 stated that “The advertised salary for technical professionals with Big Data expertise is $104,850 net of bonuses and additional compensations.
* Sample jobs in this category include Big Data Solution Architect, Linux Systems and Big Data Engineer, Big Data Platform Engineer, Lead Software Engineer, Big Data (Java, Hadoop, SQL) and others. With the Big Data market growing vigorously and as the demand for big data jobs overtakes the supply of talent, salary packages will remain alluring.
* Big Data Analytics is used everywhere. Based on an analysis by Wanted analytics it was found out that the top five industries hiring Big Data related expertise include Professional, Scientific and Technical Services (25%), Information Technology (17%), Manufacturing (15%), Finance and Insurance (9%) and Retail Trade (8%).

**LEARNING OUTCOMES**

* Through this training I get to learn a very new technology of the world i.e Big Data. Moreover Big Data isn’t a technology but a serious problem. People generating a huge amount of data daily that has to be efficiently analysed so that business companies can take the advantage of this powerful technology.
* Today, data is in fact everywhere. From every search on Google to each second spent on Facebook, everything is converted into data. Now with smartwatches, glasses and even smart clothes the world will become a data collection mechanism.

**Why this Big Data ?**

* Data is growing at a rapid pace. By 2020 the new information generated per second for every human being will approximate amount to 1.7MB.
* By 2020, the accumulated volume of big data will increase from 4.4 zettabytes to roughly 44 ZetaBytes or 44 trillion GB.
* Originally, data scientists maintained that the volume of data would double every two years thus reaching the 40 ZB point by 2020. That number was later bumped to 44ZB when the impact of IoT was brought into consideration.
* The rate at which data is created is increased exponentially. For instance, 40,000 search queries are performed per second (on Google alone), which makes it 3.46 million searches per day and 1.2 trillion every year.
* Every minute Facebook users send roughly 31.25 million messages and watch 2.77 million videos.
* The data gathered is no more text-only. An exponential growth in videos and photos is equally prominent. On YouTube alone,300 hours of video are uploaded every minute.
* Globally, the number of smartphone users will grow to 6.1 billion by 2020 (this will overtake the number of basic fixed phone subscriptions).
* In just 5 years the number of smart connected devices in the world will be more than 50 billion – all of which will create data that can be shared, collected and analyzed.

**Future of Big Data :**

## Machine Learning Will Be the Next Big Thing in Big Data :

## One of the hottest technology trends today is machine learning and it will play a big part in the future of big data as well.

## According to Ovum, Machine learning will be at the forefront of the big data revolution. It will help businesses in preparing data and conduct predictive analysis so that businesses can overcome future challenges easily.

## Privacy will be the biggest challenge :

## Whether it is the internet of things or big data, the biggest challenge for emerging technologies has been security and privacy of data.

## The volume of data we are creating right now and the volume of data that will be created in the future will make privacy even more important as stakes will be much higher.

## According to Gartner, more than 50% of business ethics violation by 2018 will be data related.

## Data security and privacy concerns will be the biggest hurdle for big data industry and if it fails to cope with it in an effective manner, we will see a long list of technology trends that became a fad very quickly.

## Chief Data Officer – A new position will emerge :

## According to Forrester, we will see the emergence of chief data officer as the new position and businesses will appoint chief data officers.

## Although, the appointment of chief data officer solely depend on the type of business and its data needs but the wider adoption of big data technologies across enterprises, hiring a chief data officer will become the norm.

## Data Scientists will be in high demand :

## As the volume of data grows and big data grows bigger, demand for data scientists, analysts and data management experts will shoot up.

## The gap between the demand for data professionals and the availability will widen. This will help data scientists and analysts draw higher salaries.

## Investments in in Big Data technology will be raised :

## According to IDC analysts, “Total revenues from big data and business analytics will rise from $122 billion in 2015 to $187 billion in 2019.” Business spending on big data will surpass $57 billion dollars this year.

## Although, the business investments in big data might vary from industry to industry, the increase in big data spending will remain consistent overall.

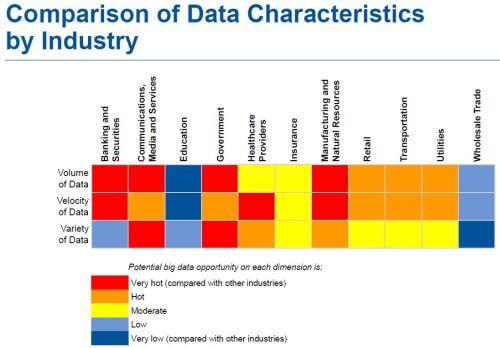
## Manufacturing industry will spend the most on big data technology while health care, banking, and resource industries will be the fastest to adopt.

## More developers will join the Big Data technology :

## According to statistics, there are six million developers currently working with big data and using advanced analytics. This makes up more than 33% of developers in the world.

## What’s even more amazing is that big data is just getting starting so will see a surge in a number of developer developing applications for big data in years to come.

## With the financial rewards in terms of higher salaries involved, developers will love to create applications that can play around with big data.



**GANTT CHART**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Activity** | **Week-1** | **Week-2** | **Week-3** | **Week-4** | **Week-5** | **Week-6** |
| 1. **Core Java** |  |  |  |  |  |  |
| 1. **HDFS** |  |  |  |  |  |  |
| 1. **Hive** |  |  |  |  |  |  |
| 1. **Sqoop** |  |  |  |  |  |  |
| 1. **MySQL** |  |  |  |  |  |  |
| 1. **MapReduce** |  |  |  |  |  |  |
| 1. **Pig** |  |  |  |  |  |  |
| 1. **HBase** |  |  |  |  |  |  |

**BIBLIOGRAPHY**

[**www.google.com**](http://www.google.com)

[**www.wikipedia.org**](http://www.wikipedia.org)

[**www.simplilearn.com/big-data-applications-in-industries-article**](http://www.simplilearn.com/big-data-applications-in-industries-article)

[**www.datamation.com**](http://www.datamation.com)

[**www.smartdatacollective.com**](http://www.smartdatacollective.com)

[**www.multisoftsystems.com**](http://www.multisoftsystems.com)

[**www.newgenapps.com**](http://www.newgenapps.com)

[**www.youtube.com**](http://www.youtube.com)

[**www.intellipaat.com**](http://www.intellipaat.com)

[**www.digitalocean.com**](http://www.digitalocean.com)

[**www.cloudxlab.com**](http://www.cloudxlab.com)