# Technologies for Handling Big Data

# Chapter Index

# Learning Objectives

- Explain distributed and parallel computing for Big Data
- Recognise Big Data technologies
- Describe cloud computing in reference to Big Data
- Discuss in-memory technology for Big Data
- Elucidate Big Data techniques

# 1. Distributed and Parallel Computing for Big Data

- Distributed computing is a method in which multiple computing resources are connected in a network and computing tasks are distributed across the resources, thereby increasing the computing power. Distributed computing is faster and more efficient than traditional computing, and, hence, of immense value when it comes to processing a huge amount of data in a limited time.
- To carry out complex computations, the processing power of a standalone personal computer can also be enhanced by adding multiple processing units, which can carry out the processing of a complex task by breaking it up into sub-tasks, and carrying out individual sub-tasks simultaneously. Such systems are often termed as parallel systems. The greater the processing power, the faster the computing.

# 2. Distributed and Parallel Computing for Big Data

- Following Table differentiates between distributed and parallel computing systems:

| Distributed Computing System | Parallel Computing System |
|---|---|
| An independent, autonomous system connected in a network for accomplishing specific tasks | A computer system with several processing units attached to it |
| Coordination is possible between connected computers that have their own memory and CPU | A common shared memory can be directly accessed by every processing unit in a network |
| Loose coupling of computers connected in a network that provides access to data and remotely located Resources | Tight coupling of processing resources that are used for solving a single, complex problem |

# 1.Introduction to Big Data Technologies

- A Big Data system is vastly different from other solution-providing systems and is based on the seven Vs, as described in previous chapter, namely: Volume, Velocity, Variety, Veracity, Variability, Value and Visualisation.
- A system that complies with these properties and happens to be robust to withstand unexpected events and scalable enough to accommodate future methodologies is qualified to be called as a Big Data system.
- A typical Big Data system consists of a setup that adheres to these seven Vs and provides a great infrastructure that can withstand the influx of huge datasets with high velocity, meanwhile providing effective mechanism to process the datasets by cleansing, shaping, filtering and sorting into meaningful information aimed towards making the data both user- and machine-friendly.

# 2.Introduction to Big Data Technologies

Hadoop

- Hadoop is an open-source platform that provides analytical technologies and computational power required to work with such large volumes of data.
- Hadoop platform provides an improved programming model, which is used to create and run distributed systems quickly and efficiently.
- A Hadoop cluster consists of single MasterNode and multiple worker nodes. The MasterNode contains a NameNode and JobTracker and a slave or worker node acts as both a DataNode and TaskTracker. Hadoop requires Java Runtime Environment (JRE) 1.6 or a higher version of JRE.
- There are two main components of Apache Hadoop – the Hadoop Distributed File System (HDFS) and the MapReduce parallel processing framework.

# 3.Introduction to Big Data Technologies

Hadoop

- Hadoop distributed file system (HDFS) is a fault-tolerant storage system in Hadoop. It stores large size files from terabytes to petabytes across different terminals.
- Data is replicated on three nodes: two on the same rack and one on a different rack. The file in HDFS is split into large blocks size of 64 MB by default (typically 64 to 128 megabytes) and each block of the file is independently replicated at multiple data nodes.
- The NameNode actively monitors the number of replicas of a block (by default 3 times). When a replica of a block is lost due to a DataNode failure or disk failure, the NameNode creates another replica of the block.

# 4.Introduction to Big Data Technologies

R

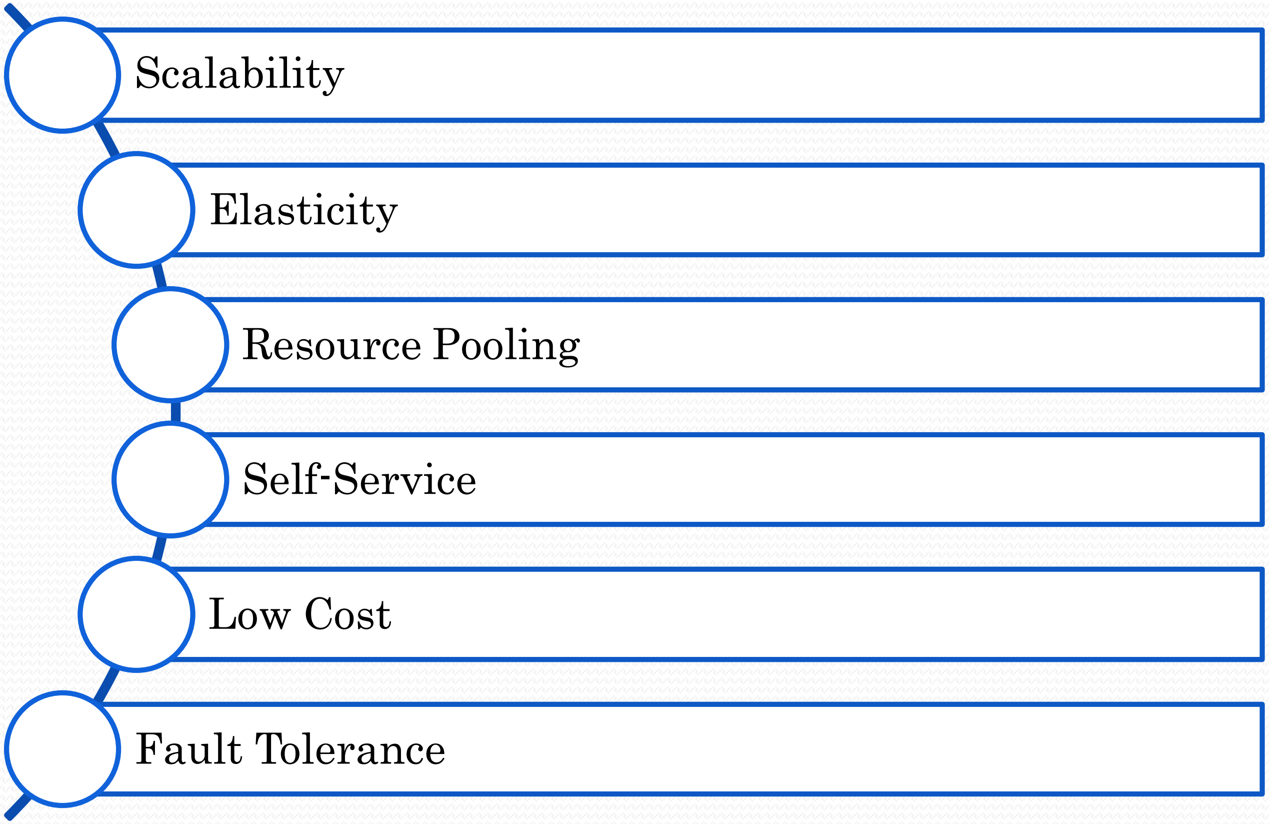
- R is an open source programming language and an application environment for statistical computing with graphics, developed by R Foundation for Statistical Computing.
- It is an interpreted language like Python and uses a command line interpreter. It supports procedural as well as generic functions with OOP.
- R is extensively used by data miners and statisticians, providing a vast variety of graphical and statistical techniques, with linear and nonlinear modelling, time-series analysis, classical statistical tests, clustering, classification and others.
- R is easily extendable and implementable through functions and available extensions.

# 1. Cloud Computing and Big Data

- One of the vital issues that organisations face with the storage and management of Big Data is the huge amount of investment to get the required hardware setup and software packages.
- Some of these resources may be overutilised or underutilised with varying requirements overtime. We can overcome these challenges by providing a set of computing resources that can be shared through cloud computing.
- The cloud computing environment saves costs related to infrastructure in an organisation by providing a framework that can be optimised and expanded horizontally.

# 2. Cloud Computing and Big Data

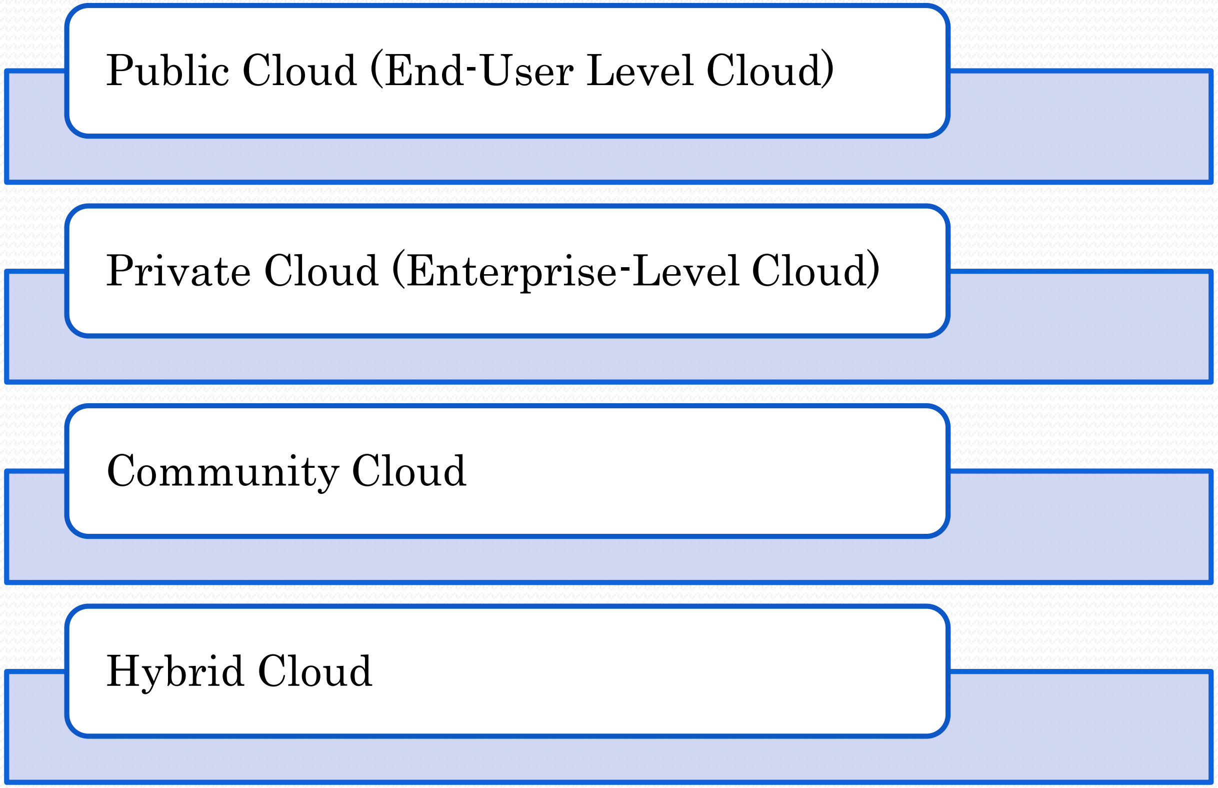
Features of Cloud Computing



- Scalability
- Elasticity
- Resource Pooling
- Self-Service
- Low Cost
- Fault Tolerance

# 3. Cloud Computing and Big Data

Cloud Deployment Models



Public Cloud (End-User Level Cloud)

Private Cloud (Enterprise-Level Cloud)

Community Cloud

Hybrid Cloud

# 4. Cloud Computing and Big Data

## Cloud Delivery Models

**Infrastructure as a Service (IaaS)**

- It is one of the categories of cloud computing services, which makes available virtualised computing resources on Internet.

**Platform as a Service (PaaS)**

- It is built above IaaS and is the layer that interacts with the users, allowing them to deploy and use applications created using programming and run-time environment platforms that are supported by the provider.

**Software as a Service (SaaS)**

- SaaS is one of the most popular cloud-based models and comprises applications provided by the service provider.

# 5. Cloud Computing and Big Data

## Cloud Providers in Big Data Market

Big Data cloud providers have been gearing up to bring the most advanced technologies at competitive prices in the market. Some providers are established, whereas some of them are relatively new to the field of cloud services. Some of these providers are rendering services that are relevant to Big Data analytics only. Some such providers are as follows:

- Amazon
- Google
- Windows Azure

# In-Memory Technology for Big Data

- Hardware obstructions and limitations, lag of memory indifferences have to be side-lined and streamlined with something faster like a cache memory or dynamic access memory so that the data is readily available for disposal.
- The in-memory big data computing tool supports processing of high velocity data in real-time and also faster processing of the stationary data.
- Technologies like event-streaming platforms, in-memory databases and analytics, and high-level messaging structures are witnessing massive growth that resonates with the organisational needs for better understandings achievable by a wider and deeper data assessment.

# 1. Big Data Techniques

- To analyse the datasets, there are many techniques available, some of which are as follows:
  - Massive Parallelism
  - Data Distribution
  - High-Performance Computing
  - Task and Thread Management
  - Data Mining and Analytics
  - Data Retrieval

# 2. Big Data Techniques

## Massive Parallelism

- According to the simplest definition available, a parallel system is a system where multiple processors are involved and associated to carry out the concurrent computations.
- Massive parallelism refers to a parallel system where multiple systems interconnected with each other pose as a single mighty conjoint processor and carry out tasks received from the data sets parallelly.
- In terms of Big Data dynamics, the systems can not only be processor, but also memory, hardware and even network conjoint to scale up the operational efficiency posing as a massive system that can eat humongous datasets parallelly without breaking a sweat.

# 3. Big Data Techniques

## Data Distribution

- There are approaches to data distribution in a Big Data system described as follows:
  - **Centralised Approach:** A central repository is used to store and download the essential dataset by virtual machines.
  - **Semi-Centralised Approach:** Semi-centralised approach reduces the stress on the networking infrastructure.
  - **Hierarchical Approach:** In a hierarchical approach, the data is fetched from the parent node, i.e., the virtual machine, in the hierarchy.
  - **P2P Approach:** P2P streaming connections are based on hierarchical multi-trees.

# 4. Big Data Techniques

## High-Performance Computing

- High-performance computing is the simultaneous use of supercomputers and parallel processing techniques for solving intricate computation problems.
- It emphasises on making parallel processing systems and algorithms by joining both parallel and administrative computational methods.
- The words supercomputing and high-performance computing are often used to resemble each other.
- High-performance computing is used for performing research activities and cracking advanced problems through computer simulation, modelling and analysis.

# 5. Big Data Techniques

## Task and Thread Management

- Threads are simply the OS-based feature with their own Kernel and memory resources, and allow an application logic to be segregated into concurrent multiple execution paths. It is a useful feature when complex applications having multiple tasks need to be performed at the same time.
- Task parallelism refers to the execution of computer programmes throughout the multiple processors on the different or same machines. It emphasises on performing diverse operations in parallel to best utilise the accessible computing resources like memory and processors.
- Data parallelism focusses on effective distribution of datasets throughout the multiple calculation programs.

# 6. Big Data Techniques

Data Mining and Analytics

- Data mining is the process of data extraction, evaluating it from multiple perspectives and then producing the information summary in a meaningful form that identifies one or more relationships within the dataset.
- Data analysis is an experiential activity, where the data scouring gives out some insight.
- Data analytics is about applying an algorithmic or logical process to derive the insights from a given dataset. For example, looking at the past year's weather and pest data, for the current month, we can determine that a particular type of fungus grows often when the humidity levels reach a definite point.

# 7. Big Data Techniques

Data Retrieval

- Big Data refers to the large amounts of multi-structural data that continuously flows around and within the organisations, and includes text, video, transactional records and sensor logs.
- Big Data systems utilise the Hadoop and the HDFS architecture to retrieve the data using MapReduce – a distributed processing framework.
- It helps programmers in solving parallel data problems where the dataset can be divided into small chunks and handled autonomously.
- MapReduce is important step as it allows normal developers to utilise parallel programming concepts irrespective of cluster communication details, failure handling and task monitoring.

# 8. Big Data Techniques

## Machine Learning

- Machine leaning formally focusses on the performance, theory and properties of learning algorithms and systems. Machine learning is considered an ideal research field for taking advantage of the opportunities available in big data.
- It delivers on the potential of mining the value from huge and different data sources with less dependence on human instructions. It is data-driven and runs at machine scale and well-suited to the complication of dealing with different data sources and the enormous range of variables and quantities of data involved.
- Machine learning systems utilise multiple algorithms to discover and show the patterns hidden in the datasets.

# 9. Big Data Techniques

Data Visualisation

- Data visualisation is a valuable means through which the larger datasets after being combined may appear practical, sensible and open to most people. Data visualisation is a trail-blazing method that not only keeps you enlightened but helps others with the attributes of a typical statistical and computational result that would've otherwise appeared intimidating for normal minds.
- Visual representation is often considered the most effective medium of information and communication channel. As the saying goes, a picture is worth thousand words, data visualisation is a great example of that saying. When properly aligned, it can convey critical information of data analysis in probably the easiest way possible.

# Let’s Sum Up

- The distributed computing works on the rules of divide and conquer, performing modules of the parent tasks on multiple machines and then combining the results.
- Parallel computing refers to the utilisation of a single CPU present in a system or a group of internally coupled systems by the means of efficient and clever multi-threading operations.
- Distributed computing is considered as the subset of parallel computing, which further is the subset of concurrent computing.
- A Big Data system is vastly different from other solution-providing systems and is based on the seven Vs, namely: Volume, Velocity, Variety, Veracity, Variability, Value and Visualisation.