Fundamentals of Data Engineering

Assignment - 1

Dim.

23200304-

1A) Illustrate the main objective of data Engineering in a data-driven organization

12

~~Ans) In a data driven Organization~~

Ans: In a data driven Organization, the primary objective of data Engineering is to develop, implement and maintain systems and processes that transform raw data into high-quality, consistent and usable information. This serves as the foundation for all analytics, reporting and machine learning activities

- Transforming raw ~~this~~ into usable information:

→ Data engineers design pipelines to collect data from multiple sources.

→ The data engineer cleans and normalizes the messy data and transforms it into a reliable format

- supporting downstream usecases:

→ The refined data is served to analysts, BI tools, and ML models.

→ without high quality input, analytics and AI outcomes will be inaccurate or misleading

- Integrating multiple disciplines:

→ Data Engineering combines Security, data management, data architecture, Orchestration and software engineering.

ensures not just correctness of data but also scalability

performance and security.

- lifecycle approach:

→ The lifecycle is generation → storage → ingestion → transformation → serving.

→ The goal is to keep this flow smooth, automated and optimized for real time.

- value of organization:

→ Enables real time insights from large datasets

→ Allows decision makers to trust their data.

B) Summarize the role of a data engineer has evolved from a traditional ETL developer to supporting modern big data pipelines.

Ans:- The role of a data engineer has shifted significantly over time.

Initially in the ETL developer era (1980-2005), the focus was on moving structured data from operational systems into systeme centralized warehouse for batch processing and BI reporting, using relational databases and ETL tools.

with the rise of the big data era data engineers began handling massive fast and varied datasets from web, IOT d social media They adpoited distributed processing tools ke hadoop, spark, kafka moving beyond batch to include eal time streaming.

oday, the modurn data engineer supports the full data ngineering life cycle, working with cloud platforms, rchestration tools, and DataOps practices.

7) Outline the two features of modurn data platforms that support scalability and flexibility.

ns) Two featurs of modurn data platforms that support scalability and flexibility:

① cloud computing - modurn data platforms often leverage cloud services such as AWS, Azure and google cloud. These platforms provide on demand infrastructure that can scale up or down based on workload requirements, They also offer a variety of storage, processing and pipeline automation tools, enabling organizations to adapt quickly to changing businesses and data needs while maintaining cost efficiency.

② Distributed system architecture:

many modurn platforms are built on distributed systems. which allow massive databases to be processed across multiple nodes simultaneously - This architecture ensures fault

8) Compare skillsets of early data engineers and modern data Engineers in industry.

Ans:- a) Early data Engineers:-

- primary role - focused mainly on ETL proces to move structured data from operational systems into centralized data warehouse for business intelligence reporting.

- skills -
  → strong SQL for querying relational databases.
  → proficiency with traditional ETL tools.
  → Data modeling for structured relational database.
  → basic scripting for automation.

b) modern data engineers:

- primary role - responsible for entire data engineering life cycle, handling structured and unstructured data at scale for analytics, BI and machine learning.

- skills -
  → programming in python, SQL and sometimes scala or Java.
  → big technologies for large scale and real time data processing
  → cloud computing skills
  → No SQL database for flexible storage
  → Distributed systems archicture and orchestration tools

) Contrast the concept of data engineering by Exp
role in enabling data driven decision making in an Organization.

ans) Data engineering is the practice of desigining, building and maintaining system that collect, store, process and deliver data in a usable form. It converts the full data lifecycle - from generation and ingestion, through transformation to serving. ensuring the raw, messy data becomes high quality, consistent and accessible for various use cases

Role of decision making:

- Foundation for analytics - data engineers create pipelines and platforms that supply data analysts, BI tools and data scients with clean, reliable data. without this foundation, insights would be inaccurate or incomplete
- Real time insights - by enabling both batch and streaming pipelines they ensure decision making can act quickly based on current information.
- scalability and flexibility - data engineering designs systems that adapt to growing data volumes and new data.

Integration across sources - data engineers consolidate information from multiple internal and External systems, providing a unified view for strategic decisions.

Q) Illustrate the stages of the data engineering life cycle using a practical example such as a analytics or healthcare systems.

Ans) Stages of data engineering Lifecycle - eg sales analytics system

1) Generation - sales data is created from multiple sources:

→ point of sale systems in retail source.

→ E commerce transactions from the company's website.

→ customer interactions from a mobile shopping app.

2) Storage - The raw sales data is storage in secure systems.

→ Transation logs saved in relational databases

→ webapp interactions from a mobil shopping app.

3) Ingestion - data pipelines pull in data from different sources:

→ Batch ingestion: daily upload of store sales records into the central data platform.

→ streaming ingestion: real-time order data from the e-commerce site using Apache kafka.

5) Serving - processed data is made available for use:

→ sales dashboards in BI tools like power BI or tableau slow daily revenue trends.

→ predictive analytics models forecast future sale demand

→ Management receives automated reports for decision making.