What are the different vendor-specific distributions of Hadoop?

The different vendor-specific distributions of Hadoop are Cloudera, MAPR, Amazon EMR, Microsoft Azure, IBM InfoSphere, and Hortonworks (Cloudera).

What are the different Hadoop configuration files?

The different Hadoop configuration files include:

- hadoop-env.sh
- mapred-site.xml
- core-site.xml
- yarn-site.xml
- hdfs-site.xml
- Master and Slaves

What are the three modes in which Hadoop can run

The three modes in which Hadoop can run are :

1.Standalone mode: This is the default mode. It uses the local FileSystem and a single Java process to run the Hadoop services.

2.Pseudo-distributed mode: This uses a single-node Hadoop deployment to execute all Hadoop services.

3.Fully-distributed mode: This uses separate nodes to run Hadoop master and slave services.

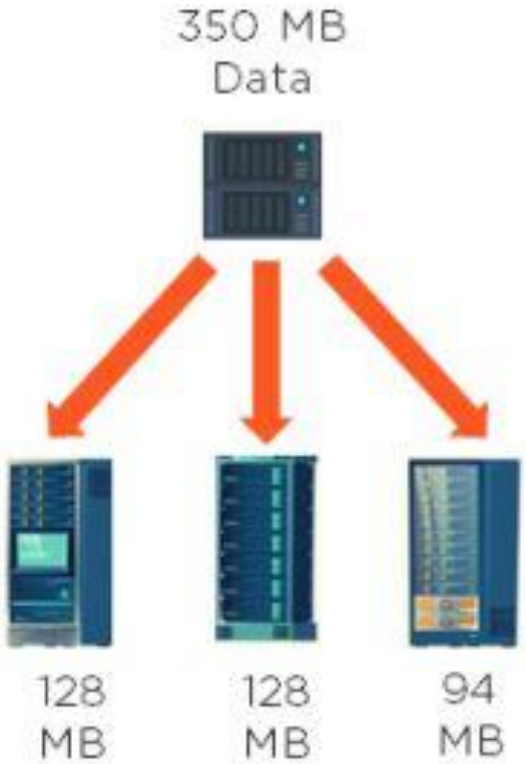
What are the differences between regular FileSystem and HDFS?

1.Regular FileSystem: In regular FileSystem, data is maintained in a single system. If the machine crashes, data recovery is challenging due to low fault tolerance. Seek time is more and hence it takes more time to process the data.

2.HDFS: Data is distributed and maintained on multiple systems. If a DataNode crashes, data can still be recovered from other nodes in the cluster. Time taken to read data is comparatively more, as there is local data read to the disc and coordination of data from multiple systems.

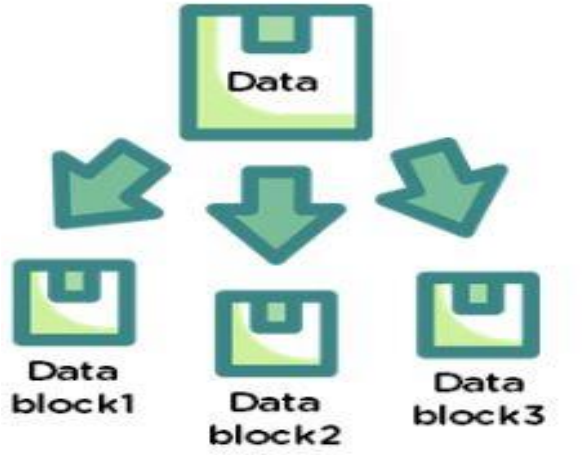
If you have an input file of 350 MB, how many input splits would HDFS create and what would be the size of each input split?

By default, each block in HDFS is divided into 128 MB. The size of all the blocks, except the last block, will be 128 MB. For an input file of 350 MB, there are three input splits in total. The size of each split is 128 MB, 128MB, and 94 MB.

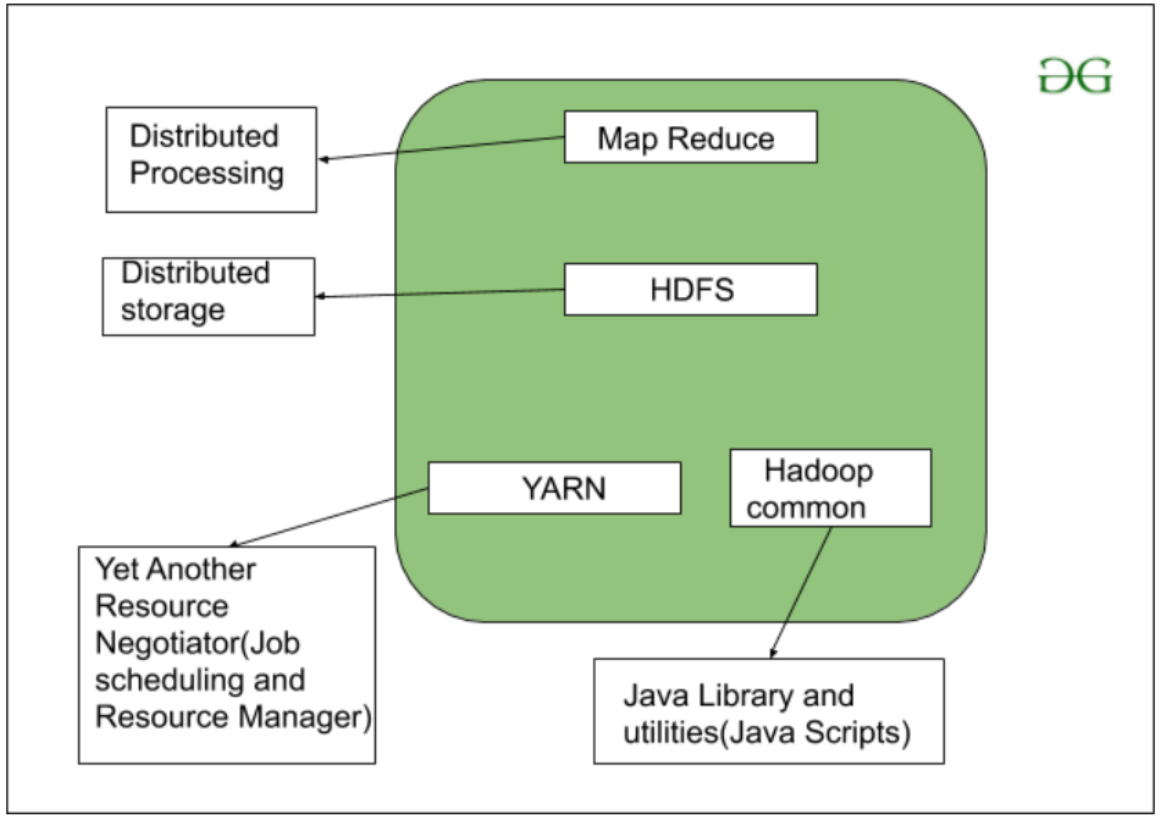


Why is HDFS fault-tolerant?

HDFS is fault-tolerant because it replicates data on different DataNodes. By default, a block of data is replicated on three DataNodes. The data blocks are stored in different DataNodes. If one node crashes, the data can still be retrieved from other DataNodes.



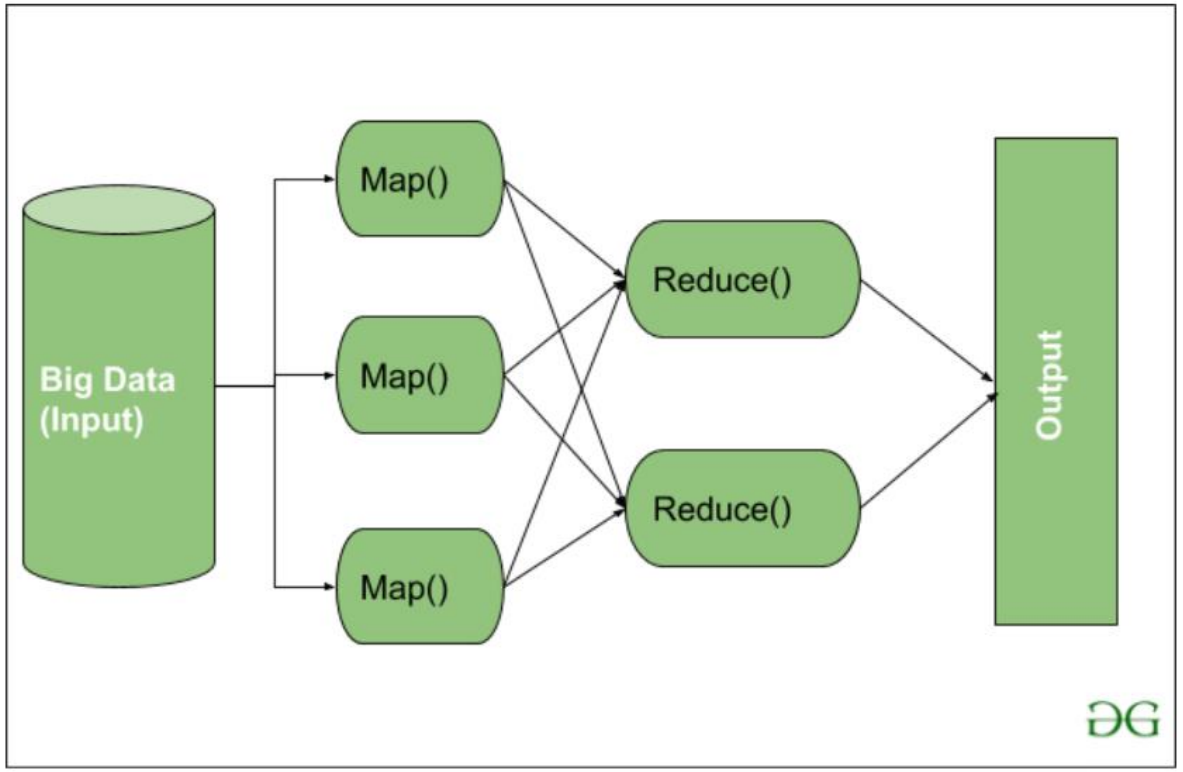
Explain the architecture of HDFS.



MapReduce

MapReduce nothing but just like an Algorithm or a data structure that is based on the YARN framework. The major feature of MapReduce is to perform the distributed processing in parallel in a Hadoop cluster which Makes Hadoop working so fast. When you are dealing with Big Data, serial processing is no more of any use. MapReduce has mainly 2 tasks which are divided phase-wise:

In first phase, Map is utilized and in next phase Reduce is utilized.



As we can see that an Input is provided to the Map(), now as we are using Big Data. The Input is a set of Data. The Map() function here breaks this DataBlocks into **Tuples** that are nothing but a key-value pair. These key-value pairs are now sent as input to the Reduce(). The Reduce() function then combines this broken Tuples or key-value pair based on its Key value and form set of Tuples, and perform some operation like sorting, summation type job, etc. which is then sent to the final Output Node. Finally, the Output is Obtained.

HDFS

HDFS(Hadoop Distributed File System) is utilized for storage permission. It is mainly designed for working on commodity Hardware devices(inexpensive devices), working on a distributed file system design. HDFS is designed in such a way that it believes more in storing the data in a large chunk of blocks rather than storing small data blocks.

HDFS in Hadoop provides Fault-tolerance and High availability to the storage layer and the other devices present in that Hadoop cluster. Data storage Nodes in HDFS.

- NameNode(Master)
- DataNode(Slave)

NameNode: NameNode works as a Master in a Hadoop cluster that guides the Datanode(Slaves). Namenode is mainly used for storing the Metadata i.e. the data about the data. Meta Data can be the transaction logs that keep track of the user's activity in a Hadoop cluster.

Meta Data can also be the name of the file, size, and the information about the location(Block number, Block ids) of Datanode that Namenode stores to find the closest DataNode for Faster Communication. Namenode instructs the DataNodes with the operation like delete, create, Replicate, etc.

DataNode: DataNodes works as a Slave DataNodes are mainly utilized for storing the data in a Hadoop cluster, the number of DataNodes can be from 1 to 500 or even more than that. The more number of DataNode, the Hadoop cluster will be able to store more data. So it is advised that the DataNode should have High storing capacity to store a large number of file blocks.

YARN(Yet Another Resource Negotiator)

YARN is a Framework on which MapReduce works. YARN performs 2 operations that are Job scheduling and Resource Management. The Purpose of Job schedular is to divide a big task into small jobs so that each job can be assigned to various slaves in a Hadoop cluster and Processing can be Maximized. Job Scheduler also keeps track of which job is important, which job has more priority, dependencies between the jobs and all the other information like job timing, etc. And the use of Resource Manager is to manage all the resources that are made available for running a Hadoop cluster.

Features of YARN

Multi-Tenancy
Scalability
Cluster-Utilization
Compatibility

Hadoop common or Common Utilities

Hadoop common or Common utilities are nothing but our java library and java files or we can say the java scripts that we need for all the other components present in a Hadoop cluster. these utilities are used by HDFS, YARN, and MapReduce for running the cluster. Hadoop Common verify that Hardware failure in a Hadoop cluster is common so it needs to be solved automatically in software by Hadoop Framework.

Explanation

HDFS is a system that allows you to distribute big data storage across a group of computers.

- It also keeps redundant copies of data. So, if one of your computers randomly bursts into flames or if some technical issues arise,
- HDFS can recover by creating a backup from a copy of the data it had automatically saved, and you won't even know what happened.

YARN: It comes next in the Hadoop ecosystem (Yet Another Resource Negotiator). It is the location where Hadoop's data processing is put to use.

- The system that controls the resources on your computing cluster is called YARN.
- It is the one that chooses who gets to perform the duties, as well as when, which nodes are open for more work, and which nodes are not.

Mapreduce: It is another part of the Hadoop ecosystem called MapReduce.

- It is essentially a programming model that lets you process data across an entire cluster.
- It mainly comprises Mappers and Reducers, which are several scripts or functions that you might write when creating a MapReduce programme.

Hadoop: It is a distributed, open-source, multidimensional, scalable NoSQL database. Based on Java, HBase runs on HDFS and gives Hadoop capabilities and functionalities akin to those of Google Bigtable.

If one of your computers randomly bursts into flames or if some technical issues arise, it can recover by creating a backup from a copy of the data that it had automatically saved, and you won't even know what happened. That is___

1. YARN
2. HDFS
3. Hadoop
4. Mapreduce

What is Hadoop?

1.A programming language for big data analytics

2.A distributed file system for storing big data

3.A framework for distributed storage and processing of big data

4.A database management system for big data

Which Phase in MapReduce is Responsible for Data Aggregation?

1.Map phase

2.Shuffle phase

3.Reduce phase

4.Merge phase

What is MapReduce in Hadoop Fundamentals?

1.A programming language

2.A distributed computing model for processing big data

3.A database query language

4.A data visualization tool

What is the Function of Apache Spark in Processing Big Data?

1.Data storage

2.Data visualization

3.Data processing and analytics

4.Data security

Which of the following is a Big Data Tool Employed for Real-Time Stream Processing?

1.Hadoop

2.Apache Kafka

3.MySQL

4.MongoDB

What is the Purpose of a Data Warehouse in Big Data Analytics?

1.To store real-time data streams

2.To store structured data in relational databases

3.To integrate and analyze data from multiple sources for reporting and analysis

4.To store unstructured data such as images and videos

Which of the following is a NoSQL Database Most Often Used to Deal with Huge Data Volumes?

1.MySQL

2.PostgreSQL

3.MongoDB

4.SQLite

Which of the following is a Batch-Processing Framework Commonly Used in Big Data Analytics?

1.Apache Spark

2.Apache Kafka

3.Apache Storm

4.Apache Flink

Which of the following Tools is Commonly used for Interactive Data Visualization in Big Data Analytics?

1.Tableau

2.Microsoft Excel

3.Power BI

4.MATLAB

What is Data Transformation in Big Data Processing?

1.To store raw data in a distributed file system

2.To convert data into a structured format

3.To summarize and aggregate data for analysis

4.To visualize data using charts and graphs

Which Technology Feature is Normally Used in an Integrating System of a Big Data Environment?

1.ETL (Extract, Transform, Load) tools

2.Apache Kafka

3.NoSQL databases

4.Hadoop Distributed File System (HDFS)

Which of the following is a Common Challenge in Big Data Quality Management?

1.Lack of data variety

2.Low data volume

3.Data duplication and inconsistency

4.High data velocity

1. **What is Big Data?**

1.Data with a large file size

2.Data with high velocity and variety that exceeds traditional data processing capabilities

3.Data with high-security requirements

4.Data with a high level of accuracy

2. Which of the following is a Characteristic of BigDdata?

1.Low volume

2.Structured format

3.Low velocity

4.Predictable variety

3. What do you mean by Big Data Analytics?

1.It is a process of gathering huge amounts of data.

2.The process of analysis of large and complicated data helps in revealing hidden patterns, trends, and insights.

3.The process of securing big data

4.The process of deleting unnecessary data

4. Which Techniques are used in Big Data Analytics?

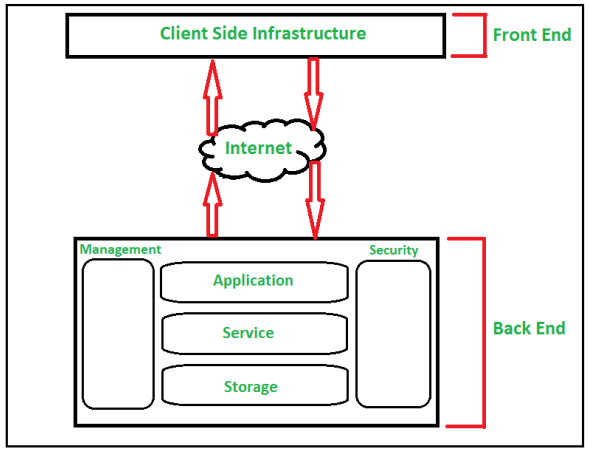
1.Regression analysis

2.Linear programming

3.Gradient descent

4.Machine learning

Category	Cloud Computing	Cluster Computing
Goal	Providing on demand IT resources and services.	Performing a complex task in a modular approach.
Resource Sharing	Specific assigned resources are not shareable.	Specific assigned resources are not shareable.
Resource type	In cloud computing there is heterogeneous resource type.	In Cluster Computing there is homogeneous resource type.
Virtualization	Virtualization hardware and software resources.	No virtualization resources.
Security	Security through isolation can be achieved.	Security through node credential can be achieved.
Initial Cost	Initial capital cost for setup is very low.	Initial capital cost for setup is very high.



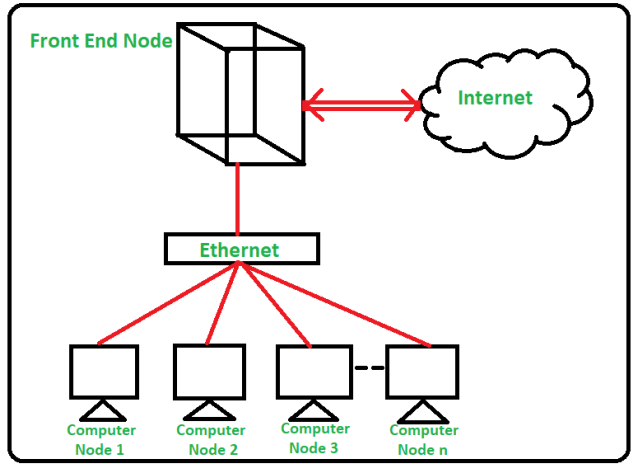
1.Cloud Computing :

Cloud Computing refers to the on demand delivery of the IT resources especially computing power and data storage through the internet with pay per use pricing. It generally refers to the data centers available to the users over internet. Cloud Computing is the virtualized pool of resources. It allows us to create, configure and customize our applications online. The user can access any resource at any time and any where with out worrying about the management and maintenance of actual resources. Cloud computing delivers both a combination of hardware and software based computing resources over network.

2. Cluster Computing :

Cluster computing refers to the process of sharing the computation task to multiple computers of the cluster. The number of computers are connected on a network and they perform a single task by forming a Cluster of computers where the process of computing is called as cluster computing.

Cluster Computing is a high performance computing framework which helps in solving more complex operations more efficiently with a faster processing speed and better data integrity. Cluster Computing is a networking technology that performs its operations based on the principle of distributed systems.



S.No.	Cloud Computing	Hadoop
1	Data is stored on cloud servers situated at different locations.	Large data is processed and stored as volumes of data in a HDFS environment.
2	Constitutes complex computer concepts, involves large number of computers which are connected in real time.	It is a framework with simple programming models to process data.
3	Data is stored and processed in remote servers up next accessed from any preferred location.	The processed data yields new patterns hidden in the data.
4	Requires low maintenance, backup and recovery of data is available.	Need more maintenance when compared and difficult to retrieve lost data.
5	Internet is used to provide cloud based services.	Distributed computing is used for processing the data.
6	On demand services are provided by cloud platforms.	Different formats of data is being processed and analysed.
7	Computing behaviour like Performance, scalability are analysed.	Processed data will be analysed and stored.
8	No need to purchase expensive hardware .	Business organizers can apply the predicted outcomes of the processed data in their businesses.

Hadoop:

Hadoop is a software framework which allow users to process large data sets in a distributed environment. Depending upon the size of the data set computers are clustered in a distributed file system (DFS) manner. In Hadoop Distributed File System (HDFS) each file is divided into blocks of equal size, replicated thrice and stored randomly in Data Nodes. Many organizations started using Hadoop as their data warehouse since it can process data of different formats.

Cloud Computing:

Computing services such as storage, networking, databases, servers provided over the internet is known as Cloud computing. It is widely used because it saves the hardware costs for organizations, more secured with the latest technologies, less time taken for the sender and receiver communications i.e reduced network latency. Cloud data backup will be done by the cloud providers and can be accessed by users from anywhere using the Internet. The three different cloud computing architecture are :

•**Public Cloud** – operated by third-party cloud providers for example google cloud.

•**Private cloud** – computing resources are used by the single organization for their own businesses needs.

•**Hybrid cloud** – a combination of both public and private cloud features.

Cluster Computing	Grid Computing
Nodes must be homogeneous (same hardware and OS)	Nodes can be homogeneous or heterogeneous
Computers are dedicated to the same task	Computers contribute unused resources
Computers are located close to each other.	Computers may be located at a huge distance from one another.
Computers are connected by a high speed local area network bus.	Computers are connected using a low speed bus or the internet.
Computers are connected in a centralized network topology.	Computers are connected in a distributed or de-centralized network topology.
Scheduling is controlled by a central server.	It may have servers, but mostly each node behaves independently.
Whole system has a centralized resource manager.	Every node manages it's resources independently.
Whole system functions as a single system.	Every node is autonomous, and anyone can opt out anytime.
Cluster computing is used in areas such as WebLogic Application Servers, Databases, etc.	Grid computing is used in areas such as predictive modeling, Automation, simulations, etc.
It has Centralized Resource management.	It has Distributed Resource Management.

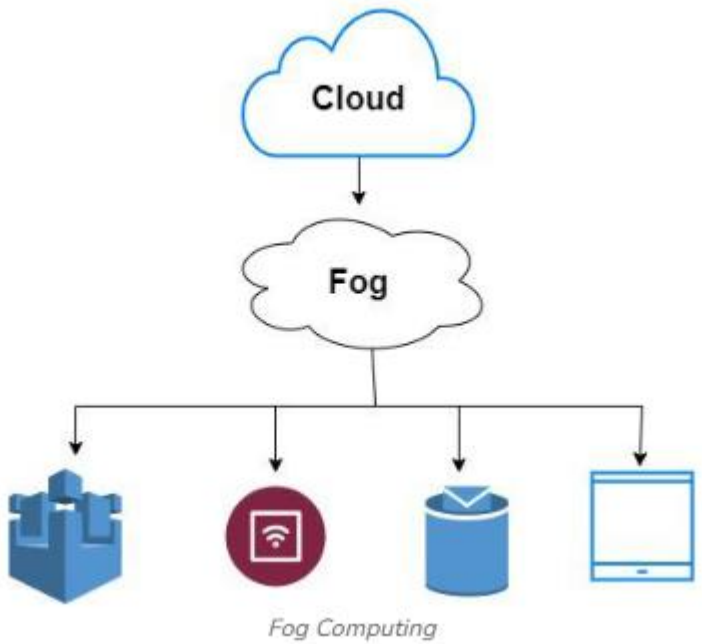
Cloud Computing	Grid Computing
Cloud computing is a Client-server computing architecture.	While it is a Distributed computing architecture.
Cloud computing is a centralized executive.	While grid computing is a decentralized executive.
In cloud computing, resources are used in centralized pattern.	While in grid computing, resources are used in collaborative pattern.
It is more flexible than grid computing.	While it is less flexible than cloud computing.
In cloud computing, the users pay for the use.	While in grid computing, the users do not pay for use.
Cloud computing is a high accessible service.	While grid computing is a low accessible service.
It is highly scalable as compared to grid computing.	While grid computing is low scalable in comparison to cloud computing.
It can be accessed through standard web protocols.	While it is accessible through grid middleware.
Cloud computing is based on service-oriented.	Grid computing is based on application-oriented.
Cloud computing uses service like IAAS, PAAS, SAAS.	Grid computing uses service like distributed computing, distributed pervasive, distributed information.

In Conclusion both cloud computing and grid computing use distributed computing resources. Cloud computing provides scalable, on-demand services with a focus on accessibility and cost efficiency, making it suitable for a wide range of applications, from personal to enterprise.

Grid computing, on the other hand, performs well in contexts that require high-performance computing and resource sharing across different domains, such as scientific research and complex computations. Understanding these characteristics helps in selecting the best technology that meets certain computing needs and goals.

What is Fog Computing?

Fog Computing is the term introduced by Cisco that refers to extending cloud computing to an edge of the enterprise's network. Thus, it is also known as Edge Computing or Fogging. It facilitates the operation of computing, storage, and networking services between end devices and computing data centers.



Fog Computing

•The devices comprising the fog infrastructure are known as fog nodes.

•In fog computing, all the storage capabilities, computation capabilities, data along with the applications are placed between the cloud and the physical host.

•All these functionalities are placed more towards the host. This makes processing faster as it is done almost at the place where data is created.

•It improves the efficiency of the system and is also used to ensure increased security.

When to Use Fog Computing?

•It is used when only selected data is required to send to the cloud. This selected data is chosen for long-term storage and is less frequently accessed by the host.

•It is used when the data should be analyzed within a fraction of seconds i.e Latency should be low.

•It is used whenever a large number of services need to be provided over a large area at different geographical locations.

•Devices that are subjected to rigorous computations and processings must use fog computing.

•Real-world examples where fog computing is used are in IoT devices Devices with Sensors, Cameras (IIoT-Industrial Internet of Things), etc.

Advantages of Fog Computing

•This approach reduces the amount of data that needs to be sent to the cloud.

•Since the distance to be traveled by the data is reduced, it results in saving network bandwidth.

•Reduces the response time of the system.

•It improves the overall security of the system as the data resides close to the host.

•It provides better privacy as industries can perform analysis on their data locally.

Disadvantages of Fog Computing

•Congestion may occur between the host and the fog node due to increased traffic (heavy data flow).

•Power consumption increases when another layer is placed between the host and the cloud.

•Scheduling tasks between host and fog nodes along with fog nodes and the cloud is difficult.

•Data management becomes tedious as along with the data stored and computed, the transmission of data involves encryption-decryption too which in turn release data.

Applications of Fog Computing

•It can be used to monitor and analyze the patients' condition. In case of emergency, doctors can be alerted.

•It can be used for real-time rail monitoring as for high-speed trains we want as little latency as possible.

•It can be used for gas and oils pipeline optimization. It generates a huge amount of data and it is inefficient to store all data into the cloud for analysis.

Difference Between Edge Computing and Fog Computing

Edge Computing	Fog Computing
Less scalable than fog computing.	Highly scalable when compared to edge computing.
Millions of nodes are present.	Billions of nodes are present.
Nodes are installed far away from the cloud.	Nodes in this computing are installed closer to the cloud(remote database where data is stored).
Edge computing is a subdivision of fog computing.	Fog computing is a subdivision of cloud computing.
The bandwidth requirement is very low. Because data comes from the edge nodes themselves.	The bandwidth requirement is high. Data originating from edge nodes is transferred to the cloud.
Operational cost is higher.	Operational cost is comparatively lower.
High privacy. Attacks on data are very low.	The probability of data attacks is higher.
Edge devices are the inclusion of the IoT devices or client's network.	Fog is an extended layer of cloud.

Feature	Cloud Computing	Fog Computing
Latency	Cloud computing has high latency compared to fog computing	Fog computing has low latency
Capacity	Cloud Computing does not provide any reduction in data while sending or transforming data	Fog Computing reduces the amount of data sent to cloud computing.
Responsiveness	Response time of the system is low.	Response time of the system is high.
Security	Cloud computing has less security compared to Fog Computing	Fog computing has high Security.
Speed	Access speed is high depending on the VM connectivity.	High even more compared to Cloud Computing.
Data Integration	Multiple data sources can be integrated.	Multiple Data sources and devices can be integrated.
Mobility	In cloud computing mobility is Limited.	Mobility is supported in fog computing.
Location Awareness	Partially Supported in Cloud computing.	Supported in fog computing.
Number of Server Nodes	Cloud computing has Few number of server nodes.	Fog computing has Large number of server nodes.
Geographical Distribution	It is centralized.	It is decentralized and distributed.

Cloud Computing: The delivery of on-demand computing services is known as cloud computing. We can use applications to storage and processing power over the internet. It is a pay as you go service. Without owning any computing infrastructure or any data centers, anyone can rent access to anything from applications to storage from a cloud service provider.
We can avoid the complexity of owning and maintaining infrastructure by using cloud computing services and pay for what we use.
In turn, cloud computing services providers can benefit from significant economies of scale by delivering the same services to a wide range of customers.

Fog Computing: Fog computing is a decentralized computing infrastructure or process in which computing resources are located between the data source and the cloud or any other data center. Fog computing is a paradigm that provides services to user requests at the edge networks. The devices at the fog layer usually perform operations related to networking such as routers, gateways, bridges, and hubs. Researchers envision these devices to be capable of performing both computational and networking operations, simultaneously. Although these devices are resource-constrained compared to the cloud servers, the geological spread and the decentralized nature help in offering reliable services with coverage over a wide area. Fog computing is the physical location of the devices, which are much closer to the users than the cloud servers.

Cloud Computing and Green Computing

1. Cloud Computing :
Cloud Computing, as name suggests, is basically a service-oriented architecture that involves delivering hosted services over internet. It delivers faster and accurate retrievals of applications and data. It is most efficient and better for promoting strong workflow and is more cost effective than traditional computing solutions.

2. Green Computing :
Green Computing, as name suggests, is basically study of designing, manufacturing, using and disposing computing devices in way that reduces their hazardous impact on environment. It is mostly used to promote energy efficiently in different applications such as washers, dryers, laptops, and refrigerators.

Cloud Computing	Green Computing
It is all about delivery of computing services including servers, storage, databases, networking, etc., over internet.	It is all about utilizing energy to perform operations in most efficient way possible.
It offers utility-oriented IT services to users worldwide.	It helps in using least amount of computing resources for doing most amount of work.
Its main goal is to provide magnitude improvement in cost effective, dynamic provisioning of IT services.	Its main goal is to attain economic viability and improve way of how computing devices are used.
It reduces energy consumption, waste, and carbon emissions, reduce carbon foot print, etc.	It reduces use of hazardous materials, increase energy efficiency during product's lifetime, manage power and energy efficiency, create sustainable business processes, etc.
It increases revenue of business organizations and help them to achieve business goals, provide faster communication, secure network collaboration, promote efficient utilization of existing resources, etc.	It reduces carbon footprint of business and provide a reputation boost, help business responsibly use energy and keep business running on energy-lean diet.
It is internet service that provides computing needs to computer users.	It is that a computer and technology is how much responsible for environmental change.
It allows company to diversity its network and server infrastructure.	It allows companies to improve disposal and recycling procedures.
It lowers IT costs, maintain business continuity, provide scalability, allows automatic software integrations, etc.	It lowers energy bills, lower overall power usage, cost-effective due to less energy usage and cooling requirements, etc.

For an HDFS service, we have a NameNode that has the master process running on one of the machines and DataNodes, which are the slave nodes.

NameNode

NameNode is the master service that hosts metadata in disk and RAM. It holds information about the various DataNodes, their location, the size of each block, etc.

DataNode

DataNodes hold the actual data blocks and send block reports to the NameNode every 10 seconds. The DataNode stores and retrieves the blocks when the NameNode asks. It reads and writes the client's request and performs block creation, deletion, and replication based on instructions from the NameNode.

How does rack awareness work in HDFS?

HDFS Rack Awareness refers to the knowledge of different DataNodes and how it is distributed across the racks of a Hadoop Cluster.