ACADGILD

UNDERSTANDING BIG DATA

A BEGINNER'S GUIDE

LEARN. DO. EARN

~ Table of Contents ~

About ACADGILD

ACADGILD is a technology education startup that aims to create an ecosystem for skill development in which people can learn from mentors and from each other. We believe that software development requires highly specialized skills that are best learned with guidance from experienced practitioners. Online videos or classroom formats are poor substitutes for building real projects with help from a dedicated mentor. Our mission is to teach hands-on, job-ready software programming skills, globally, in small batches of 8 to 10 students, using industry experts.

ACADGILD offers courses in

Enroll in our programming course
& Boost your career

ANDROID DEVELOPMENT

DIGITAL MARKETING

MACHINE LEARNING WITH R

BIG DATA ANALYSIS

JAVA FOR FRESHER

BIG DATA & HADOOP ADMINISTRATION

FULL STACK WEB DEVELOPMENT

NODE JS

CLOUD COMPUTING

FRONT END DEVELOPMENT (WITH ANGULARJS)

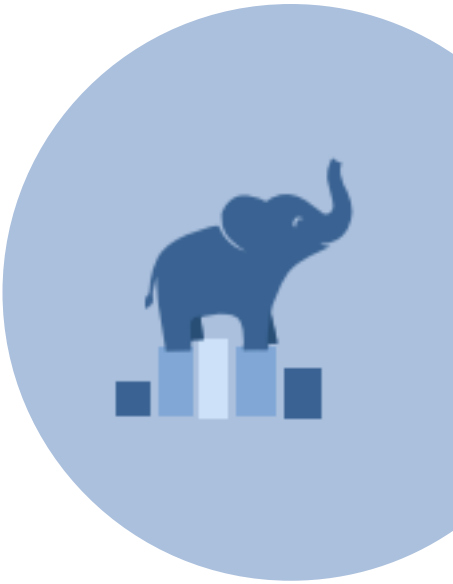
Watch this short video to know more about ACADGILD.

Disclaimer

This material is intended only for the learners and is not intended for any commercial purpose. If you are not the intended recipient, then you should not distribute or copy this material. Please notify the sender immediately or click here to contact us.

Published by
ACADGILD,
support@acadgild.com

1. Introduction to **Big Data**



This EBook is about Big Data, its meaning and current industry applications.
It's an accepted fact that Big Data has taken the world by storm and become a popular buzz phrase.

Eric Schmidt, ex-Google CEO, said in 2010,
"There were 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."

This shows how enormous Big Data has become. The interesting part is that Big Data can be implemented in almost every industry sector – banking, logistics, retail, e-commerce and social media. All these industries have adopted Big Data practices and have had outstanding success.

Having said that, let's look at some success stories from industries and sectors using Big Data technologies:

Banks across the world have used Big Data technology to ensure customer loyalty.

When issuing credit cards, banks have the luxury of analyzing loads of information to reduce the chance of credit-card fraud.

Facebook, a game-changer in today's world, has been able to predict political opinion, intelligence and the emotional stability of its users based on their activity on Facebook. Facebook has been able to feed us the news and ads we require based on our past behavior on Facebook, and this has been made possible only because of Big Data analytics.

Cricket websites like ESPN and Cricbuzz have been able to predict how the bowler will bowl or what kind of shot a batsman will play. These predictive capabilities have been enabled by Big Data analytics.

Retail and E-commerce industries have used Big Data to predict product demand, to analyze consumer behavior patterns, etc.

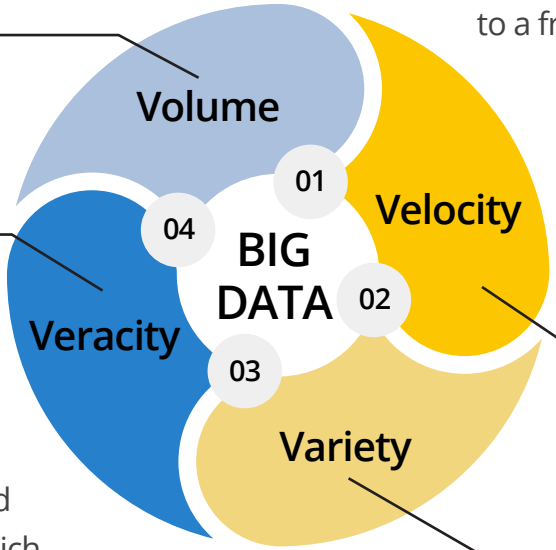
Companies have been able to understand the buying patterns of customers and have started to base their production on the results. Business decisions have been driven by Big Data analytics to predict product demand, consumer behavior patterns & supply chain mechanics. This information helps retailers sell exactly what the customer needs!

2. Characteristics of **Big Data**

Let's examine the characteristics of Big Data."Any data which has the four Vs, i.e. Volume, Variety, Veracity and Velocity can be termed as Big Data."

The four Vs are:

This represents the amount of data and is one of the main characteristics that makes data "big". This refers to the mass quantity of data that organizations have been trying to harness to improve decision- making across the enterprise.



This characteristic represents the speed of the data. It has changed the mindset of the past in that the data of yesterday, past hour or minute is now termed "recent" data.Data movement is now almost real time and the update window has been reduced to a fraction of a second. Because of the real-time nature of data creation, enterprises have invested a lot to develop Big Data solutions that can incorporate streaming data into business processes and decision-making.

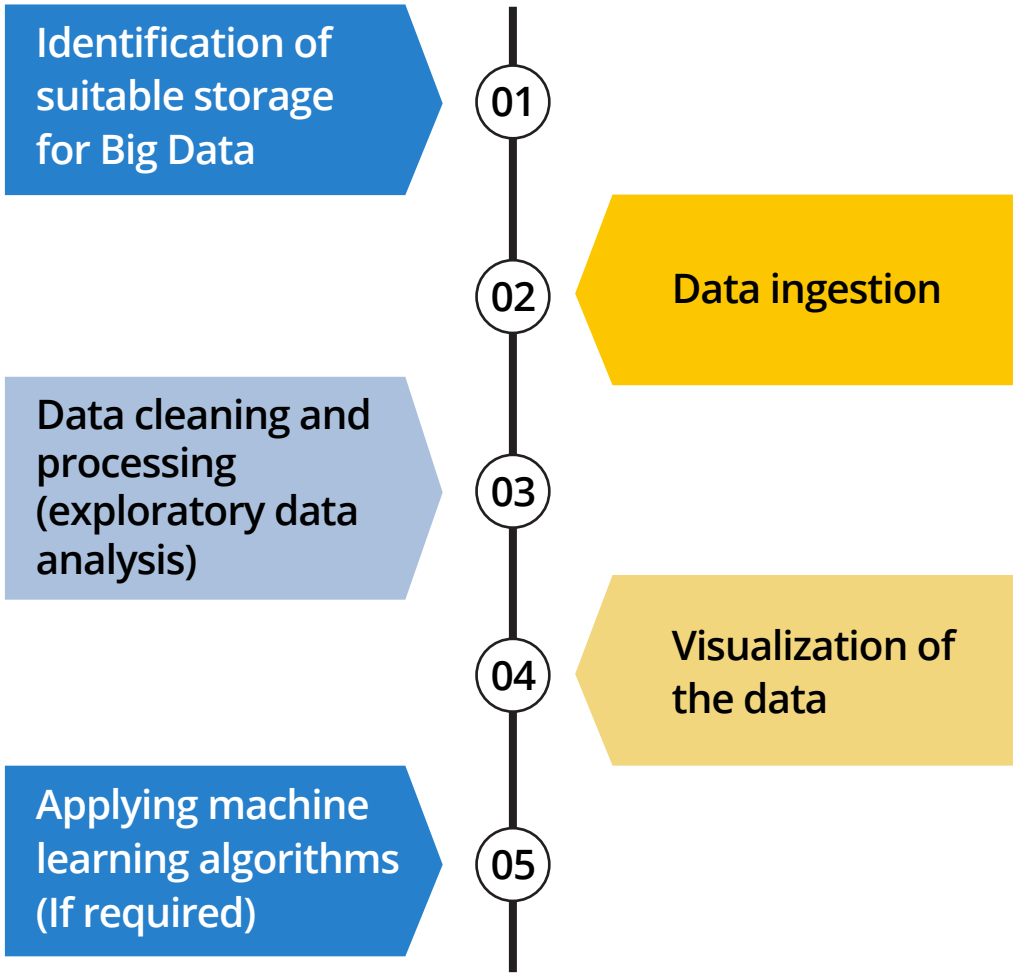
This refers to the different types of data and data resources.The world has moved beyond traditional means of structured data (like bank statements, which included information like date, amount, and time). New categories have been added to the list of data types.

Unstructured data, i.e., data that does not have a well-defined set of rules (for example, Twitter feeds, audio files, MRI images, web pages, web logs) has contributed immensely to the rise of Big Data.

This describes the trustworthiness of the data, i.e., calculations of noise, biases and abnormalities in the data.We may also define veracity as the level of reliability associated with certain types of data.

3. Processing Big Data

Let us now look at how Big Data is processed. The following are the steps involved:

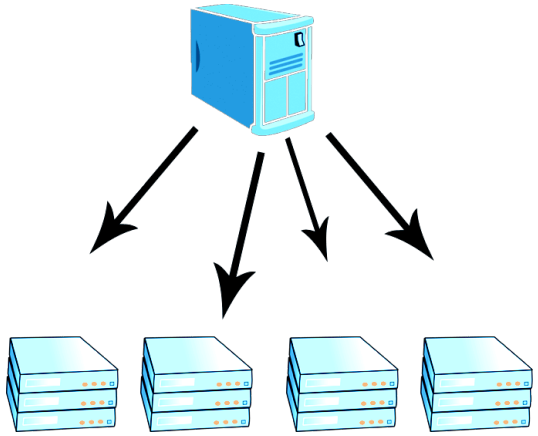


3.1 Identification of Suitable Storage for Big Data

The first step of Big Data analysis starts with the identification of appropriate storage. In the Big Data world, HDFS is one of the most-preferred file systems for storing Big Data.

Hadoop Distributed File System (HDFS)

This is a distributed file system that provides high-throughput access to application data. Data in a Hadoop cluster is broken down into smaller pieces (called blocks) and distributed throughout the cluster.



HDFS has a Master-Slave architecture, because there is a Master that takes control of all the Slaves. Here the Master is named as NameNode; the Slaves are DataNodes.

Listed below are the reasons why organizations prefer HDFS as an underlying storage for Big Data:

- HDFS is made up of commodity hardware, which makes it cost-effective.
- HDFS is a fault-tolerant file system and can store the same copy of data multiple times (replication of data). So, even if one copy is unavailable, the same copy can be retrieved from another location on the HDFS.
- HDFS can be easily used by many processing frameworks, such as:

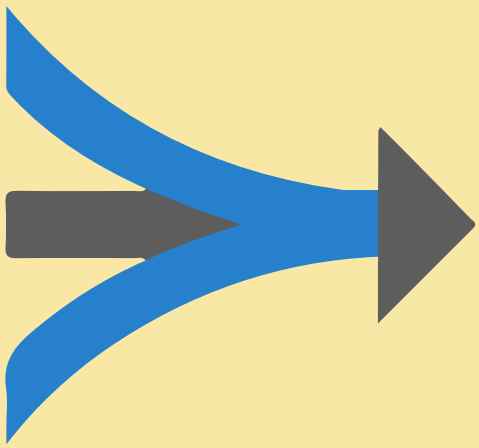
For detailed information on HDFS, please refer to our blog:

https://acadgild.com/blog/beginners-guide-for-hdfs/

3.2 Data Ingestion

Data ingestion refers to taking data from the source and placing it in a location where it can be processed. Since we're using the Hadoop HDFS as our underlying framework for storage and related echo systems for processing, let's look into the following data-ingestion options:

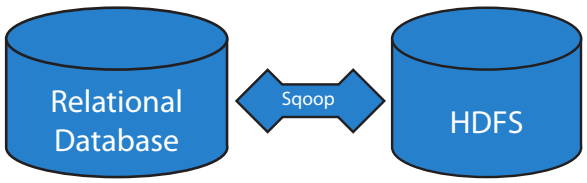
- Batch load from RDBMS using Sqoop
- Data loading from files
- Real-time data ingestion



Batch Load from RDBMS using Sqoop

Enterprises that use Hadoop are finding it necessary to transfer some of their data from traditional Relational Database Management Systems (RDBMS) to the Hadoop ecosystem.

Sqoop, an integral part of Hadoop, can perform this transfer in an automated fashion. Moreover, the data imported into Hadoop can be transformed with MapReduce before exporting them back to the RDBMS. Sqoop can also generate Java classes for programmatically interacting with the imported data.



Sqoop uses a connector-based architecture that allows it to use plugins for connecting to external databases.

Data loading from files

Use File Transfer Protocol (FTP) to transfer the data to client nodes, and then load the data using the ETL tool. Some ETL tools, like Informatica and Talent can be integrated.

Real-time data ingestion

Below is a list of some tools that enable real-time ingestion in HDFS:

Flume

Flume is a service for streaming logs into Hadoop. Apache Flume is a distributed, reliable and available service for efficiently collecting, aggregating and moving large amounts of streamed data into the Hadoop Distributed File System (HDFS).

Kafka

Apache Kafka supports a wide range of use cases as a general-purpose messaging system for scenarios where high throughput, reliable delivery, and horizontal scalability are important. Apache Storm and Apache HBase both work very well in combination with Kafka.

Storm

Storm is a distributed real-time computation system for processing large volumes of high-velocity data. Storm is extremely fast, as it has the ability to process over a million records per second, per node, on a cluster of modest size. Enterprises harness this speed and combine it with other data-access applications in Hadoop to prevent undesirable events or to optimize positive outcomes.

3.3 Data Cleaning and Processing (Exploratory Data Analysis)

After getting the data into HDFS, we should clean the data and convert it to a format that can be processed.

A common traditional approach is to use a sample of the large dataset that could fit in memory. But with the arrival of Big Data, processing tools like Hadoop can now be used to run many exploratory data-analysis tasks on full datasets, without sampling. Just write a MapReduce job, PIG or HIVE script, launch it directly on Hadoop over the full dataset, and get the results right back on your laptop.

Here's a more detailed discussion of various processing and cleaning methodologies provided by Hadoop:

Java MapReduce

Java MapReduce is a native MapReduce in Java. We write code in Java as map and reduce. This is suitable for data which has no structure or is semi-structured.

Pig

Pig is a data-flow language that allows users to write complex MapReduce operations in a simple scripting language. Then Pig transforms those scripts into a MapReduce job.

Hive

The Apache Hive data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time, this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it's inconvenient or inefficient to express this logic in HiveQL.

Impala

Cloudera Impala provides high-performance, low-latency SQL queries on data stored in popular Apache Hadoop file formats. The fast query responses enable interactive exploration and fine-tuning of analytic queries, rather than long batch jobs traditionally associated with SQL-on-Hadoop technologies. (You will often see the term "interactive" applied to fast queries with human-scale response times.)

3.4 Visualization of the Data

Data visualization is the presentation of processed data in a pictorial or graphical format. It enables decision-makers to see analytics presented visually so that they can grasp difficult concepts or identify new patterns. Using interactive visualization, you can take the concept a step further by using technology to drill down into charts and graphs for more detail, interactively changing what data you see and how it's processed.

There are multiple tools for visualizing processed data:

Tableau

Pig is a data-flow language that allows users to write complex MapReduce operations in a simple scripting language. Then Pig transforms those scripts into a MapReduce job.

QlikView

QlikView is a wonderful tool for data discovery that provides powerful tools to easily navigate a sea of data in an intuitive, easy and clear way and allow one to proceed from facts to Key Performance Indicators (KPI), and vice versa. QlikView can be used both as an advanced reporting tool and as a Business Intelligence KPI tool, making it a suitable base for continuous process improvements.

3.5 Application of the Machine Learning Algorithms

Machine learning explores the study and construction of algorithms that can learn from and make predictions about data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions, or decisions, rather than by following strict static program instructions.

Modern-day processing and visualization of Big Data has provided a strong platform for Machine learning algorithms to achieve better results for companies using techniques such as clustering, classifications, outlier detection and product recommenders. Historically, large datasets were not available or too expensive to acquire and store, and so machine-learning practitioners had to find innovative ways to improve models using rather limited datasets. With Hadoop as a platform that provides linearly scalable storage and processing power, you can now store ALL of the data in RAW format and use the full dataset to build better, more accurate models.

Conclusion