**Group Members:**

* Omer Abid (14922)
* Syed Huzaifa Ejaz (14905)

**Component 4**: Data Lake

**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).**

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale.

There are four core components of a data lake

**Data Ingestion**

Really powerful business insights can be drawn from data extracted from a combination of sources. Therefore,this component allows users to extract data from a variety of sources and load it into the data lake. A good data ingestion component should support data from a variety of sources including databases, emails, files, IoT, websites and social media. Furthermore, it should allow for different modes of loading data into the data lake e.g. batch loading, real-time loading and one-time load.Finally, it should support data in any form be it structured, semi-structured on unstructured.

For AWS, the services like Lake Formation and CloudFormation, an organization can establish the necessary infrastructure needed by a data lake. CloudFormation provides templates of resources and their dependencies which can be easily deployed through Lake Formation

For IBM, for ingestion of data IBM provides continuous real time flow of data to the data lake from our data sources. IBM provide these services for data ingestion:

* IBM Infosphere DataStage
* IBM Infosphere Data Replication
* IBM BigIntegrate
* IBM BigReplicate

For Azure, the data is ingested into Azure through Azure Data Factory in batches, or streamed near real-time using Apache Kafka, Event Hub, or IoT Hub

**Data Storage**

This component is responsible for making optimal storage arrangements for data in a data lake. This component should provide highly scalable storage in a data lake because by definition, the data in the data lake grows at a higher rate. Additionally, a low cost storage is also needed for the data. The data storage component should also allow users to store raw data, in-process data as well as curated data ( organized and curated data). It should also support data in any format. Finally, it should also provide fast access to data for our workloads.

For AWS, the storage component is provided by Amazon S3 which is secure and scalable by design. Additionally, it offers 99.999999999% of data durability.

For IBM, it provides IBM spectrum scale service, which is a ‘scale-out’ storage software.

For Azure, a data lake account can store trillions of files where a single file can be petabyte in size. The storage functionality is built on Azure Blob Storage which also provides 16 9s of data durability. Azure data lake is part of the Hadoop ecosystem using HDFS and YARN.

**Data Management/Governance**

To extract proper insights from the data and explore data, the data need to be organized and integrated. This is where the data governance component comes in. It allows enforcement of policies so that the data lake can be used the way it was meant to be used to drive the business forward. It also allows users to deal with any errors in the data ingestion or storage components. It enables the users to do all these by collecting metadata about what goes in the data lake and making relationships between different data sets. This protects the data lake from becoming a ‘data swamp’.

Data governance component is also responsible for the security of data when ingesting, storing or processing it. It provides security by making sure only trusted users and services can access the data inside the data lake. This means providing authentication, authorization and accountability and data protection services.

For AWS, the Amazon DynamoDB provides the necessary databases to maintain the metadata of the data lake while AWS Lambda allows you to run necessary microservices for better data governance. AWS Glue also provides indexing and cataloging services for our data lake.

For IBM, the data governance is advertised to be a very important component which is based on IBM Information Governance Catalog and etc.

For Azure, data governance is part of Azure Data Lake Storage which allows for data management and data security.

**Data Discovery, Exploration and Analytics**

This is the crucial component of the data lake because the reason we create data lakes is to derive business insights that can help in research and development and quality assurance of our products. The services that this component provides is feature extraction ( reducing the dimensionality of our data to the most important features) and training machine learning models on your data.

For IBM we have two important services IBM Db2® Big SQL and IBM Watson Studio. The former helps with advanced data queries and the latter builds and trains AI and ML models on data from data lake.

For AWS, AWS Glue provides the necessary services to prepare data for analysis. It also maintains a Data Catalog of the data lake so that users can find appropriate datasets for data exploration. We can also create and run ETL jobs through AWS Glue. Amazon Athena also allows users to query data and get results quickly. AWS AI services such as Amazon Comprehend, Amazon Forecast, Amazon Personalize, and Amazon Rekognition help discover insights from your unstructured datasets, get accurate forecasts, create recommendation machines, and analyze images and videos stored in S3. You can also deploy Amazon Sagemaker to build, train, and deploy ML models quickly with your datasets stored in S3.

For Azure data lake, Azure HDInsight provides analytics services in the cloud for the users. Data visualization is handled by Microsoft Power BI. Azure Databricks is yet another service that allows users to run SQL queries on the data lake.

**References:**

[**https://www.yash.com/blog/data-lake-architectural-components/**](https://www.yash.com/blog/data-lake-architectural-components/)

[**https://www.youtube.com/watch?v=LxcH6z8TFpI**](https://www.youtube.com/watch?v=LxcH6z8TFpI)

[**https://www.ibm.com/analytics/use-cases/governing-data-lake**](https://www.ibm.com/analytics/use-cases/governing-data-lake)

[**https://www.ibm.com/analytics/data-lake#2608990**](https://www.ibm.com/analytics/data-lake#2608990)

[**https://aws.amazon.com/cloudformation/**](https://aws.amazon.com/cloudformation/)

[**https://en.wikipedia.org/wiki/Azure\_Data\_Lake**](https://en.wikipedia.org/wiki/Azure_Data_Lake)

[**https://azure.microsoft.com/en-us/services/storage/data-lake-storage/**](https://azure.microsoft.com/en-us/services/storage/data-lake-storage/)