CPSC 4330 Big Data Analytics

In-class Exercise 12 – Use Spark SQL for ETL

In this exercise, you will use Spark SQL to load data from HDFS, parse and analyze it, and store the analysis result in a different data format to HDFS.

Use Spark SQL to Parse and Analyze Data

The dataset from KDD Cup is used in this exercise. You can download the full data set "kddcup.data.gz" from Canvas. The dataset contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment. The task of this exercise is to parse and analyze the data.

1. Upload the data set to your bucket in S3
2. In Notebook, write python code to do the following things.
    - Load and parse the data. Every entry in the dataset is a comma-delimited line. The data consists of different attributes captured from connection data. For this exercise, you only need the 1st attribute (duration), the 2nd attribute (protocol-type), and the last attribute (label).
        - Duration: length (number of seconds) of the connection
        - Protocol-type: type of the protocol, e.g. tcp, udp, etc.
        - Label: types of the connection, e.g. normal, guess-passwd, etc.
    - Get the total number of connections based on the type of connectivity protocol.
    - Find out the total duration time for each protocol.
    - Find out which protocol is most vulnerable to attacks. In other words, which protocol has the highest number of attacks. "normal" is no attack; other values of the attribute *Label* are considered as attack.
    - Save the analysis results as JSON files in S3.

When you finish the exercise, remember to terminate your cluster and stop the Notebooks.