Data Engineer Interview Preparation

Table of Contents

1. Introduction to Data Engineering Interviews
   - Overview of Data Engineering Roles
   - Key Skills Assessed in Interviews
2. General Data Engineering Concepts
   - Data Warehousing
   - ETL/ELT Processes
   - Data Modeling (Star vs. Snowflake Schema)
   - Data Pipelines and Workflow Orchestration
3. Programming and Coding
   - SQL Interview Questions and Answers
   - Python Interview Questions and Answers
   - Other Languages: Java, Scala
4. Big Data Ecosystem
   - Hadoop and HDFS
   - Apache Spark and PySpark
   - Hive and Impala
5. Cloud Technologies
   - AWS: S3, Redshift, EMR, Glue
   - Azure: Data Factory, Synapse Analytics
   - Google Cloud Platform (GCP): BigQuery
6. Database Concepts
   - Relational Databases: Postgres, MySQL
   - NoSQL Databases: MongoDB, Cassandra
7. Data Engineering Case Studies and Scenarios
   - Building an ETL Pipeline
   - Optimizing Query Performance
   - Designing a Scalable Data Pipeline
8. Distributed Systems and Data Processing
   - Kafka and Real-Time Data Streaming

  - Batch vs. Stream Processing

9. Data Governance and Security

  - Data Privacy (GDPR, CCPA)

  - Data Integrity and Auditing

10. Soft Skills and Behavioral Questions

  - Communication and Teamwork

  - Problem-Solving Frameworks

11. Additional Resources

  - Online Courses and Certifications

  - Books and Blogs

Introduction to Data Engineering Interviews

- Data engineers are responsible for designing, building, and maintaining the systems that store and process large-s
- Key skills include programming, data modeling, ETL development, and familiarity with big data tools.

General Data Engineering Concepts

- Data Warehousing: Centralized repository for structured data.
- ETL/ELT Processes: Extract, Transform, Load vs. Extract, Load, Transform.
- Data Modeling: Difference between Star and Snowflake schema.
- Data Pipelines: Automation of data flow between systems.

Programming and Coding

- SQL: Writing complex queries for data extraction and transformation.
- Python: Implementing ETL scripts and working with libraries like pandas, numpy.
- Java/Scala: Frequently used for big data processing in Hadoop/Spark ecosystems.