**Javairia Junaid(13014) and Mustafa Poonjla(13220)**

**DATA LAKE**

A data lake is a central storage repository that holds big data from many sources in a raw, granular format. It can store structured, semi-structured, or unstructured data, which means data can be kept in a more flexible format for future use. When storing data, a data lake associates it with identifiers and metadata tags for faster retrieval.

**COMPONENTS OF DATA LAKE**

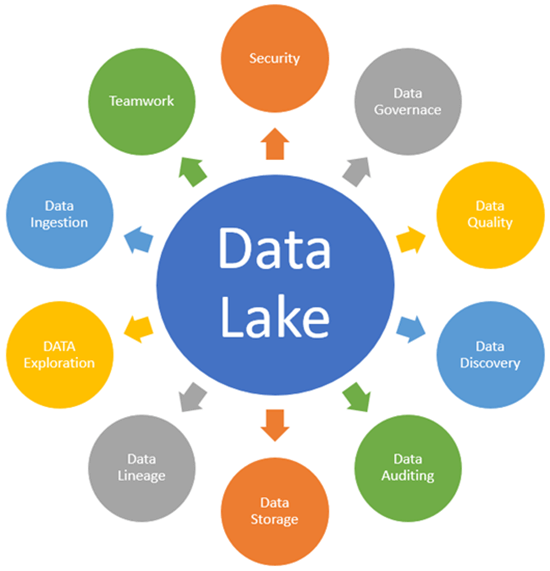


Figure 1: Key Components of Data Lake

* **Data Ingestion**

Data Ingestion allows connectors to get data from a different data sources and load into the Data lake.

Data Ingestion supports:

1. All types of Structured, Semi-Structured, and Unstructured data.
2. Multiple ingestions like Batch, Real-Time, One-time load.
3. Many types of data sources like Databases, Webservers, Emails, IoT, and FTP.

* **Data Storage**

Data storage should be scalable, offers cost-effective storage and allow fast access to data exploration. It should support various data formats.

* **Data Governance**

Data governance is a process of managing availability, usability, security, and integrity of data used in an organization.

* **Security**

Security needs to be implemented in every layer of the Data lake. It starts with Storage, Unearthing, and Consumption. The basic need is to stop access for unauthorized users. It should support different tools to access data with easy to navigate GUI and Dashboards.

Authentication, Accounting, Authorization and Data Protection are some important features of data lake security.

* **Data Quality**

Data quality is an essential component of Data Lake architecture. Data is used to exact business value. Extracting insights from poor quality data will lead to poor quality insights.

* **Data Discovery**

Data Discovery is another important stage before you can begin preparing data or analysis. In this stage, tagging technique is used to express the data understanding, by organizing and interpreting the data ingested in the Data lake.

* **Data Auditing**

Two major Data auditing tasks are tracking changes to the key dataset.

1. Tracking changes to important dataset elements
2. Captures how/ when/ and who changes to these elements.

Data auditing helps to evaluate risk and compliance.

* **Data Lineage**

This component deals with data's origins. It mainly deals with where it movers over time and what happens to it. It eases errors corrections in a data analytics process from origin to destination.

* **Data Exploration**

It is the beginning stage of data analysis. It helps to identify right dataset is vital before starting Data Exploration.All given components need to work together to play an important part in Data lake building easily evolve and explore the environment.

* **Team work**

Data lakes work best in a collaborative environment where analysis and findings of one group of users can be shared with other users (or groups), avoiding the need for duplicate effort and improve the overall business outcome.

All above-discussed components work together and play a vital role in Data Lake to create an environment where end users can discover and explore valuable insights out of the data in a secured and managed environment.

**DATA LAKE OFFERINGS**

* **Data lakes on AWS**

AWS has an exhaustive suite of product offerings for its data lake solution.

Amazon Simple Storage Service (Amazon S3) is at the center of the solution providing storage function. Kinesis Streams, Kinesis Firehose, Snowball, and Direct Connect are data ingestion tools that allow users to transfer massive amounts of data into S3. There is also a database migration service that helps migrate existing on-premises data to the cloud.

In addition to S3, there is DynamoDB, a low-latency No-SQL database, and Elastic Search, a service that provides a simplified mechanism to query the data lake. Cognito User Pools define user authentication and access to the data lake. Services such as Security Token Service, Key Management Service, CloudWatch, and CloudTrail ensure data security. For processing and analytics, there are tools such as RedShift, QuickSight, EMR, and Machine Learning.

The huge list of products offerings available from AWS come with a steep initial learning curve. However, the solution’s comprehensive functionalities find extensive use in business intelligence applications.

**Strengths:**

* Exhaustive and feature-rich product suite
* Flexibility to pick and choose products based unique requirements
* Low costs
* Strong security and compliance standards
* Separation of compute and storage to scale each one as needed
* Collaboration with APN (AWS Partner Network) firms such as Talend ensures seamless AWS onboarding

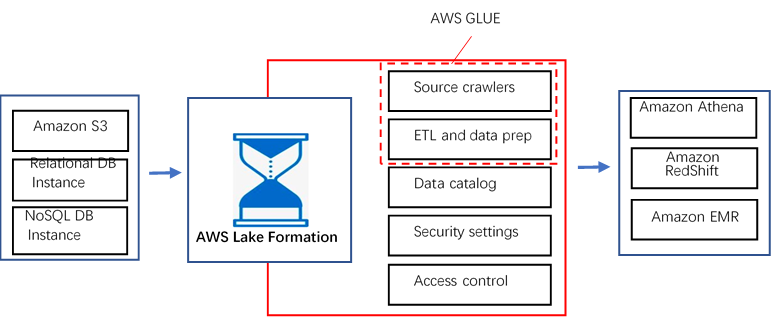


Figure 2: Data Lake Solution Provided by AWS

* **Data lakes on Azure**

[Azure](https://azure.microsoft.com/) is a data lake offered by Microsoft. It has a storage and an analytics layer; the storage layer is called as Azure Data Lake Store (ADLS) and the analytics layer consists of two components: Azure Data Lake Analytics and HDInsight.

ADLS is built on the HDFS standard and has unlimited storage capacity. It can store trillions of files with a single file larger than one petabyte in size. ADLS allows data to be stored in any format and is secure and scalable. It supports any application that uses the HDFS standard. This makes migration of existing data easier, and also facilitates plug-and-play with other compute engines.

HDInsight is a cloud-based data lake analytics service. Built on top of Hadoop YARN, it allows data to be accessed using tools such as Spark, Hive, Kafka, and Storm. It supports enterprise-grade security due to integration with Azure Active Directory.

Azure Data Lake Analytics is also an analytics service, but its approach is different. Rather than using tools such as Hive, it uses a language called U-SQL, a combination of SQL and C#, to access data. It is ideal for big data batch processing as it provides faster speed at lower costs (pay only for the jobs used).

**Strengths:**

* Both storage and compute in the cloud makes it simple to manage.
* Strong analytical services with powerful functionalities
* Easy to migrate from an existing Hadoop cluster
* Many big data experts are familiar with Hadoop and its tools, so it is easy to find skilled manpower.
* Integration with Active Directory ensures no separate effort to manage security

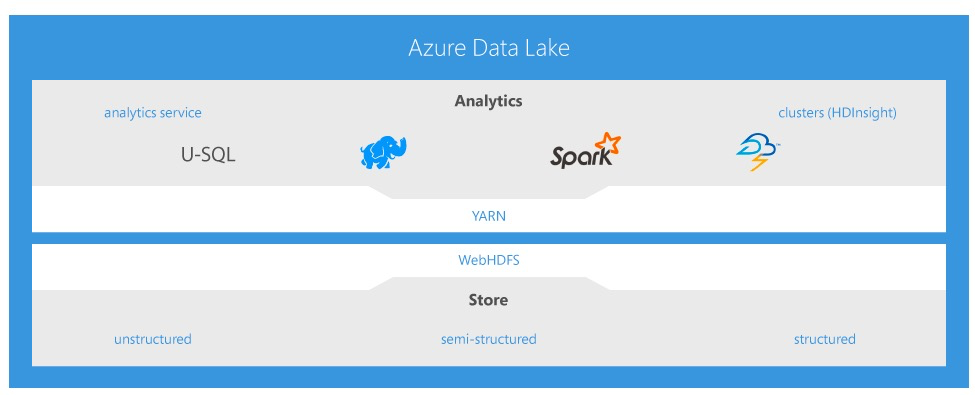


Figure 3: Architecture of Azure's Data Lake Solution

* **Data Lakes on IBM**

IBM Analytics have a strong portfolio of products to implement and support a Data Lake.

* InfoSphere Information Server is an integrated suite of capabilities for data quality management and data integration, of which the Information Governance Catalog is a part. It includes capabilities for data integration on Hadoop.
* DataWorks family provide SaaS-based integration and governance capabilities.
* InfoSphere MDM, and complementary entity matching technologies, for management of master data and analytical reference data of all types.
* Optim and Guardium for information lifecycle management, security and privacy.
* Aspera for file transfer where a cloud or hybrid solution is deployed.
* BigInsights provides native Hadoop with IBM enterprise integration, ease-of-use and integrated analytical tools for use within a Data Lake, including as a service.
* DB2 (including BLU acceleration, IDAA, PureData for Analytics/Operational Analytics) for storage and management of relational data within the Lake.
* dashDB, Cloudant and Bluemix database as a service offering for cloud-based data management.

In addition, IBM Global Business Services and IBM Software Services offer:

* Quick-start services to rapidly deploy infrastructure and software
* Proof-of-concept delivery
* “Big Data Stampede” delivery, to rapidly deliver pilot implementations and prove the value of a newly established capability
* Strategy and Analytics Consulting services to assist clients in defining their business and technical strategy & implementation roadmap.

IBM Systems and Technology Group offer specific infrastructure solutions designed and tailored for “Big Data” and Analytics workloads.

* **Data Lakes on 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.

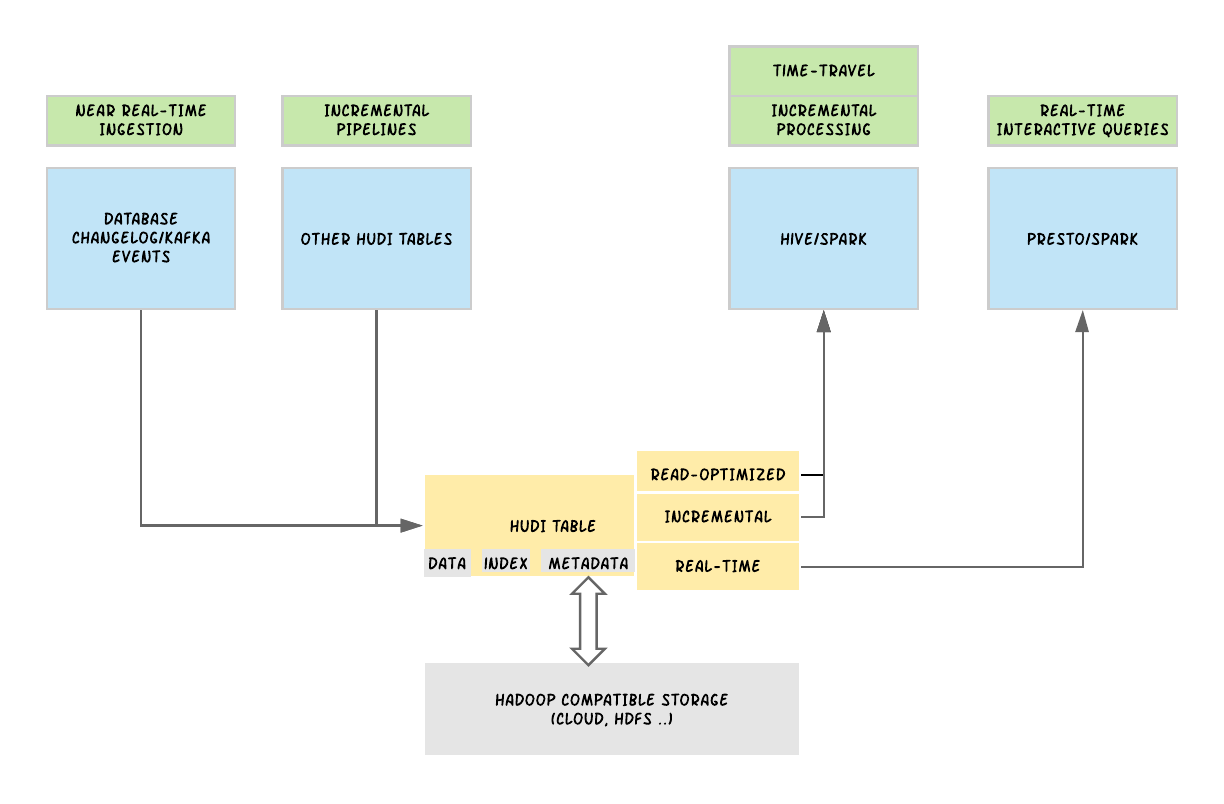


Figure 4: Data Lake Solution Provided by Apache Hudi

At a high level, Hudi is conceptually divided into 3 main components: the raw data that needs to be stored, the data indexes that are used to provide upsert capability, and the metadata used to manage the dataset. At its core, Hudi maintains a timeline of all actions performed on the table at different points in time, referred to as instants in Hudi. This offers instantaneous views of the table, while also efficiently supporting retrieval of data in the order of arrival. Hudi guarantees that the actions performed on the timeline are atomic & consistent based on the instant time, in other words, the time at which the change was made in the database. With this information, Hudi provides different views of the same Hudi table, including a read-optimized view for fast columnar performance, a real-time view for fast data ingestion, and an incremental view to read Hudi tables as a stream of changelogs, depicted in Figure 1, above.

Hudi organizes data tables into a directory structure under a base path on a distributed file system. Tables are broken up into partitions, and within each partition, files are organized into file groups, uniquely identified by a file ID. Each file group contains several file slices, where each slice contains a base file (\*.parquet) produced at a certain commit/compaction instant time, along with set of log files (\*.log.\*) that contain inserts/updates to the base file since the base file was produced. Hudi adopts Multiversion Concurrency Control (MVCC), where compaction action merges logs and base files to produce new file slices and cleaning action gets rid of unused/older file slices to reclaim space on the file system.

Hudi supports two table types: copy-on-write and merge-on-read. The copy-on-write table type stores data using exclusively columnar file formats (e.g., Apache Parquet). Via copy-on-write, updates simply version and rewrite the files by performing a synchronous merge during write.

The merge-on-read table type stores data using a combination of columnar (e.g., Apache parquet) and row based (e.g., Apache Avro) file formats. Updates are logged to delta files and later compacted to produce new versions of columnar files synchronously or asynchronously.

Hudi also supports two query types: snapshot and incremental queries. Snapshot queries are requests that take a “snapshot” of the table as of a given commit or compaction action. When leveraging snapshot queries, the copy-on-write table type exposes only the base/columnar files in the latest file slices and guarantees the same columnar query performance compared to a non-Hudi columnar table. Copy-on-write provides a drop-in replacement for existing Parquet tables, while offering upsert/delete and other features. In the case of merge-on-read tables, the snapshot queries expose near-real time data (order of minutes) by merging the base and delta files of the latest file slice on-the-fly is provided. For copy-on-write tables, incremental queries provide new data written to the table since a given commit or compaction, providing change streams to enable incremental data pipelines.

* **Data Lakes On Huawei**

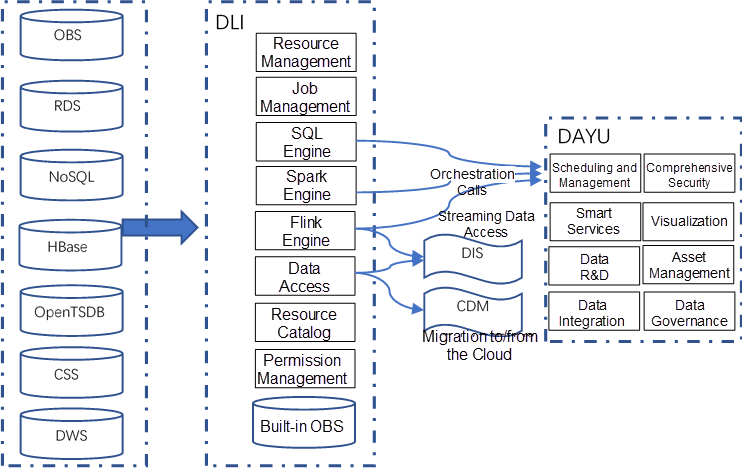


Figure 5: Data Lake Solution Provided by Huawei

The information about Huawei's data lake solution provided here was taken from the Huawei website. Data lake-related products available on the Huawei website include Data Lake Insight (DLI) and an intelligent data lake operations platform called DAYU. DLI can be viewed as a combination of AWS Lake Formation, AWS Glue, Amazon Athena, and Amazon EMR (based on Flink and Spark.) I could not find a diagram of DLI's overall architecture on the Huawei website, so I tried to draw one based on my knowledge. For the architecture, I used a format similar to that of AWS's data lake solution, so we can intuitively compare the two solutions. If you understand DLI differently, you are welcome to share your views.

Huawei's data lake solution provides a full range of functions. DLI implements the core functions of data lake setup, data processing, data management, and data application. DLI provides a complete set of analytics engines, including the SQL-based interactive analytics engine and the stream-batch integrated processing engine based on Spark and Flink. DLI uses its built-in OBS as the core storage engine, which provides capabilities comparable to those of Amazon S3. Huawei's data lake solution provides an ecosystem with a more sophisticated upstream-downstream relationship than that of AWS's data lake solution. In terms of external data sources, Huawei's data lake solution supports almost all data source services provided by Huawei Cloud.

DLI can be connected to Huawei Cloud Data Migration (CDM) and Data Ingestion Service (DIS). With DIS, DLI defines all types of data points, which can be used by Flink jobs as sources or sinks. With CDM, DLI can access the data of Internet Data Centers (IDCs) and third-party cloud services.

Huawei Cloud provides the DAYU platform to better support advanced data lake functions, such as data integration, data development, data governance, and quality management. The DAYU platform is an implementation of Huawei's methodology for data lake governance and operations. DAYU covers the core processes of data lake governance and provides tool support. Huawei's official documentation offer suggestions for building a data governance organization.