**Final Project**

Kafka for Stream Processing

**Problem Statement:**

Data is constantly being produced meaning the amount of data is no longer an issue. The problem we face now is the need for immediate access that data as it is produced and the ability sort through the data to separate the signal from the noise. The use case that will be presented as an example is Police Dispatch Calls. With many calls coming in, not all officers need to be notified of every call. For example, a police officer on traffic duty may only need to know the calls that come in regarding traffic. My project builds a data infrastructure using three Kafka topics to ingest police log files from the Rockford Police department in Illinois. Spark Streaming and Cassandra applications are used to connect to the needed data to process and store.

**Overview of the Technology:**

To solve the above problem, I created a Kafka Producer program that read in one line of data at a time to create a stream. The data was sent to one of three Kafka topics depending in key words present in the data. This allows end user applications to connect only to the streaming data that is important for their uses. Then I demonstrate the ingestion of the streaming data for two of the topics created using Spark Streaming. The data is then sent to a Cassandra database to be stored for later analysis. The use of these echnologies allows for flexible and fast data processing. One challenge was connecting Spark to other applications. Spark streaming only supports RDD’s and connections using python API’s can be difficult since Spark is supported mostly by Scala and Java.

**High Level Steps:**

1. Download data and pre-process
2. Install and configure software
3. Implement Kafka Producer application in testing environment
4. Implement Kafka Consumer application in testing environment
5. Run process in production environment

**Data Source:**

Rockford Police Department Dispatch Call Log

https://data.illinois.gov/dataset/639rockford\_police\_department\_72\_hour\_dispatch\_call

**Hardware Used:**

Windows 10 64 bit processor laptop running VMWare

**Software Used:**

VMWare Workstation Pro with CentosOS7.4 installed (<https://www.centos.org/download/>)

Python 2.7.5 (<https://www.python.org/downloads/>)

Kafka 0.11.0.1 (<https://kafka.apache.org/downloads>)

Cassandra 2.2.11 (http://cassandra.apache.org/download/)

Spark 2.2.0 (<https://spark.apache.org/downloads.html>)

**YouTube Links:**

2 Min: <https://www.youtube.com/watch?v=RlERTAGvqBk>

15 Min: <https://www.youtube.com/watch?v=bcin-AM4354c>