Indexing & Searching for Hot Topics with ElasticSearch, Apache Spark, Kafka, Count-Min, and Heavy Hitters

Instructions

Register for CSC591_ADBI_v3 VCL from https://vcl.wordpress.ncsu.edu

Setting up Elasticsearch

Elasticsearch requires at least Java 7. Once you have Java set up, you can then download and run Elasticsearch

First update your package index

sudo apt-get update

Download the Elasticsearch version 2.3.4

wget https://download.elastic.co/elasticsearch/release/org/elasticsearch/distribution/deb/elasticsearch/2.3.4/elasticsearch-2.3.4.deb

Install it with Ubuntu way with dpkg

sudo dpkg -i elasticsearch-2.3.4.deb

To make sure Elasticsearch starts and stops automatically with the server, Run Below command

sudo systemctl enable elasticsearch.service

To check service status of Elasticsearch, follow below commands:

To Check running status:

sudo service elasticsearch status

To start the service:

sudo service elasticsearch start

To stop the service:

sudo service elasticsearch stop

Setting up connection between Kafka and Elasticsearch

To continue this setup, we should have one Zookeeper instance, one Kafka broker, and one Elasticsearch node, all installed on a single machine and listening on the following ports:

Zookeeper: 2181

Kafka: 9092

Elasticsearch: 9200

Start Zookeeper service

$KAFKA_HOME/bin/zookeeper-server-start.sh $KAFKA_HOME/config/zookeeper.properties

Start Kafka service

$KAFKA_HOME/bin/kafka-server-start.sh $KAFKA_HOME/config/server.properties

Create a topic named twitterstream in kafka:

$KAFKA_HOME/bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic twitterstream

Setup Kafka connector

Follow below commands for setup:

sudo apt install maven

git clone -b 0.10.0.0 https://github.com/confluentinc/kafka-connect-elasticsearch.git

cd kafka-connect-elasticsearch

mvn clean package

cd target/kafka-connect-elasticsearch-3.2.0-SNAPSHOT-package/share/java/kafka-connect-elasticsearch/

Copy all the libraries from above folder to Kafka libs directory

cp * $KAFKA_HOME/libs/.

Create Index using below command

curl -X PUT "localhost:9200/twitterstream"

```
mdutta3@vml7-4:~$
mdutta3@vml7-4:~$ curl -X GET "localhost:9200/_cat/indices?v"
health status index pri rep docs.count docs.deleted store.size pri.store.size
mdutta3@vml7-4:~$ curl -X PUT "localhost:9200/twitterstream"
{"acknowledged":true}mduttGET "localhost:9200/_cat/indices?v"00/twitterstream"
health status index         pri rep docs.count docs.deleted store.size pri.store.size
yellow open   twitterstream   5   1          0            0       260b           260b
mdutta3@vml7-4:~$
```

copy below files from config folder of the submit folder to Kafka config Directory

```
-rwxrwxrwx 1 root root  530 May  2 15:24 connect-standalone.properties
-rwxrwxrwx 1 root root  265 May  3 19:56 elasticsearch-connect.properties
```

Destination path: **$KAFKA_HOME/config**

Note: Use sudo command to copy these files

Run the Connector using below commands

cd $KAFKA_HOME

bin/connect-standalone.sh config/connect-standalone.properties config/elasticsearch-connect.properties

Run below command

sudo pip install tweepy

Run below script to stream tweets

python twitter_to_kafka.py

To check data is landing in kafka

$KAFKA_HOME/bin/kafka-console-consumer.sh –zookeeper localhost:2181 –topic twitterstream –from-beginning

Create if data is coming in Index using below command

curl -X GET "localhost:9200/_cat/indices?v"

```
mdutta3@vm17-4:~$
mdutta3@vm17-4:~$ curl -X GET "localhost:9200/_cat/indices?v"
health status index         pri rep docs.count docs.deleted store.size pri.store.size
yellow open   twitterstream   5   1        483            0     10.8mb         10.8mb
mdutta3@vm17-4:~$
```

Test DRPC queries by running below command

python queriestest.py

```
mdutta3@vm17-4:/afs/unity.ncsu.edu/users/m/mdutta3/capstone$ python queriestest.py
**********Check For Query******************************
1. Term Query
2. Match Query
3. Range Query
4. EXIT
Enter Your Input:
```

Press 1 to test Term query, 2 to test Match query, 3 to test range query and 4 to exit