**What is Apache Hadoop?**

Hadoop is an open-source framework based on Java that manages the storage and processing of large amounts of data for applications. Hadoop uses distributed storage and parallel processing to handle big data and analytics jobs, breaking workloads down into smaller workloads that can be run at the same time.

**Hadoop Ecosystem:**

1. **Hadoop Distributed File System (HDFS):**

As the primary component of the Hadoop ecosystem, HDFS is a distributed file system in which individual Hadoop nodes operate on data that resides in their local storage. This removes network latency, providing high-throughput access to application data. In addition, administrators don’t need to define schemas up front.

1. **Yet Another Resource Negotiator (YARN):**

YARN is a resource-management platform responsible for managing compute resources in clusters and using them to schedule users’ applications. It performs scheduling and resource allocation across the Hadoop system.

1. **MapReduce:**

MapReduce is a programming model for large-scale data processing. In the MapReduce model, subsets of larger datasets and instructions for processing the subsets are dispatched to multiple different nodes, where each subset is processed by a node in parallel with other processing jobs. After processing the results, individual subsets are combined into a smaller, more manageable dataset.

1. **Hadoop Common:**

Hadoop Common includes the libraries and utilities used and shared by other Hadoop modules.

**How does Hadoop work?**

Hadoop allows for the distribution of datasets across a cluster of commodity hardware. Processing is performed in parallel on multiple servers simultaneously.

Software clients input data into Hadoop. HDFS handles metadata and the distributed file system. MapReduce then processes and converts the data. Finally, YARN divides the jobs across the computing cluster.

All Hadoop modules are designed with a fundamental assumption that hardware failures of individual machines or racks of machines are common and should be automatically handled in software by the framework.

**What are the benefits of Hadoop?**

1. **Scalability**

Hadoop is important as one of the primary tools to store and process huge amounts of data quickly. It does this by using a distributed computing model which enables the fast processing of data that can be rapidly scaled by adding computing nodes.

1. **Low cost**

As an open-source framework that can run on commodity hardware and has a large ecosystem of tools, Hadoop is a low-cost option for the storage and management of big data.

1. **Flexibility**

Hadoop allows for flexibility in data storage as data does not require preprocessing before storing it which means that an organization can store as much data as they like and then utilize it later.

1. **Resilience**

As a distributed computing model, Hadoop allows for fault tolerance and system resilience, meaning if one of the hardware nodes fail, jobs are redirected to other nodes. Data stored on one Hadoop cluster is replicated across other nodes within the system to fortify against the possibility of hardware or software failure.

**What is Apache Hadoop used for?**

1. **Analytics and big data**

A wide variety of companies and organizations use Hadoop for research, production data processing, and analytics that require processing terabytes or petabytes of big data, storing diverse datasets, and data parallel processing.

1. **Data storage and archiving**

As Hadoop enables mass storage on commodity hardware, it is useful as a low-cost storage option for all kinds of data, such as transactions, click streams, or sensor and machine data.

1. **Data lakes**

Since Hadoop can help store data without preprocessing, it can be used to complement to data lakes, where large amounts of unrefined data are stored.

1. **Marketing analytics**

Marketing departments often use Hadoop to store and analyse customer relationship management (CRM) data.

**What is Amazon S3?**

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. Customers of all sizes and industries can use Amazon S3 to store and protect any amount of data for a range of use cases, such as data lakes, websites, mobile applications, backup and restore, archive, enterprise applications, IoT devices, and big data analytics. Amazon S3 provides management features so that you can optimize, organize, and configure access to your data to meet your specific business, organizational, and compliance requirements.

**How Amazon S3 works?**

Amazon S3 is an object storage service that stores data as objects, hierarchical data, or tabular data within buckets. An object is a file and any metadata that describes the file. A bucket is a container for objects.

To store your data in Amazon S3, you first create a bucket and specify a bucket name and AWS Region. Then, you upload your data to that bucket as objects in Amazon S3. Each object has a key (or key name), which is the unique identifier for the object within the bucket.

S3 provides features that you can configure to support your specific use case. For example, you can use S3 Versioning to keep multiple versions of an object in the same bucket, which allows you to restore objects that are accidentally deleted or overwritten.

Buckets and the objects in them are private and can be accessed only if you explicitly grant access permissions. You can use bucket policies, AWS Identity and Access Management (IAM) policies, access control lists (ACLs), and S3 Access Points to manage access.

**How Hadoop is used as data lake in AWS S3?**

Hadoop, particularly when used with Amazon EMR (Elastic MapReduce), can be integrated with Amazon S3 to create a data lake. Hadoop provides the processing engine, while S3 offers scalable and durable storage. EMRFS, optimized for Hadoop on S3, allows direct read/write access to S3, and HDFS can be used for on-cluster storage of intermediate data during processing.

**Hadoop as a Processing Engine:**

Hadoop is a framework for distributed storage and processing of large datasets. In the context of a data lake on AWS, Hadoop (often within Amazon EMR) provides the tools for analysing and processing the data stored in S3.

**Amazon S3 as the Storage Layer:**

Amazon S3 acts as the central repository for the data lake, storing data in its raw or native format.

**EMRFS for Optimized Access:**

EMRFS (Amazon Elastic MapReduce File System) is a Hadoop file system that is optimized to work with Amazon S3. It allows Hadoop to directly read and write data to S3, enabling efficient processing of data stored there.

**Decoupling Storage and Compute:**

By using S3 as the storage layer, the compute and storage resources can be scaled independently, allowing for cost optimization and flexible data processing workflows.

**HDFS for Intermediate Storage:**

While S3 is the primary storage, Hadoop's HDFS can be used for storing intermediate data generated during processing on the EMR cluster.

**Data Ingestion:**

AWS offers tools like AWS DataSync to migrate data from on-premises Hadoop clusters to S3, and other methods like AWS Snowball Edge can also be used.

**Data Lake Architecture:**

In a data lake architecture, data is stored in its raw format, allowing for diverse analytical tools and machine learning models to be applied to the same data.

**Benefits of this approach:**

This approach allows for cost-effective storage, scalability, and flexible processing of large datasets, making it suitable for various big data and analytics applications.