**Top 3 Python, SQL, and Databricks Projects for Data Engineering Jobs**

---

## 1. ETL Pipeline for E-commerce Data Analytics

**Objective:** Build a fully automated ETL (Extract, Transform, Load) pipeline for e-commerce transaction and user data to provide business insights.

**Technologies/Skills:** - Python: Pandas, PySpark, Airflow - SQL: Data cleaning, joins, aggregations - Databricks: PySpark notebooks, Delta Lake - Others: AWS S3, Tableau/Power BI

**Steps:** 1. Extract raw data from APIs, CSV files, or databases. 2. Transform data: - Clean missing/inconsistent data - Join datasets (user, order, product) - Aggregate sales by product, category, and time 3. Load cleaned data into Delta Lake in Databricks. 4. Analytics & Dashboard: - Top-selling products - Customer retention rates - Sales trends by category/geography

**Learning Outcome:** - End-to-end ETL process - Real-time and batch processing in Databricks - SQL-based data analysis and reporting

---

## 2. Real-time Streaming Pipeline for Social Media Data

**Objective:** Build a streaming pipeline to process Twitter/Reddit posts in real-time and detect trending topics.

**Technologies/Skills:** - Python: Tweepy, Pandas, PySpark Structured Streaming - SQL: Store aggregated results in Databricks SQL/PostgreSQL - Databricks: Real-time streaming, Delta tables, ML integration - Others: Kafka or Kinesis for streaming

**Steps:** 1. Ingest data: Stream tweets/posts via API → Kafka/Kinesis → Databricks 2. Transform data in real-time: - Text cleaning - Sentiment analysis - Identify hashtags/keywords 3. Store results in Delta tables or SQL warehouse 4. Analytics: - Trending topics by hour/day - Sentiment trends over time

**Learning Outcome:** - Real-time data engineering - Integration of Python scripts with Databricks streaming - SQL queries on streaming data

---

## 3. Data Warehouse & Analytics for Retail Store

**Objective:** Create a Data Warehouse for a retail store and perform analytics on sales, inventory, and customer behavior.

**Technologies/Skills:** - Python: Pandas, PySpark - SQL: ETL queries, fact/dimension tables, joins, aggregate functions - Databricks: Delta Lake, notebooks - Others: Tableau/Power BI for reporting

**Steps:** 1. Collect sales, inventory, and customer datasets. 2. Data modeling: - Build Star Schema with Fact tables (sales) and Dimension tables (customer, product, store) 3. ETL pipeline: - Clean/transform data using Python/PySpark - Load into Databricks Delta tables 4. Analytics: - Best-selling products and stores - Inventory optimization - Customer segmentation

**Learning Outcome:** - Data warehouse modeling - SQL with big data - Scalable ETL pipeline using Databricks

---

**Why These Projects Help Get Data Engineering Jobs:** - Cover Python, SQL, and Databricks skills - Include ETL, data warehousing, and real-time streaming - Demonstrate end-to-end project ownership

---