In the previous chapters, we have already learned that the huge amounts of data generated by different sources need to be managed properly, if it is to be utilized in any productive manner. Now, such voluminous, varied, and scattered data cannot be handled by traditional data storage and processing systems. This requirement prompted the development of various technologies used for processing Big Data. Among the technologies that are used to handle, process, and analyze Big Data, the most effective and popular innovations have been in the fields of distributed and parallel processing, Hadoop, In-Memory Computing (IMC), Big Data cloud, etc. Hadoop, which is an open-source platform, has been by far the most popular technology associated with Big Data storage and processing different types of data. Hadoop is commonly used by data-driven organizations to extract maximum output from their normal data-usage practices at a rapid pace. Besides Hadoop, the other techniques used for Big Data processing are cloud computing and IMC.

These technologies help organizations to analyze data in different ways under varying circumstances. Cloud computing helps businesses to save cost and better manage their resources by allowing them to use resources as a service on the basis of specific requirements and paying for only those services that have actually been used. IMC helps you to organize and complete your tasks faster by carrying out all the computational activities from the main memory itself. You can use all these techniques, as per specific requirements, for analyzing Big Data.

This chapter mainly focuses on explaining the basics and exploring the relevance and role of various technologies that are used for storing, processing, and analyzing Big Data. Here, you will first learn about the concept of distributed and parallel computing followed by a discussion on various technologies, such as Hadoop, cloud computing, and IMC, used to handle Big Data.

SCENARIO

Mr. Richard Stephens is the global business head of Argon Technology and wants to use some latest technologies while analyzing Big Data. He carried out some research and found that as compared to traditional methods of computing, distributed and parallel computing technologies are more suitable to handle Big Data. The following section discusses distributed and parallel computing technologies, especially for Big Data analytics.

Distributed and Parallel Computing for Big Data

In distributed computing, multiple computing resources are connected in a network and computing tasks are distributed across these resources. This sharing of tasks increases the speed as well as the efficiency of the system. Because of reason, the distributed computing is considered faster and much more efficient than traditional methods of computing. It is also more suitable to process huge amounts of data in a limited time.

Another way to improve the processing capability of a computer system is to add additional computational resources to it. This will help in dividing complex computations into subtasks, which can be handled individually by processing units that are running in parallel. We call such systems as parallel systems in which multiple parallel computing resources are involved to carry out calculations simultaneously. The concept behind involving multiple parallel resources is that the processing capability will increase with the increase in the level of parallelism.

Figure 3.1 shows a comparison between distributed and parallel processing techniques:

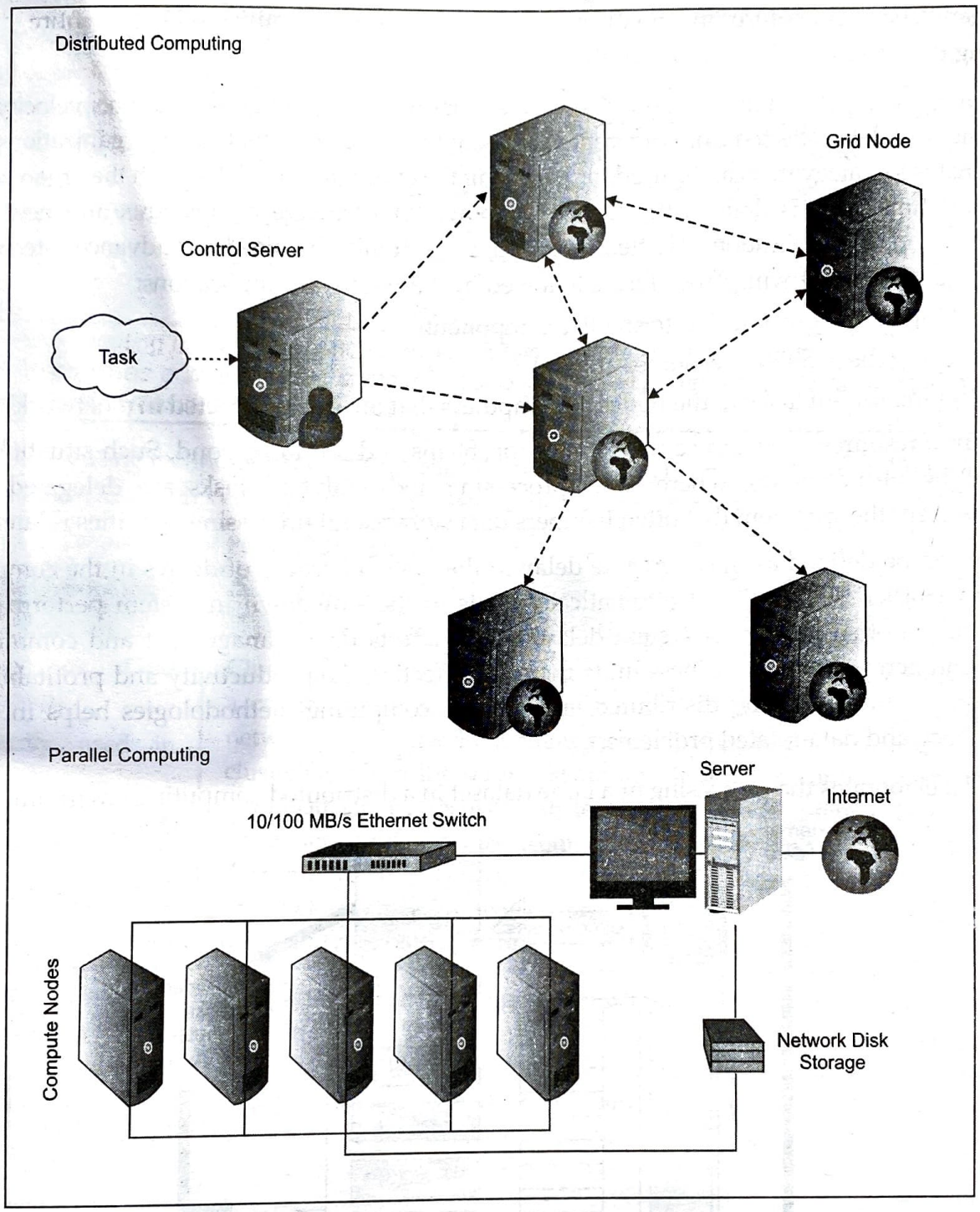


Figure 3.1: Distributed Computing and Parallel Computing

Organizations use both parallel and distributed computing techniques to process Big Data. The most important constraint for businesses today is time. In case there was no restriction on time, every organization would hire outside (or third-party) sources to perform the analysis of its complex data. The direct benefit of adopting this method is that the organization would not require any resources and data sources to process and analyze complex data. These third parties are usually specialized agencies in the field of data manipulation, processing, and analysis. Apart from being effective, hiring third party agencies also reduces the storage and processing costs of handling large amounts of data.

Moreover, data processing and analysis activities carried out within an organization would require the organization to capture and analyze only a sample dataset rather than the entire data thus, reducing the processing load significantly.

The growing competition in the market and the astronomical increase in the volume, velocity, variety, and veracity of data collected from different sources, at the same time, are forcing organizations to adopt a data analysis strategy that can be used for analysing the entire data available with the organization in a very short time. This is done with the help of powerful hardware components and new software programs and/or applications written specifically to fulfill the need of advanced technological developments. The following procedure is followed by these software applications:

1. Breaking up the given task into smaller components
2. Surveying the available resources at hand
3. Assigning the subtasks to the nodes or computers that are interconnected in a network

Sometimes, resources develop some technical problems and fail to respond. Such situations can be handled by virtualization, where some processing and analytical tasks are delegated to other resources. Another problem that often hampers data storage and processing activities is latency.

Latency can be defined as the aggregate delay in the system because of delays in the completion of individual tasks. Such a delay automatically leads to the slowdown in system performance as a whole and is often termed as system delay. It also affects data management and communication within and across various business units there by, affecting the productivity and profitability of an organization. Implementing distributed and parallel computing methodologies helps in handling both latency and data-related problems.

Figure 3.2 elaborates the processing of a large dataset in a distributed computing environment:

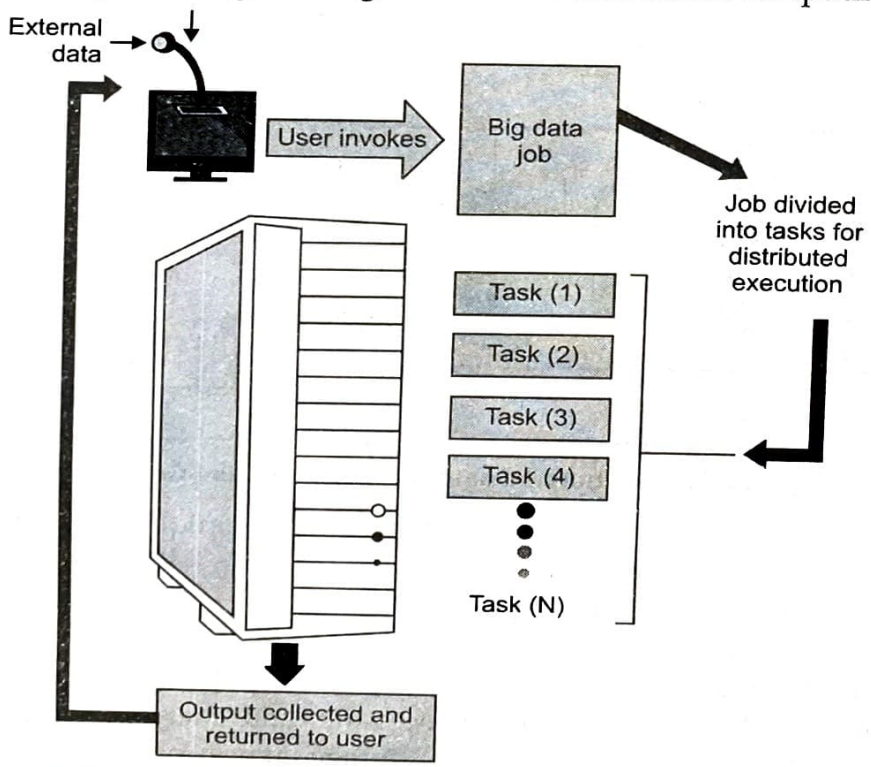


Figure 3.2: Distributed Computing Technique for Processing Large Data

As you can notice in Figure 3.2, the nodes are arranged within a system along with the elements that form the core of computing resources. These resources include CPU, memory, disks, etc. Big Data systems, usually, have higher scaling requirements. Therefore, these nodes are more beneficial for adding scalability to the Big Data environment, as and when required. The system, with added scalability, can accommodate the growing amounts of data more efficiently and flexibly. Distributed computing technique also makes use of virtualization and load balancing features. The sharing of workload across various systems throughout the network to manage the load is known as load balancing. The virtualization feature creates a virtual environment in which hardware platform, storage device, and Operating System (OS) are included.

NOTE

Distributed computing has been around for almost 50 years. Initially, the technology was used in computer science research as a way to scale computing tasks and solve complex problems without incurring the expenses of using massive computing systems.

Parallel computing technology uses a number of techniques to process and manage huge amounts of data produced at a high velocity. Some of these techniques are shown in Table 3.1:

Table 3.1: Techniques of Parallel Computing

Parallel Computing Method	Description	Uses
Cluster or Grid Computing (Primarily used in Hadoop)	Cluster or grid computing is based on a connection of multiple servers in a network. This network is known as a cluster in which the servers share the workload among them. A cluster can be either homogeneous (comprising the same type of commodity hardware) or heterogeneous (consisting of different types of hardware).	A cluster can be created even by using hardware components that were acquired a long time back to provide cost-effective storage options. The overall cost may be very high in cluster computing.
Massively Parallel Processing (MPP) (Used in data warehouses)	A single machine working as a grid is used in the MPP platform, which is capable of handling the activities of storage, memory, and computing. Software written specifically for MPP platform is used for the optimization of MPP capabilities.	MPP platforms, such as EMC Greenplum, and ParAccel are most suited for high-value use cases.
High-Performance Computing (HPC)	HPC environments are known to offer high performance and scalability by using IMC. This technology is especially suitable for processing floating-point data at high speeds.	HPC environments can be used to develop specialty and custom applications for research and business organizations where the result is more valuable, than the cost or where strategic importance of the project is of high priority.

After learning about both distributed and parallel computing techniques, we can compare these in certain aspects. Table 3.2 shows how parallel systems are different from distributed systems:

Table 3.2: Difference Between Parallel Systems and Distributed Systems

Distributed System	Parallel 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, providing access to data and remotely located resources	Tight coupling of processing resources that are used for solving a single, complex problem

SCENARIO

During the discussion about distributed and parallel computing for Big Data analytics, Mr. Stephens learned that Hadoop is a technology that works on the distributed computing technique and is vastly used by large organizations to manage their Big Data requirements. He concludes that implementing Hadoop in the company would help them handle their business problems in a better way.

How Data models and Computing models are different?

As we know Hadoop is a distributed system like distributed databases; however, there are several key differences between the two infrastructures with respect to computing model and data model in a distributed architecture. Figure 3.3 shows the distributed databases and Figure 3.4 shows Hadoop:

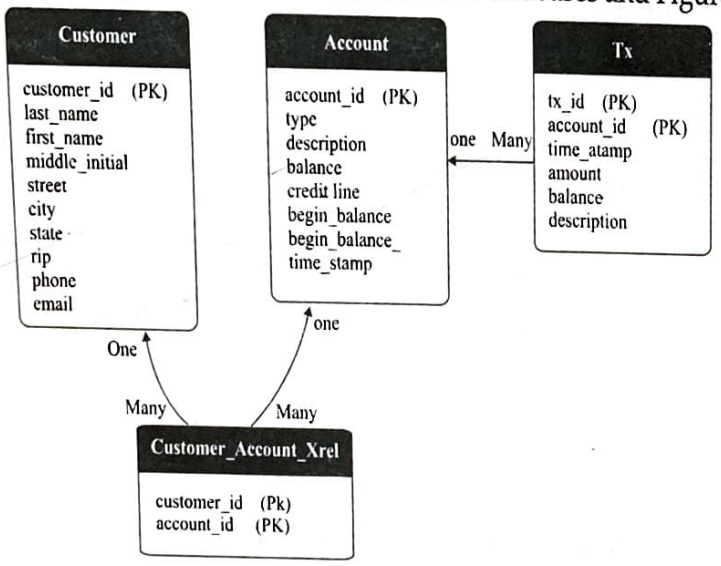


Figure 3.3: Distributed Databases

- Distributed Databases:
  - Deal with tables and relations
  - Must have a schema for data
  - Implements data fragmentation and partitioning

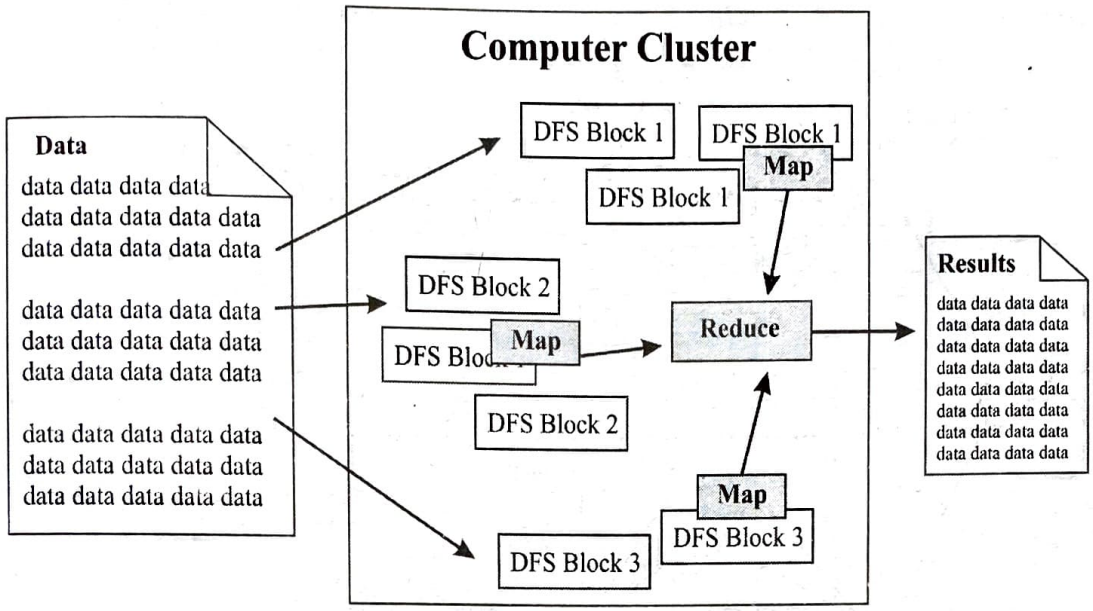


Figure 3.4: Hadoop

- Hadoop:
  - Deals with flat files in any format
  - Operates on no schema for data
  - Divides files automatically into blocks

The computing models of a distributed database and Hadoop are shown in figure 3.5 and figure 3.6, respectively:

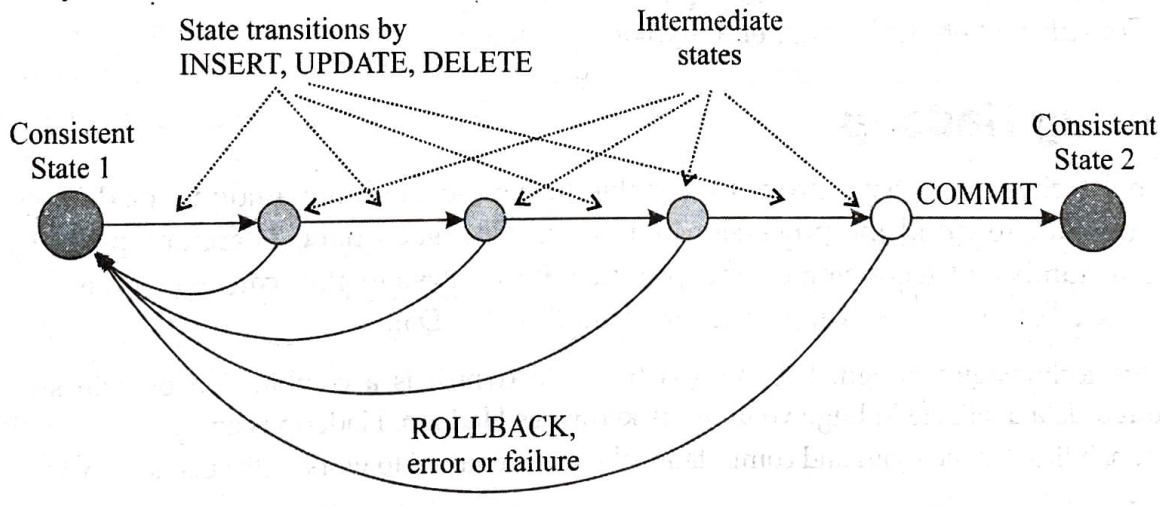


Figure 3.5: Computing Model of Distributed Database

Distributer Databases:

- Generate notations of a transaction
- Implement ACID transaction properties
- Allow distributed transactions

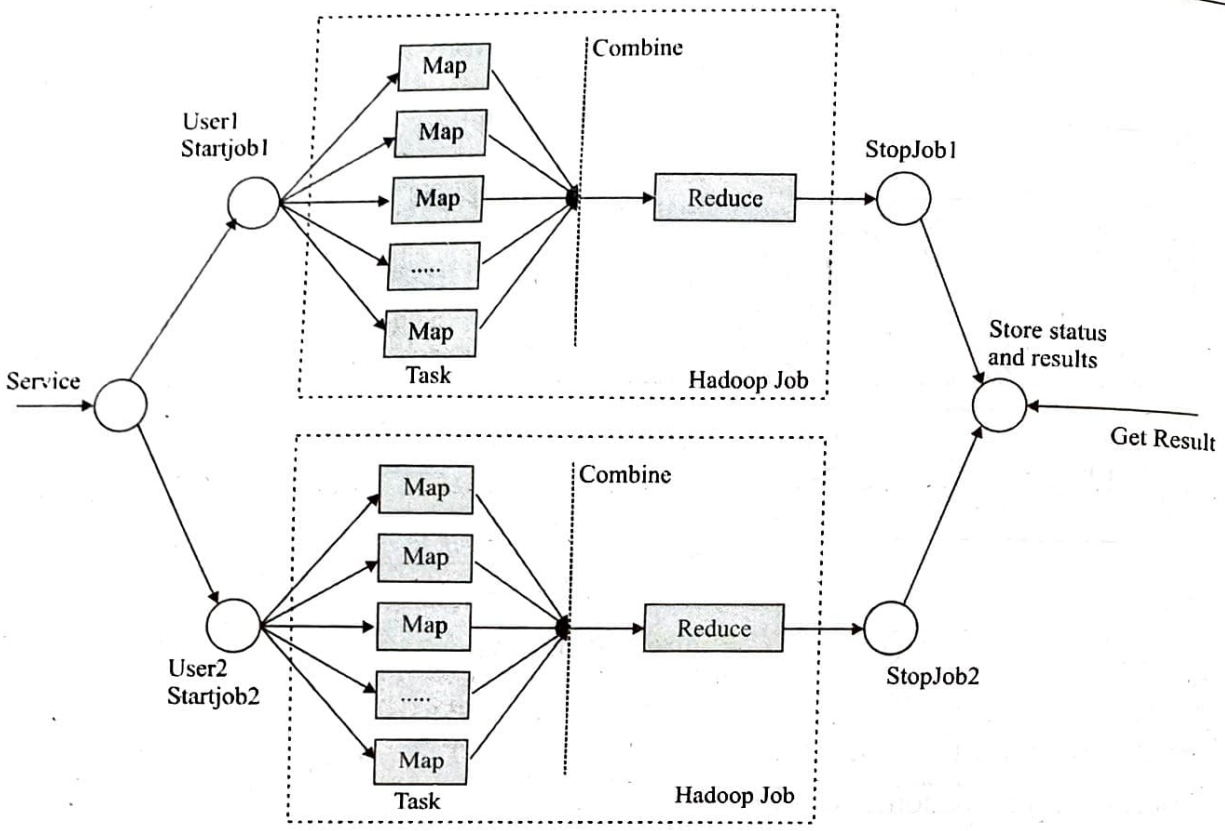


Figure 3.6: Computing Model of Hadoop

- Generates notations of a job divided into tasks
- Implements MapReduce computing model
- Considers every task as either a map or a reduce

Introducing Hadoop

Traditional technologies have proved incapable to handle the huge amounts of data generated in organizations or to fulfill the processing requirements of such data. Therefore, a need was felt to combine a number of technologies and products into a system that can overcome the challenges faced by the traditional processing systems in handling Big Data.

One of the technologies designed to process Big Data (which is a combination of both structured and unstructured data available in huge volumes) is known as Hadoop. Hadoop is an open-source platform that provides analytical technologies and computational power required to work with such large volumes of data.

Earlier, distributed environments were used to process high volumes of data. However, multiple nodes in such an environment may not always cooperate with each other through a communication system, leaving a lot of scope for errors. 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 master node 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. The standard start-up and shutdown scripts require Secure Shell to be set up between nodes in the cluster. In a larger cluster, the HDFS is managed through a NameNode server to host the file system index and a secondary NameNode that keeps snapshots of the NameNodes and at the time of failure

of NameNode the secondary NameNode replaces the primary NameNode, thus preventing file-system from getting corrupt and reducing data loss. Figure 3.7 shows Hadoop multinode cluster architecture:

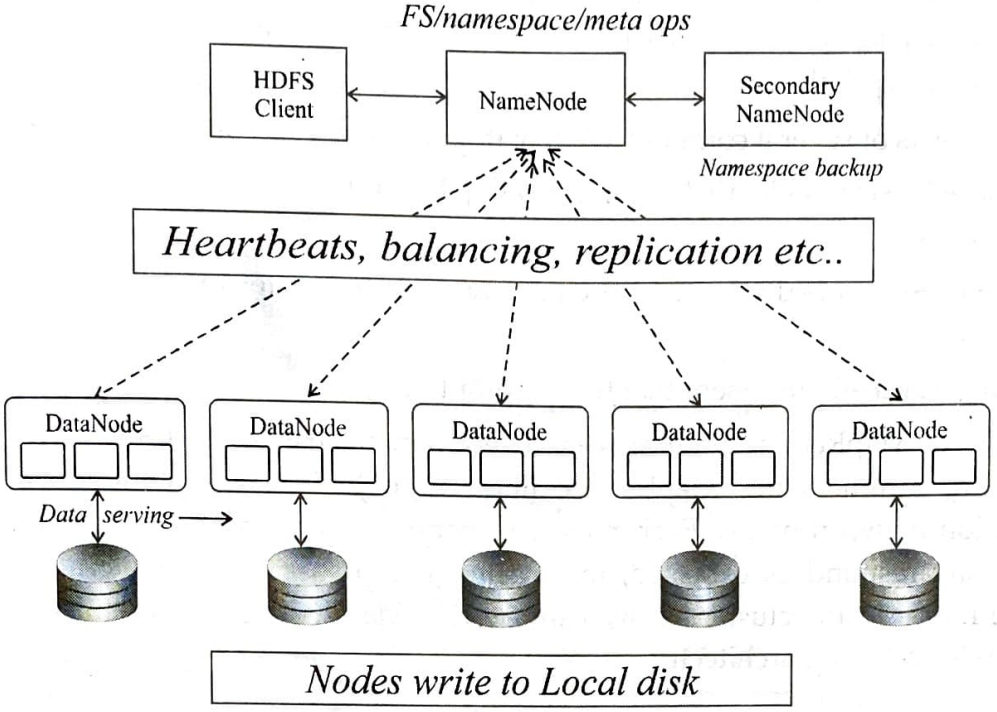


Figure 3.7: Hadoop multinode cluster architecture

The secondary NameNode takes snapshots of primary NameNode directory information after a regular interval of time, which is saved in local or remote directories. These checkpoint images can be used in the place of the primary NameNode to restart a failed primary NameNode without replaying the entire journal of file-system actions and without editing the log to create an up-to-date directory structure. NameNode is the single point for storage and management of metadata. To process the data, Job Tracker assigns tasks to the Task Tracker. Let us assume that a DataNode cluster goes down while the processing is going on, then the NameNode should know that the some DataNode is down in the cluster, otherwise it cannot continue processing. Each DataNode sends a "Heart Beat Signal" to NameNode after every few minutes (as per Default time set) to make NameNode aware of the active / inactive status of DataNodes. This system is called as Heartbeat mechanism.

HDFS and MapReduce

There are two main components of Apache Hadoop-the Hadoop Distributed File System (HDFS) and the MapReduce parallel processing framework. Both of these open source projects, HDFS is used for storage and MapReduce is used of processing.

Hadoop includes a fault-tolerant storage system called Hadoop distributed file system (HDFS). It stores large size files from terabytes to petabytes across different terminals. HDFS attains reliability by replicating the data over multiple hosts. The default replication value is 3. 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.

MapReduce is a framework that helps developers to write programs to process large volumes of unstructured data parallel over a distributed architecture/standalone architecture which produces result in useful aggregated form.

MapReduce consists of several components, few important ones are mentioned here:

- **JobTracker**—Master node that manages all jobs and resources in a cluster of commodity computers
- **TaskTrackers**—Agents deployed at each machine in the cluster to run the map and reduce task at the terminal
- **JobHistoryServer**—Component that tracks completed jobs

We can write MapReduce programs in several languages like C, C++, Java, Ruby, Perl, and Python. Programmers use MapReduce libraries to build tasks, without communication or coordination between nodes. Each node will periodically report its status to master node, if a node doesn't respond as expected, the master node re-assigns that piece of the job to other available nodes in the cluster, so we can say that MapReduce is also fault-tolerant. Figure 3.8 shows the MapReduce architecture.

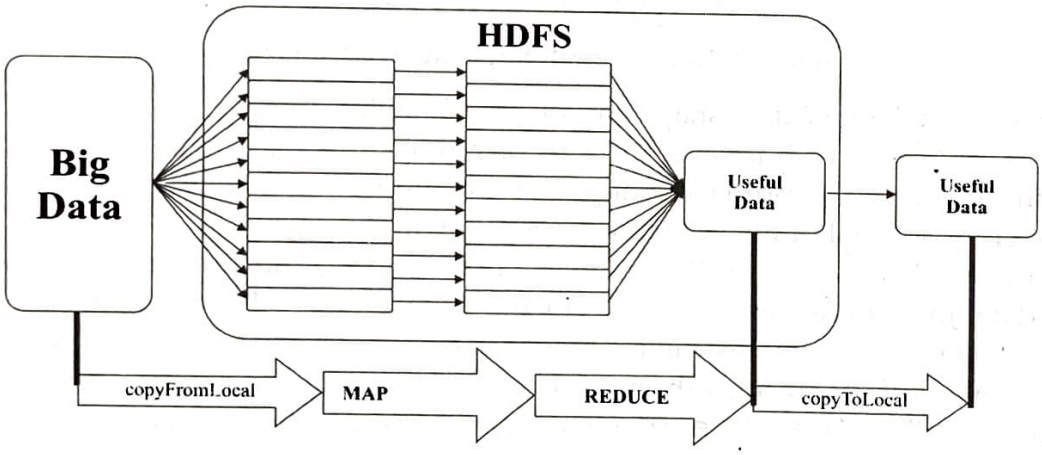


Figure 3.8: Hadoop MapReduce architecture

The following are some of the important features of Hadoop:

- Hadoop performs well with several nodes without requiring shared memory or disks among them. Hence, the efficiency-related issues in context of storage and access to data get automatically solved.
- Hadoop follows the client-server architecture in which the server works as a master and is responsible for data distribution among clients that are commodity machines and work as slaves to carry out all the computational tasks. The master node also performs the tasks of job controlling, disk management, and work allocation.
- The data stored across various nodes can be tracked in Hadoop NameNode. It helps in accessing and retrieving data, as and when required.

- Hadoop improves data processing by running computing tasks on all available processors that are working in parallel. The performance of Hadoop remains up to the mark both in the case of complex computational questions and of large and varied data.
- Hadoop keeps multiple copies of data (data replicas) to improve resilience that helps in maintaining consistency, especially in case of server failure. Usually, three copies of data are maintained, so the usual fault-replication factor in Hadoop is 3.

How does Hadoop Function?

The first thing to know about the functioning of Hadoop is the way it utilizes multiple computing resources for executing a particular task. The core components of Hadoop include the following:

- **Hadoop Distributed File System (HDFS)**—It is a cluster of storage solutions that is highly reliable, more efficient, and economical and provides facilities to manage files containing related data across machines.
- **Hadoop MapReduce**—It is a computational framework used in Hadoop to perform all the mathematical computations. It is based on a parallel and distributed implementation of MapReduce algorithm that provides high performance.

Hadoop facilitates the processing of large amounts of data present in both structured and unstructured forms. Hadoop clusters are created from the racks of commodity machines. Tasks are distributed across these machines (also known as nodes), which are allowed to work independently and provide their responses to the starting node. Moreover, it is possible to add or remove nodes dynamically in a Hadoop cluster on the basis of varying workloads. Hadoop has an ability to detect changes (which also include server failure) in the cluster and adjust to them, without causing any interruption in the system.

Hadoop accomplishes its operations (of dividing the computing tasks into subtasks that are handled by individual nodes) with the help of the MapReduce model, which comprises two functions, namely, a mapper and a reducer. The mapper function is responsible for mapping the computational subtasks to different nodes, and the reducer function takes the responsibility of reducing the responses from compute nodes, to a single result. The MapReduce model implements the MapReduce algorithm, as discussed earlier, to incorporate the capability of breaking data into manageable subtasks, processing the data on the distributed cluster simultaneously, and making the data available for additional processing or user consumption.

In the MapReduce algorithm, the operations of distributing task across various systems, handling the task placement for load balancing, and managing the failure recovery are accomplished by the map component (or the mapper function). The reduce component (or the reducer function), on the other hand, has the responsibility to aggregate all the elements together after the completion of the distributed computation.

When an indexing job is provided to Hadoop, it requires the organizational data to be loaded first. Next, the data is divided into various pieces, and each piece is forwarded to different individual servers. Each server has a job code with the piece of data it is required to process. The job code helps Hadoop to track the current state of data processing. Once the server completes operations on the data provided to it, the response is forwarded with the job code being appended to the result.

In the end, results from all the nodes are integrated by the Hadoop software and provided to the user, as shown in Figure 3.9:

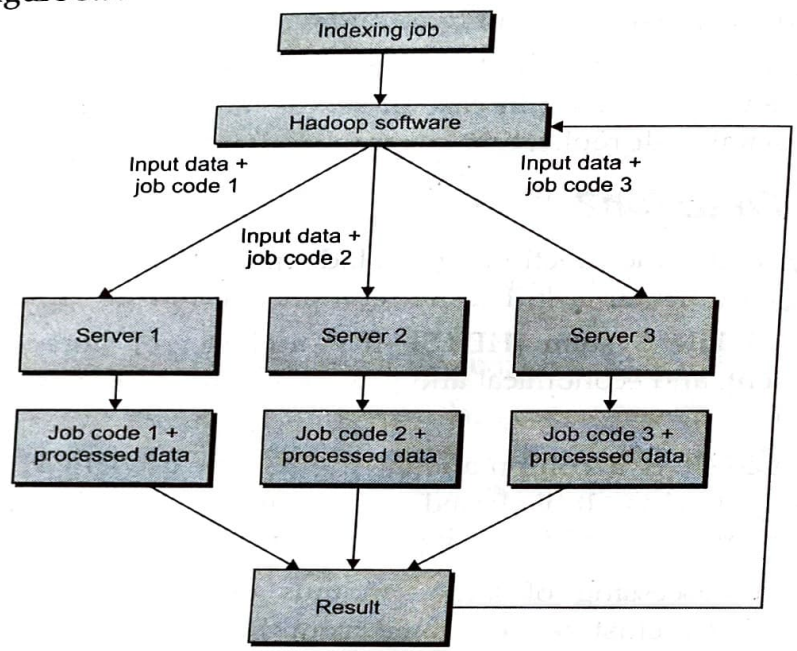


Figure 3.9: Job Tracking Process in MapReduce

We can understand the working of Hadoop by referring to the following example:

The call records of all the telephones in a city are being examined by a researcher who wants to know about those calls that are made by college students on the occasion of an event. The fields required as a result of the analysis carried out by the researcher include the timing of the event and the relevant information of the user. The query is fired on every machine to search the results from the call records stored with the machine, which will return the relevant results. Finally, a single result will be generated by aggregating individual results obtained from all the machines. We are considering that the records are collected in a Comma Separated Value (CSV) file.

- The processing starts by first loading the data into Hadoop and then applying the MapReduce programming model. Let us consider the following five columns to be contained in the CSV file:
  - `u_id`
  - `u_name`
  - `c_name`
  - `sp_name`
  - `call_time`

To identify individual users who made phone calls at a specific time, the `u_id` field is used. It helps us to determine the number of users who have called at the time. We can, thus, get the final output in terms of the number of users by whom calls were made during the specified time.

To obtain the final output, each mapper receives data line by line. Once the mapper completes its job, the results are shuffled or sorted by the Hadoop framework, which then combines the data in groups that are forwarded to the reducer. Ultimately, the final output is obtained from the reducer.

- Data in Hadoop can be stored across multiple machines. Businesses can take the advantage of this storage facility to use multiple commodity machines. These machines are capable of hosting the Hadoop software. In this case, businesses will not need to create integrated systems.

  Today, Hadoop is the most popular and used platform in businesses for Big Data processing. It helps in Big Data analytics by overcoming the obstacles that are usually faced in handling Big Data. In Hadoop, we can break down large computational problems into smaller tasks as smaller elements can be analyzed economically and quickly. All these parts are analyzed in parallel, and the results of the analysis are regrouped to produce the final output.

SCENARIO

Big Data is a vast field that needs involvement of various technologies and equipment. The hardware and infrastructure required for storing and handling Big Data is usually much higher than the existing infrastructure or hardware. To deal with this issue, Mr. Richard Stephens decides to utilize the services of cloud computing. He tells his team that cloud computing is a technique that helps organizations to hire infrastructure and other resources on rent, thereby saving a lot of initial investment.

Cloud Computing and Big Data

One of the vital issues that organizations 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 over utilized or underutilized with the varying requirements overtime. We can overcome these challenges by providing a set of computing resources that can be shared through cloud computing. These shared resources comprise applications, storage solutions, computational units, networking solutions, development and deployment platforms, business processes, etc. The cloud computing environment saves costs related to infrastructure in an organization by providing a framework that can be optimized and expanded horizontally. In order to operate in the real world, cloud implementation requires common standardized processes and their automation.

Figure 3.10 shows the cloud computing model:

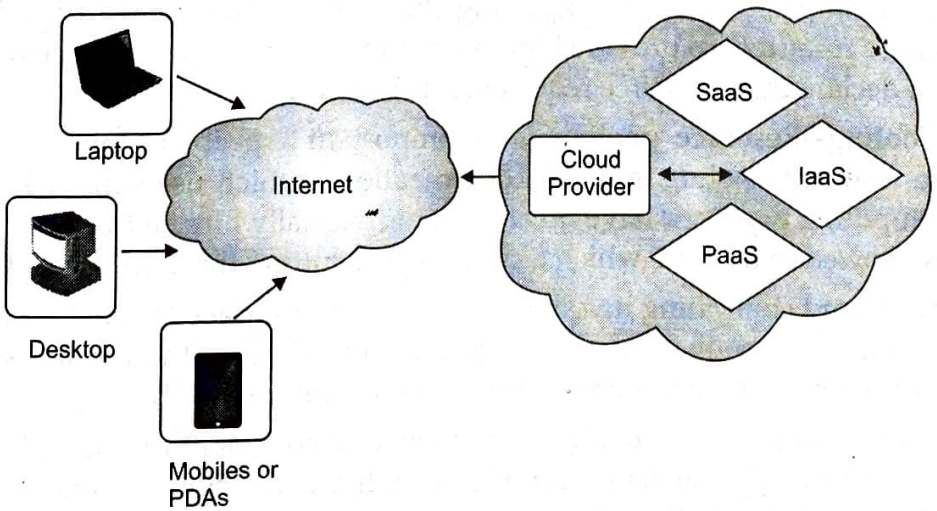


Figure 3.10: Cloud Computing Model

In cloud-based platforms, applications can easily obtain the resources to perform computing tasks. The costs of acquiring these resources need to be paid as per the acquired resources and their use. In cloud computing, this feature of resource acquisition is in accordance with the requirements and payment of cost and is known as elasticity. Cloud computing makes it possible for organizations to

dynamically regulate the use of computing resources and access them as per the need while paying only for those resources that are used. This facility of dynamic use of resources provides flexibility; however, an organization needs to plan, monitor, and control its resource utilization carefully. Careless resource monitoring and control can result in unexpectedly high costs.

The cloud computing technique uses data centers to collect data and ensures that data backup and recovery are automatically performed to cater to the requirements of the business community. Both cloud computing and Big Data analytics use the distributed computing model in a similar manner and hence, are complementary to each other.

CASELET

Companies, such as Google and Amazon, require systems that could provide massive capabilities for managing large amounts of data in order to take their business forward. This requires strong infrastructure and advanced technologies to assist their applications at a large scale. For example, Google processes millions of Gmail messages every minute through an optimized Linux OS and its software environment, and Amazon handles massive workloads through its Infrastructure as a Service (IaaS) data centers. These companies also offer cloud-based services to contribute in the field of Big Data processing.

Features of Cloud Computing

The following are some features of cloud computing that can be used to handle Big Data:

- **Scalability**—Scalability means addition of new resources to an existing infrastructure. The increase in the amount of data being collected and analyzed requires organizations to improve their hardware components' processing ability. These organizations may, at times, need to replace the existing hardware with a new set of hardware components in order to improve data management and processing activities. The new hardware may not provide complete support to the software that used to run properly on the earlier set of hardware. We can solve such issues by using cloud services that employ the distributed computing technique to provide scalability to the architecture.
- **Elasticity**—Elasticity in cloud means hiring certain resources, as and when required, and paying for the resources that have been used. No extra payment is required for acquiring specific cloud services. For example, a business expecting the use of more data during in-store promotion could hire more resources to provide high processing power. Moreover, a cloud does not require customers to declare their resource requirements in advance.
- **Resource Pooling**—Resource pooling is an important aspect of cloud services for Big Data analytics. In resource pooling, multiple organizations, which use similar kinds of resources to carry out computing practices, have no need to individually hire all the resources. The sharing of resources is allowed in a cloud, which facilitates cost cutting through resource pooling.
- **Self Service**—Cloud computing involves a simple user interface that helps customers to directly access the cloud services they want. The process of selecting the needed services requires no intervention from human beings and can be accessed automatically.
- **Low Cost**—A careful planning, use, management, and control of resources help organizations to reduce the cost of acquiring hardware significantly. Also, cloud offers customized solutions, especially to organizations that cannot afford too much initial investment in purchasing the resources that are used for computation in Big Data analytics. The cloud provides them the pay-as-you-use option in which organizations need to sign for those resources only that are essential. This also helps the cloud provider in harnessing benefits of economies of scale and providing benefit to their customers in terms of cost reduction.

- **Fault Tolerance**—Cloud computing provides fault tolerance by offering uninterrupted services to customers, especially in cases of component failure. The responsibility of handling the workload is shifted to other components of the cloud.

Cloud Deployment Models

Depending upon the architecture used in forming the network, services and applications used, and the target consumers, cloud services are offered in the form of various deployment models. The following are the most-commonly used cloud deployment models:

- Public Cloud
- Private Cloud
- Community Cloud
- Hybrid Cloud

Public Cloud (End-User Level Cloud)

A cloud that is owned and managed by a company than the one (which can be either an individual user or a company) using it is known as a public cloud. In this cloud, there is no need for the organizations (customers) to control or manage the resources; instead, they are being administered by a third party. Some examples of public cloud providers are Savvis, Verizon, Amazon Web Services, and Rackspace. You should understand that in case of a public cloud, the resources are owned or hosted by the cloud service providers (a company), and the services are sold to other companies. Companies or individuals can obtain various services in a public cloud. The workload is categorized on the basis of service category, and therefore, in this cloud, the hardware customization is possible to provide optimized performance. The process of computing becomes very flexible and scalable through customized hardware resources. For example, a cloud can be used specifically for video storage that can be streamed live on YouTube or Vimeo. You can also optimize this cloud for handling large traffic volumes.

Businesses can obtain economical cloud storage solutions in a public cloud, which provides efficient mechanisms for complex data handling. The primary concerns with a public cloud include security and latency, which can be overlooked citing the benefits of this cloud.

Figure 3.11 demonstrates the use of a public cloud:

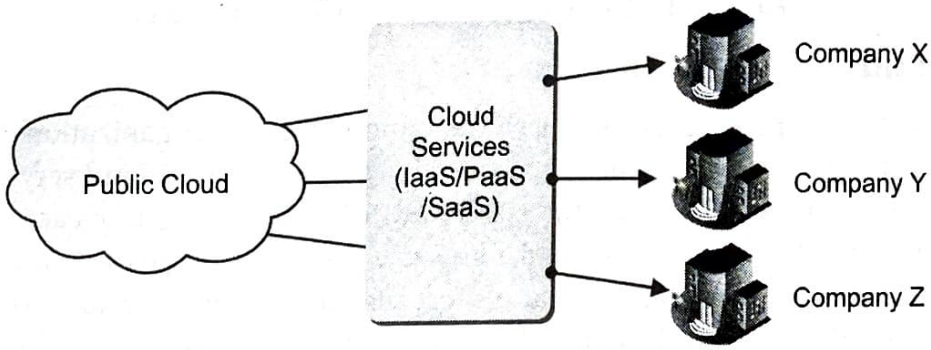


Figure 3.11: Level of Accessibility in a Public Cloud

vate Cloud (Enterprise Level Cloud)

The cloud that remains entirely in the ownership of the organization using it is known as a private cloud. In other words, in this cloud, the cloud computing infrastructure is solely designed for a single organization and cannot be accessed by other organizations. However, the organization may allow this cloud to be used by its employees, partners, and customers. The primary feature of a private cloud is that an organization installs the cloud for its own requirements. These requirements are customary to the organization that plans and manages the resources and their use. A private cloud integrates all the processes, systems, rules, policies, compliance checks, etc. of the organization at a place. In a private cloud, you can automate several processes and operations that require manual handling in a public cloud. Moreover, you can also provide firewall protection to the cloud; thereby, solving many latency and security concerns. A private cloud can be either on-premises or hosted externally. In case of on-premises private clouds, the service is exclusively used and hosted by a single organization. However, the private clouds that are hosted externally are used by a single organization and are not shared with other organizations. Moreover, the cloud services are hosted by a third party that specializes in cloud infrastructure. Note that on-premises private clouds are costlier as compared to the externally hosted private clouds. In case of a private cloud, security is kept in mind at every level of design. The general objective of a private cloud is not to sell the cloud services (IaaS/PaaS/SaaS) to the external organizations but to get the advantages of cloud architecture by not providing the privilege to manage your own data center.

Figure 3.12 demonstrates the use of a private cloud:

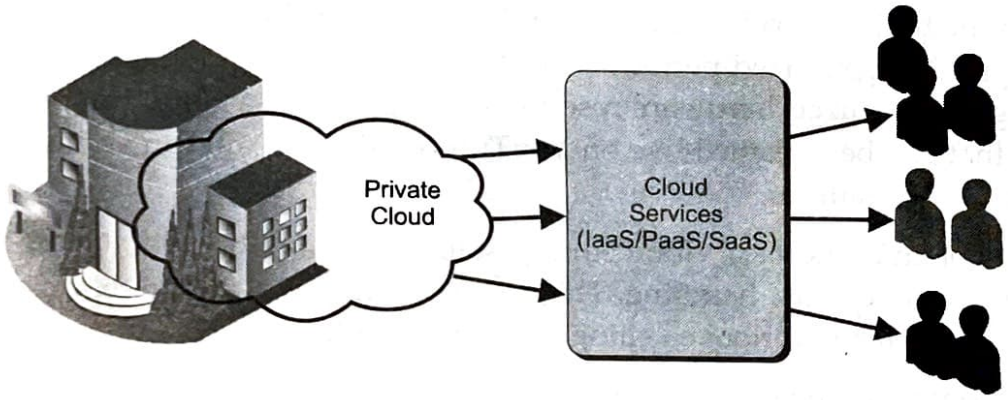


Figure 3.12: Level of Accessibility in a Private Cloud

ommunity Cloud

Community cloud is a type of cloud that is shared among various organizations with a common tie. This type of cloud is generally managed by a third party offering the cloud service and can be made available on or off premises. To make the concept of community cloud clear and to explain when community clouds can be designed, let's take an example. In any state or country, say England, the community cloud can be provided so that almost all government organizations of that state can share the resources available on the cloud. Because of the sharing of cloud resources on community cloud, the data of all citizens of that state can be easily managed by the government organizations.

Figure 3.13 shows the use of community clouds:

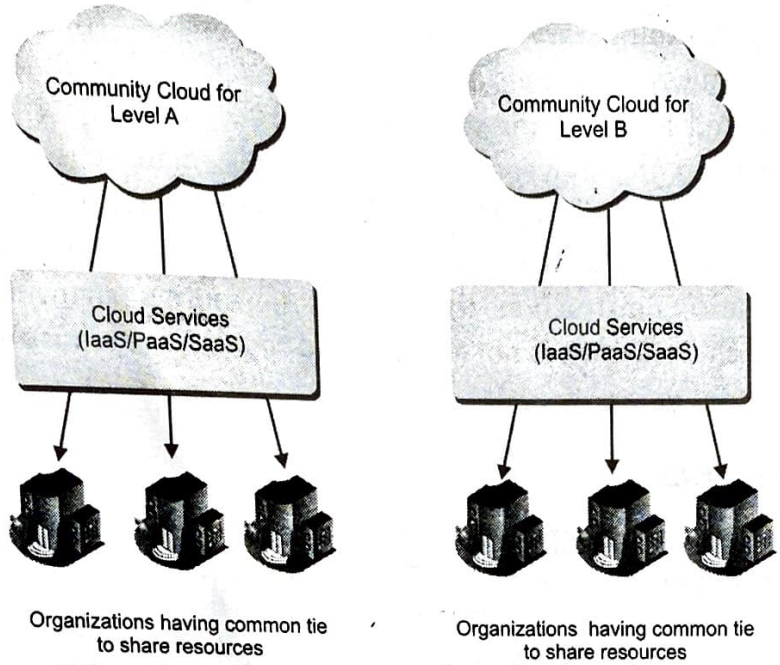


Figure 3.13: Level of Accessibility in Community Clouds

Hybrid Cloud

The cloud environment in which various internal or external service providers offer services to many organizations is known as a hybrid cloud. Generally, it is observed that an organization hosts applications, which require high level of security and are critical, on the private cloud. It is also possible that the applications that are not so important or confidential can be hosted on the public cloud. In hybrid clouds, an organization can use both types of cloud, i.e. public and private together. Such type of cloud is generally used in situations such as cloud bursting. In case of cloud bursting, an organization generally uses its own computing infrastructure; however, in high load requirements, the organization can access clouds. In other words, the organization using the hybrid cloud can manage an internal private cloud for general use and migrate the entire or a part of an application to the public cloud during the peak periods.

Figure 3.14 shows a hybrid cloud:

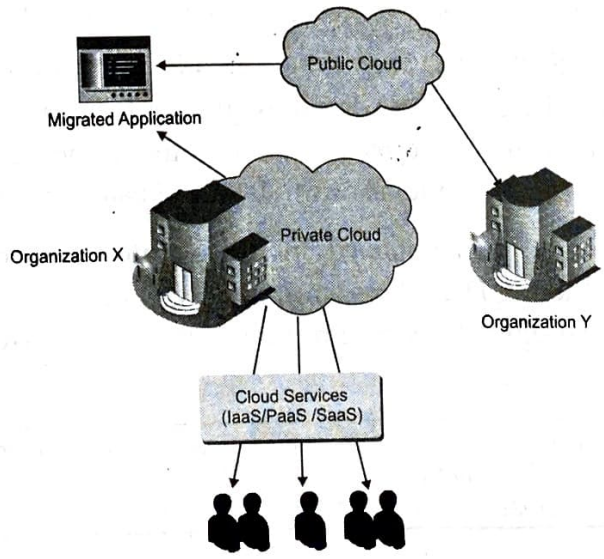


Figure 3.14: Implementation of a Hybrid Cloud

Cloud is a multipurpose platform that helps in, not only helps in handling Big Data analytics operations but also performing various tasks, including data storage, data backup, and customer service. Nowadays, business operations are performed mostly by using laptops, tablets, and mobile devices, which are suited for accessing cloud services, because most people today want to access computers even when on the move. In addition to this, many customers use the Internet for purchasing some product or service. These online orders are taken from customers by product stores, which send instructions to the warehouse for delivering the product. The entire process of receiving orders, forwarding instructions to warehouses, handling payments, and tracking deliveries can be assisted by the cloud, which is not essential but reduces the infrastructure cost and improves scalability in content storage.

Cloud Delivery Models

Cloud environment provides computational resources in the form of hardware, software, and platform, which are deployed as services. Therefore, we can categorize these services in the following manner:

- **Infrastructure as a Service (IaaS)**—Infrastructure in cloud computing can be defined as a combination of various elements. These elements include network, storage, and hardware components. A public IaaS is used when a user selects cloud for storing one's data, which may include photographs captured at an event or the favorite videos of the user. On the other hand, the server backups used by the employees of an organization for saving the daily status report of the work come under private IaaS. Virtual machines, load balancers, and network-attached storage are some examples of IaaS that include network, hardware, and storage as services. In a public cloud, IaaS may help businesses to save their investments required for purchasing physical infrastructure. A business can select the OS of its choice to create a virtual machine that includes scalable storage and processing abilities.
- **Platform as a Service (PaaS)**—User applications are provided a platform for writing code and executing it through the PaaS cloud. In terms of cloud, platform means the OS that combines software development and deployment tools along with middleware services. Windows Azure and Google App Engine (GAE) are examples of PaaS cloud. In organizations having a private PaaS, it is possible for programmers to create and deploy applications for their requirements.
- **Software as a Service (SaaS)**—The SaaS cloud provides software applications that are accessible from wherever the user is. Customers do not require purchasing the software or installing it on their own devices. They can use it directly from the cloud. These applications are hired through annual contracts, which are operational only if IaaS and PaaS clouds are already being used.

Customized applications can be maintained in the private cloud of organizations, which can interlink the applications and Big Data stored in a public cloud. Hybrid clouds allow applications to analyze the data efficiently through the strengths of both public and private clouds.

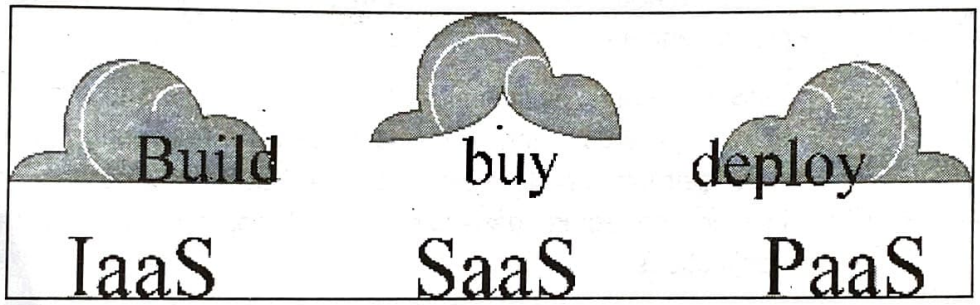
NOTE

Many organizations use both public and private clouds. Instead of using them separately, organizations combine the two cloud models in a hybrid cloud. A number of connections are formed between the two clouds, and operations are automated to improve efficiency.

EXHIBIT 1: Difference between SaaS, PaaS, IaaS

When your business has made the decision to consider cloud services for your application or infrastructure deployment, it's important that you grasp the fundamental differences between the core categories of cloud services available.

The cloud is a very broad concept, and it covers just about every possible sort of online service, but when businesses refer to cloud procurement, there are usually three models of cloud service under consideration: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). Each has its own intricacies and hybrid cloud models, but today we're going to help you develop an understanding of the high-level differences between SaaS, PaaS, and IaaS.



Software as a Service

In some ways, SaaS is very similar to the old thin-client model of software provision, where clients, in this case usually Web browsers, provide the point of access to software running on servers. SaaS is the most familiar form of cloud service for consumers. SaaS moves the task of managing software and its deployment to third-party services. Among the most familiar SaaS applications for business are customer relationship management applications like Sales force, productivity software suites like Google Apps, and storage solutions brothers like Box and Drop box.

Use of SaaS applications tends to reduce the cost of software ownership by removing the need for technical staff to install, manage, and upgrade software, as well as reduce the cost of licensing software. SaaS applications are usually provided on a subscription model.

Platform as a Service

PaaS functions at a lower level than SaaS, typically providing a platform on which software can be developed and deployed. PaaS providers abstract much of the work of dealing with servers and give clients an environment in which the operating system and server software, as well as the underlying server hardware and network infrastructure are taken care of, leaving users free to focus on the business side of scalability and the application development of their product or service. Businesses can requisition resources as they need them, scaling as demand grows, rather than investing in hardware with redundant resources. Examples of PaaS providers include Heroku, Google App Engine, and Red Hat's OpenShift.

Infrastructure as a Service

Moving down the stack, we get to the fundamental building blocks for cloud services. IaaS is comprised of highly automated and scalable compute resources, complemented by cloud storage and network capability, which can be self-provisioned, metered, and available on-demand.

IaaS providers offer these cloud servers and their associated resources via dashboard and/or API. IaaS clients have direct access to their servers and storage, just as they would with traditional servers but gain access to a much higher order of scalability. Users of IaaS can outsource and build a "virtual data center" in the cloud and have access to many of the same technologies and resource capabilities of a traditional data center without having to invest in capacity planning or the physical maintenance and management of it.

IaaS is the most flexible cloud computing model and allows for automated deployment of servers, processing power, storage, and networking. IaaS clients have true control over their infrastructure than users of PaaS or SaaS services. The main uses of IaaS include the actual development and deployment of PaaS, SaaS, and web-scale applications.

Source: https://www.computenext.com/blog/when-to-use-saas-paas-and-iaas/

Cloud Services for Big Data

Cloud services are associated with various models that are used for delivery and deployment. Cloud follows the same architecture as Big Data, both requiring distributed clusters of computing devices. Big Data systems include different specifications of cloud as an integral part; therefore, a cloud is termed as an ideal computing environment for handling Big Data.

In Big Data, the IaaS, PaaS, and SaaS clouds are used in the following manner:

- **IaaS**—The huge storage and computational power requirements for Big Data are fulfilled by the limitless storage space and computing ability obtained by the IaaS cloud.
- **PaaS**—PaaS offerings of various vendors have started adding various popular Big Data platforms that include MapReduce and Hadoop. These offerings save organizations from a lot of hassle, which may occur in managing individual hardware components and software applications.
- **SaaS**—Various organizations require identifying and analyzing the voice of customers, particularly on social media platforms. The social media data and the platform for analyzing the data are provided by SaaS vendors. In addition, private cloud facilitates the access to enterprise CRM data, which enables these analyses.

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 discussed as follows:

- **Amazon**—Amazon is one of the largest cloud service provider, and it offers its cloud services as Amazon Web Services (AWS).AWS includes some of the most popular cloud services, such as Elastic Compute Cloud (EC2), Elastic MapReduce, Simple Storage Service (S3), etc. Some of these services are discussed as follows:
  - **EC2**—Refers to a Web service that employs a large set of computing resources to perform its business operations. These resources are not properly utilized by Amazon, and therefore, they are pooled in the form of an IaaS cloud so that other organizations can take the benefit of these resources, ultimately benefitting Amazon through the rental cost. Organizations can use these resources elastically in a way that the hiring of resources is possible on an hourly basis.
  - **Elastic MapReduce**—It is a Web service that uses Amazon EC2 computation and Amazon S3 storage for storing and processing large amounts of data so that the cost of processing and storage is reduced significantly.
  - **DynamoDB**—Refers to a NoSQL database system in which data storage is done on Solid State Devices (SSDs). DynamoDB allows data replication for high availability and durability.
  - **Amazon S3**—Amazon Simple Storage Service (Amazon S3)is a Web interface that allows data storage over the Internet and makesweb-scale computing possible.
  - **High Performance Computing (HPC)**—Refers to a network that is replete with high bandwidth, low latency, and high computational abilities, which are required for processing Big Data, especially for solving issues related to education and business domains.
  - **RedShift**—Refers to a data warehouse service that is used to analyze data with the help of existing tools of business intelligence in an economical manner. You can scale Amazon RedShift for handling data up to a petabyte.
- **Google**—The cloud services that are provided by Google for handling Big Data include the following:
  - Google Compute Engine is a computing environment, which is secure, flexible, and based on virtual machine.
  - Google BigQuery is a Desktop as a Service (DaaS), which is used for searching huge amounts of data at a faster pace on the basis of SQL-format queries.
  - Google Prediction API, which is used for the identifying patterns in data, storingof patterns, and improving the patterns with successive utilizations.
- **Windows Azure**—Microsoft offers a PaaS cloud that is based on Windows and SQL abstractions and consists of a set of development tools, virtual machine support, management and media services, and mobile device services. Windows Azure PaaS is easy-to-adopt for people who are well equipped with the operations of .NET, SQL Server, and Windows. In addition, the Windows Azure HD Insight option added to the PaaS cloud makes it possible for the cloud users to address the emerging requirements to for integrating Big Data into Windows Azure

solutions. The platform used for building the Windows Azure PaaS is Horton works Data Platform (HDP) that, as stated by Microsoft, is fully compatible with Apache Hadoop. Moreover, Microsoft Excel and various other Business Intelligence (BI) tools can be connected to Windows Azure with support from HDInsight, which can be developed on the Windows Server also.

Hadoop is used as a cloud service in Windows Azure PaaS with the help of HDInsight. HDFS and MapReduce related frameworks are thus, offered economically, and in a simpler way, by the integration of Hadoop in this PaaS. The efficient management and storage of data are important features of HDInsight, which also uses Sqoop connector for importing the Windows Azure SQL data into HDFS or for exporting the data to a Windows Azure SQL database from HDFS.

SCENARIO

While discussing various Big Data handling technologies, Mr. Richard Stephens finds that a technology known as In-Memory Computing (IMC) can be used for carrying out calculations within the main memory. You will learn about IMC in the following section.

In-Memory Computing Technology for Big Data

We learned that distributed computing can help us meet requirements of storage and processing power for Big Data analytics. Another way to improve the computational speed and power of processing data is to use IMC. The representation of data in the form of rows and columns makes data processing easier and faster. Data stored in this manner is known as structured data in which a set of variables, each of which is associated with certain values, is used. Today, however, the data being generated is largely unstructured. The volume of data being generated today must be processed at a very high speed because the data is growing at a very fast rate. IMC is used to facilitate high-speed data processing. For example, IMC can help in tracking and monitoring consumers' activities and behaviors, which allow organizations to take timely actions for improving customer services and, thus, customer satisfaction.

We all know that data is stored on external devices known as secondary storage space. This data had to be accessed from the external source whenever an operation or any kind of modification on the data was required. External sources are accessed through Input/Output (I/O) channels that transfer data temporarily from secondary storage space to primary storage for processing purposes. The process of accessing external devices used to consume too much time during which the CPU could not be used for any other operation. The advantage of using external devices for data storage is that secondary storage is economical as compared to primary storage.

In the IMC technology, the RAM (or the primary storage space) is used for analyzing data. RAM helps us to increase the computing speed. Simultaneously, the reduction of primary storage cost has also made it possible to store data in the primary memory. The application finds data in the same location where it (the application) resides. Therefore, the analysis of data can be carried out in a more quick and efficient manner. Figure 3.15 shows the working of IMC:

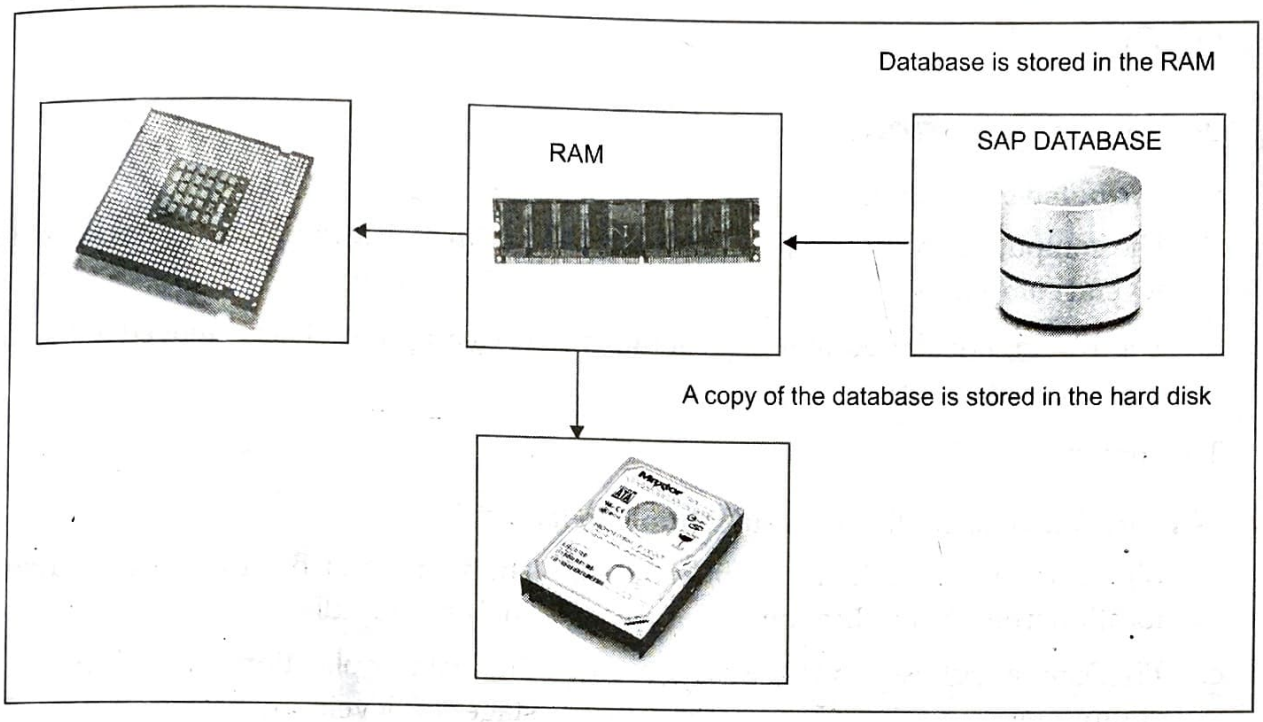


Figure 3.15: IMC Process

Relational Databases (RDBs) are used to store relational data. RDBs are the sources of information that are obtained by using SQL queries. The data stored in the RDBs is mostly structured, for example, accessing your marksheet from the college's website by filling in your roll number. Unstructured data, which comprises a wide range of text, images, videos, Web pages, blogs, business reports, press releases, emails, and text messages, provides information in the form of search results that are obtained by searching for specific keywords. Unstructured data is stored in a special kind of databases known as NoSQL databases. Entering a name in the Google search space to find the Web page, blog, or birthday video of a person is an example of accessing unstructured data.

The volume-related issues of Big Data are addressed by using IMC, and the diversity of Big Data is taken care of by NoSQL databases. Large organizations have centralized data warehouses for keeping the data safe. Users can access these warehouses only with the help of the IT department. The IMC technology helps different departments or business units of an organization to access and process the data that is relevant to them. This reduces the load on the central warehouse as every department takes care of processing its own data.

Summary

This chapter discussed the concept of distributed and parallel computing technologies and the role they play in handling Big Data. You also learned about various techniques used in parallel computing. Next, you learned about Hadoop and its functioning. The chapter further discussed cloud computing in detail, including its features, cloud deployment models, cloud delivery models, cloud services for Big Data, and cloud providers in the Big Data market. Finally, you learned about IMC and its importance in the context of Big Data.

Quick Revise

Multiple-Choice Questions

Q1. In distributed computing, _____.

a. The computing task is divided among several computers
b. Additional high-capacity disks are added to the system
c. The results are shared among several users of the network
d. The computing task is moved to the cloud

Ans. The correct option is a.

Q2. Why are Big Data applications susceptible to latency?

a. Big Data may reside in a different location from the application.
b. The volume of Big Data is too large to be analyzed rapidly.
c. Big Data cannot use in-memory computing.
d. Big Data applications are still in the early stages of development.

Ans. The correct option is b.

Q3. The three major parallel computing platforms are _____.

a. IaaS, PaaS, and SaaS
b. Clusters or grids, MPP, and HPC
c. Database, SQL, and network
d. Network, cloud, and DaaS

Ans. The correct option is b.

Q4. How does the Hadoop architecture use computing resources?

a. By distributing software to computing resources.
b. By distributing data and computing tasks to computing resources.
c. By creating shared memory for computing resources.
d. By distributing data to computing resources.

Ans. The correct option is d.

Q5. Hadoop makes the system more resilient by _____.

a. Using an effective firewall and anti-virus.
b. Keeping multiple copies of data.
c. Keeping each computing resource isolated.
d. Uploading data to a cloud for backup.

Ans. The correct option is c.

Q6. Which of the following are the disadvantages of using a public cloud?

a. Latency and risk to data security
b. Latency and software incompatibility
c. Higher cost and risk to data security
d. Higher cost and legal issues of location

Ans. The correct option is c.

Q7. Which of the following is an appropriate combination of resources in a hybrid cloud?

a. Backup in the private cloud and HR policies in the public cloud

b. Internal processes in the private cloud and Big Data in the public cloud

c. Customer communication in the private cloud and financial compliance in the public cloud

d. Big Data in the private cloud and backup in the public cloud

Ans. The correct option is b.

Q8. To be able to use accounting software on a cloud, you should have an appropriate _____ and _____.

a. Infrastructure and data

b. Platform and network

c. Infrastructure and platform

d. Platform and data

Ans. The correct option is c.

Q9. Cloud computing offers cost-effective solutions for computing resources because_____.

a. Cloud vendors negotiate for lower prices of resources.

b. Each resource has multiple users, which divides the cost.

c. All computing resources are located in low-cost regions.

d. Distributed computing lowers the cost of each resource.

Ans. The correct option is c.

Q10. A business that has customers' confidential information wants to use a public cloud for backup. Which of the following must the business ensure?

a. The cloud resources are compatible with the business' hardware and software.

b. The expected performance of the cloud is specified in detail in the service agreement.

c. The cloud allows access to government regulators and authorized third parties.

d. The cloud provides access to data only to designated persons of the business.

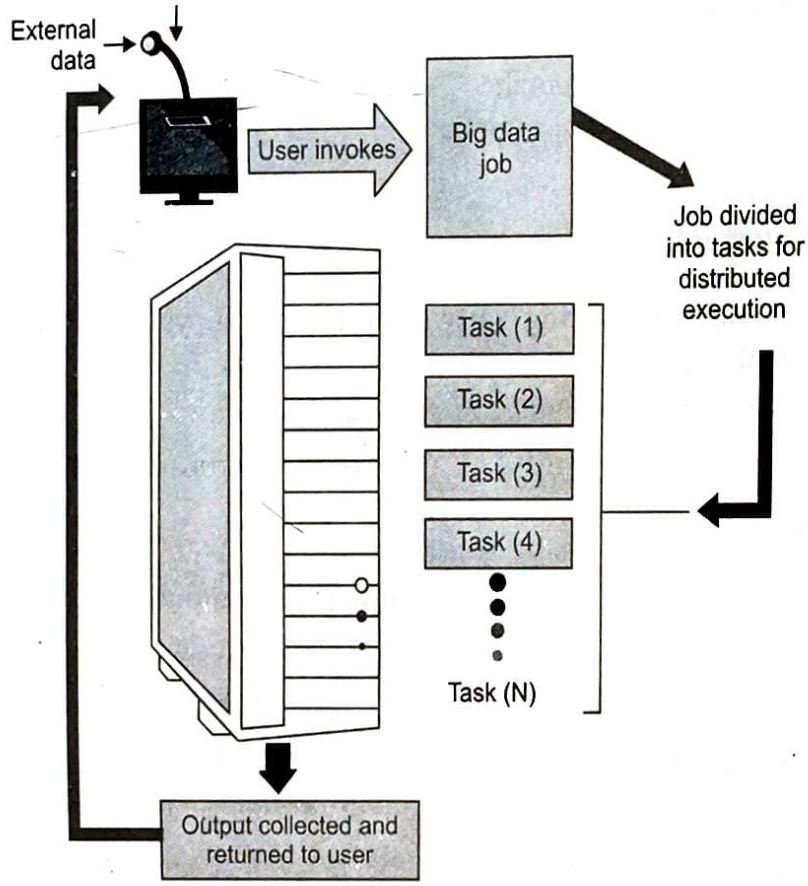
Ans. The correct option is a.

Subjective Questions

Q1. What is distributed computing? Explain the working of a distributed computing environment.

Ans. In distributed computing, multiple computing resources are connected in a network and computing tasks are distributed across these resources. This sharing of tasks increases the speed as well as the efficiency of the system. Because of this reason, distributed computing is considered faster and much more efficient than traditional methods of computing. It is also more suitable to process huge amounts of data in a limited time.

The following figure elaborates the processing of a large dataset in a distributed computing environment:



Distributed Computing Technique for Processing Large Data

As you can notice, the nodes are arranged within a system and are the elements that form the core of computing resources. These resources include CPU, memory, disks, etc. Big Data systems, usually, have higher scaling requirements. Therefore, these nodes are more beneficial for adding scalability to the Big Data environment, as and when required. The system, with added scalability, can accommodate the growing amounts of data more efficiently and flexibly. Distributed computing technique also makes use of virtualization and load balancing features.

The sharing of workload across various systems throughout the network to manage the load is known as load balancing. The virtualization feature creates a virtual environment in which hardware platform, storage device, and OS are included.

Q2. List the differences between parallel and distributed systems.

Ans. The following table lists the differences between parallel and distributed systems:

Distributed System	Parallel System
An independent, autonomous system connected in a network for accomplishing specific tasks	A computer system with several processing units attached to it

Distributed System	Parallel System
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, providing access to data and remotelylocated resources	Tight coupling of processing resources that are used for solving a single, complex problem

Q3. Discuss the techniques of parallel computing.

Ans. Some of the techniques of parallel computing are given in the table followings:

Parallel Computing Method	Description	Uses
Cluster or Grid Computing (Primarily used in Hadoop)	Cluster or grid computing is based on a connection of multiple servers in a network. This network is known as a cluster in which the servers share the workload among them. A cluster can be either homogeneous (comprising the same type of commodity hardware) or heterogeneous (consisting of different types of hardware).	A cluster can be created even by using hardware components that were acquired a long time back to provide cost-effective storage options. The overall cost may be very high in cluster computing.
Massively Parallel Processing (MPP) (Used in data warehouses)	A single machine working as a grid is used in the MPP platform, which is capable of handling the activities of storage, memory, and computing. Software written specifically for MPP platform is used for the optimization of MPP capabilities.	MPP platforms, such as EMC Greenplum, and ParAccel, are most suited for high-value use cases.
High-Performance Computing (HPC)	HPC environments are known to offer high performance and scalability by using IMC. This technology is especially suitable for processing floating-point data at high speeds.	HPC environments can be used to develop specialty and custom applications for research and business organizations where the result is more valuable, than the cost or where strategic importance of the project is of high priority.

Q4. List the important features of Hadoop.

Ans. The following are some of the important features of Hadoop:

- Hadoop performs well with several machines without requiring shared memory or disks among them. Hence, the efficiency-related issues related to storage and access to data get automatically solved.
- Hadoop uses various servers for data distribution, thus improving data loading and data storage operations.
- The data stored across various servers can be tracked by Hadoop. It helps in accessing and retrieving data, as and when required.
- Hadoop improves data processing by running computing tasks on all available processors that are working in parallel. The performance of Hadoop remains up to the mark both in the case of complex computational questions and of large and varied data.
- Hadoop keeps multiple copies of data to improve resilience that helps in maintaining consistency, especially in case of server failure.

Q5. Discuss the features of cloud computing that can be used to handle Big Data.

Ans. The following are some features of cloud computing that can be used to handle Big Data:

- **Scalability**—Scalability means addition of new resources to an existing infrastructure. The increase in the amount of data being collected and analyzed requires organizations to improve their hardware components' processing ability. These organizations may, at times, need to replace the existing hardware with a new set of hardware components in order to improve data management and processing activities. The new hardware may not provide complete support to the software that used to run properly on the earlier set of hardware. We can solve such issues by using cloud services that employ the distributed computing technique to provide scalability to the architecture.
- **Elasticity**—Elasticity in cloud means hiring certain resources, as and when required, and paying for the resources that have been used. No extra payment is required for acquiring specific cloud services. For example, a business expecting the use of more data during in-store promotion could hire more resources to provide high processing power. Moreover, a cloud does not require customers to declare their resource requirements in advance.
- **Resource Pooling**—Resource pooling is an important aspect of cloud services for Big Data analytics. In resource pooling, multiple organizations, which use similar kinds of resources to carry out computing practices, have no need to individually hire all the resources. The sharing of resources is allowed in a cloud, which facilitates cost cutting through resource pooling.
- **Self Service**—Cloud computing involves a simple user interface that helps customers to directly access the cloud services they want. The process of selecting the needed services requires no intervention from human beings and can be accessed automatically.

- **Low Cost**—A careful planning, use, management, and control of resources help organizations to reduce the cost of acquiring hardware significantly. Also, cloud offers customized solutions especially to organizations that cannot afford too much initial investment in purchasing the resources that are used for computation in Big Data analytics. The cloud provides them the pay-as-you-use option in which organizations need to sign for those resources only that are essential. This also helps the cloud provider in harnessing benefits of economies of scale and providing benefit to their customers in terms of cost reduction.
- **Fault Tolerance**—Cloud computing provides fault tolerance by offering uninterrupted services to customers, especially in cases of component failure. The responsibility of handling the workload is shifted to other components of the cloud.

Q6. Discuss the role cloud services play in handling Big Data.

Ans. Cloud services are associated with various models that are used for delivery and deployment. Cloud follows the same architecture as Big Data, both requiring distributed clusters of computing devices. Big Data systems include different specifications of cloud as an integral part; therefore, a cloud is termed as an ideal computing environment for handling Big Data.

In Big Data, the IaaS, PaaS, and SaaS clouds are used in the following manner:

- **IaaS**—The huge storage and computational power requirements for Big Data are fulfilled by the limitless storage space and computing ability obtained by the IaaS cloud.
- **PaaS**—PaaS offerings of various vendors have started adding various popular Big Data platforms that include MapReduce and Hadoop. These offerings save organizations from a lot of hassles, which may occur in managing individual hardware components and software applications.
- **SaaS**—Various organizations require identifying and analyzing the voice of customers, particularly on social media platforms. The social media data and the platform for analyzing the data are provided by SaaS vendors. In addition, private cloud facilitates the access to enterprise CRM data, which enables these analyses.

Q7. Write a short note on In-Memory Computing technology.

Ans. IMC is used to facilitate high-speed data processing. For example, IMC can help in tracking and monitoring consumers' activities and behaviors, which allow organizations to take timely actions for improving customer services and, thus, customer satisfaction.

We all know that data is stored on external devices known as secondary storage space. This data had to be accessed from the external source whenever an operation or any kind of modification on the data was required. External sources are accessed through Input/Output (I/O) channels that transfer data temporarily from secondary storage space to primary storage for processing purposes. The process of accessing external devices used to consume too much time during which the CPU could not be used for any other operation. The advantage of using external devices for data storage is that secondary storage is economical as compared to primary storage.

In the IMC technology, the RAM (or the primary storage space) is used for analyzing data. RAM helps us to increase the computing speed. Simultaneously, the reduction of primary storage cost has also made it possible to store data in the primary memory. The application finds data in the same location where it (the application) resides. Therefore, the analysis of data can be carried out in a more quick and efficient manner.