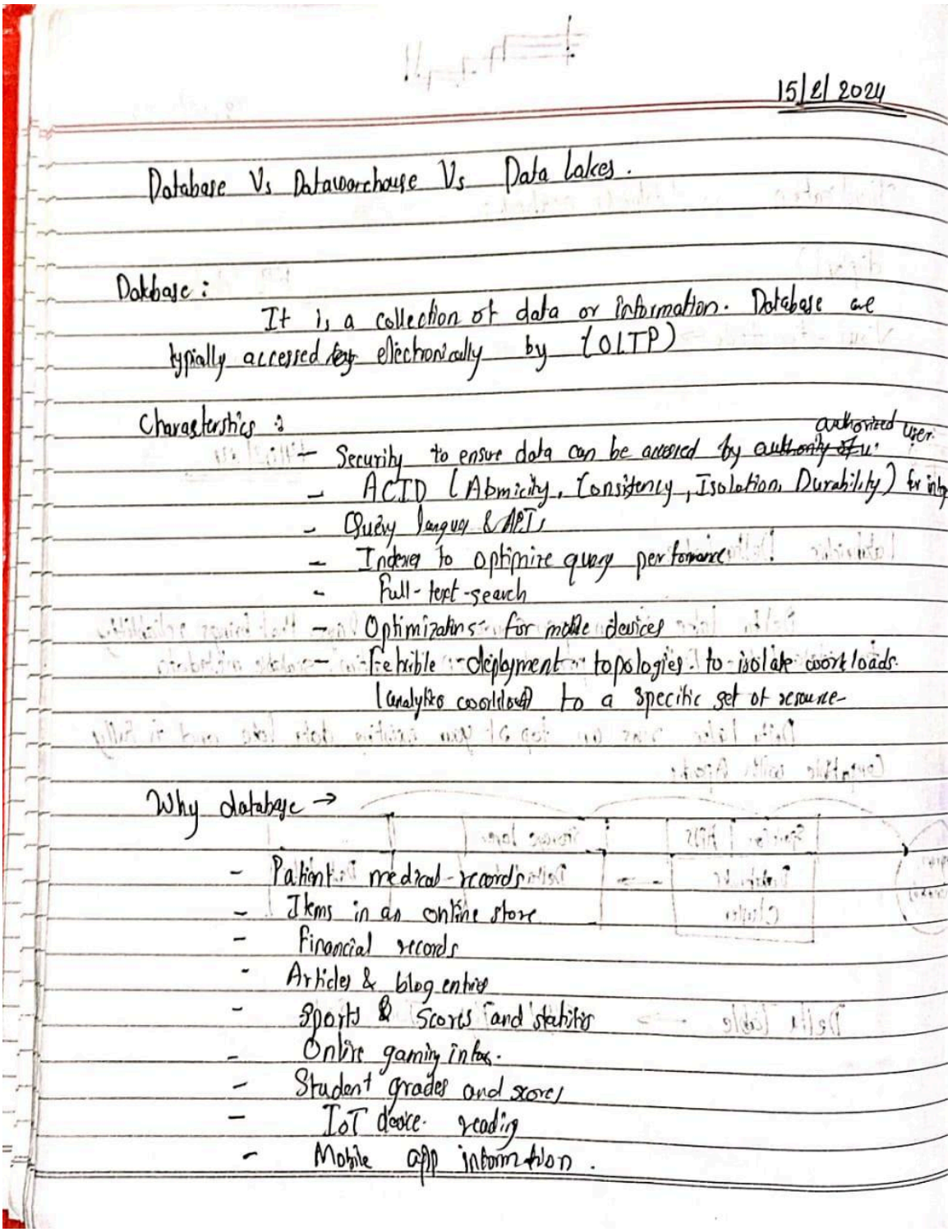
Name : Parth Nandedkar
Date : 15 Feb 2024
Topics : Azure Storage and Data Lake, ETL
Batch : Data Engineering Batch-1

Hand written notes :



15/2/2024

Database Vs Datawarehouse Vs Data lakes.

Database :
It is a collection of data or information. Database are typically accessed electronically by (OLTP)

Characteristics :
- Security to ensure data can be accessed by authorized user.
- ACID (Atomicity, Consistency, Isolation, Durability) for integrity
- Query language & API's
- Indexes to optimize query performance
- Full-text-search
- Optimizations for mobile devices
- Flexible deployment topologies to isolate workloads (analytics workload) to a specific set of resource

Why database →
- Patient medical records
- Items in an online store
- Financial records
- Articles & blog entries
- Sports & Scores and statistics
- Online gaming info.
- Student grades and scores
- IoT device reading
- Mobile app information.

Business Intelligence.

OLAP → data warehouse + data lakes:

Both data warehouses & data lakes are meant to support Online Analytical processing. (OLAP) ; OLAP systems are typically used to collect data from a of sources. The The data is further used to power a range of analytical use cases ranging from business intelligence reporting to training

RDBs → Oracle, MySQL, Microsoft SQL Server, PostgreSQL

~~What is data warehouse?~~

- Document dbs → MongoDB & CouchDB
- Key-value → Redis & DynamoDB
- Wide-column → Cassandra & HBase
- Graph → Neo4j & Amazon Neptune.

Data warehouse

- A data warehouse is a system stores highly structured information & from various sources. Data warehouses typically store current & historical data from one or more systems. The goal of using a data warehouse is to combine disparate data data sources in order to analyze data.

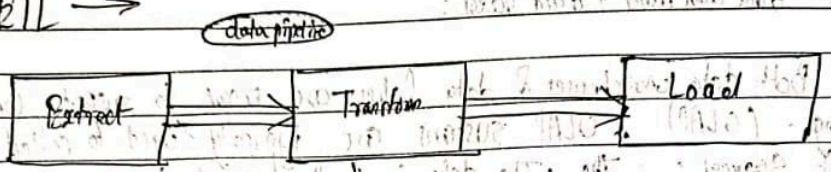
Data warehouse Characteristics:

Snowflake
teradata

Pipeline → one after another activity.

etl vs elt

ETL →



Raw Data.

highly cleansed & filtered & aggregated data.

Data Loging in DW, DBL, DATA LAKE,
- Database
- Databricks cluster

Data Lake : →

Datalake is a repository of data from disparate sources that is stored in its original, raw format

Azure storage :

Azure Storage is Microsoft's cloud storage solution that provides scalable, secure, and highly available storage for data, applications, and services. It offers various storage services catering to different data storage and access needs. Here's an overview of the key components and features of Azure Storage:

Blob Storage:

Blob Storage is designed to store large amounts of unstructured data, such as text or binary data (e.g., images, videos, documents).
It offers hot and cool access tiers to optimize storage costs based on data access patterns.
Blob Storage supports secure data transfer and access control using shared access signatures (SAS) and Azure Active Directory (Azure AD) integration.
It provides lifecycle management policies for automated data retention and deletion based on predefined rules.

File Storage:

Azure File Storage offers fully managed file shares in the cloud, accessible via the Server Message Block (SMB) protocol.
It allows organizations to migrate file-based applications to the cloud without modifying their code.
Azure File Storage supports both standard and premium performance tiers, with options for encryption and access control.

Queue Storage:

Azure Queue Storage provides a message queuing service for building scalable and distributed applications.
It enables asynchronous communication between application components, helping to decouple and scale workloads.
Queue Storage supports message expiration, visibility timeout, and message peeking for reliable message processing.

Table Storage:

Azure Table Storage is a NoSQL key-value store for semi-structured data.
It offers schema-less storage, allowing developers to store and retrieve flexible datasets.
Table Storage is well-suited for applications that require massive scale and low-latency access to data.

Disk Storage:

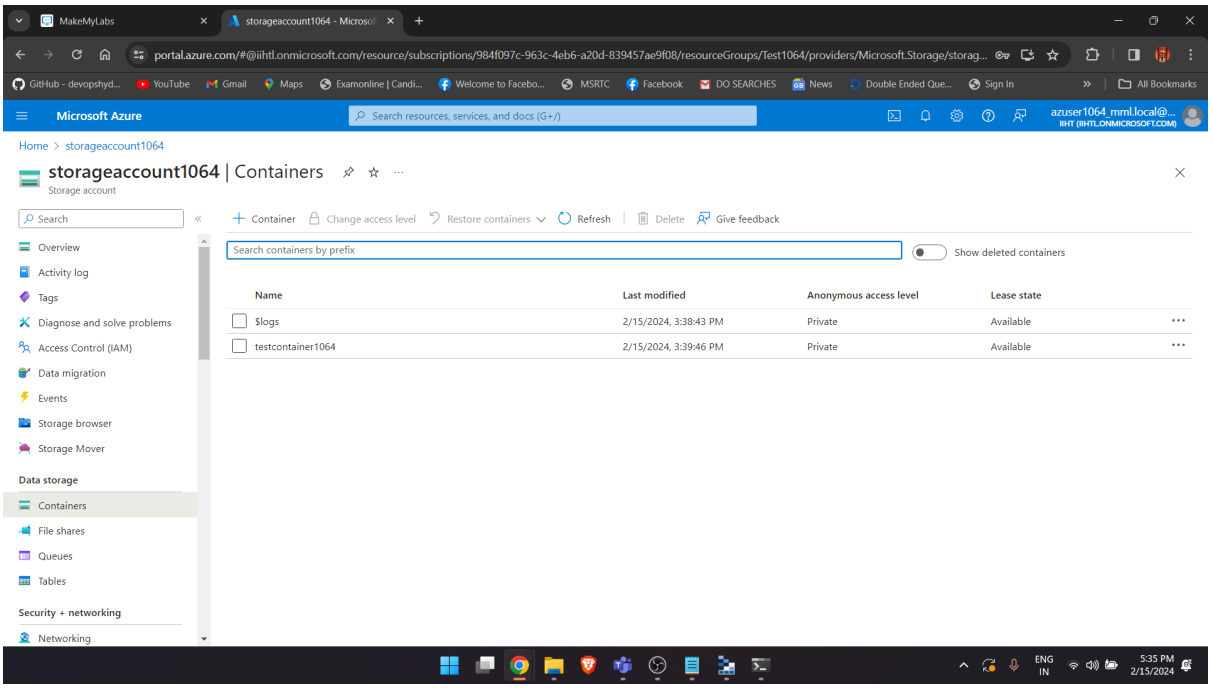
Azure Disk Storage provides persistent, high-performance block storage for Azure Virtual Machines (VMs) and other compute resources.
It offers both managed and unmanaged disk options with support for different disk types (e.g., Standard HDD, Standard SSD, Premium SSD).
Disk Storage supports features such as disk snapshots, disk encryption, and disk resizing.

Data Lake Storage:

Azure Data Lake Storage is optimized for big data analytics and large-scale data processing.
It provides a hierarchical namespace and integrated security features for managing and securing data at scale.
Data Lake Storage supports integration with Azure Synapse Analytics, Azure Databricks, and other Azure services for advanced analytics and AI workloads.

Overall, Azure Storage offers a comprehensive set of storage services to meet various data storage and access requirements, from small-scale file sharing to large-scale big data analytics. It provides scalability, reliability, security, and performance for storing and managing data in the cloud.

Screenshots of creating and adding files :



Accessing cloud storage through local :