**Project Overview**  
  
**Project Title:** Real-Time Fraud Detection with ML in Azure Databricks

**Goal:**

To build a real-time data pipeline that detects potentially fraudulent transactions using machine learning, leveraging Azure services and Databricks for analytics capabilities.

**Project Scope**

This project aims to:

* Simulate real-time financial transaction data
* Stream data into Azure using Event Hub or Storage
* Process data using **Databricks Structured Streaming**
* Detect anomalies using a **trained ML model**
* Store suspicious transactions separately for investigation
* Visualize insights and fraud statistics in Power BI or other tools

**Architecture Components**

| **Component** | **Purpose** |
| --- | --- |
| **Azure Event Hub / Storage** | Ingest simulated transaction stream |
| **Azure Databricks** | Process, transform, and analyze the data |
| **Delta Lake** | Store raw and processed data |
| **ML Model** (e.g., Logistic Regression / Isolation Forest) | Classify transactions |
| **Power BI / Azure SQL** | Visualize results and fraud trends |

**Project Planning**

| **Phase** | **Task** | **Tools/Tech** | **Duration (est.)** |
| --- | --- | --- | --- |
| 1️ | Define schema & simulate data | Python | 1 day |
| 2️ | Set up Azure Event Hub or Storage | Azure | 0.5 day |
| 3 | Ingest stream in Databricks | Spark, Databricks | 1 day |
| 4️ | Train ML model with sample data | Spark ML | 1 day |
| 5️ | Apply model to streaming data | Spark Structured Streaming | 1 day |
| 6️ | Store suspicious data in Delta | Delta Lake | 0.5 day |
| 7️ | Build Power BI dashboard | Power BI, Delta Lake | 1 day |

**Deliverables**

* Real-time streaming pipeline
* Trained ML model for fraud detection
* Suspicious transaction store
* Dashboard showing real-time fraud trends