Setting up Spark-Kafka-Spark Streaming

Running Spark on Bash Console

- curl -O https://d3kbcqa49mib13.cloudfront.net/spark-2.2.0-bin-hadoop2.7.tgz
- tar zxvf spark-2.2.0-bin-hadoop2.7.tgz
- cd ~
- sudo nano .bashrc
  - alias python=python3.7
  - export SPARK_HOME="/home/<your_username>/spark-2.2.0-bin-hadoop2.7/“
  - export PATH=$SPARK_HOME/bin:$PATH
- source ~/.bashrc
- pyspark

Running Spark on Bash Console - v3.1.1

- cd ~
- curl -O https://apachemirror.wuchna.com/spark/spark-3.1.1/spark-3.1.1-bin-hadoop2.7.tgz
- tar zxvf spark-3.1.1-bin-hadoop2.7.tgz
- mv spark-3.1.1-bin-hadoop2.7 /opt/spark311
- chmod -R a+r /opt/spark311
- chmod -R a+w /opt/spark311
- python -m venv ~/python-virtual-environments/spark_venv_311
- source ~/python-virtual-environments/spark_venv_311/bin/activate
- pip install pyspark findspark
- pyspark

Setup Jupyter (Optional)

- pip install jupyter
- jupyter notebook --ip=* --no-browser --allow-root &

Configure PySpark on Jupyter Notebook

- import findspark
- findspark.init ("/opt/spark311")
- import pyspark
- from pyspark.sql import SparkSession

- spark = SparkSession.builder.getOrCreate()

- df = spark.sql("select 'spark' as hello ")

- df

Setup Kafka

- wget https://apachemirror.wuchna.com/kafka/2.8.0/kafka_2.13-2.8.0.tgz
- tar -xzf kafka_2.13-2.8.0.tgz
- cd kafka_2.13-2.8.0

- bin/zookeeper-server-start.sh config/zookeeper.properties &
- bin/kafka-server-start.sh config/server.properties &
- bin/kafka-topics.sh --create --topic test-topic --bootstrap-server localhost:9092

- Refer:
  - https://kafka.apache.org/quickstart

Configure Spark for Kafka

- cp /opt/spark311/conf/spark-defaults.conf.template /opt/spark/conf/spark-defaults.conf
- vim /opt/spark311/conf/spark-defaults.conf
  - spark.master   spark://localhost:7077
  - spark.jars.packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.0.1

Run Spark Standalone

- cd /opt/spark311/
- ./sbin/start-master.sh
- tail -f /opt/spark/logs/spark--org.apache.spark.deploy.master.Master-1-bae5ef2081fd.out
- ./sbin/start-slave.sh spark://bae5ef2081fd:7077

Install Kafka modules for Python

- source ~/python-virtual-environments/spark_venv_311/bin/activate
- pip install kafka
- pip install kafka-python
- python kafka_producer_csv.py
- python kafka_streaming_csv_demo.py

Install Kafka modules for Python

- source ~/python-virtual-environments/spark_venv_311/bin/activate
- pip install kafka
- pip install kafka-python

Thanks