# **Apache Spark ETL Pipeline - Setup Instructions**

## **Prerequisites**

* Apache Spark (3.x)
* PostgreSQL (12+)
* Python 3.7+ with PySpark
* PostgreSQL JDBC driver (postgresql-42.3.1.jar)

## **Setup Steps**

1. **Install necessary Python packages**:  
    pip install pyspark pandas psycopg2-binary
2. **Create PostgreSQL database**:  
    psql -U postgres -c "CREATE DATABASE pursuit\_data;"
3. **Set up the ETL script**:
   * Save the provided ETL code as spark\_etl.py
   * Update the database connection details if needed
4. **Place your CSV files in the working directory**:  
   * contacts.csv
   * places.csv
   * techstacks.csv
   * customer\_mappings.csv
   * Contact\_sync\_status.csv
5. **Create execution script**:

# Create run\_spark\_etl.sh

echo '#!/bin/bash

echo "==================== SPARK ETL PIPELINE ===================="

echo "Setting up environment..."

# Create PostgreSQL database if needed

echo "Creating PostgreSQL database if it doesn't exist..."

psql -U postgres -c "CREATE DATABASE pursuit\_data;" || echo "Database already exists,

continuing..."

# Run Spark ETL pipeline

echo "Starting Spark ETL Pipeline..."

spark-submit --jars ~/spark/jars/postgresql-42.3.1.jar spark\_etl.py

echo "==================== PIPELINE COMPLETE ===================="

' > run\_spark\_etl.sh

# Make executable

chmod +x run\_spark\_etl.sh

1. **Run the pipeline**:  
     
    ./run\_spark\_etl.sh

## **Troubleshooting**

* **JDBC errors**: Ensure the PostgreSQL JDBC driver path in spark\_etl.py matches your environment
* **Database connection issues**: Verify PostgreSQL credentials and that the service is running
* **CSV loading errors**: Check that your CSV files match the expected format

The pipeline is configured to continue execution even if schema creation encounters errors, as actual table creation will happen during data insertion.