DIGITAL GFT

Big Data

PROJECT SIMULATION

GROUP-A

Md. Sabir - 11607047

Shweta L - 11607320

Akshita Johri - 11611168

Amit Srivatsa – 11606808

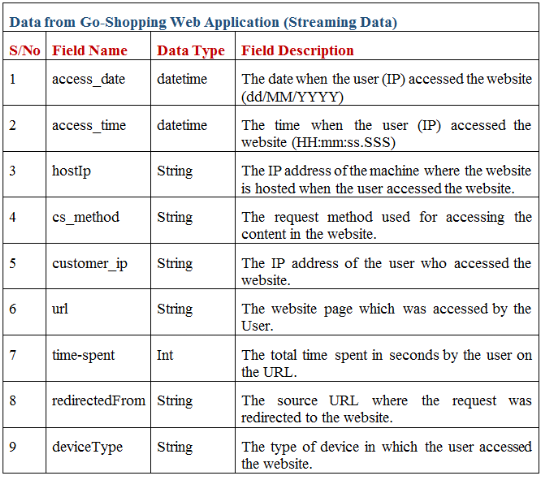
TABLE OF CONTENTS

|  |  |
| --- | --- |
| 1.0 | Abstract |
| 2.0 | Data and MetaData description |
| 3.0 | Architecture |
| 4.0 | Data-Ingestion |
| 5.0 | Data-Analysis |
| 6.0 | Data Visualization |

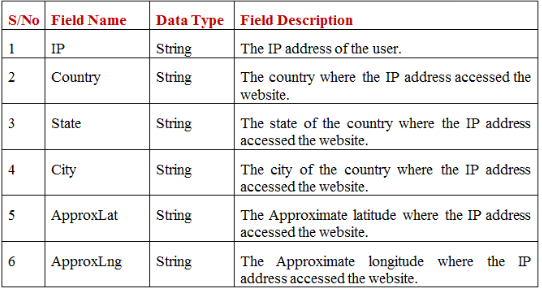
ABSTRACT

1. What dataset we used
2. How did we ingest the data
3. How was the analysis done based on the ingestion
4. How did we respresnt the analysed data
5. Application jars we built

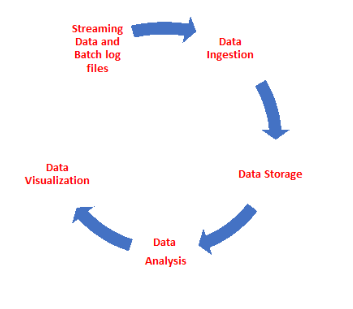
DATA AND METADATA DESCRIPTION







ARCHITECTURE



DATA INGESTION

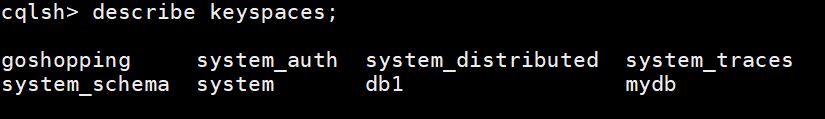
Batch Files

Cassandra :

1. Creating database named “GoShopping”

CODE :

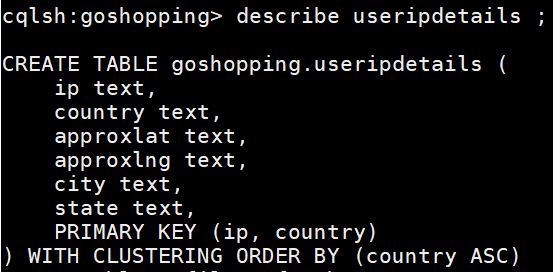
**create keyspace GoShopping with replication ={ 'class': 'SimpleStrategy', 'replication\_factor': 1 };**



1. Creating table named “UserIpDetails” and loading the data from goShopping\_IpLookup

CODE:

**create table goshopping.UserIpDetails (IP text,Country text,State text,City text,ApproxLat text,ApproxLng text,primary key(IP,Country));**



**copy UserIpDetails(IP,Country,State,City,ApproxLat,ApproxLng) from '/home/vagrant/goShopping\_IpLookup.txt';**



HDFs :

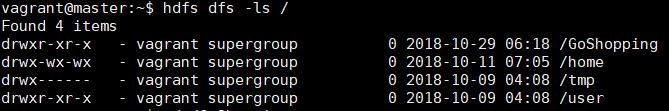
1. Creating a “GoShopping” directory

**hdfs dfs -mkdir /GoShopping**



1. Loading the goShopping\_IpLookup file to the above directory

**hdfs dfs -put /home/vagrant/GoShoppingIpLookup.txt /GoShopping**

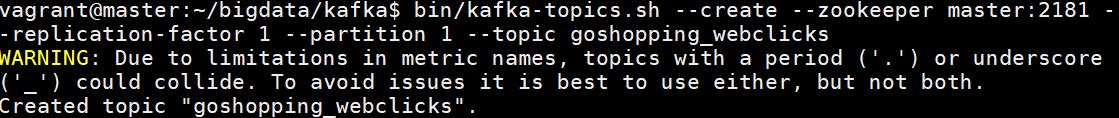


Streaming Data

Straming Source to kafka :

1. Creating “goshopping\_webclicks” topic with a single replication

**bin/kafka-topic.sh --create --zookeeper master:2181 --replication-factor 1 --partition 1 --topic goshopping\_webclicks**

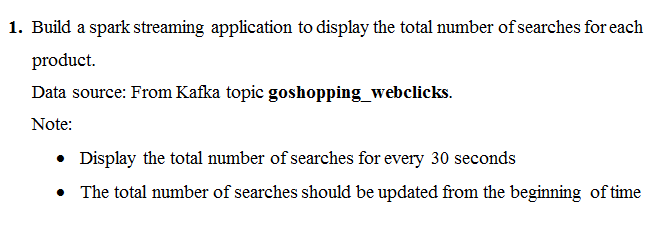


Kafka to Hdfs :

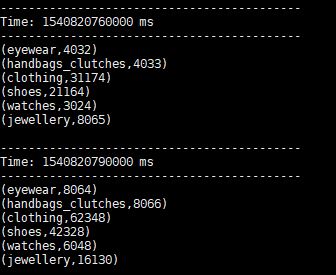
1. Spark streaming application to receive message from kafka topic and push it on to HDFS

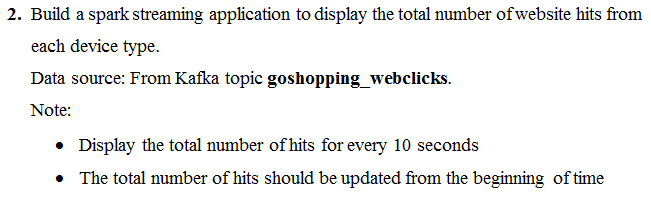
**SOURCE: Cassandra.txt**

Real Time Processing

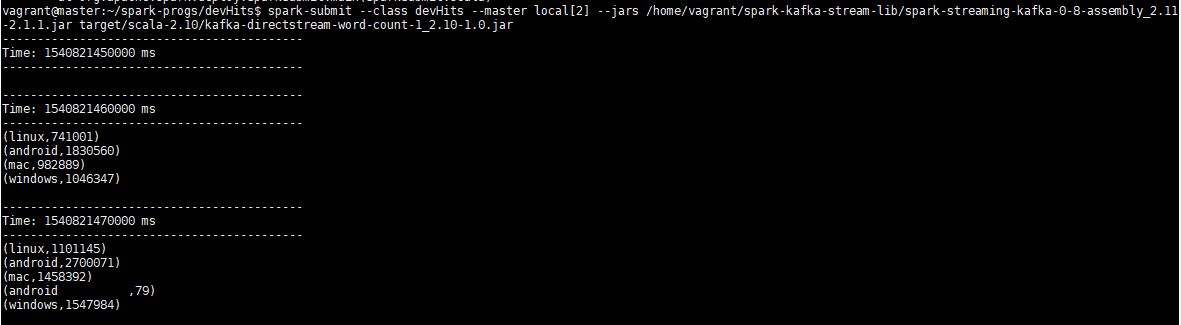


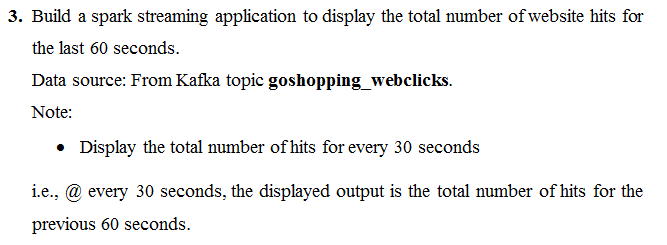
**SOURCE: Spark-Streaming.txt**



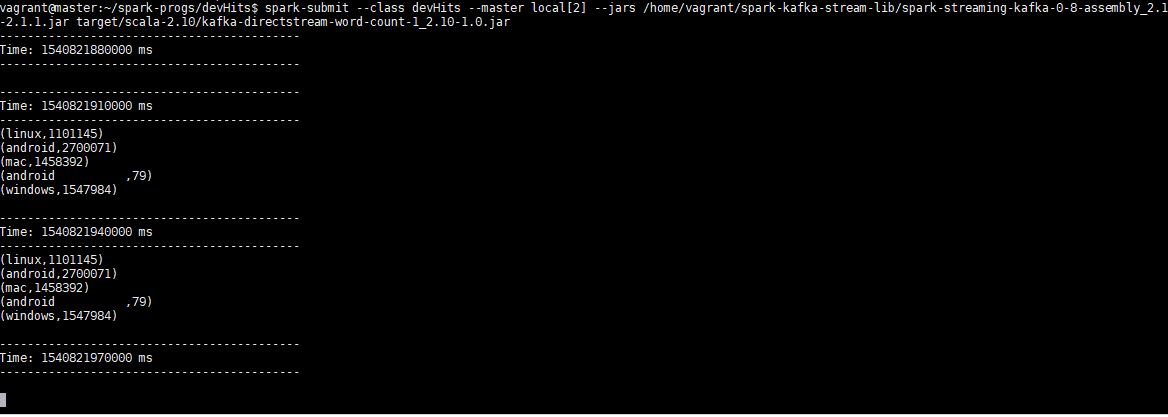


**SOURCE: Spark-Streaming.txt**

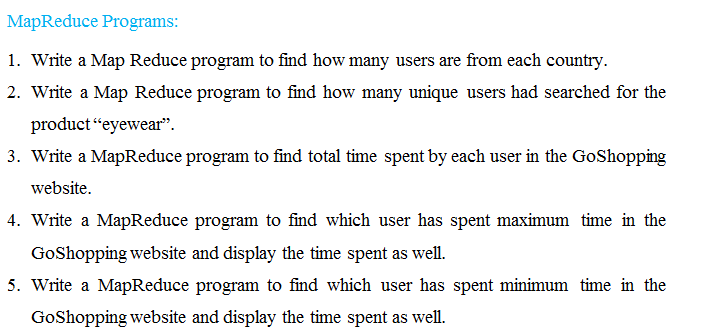




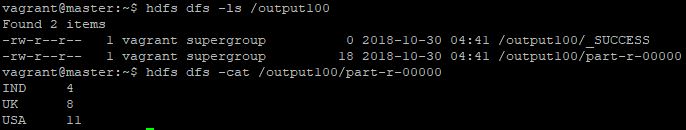
**SOURCE: Spark-Streaming.txt**



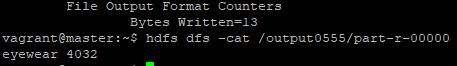
Batch Time Processing



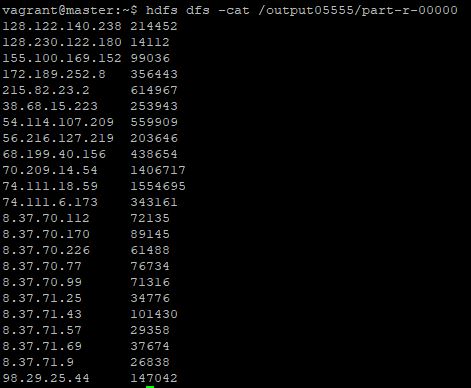
**SOURCE: Map\_reduce\_1.txt**



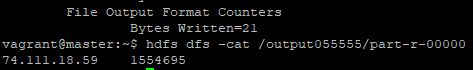
**SOURCE: Map\_reduce\_2.txt**



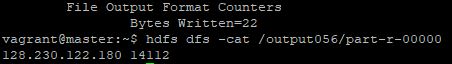
**SOURCE: Map\_reduce\_3.txt**

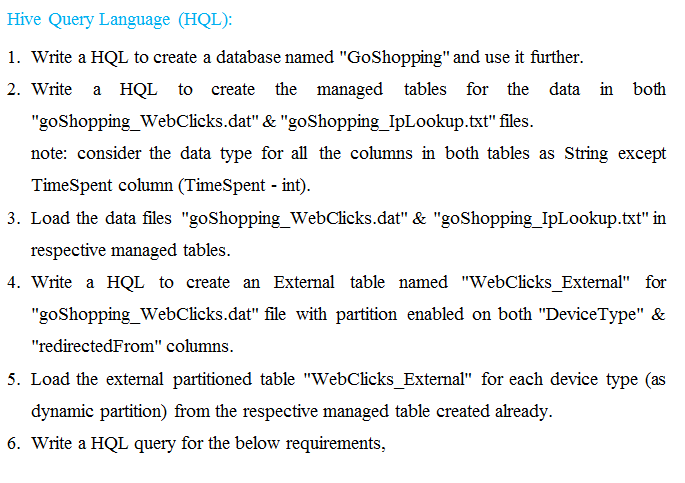


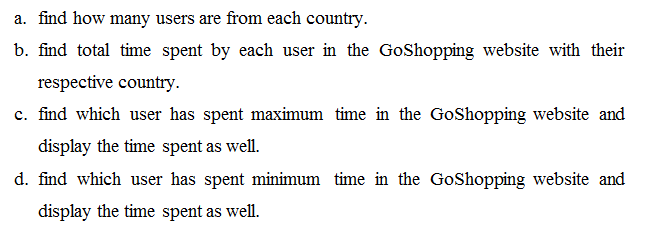
**SOURCE: Map\_reduce\_4.txt**



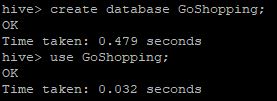
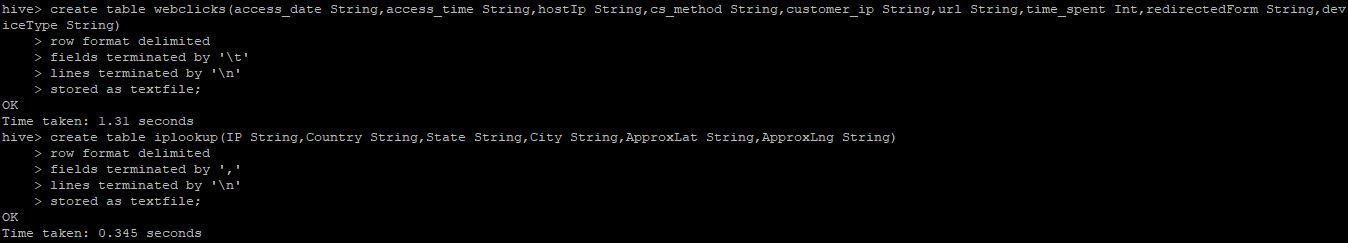
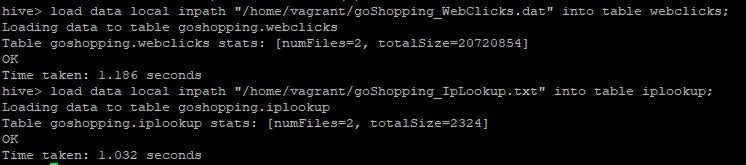
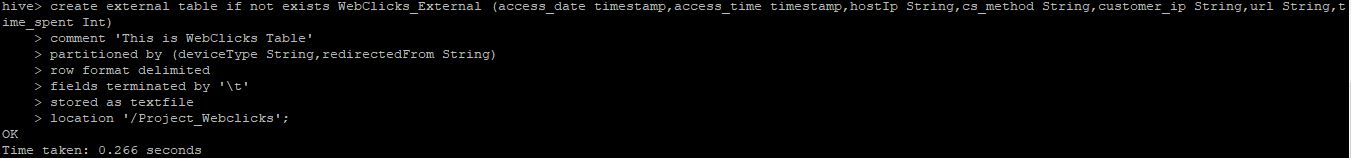
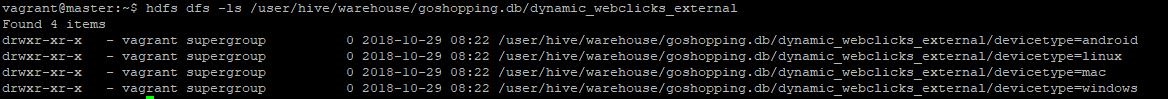
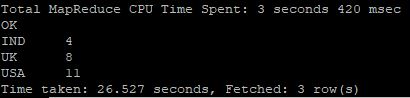
**SOURCE: Map\_reduce\_5.txt**

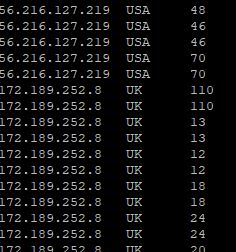


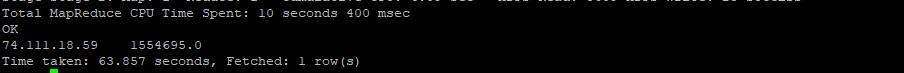


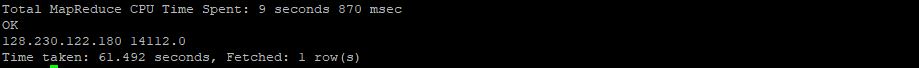


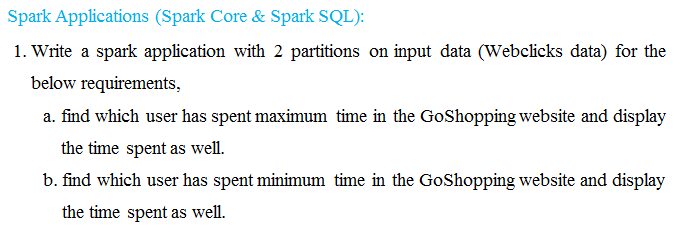
**SOURCE: HQL.txt**

1. 
2. 
3. 
4. 
5. 
6. a. 

b. 

c. 

d. 



**SOURCE: Spark-core-1.txt**

**a.**

**b.**



