Basics of Data Engineering – Part 1

What is a Data Pipeline?

A data pipeline is like a well-organized system of actions and tools that automatically gather, change, and transport data from different places to one central spot, usually a place where data can be analyzed or used for making decisions. Usually this involves moving a data from the Data generator to a data consolidator.

Data pipelines are crucial for data management, ensuring data is accessible, accurate, and timely.

Types of Data Pipelines:

- ETL (Extract, Transform, Load): Ideal for structured data, it extracts, transforms based on requirements of the target system and loads data in batches.
- ELT (Extract, Load, Transform): This is ideal of for semi-structured or unstructured data, it extracts and loads data first, then transforms it.

Building a Data Pipeline

- Define Data Sources application and fields
- Choose appropriate Tools and Technologies
- Design Data Flow and methodoloy
- Implementation and Integration, and testing
- Error Handling and Monitoring

Linkedin:@ganeshdg

Basics of Data Engineering – Part 1

Key Considerations in creating Pipelines

- Data Quality: Understand the quality of what is right data and what is not.
- Scalability: Ensuring the solution can handle increasing data volumes.
- Security: Design pipelines based on sensitivity of data throughout the pipeline.
- Monitoring and Logging: Based on requirements of real-time monitoring and detailed logging.

Few popular tools used in Pipeline creation

- Apache Airflow – Creates workflow
- Talend
- Apache Kafka- Streaming platform
- AWS Glue- Amazons ERL solution
- Google Cloud Dataflow- Google cloud offering
- Microsoft Azure Data Factory
- Informatica PowerCenter

Few best practices in creating Data pipelines

- Implement Data Quality rules during creation
- Check on Modular and Reusability consideration
- Proper Documentation explaining features
- Implement modes of Backup and Recovery
- Monitor for Performance Tuning

Linkedin:@ganeshdg