Projects

1. Money Laundering Detection System – Banking

Domain: Banking & Financial Crime

- **Goal:** Build an ETL pipeline to ingest banking transactions (structured & unstructured) from multiple sources, identify suspicious transaction patterns using Spark MLlib, and store results in a DWH for analytics.
- **Tech:**
  - **Ingest:** Kafka → Spark Structured Streaming → HDFS
  - **Process:** Spark (Scala/Python) for anomaly detection
  - **Store:** PostgreSQL (DWH) + MongoDB (case details)
  - **Modeling:** Star schema with `fact_transactions` & `dim_customer`, `dim_account`
  - **Analytics:** BI dashboard for AML officers

---

2. Credit Card Fraud Detection – Banking

Domain: Banking & Risk Analytics

- **Goal:** Create a streaming pipeline to flag fraudulent credit card usage in near real-time.
- **Key Steps:**
  - ETL from transaction logs in NoSQL & RDBMS sources
  - Spark MLlib for classification (Random Forest / Isolation Forest)

    - Store predictions in Hive for reporting
- **Modeling:** Snowflake schema for `fact_fraud_events` with `dim_card`, `dim_location`, `dim_merchant`

---

3. Manufacturing Defect Detection

Domain: Manufacturing & IoT Analytics

- **Goal:** Process sensor & quality inspection data from assembly lines to detect defects and bottlenecks.
- **Pipeline:**
    - Ingest: IoT data via Kafka → HDFS
    - Process: Spark (Scala) for anomaly detection
    - Store: PostgreSQL for structured metrics, MongoDB for image defect reports
- **Modeling:** Fact table `fact_defects` linked to `dim_machine`, `dim_operator`, `dim_product`

---

4. Predictive Maintenance for Industrial Equipment

Domain: Manufacturing

- **Goal:** Use historical maintenance logs & sensor telemetry to predict machine failures.
- **Pipeline:** Hadoop batch ingestion + Spark MLlib for time series forecasting
- **Modeling:** Fact table `fact_maintenance_events` with `dim_equipment`, `dim_location`, `dim_maintenance_team`

---

5. Patient Risk Prediction – Healthcare

Domain: Healthcare Analytics

- **Goal:** Identify patients at high risk of chronic diseases using EMR & lab test data.
- **Tech:** Spark (Scala/Python) for preprocessing + MLlib logistic regression
- **Modeling:** Star schema with `fact_patient_risk` and dimensions for patient, hospital, and diagnosis
- **Store:** PostgreSQL for DWH, MongoDB for unstructured notes

---

6. Drug Prescription Anomaly Detection

Domain: Healthcare Fraud Prevention

- **Goal:** Detect unusual prescription patterns indicating drug abuse or fraud.
- **Pipeline:** Hive + Spark SQL for trend analysis, ML model for outlier detection
- **Modeling:** Snowflake schema with `fact_prescriptions`, `dim_doctor`, `dim_pharmacy`, `dim_drug`

---

7. Real-time Retail Demand Forecasting

Domain: Retail & E-commerce

- **Goal:** Predict product demand using sales transactions, seasonality, and promotions.
- **Pipeline:** Kafka → Spark Streaming → PostgreSQL DWH

- **Modeling:** Star schema for `fact_sales` and related dimensions
- **ML:** ARIMA / Prophet forecasting in Python

---

8. E-commerce Recommendation Engine

Domain: Retail / E-commerce

- **Goal:** Build a collaborative filtering recommendation system using transaction logs and customer behavior data.
- **Tech:** Spark MLlib ALS (Alternating Least Squares) model
- **Store:** Hive (historical logs) + PostgreSQL (current recommendations)

---

9. Insurance Claim Fraud Detection

Domain: Insurance Analytics

- **Goal:** Identify fraudulent claims by correlating claim data with historical records & external datasets.
- **Pipeline:** ETL from relational + NoSQL sources → Spark MLlib classification
- **Modeling:** Snowflake schema for `fact_claims` linked to claim type, customer, and provider dimensions

---

10. Traffic Flow Prediction for Smart Cities

Domain: Transportation & IoT

- **Goal:** Predict traffic congestion using GPS + sensor data streams.

- **Pipeline:** Kafka ingestion, Spark Streaming processing, PostgreSQL storage
- **Modeling:** Fact table `fact_traffic_flow` linked with `dim_location`, `dim_time`, `dim_weather`

---

11. Financial Market Data Warehouse

Domain: Capital Markets

- **Goal:** Build a DWH for stock, forex, and commodity market data with real-time feeds.
- **Pipeline:** Hadoop batch loads + Spark for intraday aggregation
- **Modeling:** Star schema for `fact_trades` with ticker, exchange, and sector dimensions

---

12. Energy Consumption Forecasting

Domain: Utilities / Smart Grid

- **Goal:** Predict household and industrial electricity demand from smart meter data.
- **Tech:** Spark MLlib time series modeling, PostgreSQL DWH for historical data
- **Modeling:** Star schema with `fact_energy_usage`, `dim_customer`, `dim_meter`

---

13. Social Media Sentiment Analytics

Domain: Marketing Analytics

- **Goal:** Analyze customer sentiment on social media posts for brand monitoring.
- **Pipeline:** Twitter API → Spark NLP → PostgreSQL DWH
- **Modeling:** Fact table `fact_sentiment` linked with `dim_platform`, `dim_campaign`

---

14. Airline Flight Delay Prediction

Domain: Aviation Analytics

- **Goal:** Predict flight delays based on weather, historical delays, and route data.
- **Pipeline:** Hadoop (historical batch) + Spark MLlib classification
- **Modeling:** Star schema for `fact_flight_delays` linked to `dim_airline`, `dim_airport`, `dim_weather`

---

15. Healthcare Image Processing & Storage

Domain: Healthcare + AI/ML

- **Goal:** Store, process, and analyze MRI/X-ray images for diagnostic assistance.
- **Pipeline:** Images in HDFS, metadata in PostgreSQL, ML inference in Python
- **Modeling:** `fact_diagnostics` linked to patient and scan dimensions, with image paths in NoSQL

If you want, I can **create a table mapping all these 15 projects to the exact ETL stages, Data Modeling type, storage layers, and ML algorithms** so you have a ready blueprint for implementation.

That would make it very actionable for your trainees or team.

Got it — here's **another 15 advanced, domain-diverse Data Engineering + Analytics project ideas** involving **PostgreSQL, Spark, Scala, Python, Unix, Hadoop, Hive, NoSQL**, plus **ETL, Data Warehousing, and Data Modeling** concepts.

Banking & Finance

1. Anti-Money Laundering Network Analysis

- **Goal:** Build a graph-based analytics platform to detect hidden relationships between suspicious accounts.
- **Tech:** Spark GraphX for relationship analysis, PostgreSQL for transactional storage, Neo4j for visualization.
- **Modeling:** Snowflake schema with `fact_transactions` linked to `dim_account`, `dim_customer`, `dim_relationship`.

2. Loan Default Risk Prediction

- **Goal:** Predict likelihood of loan default using historical credit data and customer behavior.
- **Pipeline:** Hive staging tables → Spark MLlib (Gradient Boosted Trees) → DWH.
- **Modeling:** Star schema with `fact_loans`, `dim_customer`, `dim_branch`, `dim_product`.

Manufacturing & Industrial

3. Supply Chain Optimization

- **Goal:** Optimize delivery routes and vendor selection using cost, delay, and quality data.
- **Pipeline:** ETL from ERP → Spark for optimization → PostgreSQL DWH.
- **Modeling:** Fact table `fact_shipments` with `dim_supplier`, `dim_route`, `dim_product`.

4. Quality Assurance Image Analytics

- **Goal:** Use AI image classification to detect quality issues in assembly line product photos.
- **Pipeline:** Images in HDFS, features extracted via Python OpenCV, classification in Spark MLlib.
- **Modeling:** Star schema with image metadata in DWH, defect reports in NoSQL.

Healthcare & Life Sciences

5. Genomics Data Processing Pipeline

- **Goal:** Process large genomic datasets for disease association analysis.
- **Pipeline:** Hadoop for raw data storage, Spark for distributed sequence processing.
- **Modeling:** Star schema for `fact_gene_analysis` with `dim_gene`, `dim_patient`, `dim_study`.

6. Hospital Bed & Resource Forecasting

- **Goal:** Predict hospital occupancy rates using patient inflow and seasonal patterns.
- **Pipeline:** Hive for historical data, Spark MLlib ARIMA for forecasting.
- **Modeling:** Snowflake schema with `fact_occupancy` and multiple dimension tables.

Retail & E-commerce

7. Dynamic Pricing Engine

- **Goal:** Adjust prices in real-time based on demand, stock, and competitor data.
- **Pipeline:** Kafka (competitor feeds) → Spark Streaming → DWH.
- **Modeling:** Star schema for `fact_price_changes` with `dim_product`, `dim_store`, `dim_event`.

8. Customer Lifetime Value (CLV) Prediction

- **Goal:** Predict revenue contribution of customers over their lifecycle.
- **Pipeline:** Hive for transaction history, Spark MLlib regression models.
- **Modeling:** Fact table `fact_customer_value` with `dim_customer`, `dim_campaign`.

Transportation & Logistics

9. Fleet Fuel Optimization

- **Goal:** Reduce fuel costs by analyzing vehicle sensor and route data.
- **Pipeline:** HDFS for GPS/sensor logs, Spark SQL aggregation, Hive staging.

- **Modeling:** Star schema for `fact_fuel_usage` linked with vehicle, route, and driver dimensions.

10. Railway Delay Pattern Analysis

- **Goal:** Analyze delay causes and predict disruptions in train schedules.
- **Pipeline:** ETL from historical logs → Spark MLlib classification.
- **Modeling:** Snowflake schema with `fact_train_delays` and `dim_route`, `dim_weather`.

Energy & Utilities

11. Renewable Energy Production Forecast

- **Goal:** Forecast solar and wind energy generation using weather patterns.
- **Pipeline:** Hive for weather history, Spark MLlib for prediction.
- **Modeling:** Star schema for `fact_energy_output` with `dim_plant`, `dim_weather`.

12. Smart Water Leakage Detection

- **Goal:** Detect unusual consumption patterns to identify leaks.
- **Pipeline:** IoT sensor feeds → Spark anomaly detection → PostgreSQL DWH.
- **Modeling:** Snowflake schema with `fact_water_usage`, `dim_meter`, `dim_location`.

Telecom & Media

13. Call Detail Record (CDR) Analysis

- **Goal:** Analyze telecom usage patterns for churn prevention.
- **Pipeline:** Kafka → Spark Streaming → Hive.
- **Modeling:** Star schema for `fact_calls` with `dim_customer`, `dim_tower`, `dim_plan`.

14. OTT Platform Recommendation System

- **Goal:** Recommend movies/shows based on user watch patterns.
- **Pipeline:** Spark MLlib ALS recommendation, Hive as data source, PostgreSQL as serving layer.
- **Modeling:** Snowflake schema with `fact_views`, `dim_content`, `dim_user`.

Government & Public Sector

15. Crime Pattern Prediction

- **Goal:** Predict crime-prone areas using historical police reports and socio-economic indicators.
- **Pipeline:** Hadoop for raw datasets, Spark MLlib classification, NoSQL for geo-tagged data.
- **Modeling:** Star schema with `fact_crime_events`, `dim_location`, `dim_offense`, `dim_time`.

SAAS

Ideas

Master Project Table – Data Engineering & Analytics (PostgreSQL, Spark, Scala, Python, Unix, Hadoop, Hive, NoSQL)

#	Domain	Business Problem	ETL Flow (High-Level)	Data Modeling Approach	Tech Stack

1	Banking	Money Laundering Detection	Kafka → Spark Streaming → HDFS → PostgreSQL/MongoDB	Star schema (fact_transactions, dim_customer, dim_account)	Spark (Scala/Py), Kafka, PostgreSQL, MongoDB, Hive
2	Banking	Credit Card Fraud Detection	RDBMS/NoSQL → Spark MLlib → Hive	Snowflake (fact_fraud_events, dim_card, dim_location, dim_merchant)	Spark, Hive, PostgreSQL
3	Manufacturing	Defect Detection	Kafka IoT → Spark → PostgreSQL/MongoDB	Star (fact_defects, dim_machine, dim_operator, dim_product)	Spark, HDFS, PostgreSQL, MongoDB
4	Manufacturing	Predictive Maintenance	Hadoop batch → Spark MLlib	Star (fact_maintenance_events, dim_equipment, dim_location, dim_team)	Hadoop, Spark, PostgreSQL
5	Healthcare	Patient Risk Prediction	EMR & lab data → Spark MLlib → DWH	Star (fact_patient_risk, dim_patient, dim_hospital, dim_diagnosis)	Spark, PostgreSQL, MongoDB
6	Healthcare	Drug Prescription Anomaly Detection	Hive → Spark SQL/ML → DWH	Snowflake (fact_prescriptions, dim_doctor, dim_pharmacy, dim_drug)	Hive, Spark, PostgreSQL
7	Retail	Real-time Demand Forecasting	Kafka → Spark Streaming → PostgreSQL	Star (fact_sales, dim_product, dim_store, dim_time)	Kafka, Spark, PostgreSQL
8	Retail	E-commerce Recommendation Engine	Logs → Spark MLlib ALS → Hive/PostgreSQL	Star (fact_user_item, dim_user, dim_item)	Spark MLlib, Hive, PostgreSQL
9	Insurance	Claim Fraud Detection	RDBMS + NoSQL → Spark MLlib → Hive	Snowflake (fact_claims, dim_customer, dim_provider, dim_type)	Spark, Hive, PostgreSQL
10	Transportation	Traffic Flow Prediction	Kafka → Spark → PostgreSQL	Star (fact_traffic_flow, dim_location, dim_time, dim_weather)	Kafka, Spark, PostgreSQL
11	Finance	Financial Market DWH	Real-time API → Hadoop batch + Spark	Star (fact_trades, dim_ticker, dim_exchange, dim_sector)	Hadoop, Spark, PostgreSQL
12	Energy	Energy Consumption Forecasting	IoT → Spark MLlib → PostgreSQL	Star (fact_energy_usage, dim_customer, dim_meter)	Spark, PostgreSQL
13	Marketing	Social Media Sentiment Analysis	Twitter API → Spark NLP → PostgreSQL	Star (fact_sentiment, dim_platform, dim_campaign)	Spark NLP, PostgreSQL
14	Aviation	Flight Delay Prediction	Hadoop batch → Spark MLlib → Hive	Star (fact_flight_delays, dim_airline, dim_airport, dim_weather)	Hadoop, Spark, Hive

15	Healthcare	Healthcare Image Processing	Images → HDFS → ML inference → PostgreSQL/NoSQL	Star (fact_diagnostics, dim_patient, dim_scan)	HDFS, Spark, PostgreSQL, MongoDB
16	Banking	AML Network Analysis	RDBMS + NoSQL → Spark GraphX → Neo4j/PostgreSQL	Snowflake (fact_transactions, dim_relationship)	Spark GraphX, PostgreSQL, Neo4j
17	Banking	Loan Default Risk Prediction	Hive → Spark MLlib → PostgreSQL	Star (fact_loans, dim_customer, dim_branch, dim_product)	Hive, Spark, PostgreSQL
18	Manufacturing	Supply Chain Optimization	ERP → Spark optimization → PostgreSQL	Star (fact_shipments, dim_supplier, dim_route, dim_product)	Spark, PostgreSQL
19	Manufacturing	QA Image Analytics	HDFS images → OpenCV → Spark MLlib	Star (fact_defect_images, dim_machine, dim_product)	HDFS, OpenCV, Spark
20	Healthcare	Genomics Data Processing	Hadoop → Spark → PostgreSQL	Star (fact_gene_analysis, dim_gene, dim_patient)	Hadoop, Spark
21	Healthcare	Hospital Bed Forecasting	Hive → Spark MLlib → DWH	Snowflake (fact_occupancy, dim_hospital, dim_time)	Hive, Spark, PostgreSQL
22	Retail	Dynamic Pricing Engine	Kafka → Spark Streaming → PostgreSQL	Star (fact_price_changes, dim_product, dim_store, dim_event)	Kafka, Spark, PostgreSQL
23	Retail	CLV Prediction	Hive → Spark MLlib → DWH	Star (fact_customer_value, dim_customer, dim_campaign)	Hive, Spark, PostgreSQL
24	Logistics	Fleet Fuel Optimization	IoT logs → Spark SQL → Hive/PostgreSQL	Star (fact_fuel_usage, dim_vehicle, dim_route, dim_driver)	Spark, Hive, PostgreSQL
25	Transportation	Railway Delay Analysis	Hadoop batch → Spark MLlib → Hive	Snowflake (fact_train_delays, dim_route, dim_weather)	Hadoop, Spark, Hive
26	Energy	Renewable Energy Forecasting	Hive weather data → Spark MLlib → DWH	Star (fact_energy_output, dim_plant, dim_weather)	Hive, Spark, PostgreSQL
27	Utilities	Smart Water Leakage Detection	IoT → Spark anomaly detection → PostgreSQL	Snowflake (fact_water_usage, dim_meter, dim_location)	Spark, PostgreSQL
28	Telecom	CDR Analysis	Kafka → Spark → Hive/PostgreSQL	Star (fact_calls, dim_customer, dim_tower, dim_plan)	Kafka, Spark, Hive
29	Media	OTT Recommendation System	Logs → Spark MLlib ALS → Hive/PostgreSQL	Snowflake (fact_views, dim_content, dim_user)	Spark MLlib, Hive, PostgreSQL
30	Public Sector	Crime Pattern Prediction	Hadoop batch → Spark MLlib → NoSQL/PostgreSQL	Star (fact_crime_events, dim_location, dim_offense, dim_time)	Hadoop, Spark, PostgreSQL, MongoDB