**Name -Monika Gariya**

**Email-** [**monikagariya2023@gmail.com**](mailto:monikagariya2023@gmail.com)

**Data Engineering Batch 1**

**Date – 15-02-2024**

**Microsoft Azure Assessment 4**

**Topic -** Databases vs. Data Warehouses vs. Data Lakes, Create a storage account and connect with the Microsoft azure storage explorer

**What is a database?**

A database is a collection of data or information. Databases are typically accessed electronically and are used to support Online Transaction Processing (OLTP). Database Management Systems (DBMS) store data in the database and enable users and applications to interact with the data.

### Database characteristics

* Security features to ensure the data can only be accessed by authorized users.
* ACID (Atomicity, Consistency, Isolation, Durability) transactions to ensure data integrity.
* Query languages and APIs to easily interact with the data in the database.
* Indexes to optimize query performance.
* Full-text search.
* Optimizations for mobile devices.
* Flexible deployment topologies to isolate workloads (e.g., analytics workloads) to a specific set of resources.
* On-premises, private cloud, public cloud, hybrid cloud, and/or multi-cloud hosting options.

## What is a data warehouse?

A data warehouse is a system that stores highly structured information from various sources. Data warehouses typically store current and historical data from one or more systems. The goal of using a data warehouse is to combine disparate data sources in order to analyze the data, look for insights, and create business intelligence (BI) in the form of reports and dashboards.

### **Data warehouse characteristics**

Data warehouses store large amounts of current and historical data from various sources. They contain a range of data, from raw ingested data to highly curated, cleansed, filtered, and aggregated data.

Extract, transform, load (ETL) processes move data from its original source to the data warehouse. The ETL processes move data on a regular schedule (for example, hourly or daily), so data in the data warehouse may not reflect the most up-to-date state of the systems.

Data warehouses typically have a pre-defined and fixed relational schema. Therefore, they work well with structured data. Some data warehouses also support semi-structured data.

## What is a data lake?

A data lake is a repository of data from disparate sources that is stored in its original, raw format. Like data warehouses, data lakes store large amounts of current and historical data. What sets data lakes apart is their ability to store data in a variety of formats including JSON, BSON, CSV, TSV, Avro, ORC, and Parquet.

### **Data lake characteristics**

Data lakes store large amounts of structured, semi-structured, and unstructured data. They can contain everything from relational data to JSON documents to PDFs to audio files.

Data does not need to be transformed in order to be added to the data lake, which means data can be added (or “ingested”) incredibly efficiently without upfront planning.

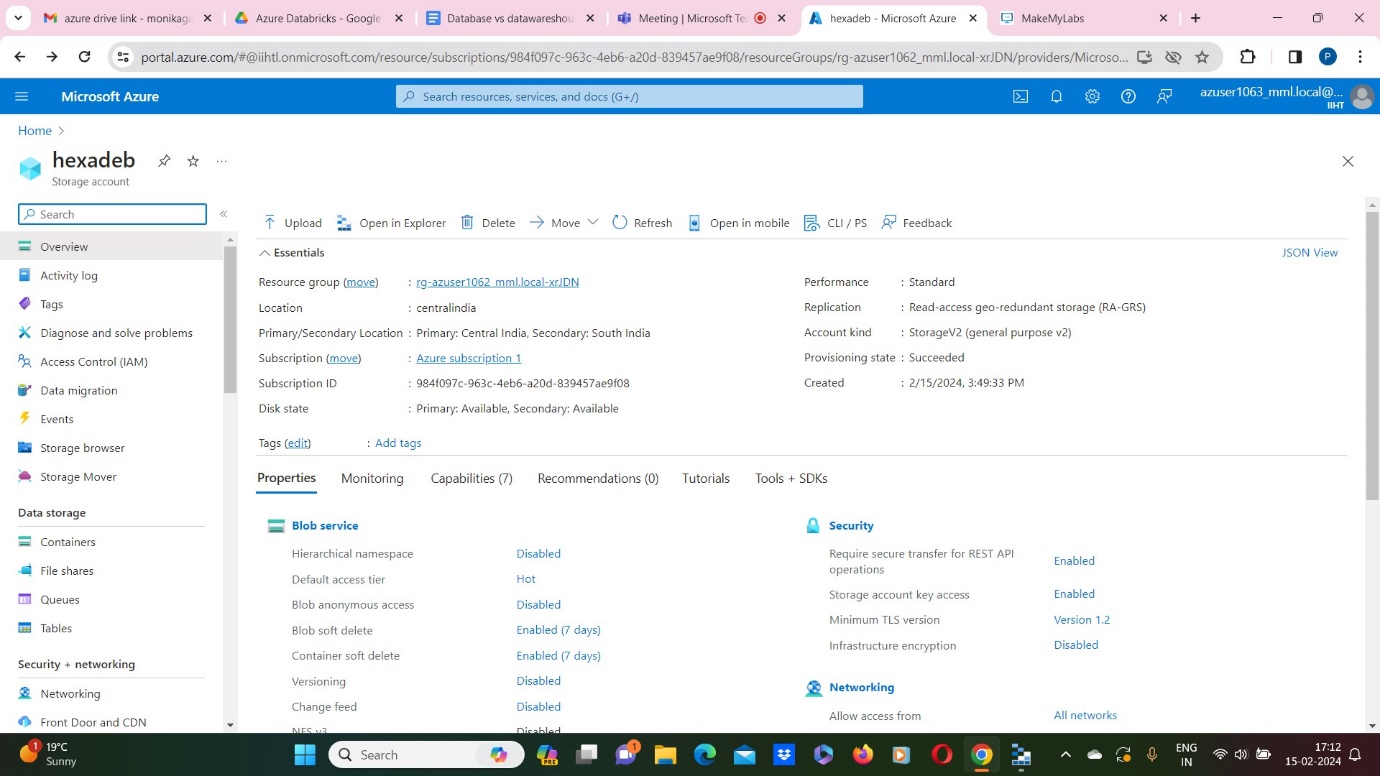
The primary users of a data lake can vary based on the structure of the data. Business analysts will be able to gain insights when the data is more structured. When the data is more unstructured, data analysis will likely require the expertise of developers, data scientists, or data engineers.

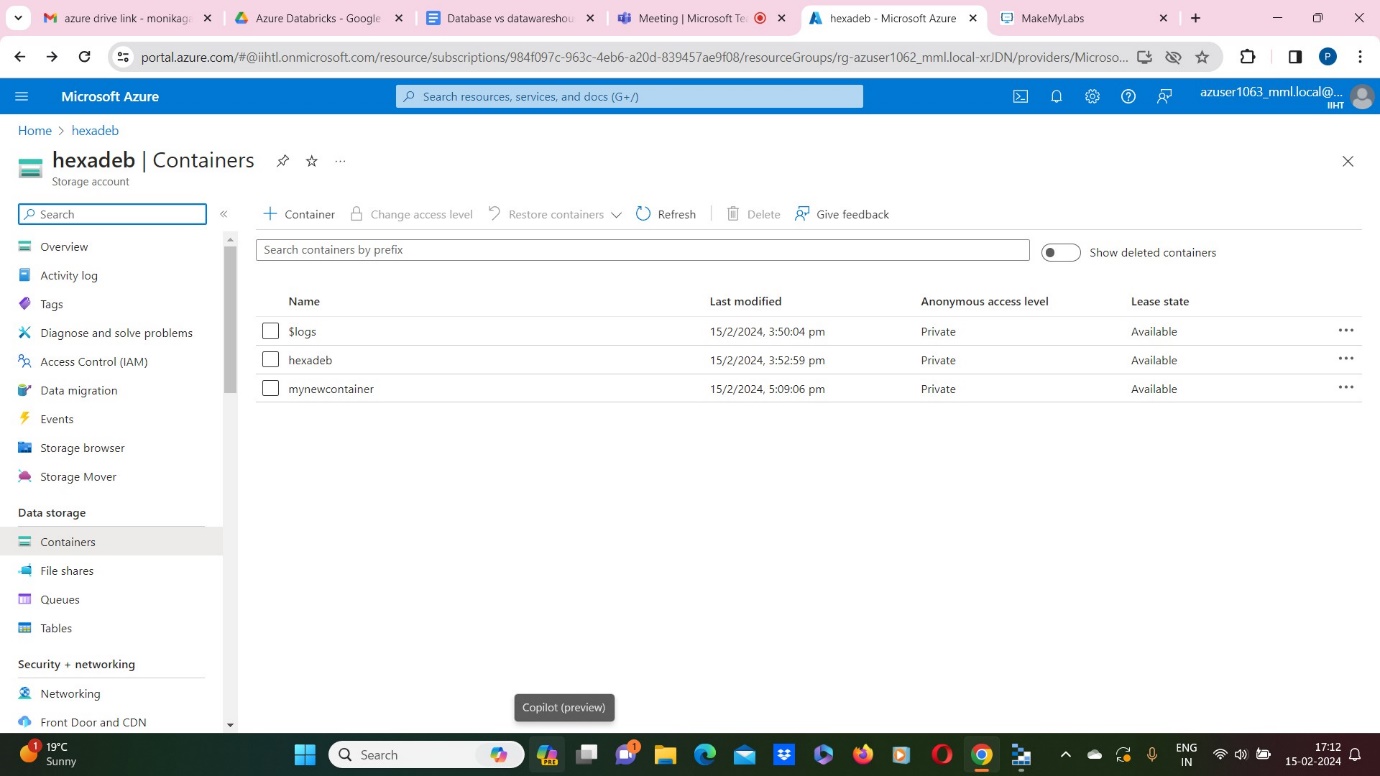
The flexible nature of data lakes enables business analysts and data scientists to look for unexpected patterns and insights. The raw nature of the data combined with its volume allows users to solve problems they may not have been aware of when they initially configured the data lake.

**DIFFERENCE BETWEEN DATABASE DATALAKE AND DATAWAREHOUSE**

|  | Database | Data Lake | Data Warehouse |
| --- | --- | --- | --- |
| Workloads | Operational and transactional | Analytical | Analytical |
| Data Type | Structured or semi-structured | Structured, semi-structured, and/or unstructured | Structured and/or semi-structured |
| Schema Flexibility | Rigid or flexible schema depending on database type | No schema definition required for ingest (schema on read) | Pre-defined and fixed schema definition for ingest (schema on write and read) |
| Data Freshness | Real time | May not be up-to-date based on frequency of ETL processes | May not be up-to-date based on frequency of ETL processes |
| Users | Application developers | Business analysts, application developers, and data scientists | Business analysts and data scientists |
| Pros | Fast queries for storing and updating data | Easy data storage simplifies ingesting raw data A schema is applied afterwards to make working with the data easy for business analysts Separate storage and compute | The fixed schema makes working with the data easy for business analysts |
| Cons | May have limited analytics capabilities | Requires effort to organize and prepare data for use | Difficult to design and evolve schema Scaling compute may require unnecessary scaling of storage, because they are tightly coupled |

2) Create a storage account and connect with the Microsoft azure storage explorer.





Now,Connect with the Microsoft azure storage explorer.

