Aws Machine Learning Engineer

① Data Ingestion and Storage

① Types of data

* **Structured data**
  - Data that is organized in defined manner
  - found in relational database, consistant structure
  - ex : database tables, Excel spread sheets

* **Unstructured data**
  - Data without predefined structure or schema
  - need preprocessing for query data
  - ex : video/audio files, images, emails or word docs text without a fixed format

* **Semi structured data**
  - not organized as structed data but has some level of structure in form of tags or hierarchies
  - more flexible than structured and less chaotic than unstructured
  - ex : XML or Json, Log files with vary format, email head

② Properties of Data

* **Volume**
  - Refers to amount/size of data at a given time
  - It helps us decide on storing, processing and analysis
  - ex : social media processing daily posts, images or videos

* **Velocity**
  - Refers to speed at which data is generated, processed & collected
  - It helps us decide on real-time/near real time processing capabilities as rapid ingestion and processing can be critical for certain applications
  - example : high frequency trading system where milli-seconds can make a difference in decision making

* **Variety**
  - refers to type, sources and structure of data
  - example : hospital collecting data from machine, forms and devices

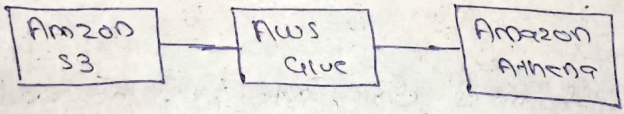
3) Data warehouse, lakes, warehouses

* Data warehouse

- Centralized repository optimized for analysis where data from different sources are stored in structured format
- Characteristics
  - ⇒ designed for complex query & analysis
  - ⇒ Data is cleaned, transformed and loaded
  - ⇒ Stored in snowflake or star schema
  - ⇒ optimized for heavy read operations
- example : Redshift [ETL], more cost

* Data Lake

- Storage repository that stores large amount of data in its native format including structured, unstructed & semi-structured
- characteristics
  - ⇒ large amount of data without predefined schema
  - ⇒ No pre-processing [loaded as it is]
  - ⇒ Supports batch, real-time & stream processing
  - ⇒ queried for data transformation or exploration purpose
- example : Amazon Simple Storage solution (S3) used as data lake [ETL], less cost



* Data Lakehouse

- A hybrid data architecture that combines best features of data lake and warehouse aiming to provide performance, reliability and capabilities of data warehouse while maintaining the flexibility, scale and low cost of data lake
- Characteristics
  - ⇒ Supports structured & unstructured data
  - ⇒ Support schema-on-write & schema-on-read
  - ⇒ detailed analytics & ML tasks
  - ⇒ built on top of cloud or distributed architecture
- example : S3 with redshift spectrum

like big blob of data in S3 but using S3 spectrum to query that data using Redshift and underlying storage is not structured

20 min

④ Data mesh

- It emphasizes decentralized ownership of data, treating data as a product and federated governance
- Individual team owns data

⑤ ETL

- ETL means extract, transform and load. It is a process used to move data from data source to data warehouse

* Extract

- retrive data from source systems, which can be database, CRMs, flat files or other data repo
- Ensure data integrity during extraction phase
- Can be done in real time or in batches

* Transform

- convert data into format suitable for target data warehouse
- data cleaning, enrichment, format changes, computation, encode-decode and handle missing values

* Load

- move transformed data into targeted warehouse
- can be done in batches or streaming manner
- ensure data integrity during load

Managing ETL pipelines

- ETL should be automated in a reliable way
- AWS Glue ⟶ helps in automation
- Orchestration service
  - Event bridge
  - AWS Step functions
  - Lambda
  - Glue workflows
  - Amazon managed workflows for Apache airflow.

(6) Data sources

* JDBC
  - java database connecti
  - Platform- independent
  - Language dependent
* ODBC
  - Open database connect
  - Platform dependent
  - language independent
* Raw logs
* APIs
* Streams

Data Formats

CSV

- comma seperated values
- text based format that represent tabular data
- Small - medium size
- interchange btw systems
- import/export from database or spreadsheet.
- Excel, Python with Pandas, SQL based database

JSON

- Javascript Object notation
- text based format with key value pair
- structured or semi-structured
- interchange btw web server and web client
- web browsers, restful apis mongodb
- used for flexible schema or nested data structure

AVRO

- Binary format that store both data and its schema, allowing it to be processed later with different systems without needing the original system context
- used with big data & real time processing system
- when schema (data structure) can be changed
- Efficient Serialization
- Apache kafka, apache spark, Hadoop

Parquet

- Columnar storage format for optimized analytics
- efficient compression & encoding schemes
- analysing large data sets with analytics engines
- use cases where reading specific colums instead of entire records is beneficial
- storing data on distributed systems where I/O operations and storage needs optimization
- Hadoop, Apache spark, Hive, Amazon Redshift Spectrum.