ETL Pipeline Guide + Top 20 Interview Questions

📌 **Prepared By:**

Ankita Gulati
linkedin.com/in/ankita-gulati-de

Pooja Jain
linkedin.com/in/pooja-jain-898253106

---

1. What is ETL?

ETL stands for **Extract, Transform, Load** — a foundational process in modern data engineering:

- **Extract:** Pull data from various source systems
- **Transform:** Clean, enrich, and structure the data
- **Load:** Store the processed data into data warehouses or lakes

It enables teams to make data analysis-ready for reporting, dashboards, and ML.

---

2. ETL vs ELT — What's the Difference?

Feature	ETL (Extract → Transform → Load)	ELT (Extract → Load → Transform)
Transformation	Before loading	After loading
Use Case	On-prem or traditional systems	Cloud-native platforms
Tools	Informatica, SSIS, Talend	dbt, BigQuery, Snowflake

---

3. Real-World ETL Pipeline — End-to-End

◇ Step 1: Data Sources

Where raw data originates:

- APIs (e.g., Stripe, GA4, Salesforce)
- SQL/NoSQL DBs (MySQL, MongoDB, PostgreSQL)
- File formats (CSV, Excel, JSON, Parquet)

- Streaming sources (Kafka, Kinesis)
- Webhooks, third-party integrations

◇ Step 2: Ingestion Layer

Mechanism to pull data in real-time or batches:

- **Tools:** Python scripts, Apache NiFi, Kafka Connect, Fivetran
- **Key Practices:**
    - Handle retries and timeouts
    - Avoid duplication
    - Deal with schema drift
    - Maintain idempotency

◇ Step 3: Raw Landing Zone

Storage of unprocessed/raw data:

- **Tools:** AWS S3, Azure Data Lake Storage (ADLS), Google Cloud Storage (GCS)
- **Best Practices:**
    - Store data immutably for reproducibility
    - Organize with partitioning by date/source
    - Maintain metadata/catalog

◇ Step 4: Data Transformation Layer

Clean, enrich, standardize and join raw data:

- **Tools:** dbt, Apache Spark, SQL, Python (Pandas), Databricks Notebooks
- **Transformations:**
    - Filtering nulls, handling missing values
    - Joining datasets, renaming columns
    - Generating derived columns (e.g., revenue)
    - Implementing Slowly Changing Dimensions (SCD Type 1 and 2)

◇ Step 5: Orchestration & Workflow Management

Schedule, monitor, and manage dependencies:

- **Tools:** Apache Airflow, Prefect, Dagster
- **Capabilities:**
    - Define DAGs and task dependencies
    - Configure retries, alerting, SLAs
    - Enable backfilling and parameterization

◇ Step 6: Load into Warehouse / Lakehouse

Push processed data to final storage:

- **Warehouses:** Snowflake, BigQuery, Redshift
- **Lakehouses:** Delta Lake (Databricks)
- **Techniques:**
    - Incremental vs full refresh
    - Clustering and partitioning
    - Materialized views and indexes

◇ Step 7: Consumption Layer

Make data accessible for stakeholders:

- **Tools:** Power BI, Looker, Tableau, Superset
- **Other Outputs:**
    - Machine Learning feature stores
    - APIs for operational systems
    - Reverse ETL to Salesforce/CRM

---

4. Key Considerations for Robust ETL

- **Data Quality:** Null checks, duplicates, referential integrity
- **Monitoring:** Real-time dashboards, alerting, pipeline health metrics
- **Scalability:** Support TB-scale processing with Spark, parallelization
- **Governance:** Lineage (e.g., using OpenLineage), cataloging, RBAC
- **Documentation:** Auto-generating lineage docs with dbt docs or Airflow UI
- **Recovery:** Backfill strategies, dead-letter queues, rerun options

20 Real ETL Interview Questions

Basics

1. What is the difference between ETL and ELT?
2. What are common data quality checks in ETL pipelines?
3. How do you handle schema evolution in ETL processes?
4. What's the role of orchestration tools like Airflow?
5. What's the difference between batch and streaming ETL?

Intermediate

6. Design an end-to-end ETL pipeline for e-commerce orders data.
7. How do you manage dependencies between ETL jobs?
8. Explain your approach for incremental vs full data loads.
9. How do you debug a failed ETL job in production?
10. What are the common causes for pipeline failures?

Advanced

11. How do you ensure data consistency across retries or multiple runs?
12. How do you monitor and optimize performance in Spark ETL jobs?
13. Explain how you've implemented SCD Type 2 in a real project.
14. What is backfilling and how do you approach it in ETL?
15. How do you handle late-arriving or out-of-order event data?

Scenario-Based

16. A job that updates sales data fails halfway. What steps would you take?
17. A stakeholder reports incorrect dashboard metrics — how do you debug?
18. You're asked to reduce a legacy ETL job's runtime by 80%. Your strategy?
19. A marketing team asks for near real-time campaign data — how would you deliver that?
20. Share a challenging ETL failure you handled and lessons learned.

📌 **Prepared By:**

Ankita Gulati
linkedin.com/in/ankita-gulati-de

Pooja Jain
linkedin.com/in/pooja-jain-898253106