# Requirements Document

## 1. Required Libraries & Dependencies

To run the Telecom Fraud Detection Pipeline successfully, ensure the following libraries are installed:

* apache-airflow
* pandas
* pyspark
* os
* logging, glob
* apache-airflow-providers-postgres
* apache-airflow-providers-apache-spark
* JDBC driver for postgres
* Additional dependencies:  
  - PostgreSQL installed and running.  
  - Airflow configured with DAGs directory set up.  
  - Apache Spark installed for processing.

## 2. Environment Setup

### PostgreSQL Configuration

* Ensure PostgreSQL service is running.
* Create a database (`airflow\_pipeline`) and required tables (created automatically by DAG).

### Apache Airflow Configuration

Airflow should be initialized and webserver/scheduler running:  
 airflow db init  
 airflow webserver --port 8080  
 airflow scheduler

### Apache Spark Configuration

Ensure Spark is installed and `SPARK\_HOME` is set.  
 The processing scripts use PySpark, so ensure the Python environment is set up properly.

## 3. File System Setup

Create the following directories before running the pipeline:

* extracted/
* processed/`
* logs/
* Archived/
* Scripts/
* Auditfiles/

I have used below for my pipeline change accordingly

SOURCE\_DIRECTORY= "/home/lokesh/Telecome\_Pipeline/SourceInputDir/”

LOG\_DIRECTORY="/home/lokesh/Telecome\_Pipeline/logs"

AUDIT\_DIRECTORY="/home/lokesh/Telecome\_Pipeline/Audit"

PROCESSED\_DIRECTORY="/home/lokesh/Telecome\_Pipeline/ProcessedDir"

ARCHIVE\_DIR="/home/lokesh/Telecome\_Pipeline/ArchiveDir"

REJECTIONS\_DIR="/home/lokesh/Telecome\_Pipeline/RejectionsDir"

SCRIPTS =”/home/lokesh/Telecome\_Pipeline/Scripts”

## 4. Airflow DAG Execution

Once everything is set up, trigger the DAG from the Airflow UI or CLI:

## 5. Common Issues & Fixes

Refer to the troubleshooting document for common errors related to Airflow, Spark, and PostgreSQL configurations.