|  |
| --- |
| Photo displaying partial image of two pie charts on a canvas-textured page |
| Fraud Transaction Prediction |
| |  |  |  | | --- | --- | --- | | **SPA ASSIGNMENT GROUP 43** |  | Stream Processing & Analytics | |

Contents

[**Introduction** 2](#_Toc43155622)

[**List of Spark Programs Developed** 2](#_Toc43155623)

[**Spark Programs Explained Step by Step** 2](#_Toc43155624)

[**PublishToKafkaTopic:** 2](#_Toc43155625)

[**What It Does?** 2](#_Toc43155626)

[**How It Does?** 2](#_Toc43155627)

[**Input Parameters to program:** 2](#_Toc43155628)

[**Code Snippets:** 3](#_Toc43155629)

[**SubscribeToKafkaTopic:** 3](#_Toc43155630)

[**What it does?** 3](#_Toc43155631)

[**How It Does?** 3](#_Toc43155632)

[**Input Params to program:** 3](#_Toc43155633)

[**Code Snippets:** 3](#_Toc43155634)

[**LogRegSparkMLPipeline:** 4](#_Toc43155635)

[**What it does?** 4](#_Toc43155636)

[**How it does?** 4](#_Toc43155637)

[**Input Params to Program:** 4](#_Toc43155638)

[**Code Snippets:** 4](#_Toc43155639)

[**Code Snippet to Evaluate Model and display Accuracy:** 5](#_Toc43155640)

[**Code Snippet to save model path:** 5](#_Toc43155641)

[**PredictRealTime** 5](#_Toc43155642)

[**What it Does?** 5](#_Toc43155643)

[**How it Does?** 5](#_Toc43155644)

[**Input Params:** 5](#_Toc43155645)

[**Code Snippets:** 5](#_Toc43155646)

[**Code Snippet to subscribe to Kafka topic:** 5](#_Toc43155647)

[**Code Snippet to Load ML Model Saved in previous step:** 6](#_Toc43155648)

[**Code Snippet to save transaction to different Kafka topics based on Prediction** 6](#_Toc43155649)

# **Introduction**

Problem Statement: **Predicting Fraud Transactions**

Tech Stack: **Kafka, Spark**

Programming APIs: **Spark Dataframes, Structured Streaming, MLLib, Kafka APIs**

Programming Language: **Scala**

Project Build Tool: **Maven**

IDE used for Development: **IntelliJ**

Git Location of Code: <https://github.com/mighty-raj/SPA_Streaming_Assignment.git>

Training Data Used: <https://www.kaggle.com/ntnu-testimon/paysim1>

Team Name: **SPA ASSIGNMENT GROUP 43**

Team Members:

Mrutyamjaya Surampudi,

Srinivas Veerabomma,

Raja Mahesh Aravapalli

# **List of Spark Programs Developed**

As part of this project we created 4 Spark Programs and given below are the list:

1. PublishToKafkaTopic
2. SubscribeToKafkaTopic
3. LogRegSparkMLPipeline
4. PredictRealTime

# **Spark Programs Explained Step by Step**

## **PublishToKafkaTopic:**

### **What It Does?**

Reads a CSV source files from source directory path as it streams and writes content to Kafka Topic in JSON Format in real time.

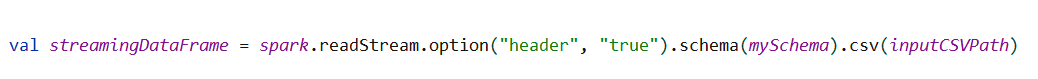
### **How It Does?**

Using Spark Structured Streaming API calls, application publishes data to Kafka topic from source directory as the new files arrives at source location.

### **Input Parameters to program:**

1. Path to the CSV Source File
2. Kafka Broker Host/IP & Port
3. Kafka Topic Name to write data to.
4. Checkpoint directory path, to recover from Failures

### **Code Snippets:**



A screenshot of a cell phone

Description automatically generated

## **SubscribeToKafkaTopic:**

### **What it does?**

Reads Kafka Topic and streams data into a Destination Path (Master Data) given.

### **How It Does?**

Using Spark Structured streaming APIs, subscribes to a kafka topic and writes data to given destination path.

### **Input Params to program:**

1. Kafka Topic Name
2. Directory location to stream data into.
3. Checkpoint directory path, to recover from Failures

### **Code Snippets:**

A screenshot of a cell phone

Description automatically generated

## **LogRegSparkMLPipeline:**

### **What it does?**

1. Trains a machine learning model that can Predict fraud transactions.
2. Uses PaySim data to train the model
3. Applies Logistic Regression ml technique.

### **How it does?**

Using Spark ML Lib Pipeline APIs, trains a logistic regression ml model, and saves trained model to given path, which can later be used to predict on unseen transaction!

### **Input Params to Program:**

1. Input Directory Path for training data
2. Directory path to save the trained model

### **Code Snippets:**

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a social media post

Description automatically generated

### **Code Snippet to Evaluate Model and display Accuracy:**

A screenshot of a cell phone

Description automatically generated

### **Code Snippet to save model path:**

A picture containing knife

Description automatically generated

## **PredictRealTime**

### **What it Does?**

1. Loads saved ML Model
2. Reads messages in real time from kafka topic
3. Applies ML Model and predicts transaction in real-time
4. Depending the prediction, weather it is a Fraud or Normal transaction, writes back the data to different kafka topics configured.

### **How it Does?**

Using Spark MLLib and Spark Structured streaming APIs, load ml model from saved directory, and then applies it to incoming kafka messages arriving into kafka in real-time

### **Input Params:**

1. Input Directory Path to ML Model to load from
2. Source Kafka topic to read incoming transactions in real-time
3. Kafka Topic Name to write Fraud Transactions
4. Kafka Topic Name to write Normal Transactions

### **Code Snippets:**

### **Code Snippet to subscribe to Kafka topic:**

A screenshot of a cell phone

Description automatically generated

### **Code Snippet to Load ML Model Saved in previous step:**

A screenshot of a cell phone

Description automatically generated

### **Code Snippet to save transaction to different Kafka topics based on Prediction**

A screenshot of a cell phone

Description automatically generated