

Instructions:

What are the core components of a data lake? For the answer, compare the data lake offerings of AWS, Azure and IBM and also Apache Hudi along with any other open source solutions you can find (plus Wikipedia and YouTube).

Answer

Data stored in a raw original format is called a data lake. It is basically one place for all sorts of data and can be used as a source to do machine learning, analysis, reporting and BI. It is not necessary for the data to be structured and can store a large amount of data.

The components of a data lake are:

- Data Ingestion
- Data Storage
- Data Governance
- Data Discovery
- Data Exploration and Visualization

AWS:

- Variety of tools and product suite
- Comparatively Low Cost
- Data privacy ensure through strong security
- Good compliance standards

The AWS Cloud provides many of the building blocks required to help customers implement a secure, flexible, and cost-effective data lake. These include AWS managed services that help ingest, store, find, process, and analyze both structured and unstructured data. To support our customers as they build data lakes, AWS offers the data lake solution, which is an automated reference implementation that deploys a highly available, cost-effective data lake architecture on the AWS Cloud along with a user-friendly console for searching and requesting datasets

Azure:

- Unlimited storage capacity on ADLS
 - ADLA useful for large data processing a faster speed
 - Simple to use
-

- Easier to migrate from Hadoop cluster

With Azure Data Lake Store your organization can analyze all of its data in a single place with no artificial constraints. Your Data Lake Store can store trillions of files where a single file can be greater than a petabyte in size which is 200x larger than other cloud stores. This means that you don't have to rewrite code as you increase or decrease the size of the data stored or the amount of compute being spun up. This lets you focus on your business logic only and not on how you process and store large datasets. Data Lake also takes away the complexities normally associated with big data in the cloud, ensuring that it can meet your current and future business needs

IBM:

- Scalable and Elastic
- Data exploration and management
- Hybrid Data Lake
- Data Preparation and transformation

IBM is committed to open source technologies and the security, interoperability and data access they bring to advanced analytics. Together, IBM and Cloudera provide a choice of integrated technologies to build, manage and use a data lake for data science at scale. IBM and Cloudera work together to deliver enterprise-class data lake solutions to help you replace data silos with an agile, scalable platform that can collect, store, govern and secure raw data from across your business, making it ready for analysis. Available on premises or on cloud, Cloudera's advanced data platform combined with IBM products, services and multivendor support positions you to unlock the value of AI

Apache Hudi:

Apache Hudi is a storage abstraction framework that helps distributed organizations build and manage petabyte-scale data lakes. Using primitives such as upserts and incremental pulls, Hudi brings stream style processing to batch-like big data. These features help surface faster, fresher data for our services with a unified serving layer having data latencies in the order of minutes, avoiding any added overhead of maintaining multiple systems. Adding to its flexibility, Apache Hudi can be operated on the Hadoop Distributed File System (HDFS) or cloud stores.

Hudi enables Atomicity, Consistency, Isolation & Durability (ACID) semantics on a data lake. Hudi's two most widely used features are upserts and incremental pull, which give users the ability to absorb change data captures and apply them to the

data lake at scale. Hudi provides a wide range of pluggable indexing capabilities in order to achieve this, along with its own data index implementation. Hudi's ability to control and manage file layouts in the data lake is extremely important not only for overcoming HDFS namenode and other cloud store limitations, but also for maintaining a healthy data ecosystem by improving reliability and query performance. To this end, Hudi supports multiple query engine integrations such as Presto, Apache Hive, Apache Spark, and Apache Impala.

Hadoop:

- Uses YARN as a resource manager and Mapreduce to build clusters
- Data lakes can be set up on cloud or on-premises
- Open source and less expensive
- Very common and ETL tools can be integrated with Hadoop
- Easy to scale

A Hadoop data lake is a data management platform comprising one or more Hadoop clusters. It is used principally to process and store nonrelational data, such as log files, internet clickstream records, sensor data, Json objects, images and social media posts.