**Project Overview**

In this project, you will be provided with a real-world dataset, extracted from Kaggle, on San Francisco crime incidents, and you will provide statistical analyses of the data using Apache Spark Structured Streaming. You will draw on the skills and knowledge you've learned in this course to create a Kafka server to produce data, and ingest data through Spark Structured Streaming.

**Development Environment**

You may choose to create your project in the workspace we provide here, or if you wish to develop your project locally, you will need to set up your environment properly as described below:

* Spark 2.4.3
* Scala 2.11.x
* Java 1.8.x
* Kafka build with Scala 2.11.x
* Python 3.6.x or 3.7.x

**Environment Setup (Only Necessary if You Want to Work on the Project Locally on Your Own Machine)**

**For Macs or Linux:**

* Download Spark from [**https://spark.apache.org/downloads.html**](https://spark.apache.org/downloads.html). Choose "Prebuilt for Apache Hadoop 2.7 and later."
* Unpack Spark in one of your folders (I usually put all my dev requirements in /home/users/user/dev).
* Download binary for Kafka from this location [**https://kafka.apache.org/downloads**](https://kafka.apache.org/downloads), with Scala 2.11, version 2.3.0. Unzip in your local directory where you unzipped your Spark binary as well. Exploring the Kafka folder, you’ll see the scripts to execute in bin folders, and config files under config folder. You’ll need to modify zookeeper.properties and server.properties.
* Download Scala from the official site, or for Mac users, you can also use brew install scala, but make sure you download version 2.11.x.
* Run below to verify correct versions:
* java -version
* scala -version
* Make sure your ~/.bash\_profile looks like below (might be different depending on your directory):
* export SPARK\_HOME=/Users/dev/spark-2.4.3-bin-hadoop2.7
* export JAVA\_HOME=/Library/Java/JavaVirtualMachines/jdk1.8.0\_181.jdk/Contents/Home
* export SCALA\_HOME=/usr/local/scala/
* export PATH=$JAVA\_HOME/bin:$SPARK\_HOME/bin:$SCALA\_HOME/bin:$PATH

**For Windows:**

Please follow the directions found in this helpful StackOverflow post: [**https://stackoverflow.com/questions/25481325/how-to-set-up-spark-on-windows**](https://stackoverflow.com/questions/25481325/how-to-set-up-spark-on-windows)

