**CS523 - Big Data Technology Project**

**TeamMembers**

Suresh Prajapati(109524)

Sovichea Cheth(985421)

Sunena Gwachha(109578)

**Part 1: Spark Streaming**

**#Spark installation**Step 1: <http://download.nextag.com/apache/spark/spark-2.3.0/spark-2>.3.0-bin-hadoop2.7.tgz  
Step 2: tar -xvzf spark-2.3.0-bin-hadoop2.7.tgz  
Step 3: mv spark-2.3.0-bin-hadoop2.7 /usr/local/spark

**Part 2: HBase through Spark SQL**

Step 1: Import library files

Step 2: Convert bytes to string

Step 3: Create temporary table YouTube

Step 4: Run spark SQL query

**Part 3: Data Visualization**

**#Zeppelin installation**Step 1: Download Zeppelin http://www.apache.org/dyn/closer.cgi/zeppelin/zeppelin-0.7.3/zeppelin-0.7.3-bin-all.tgz

Step 2: Extract and put in any directory

Step 3: Set up SPARK\_HOME and JAVA\_HOME in .bashrc

Step 4: Update Zeppelin config files

Zeppelin-env.sh zeppelin-site.xml

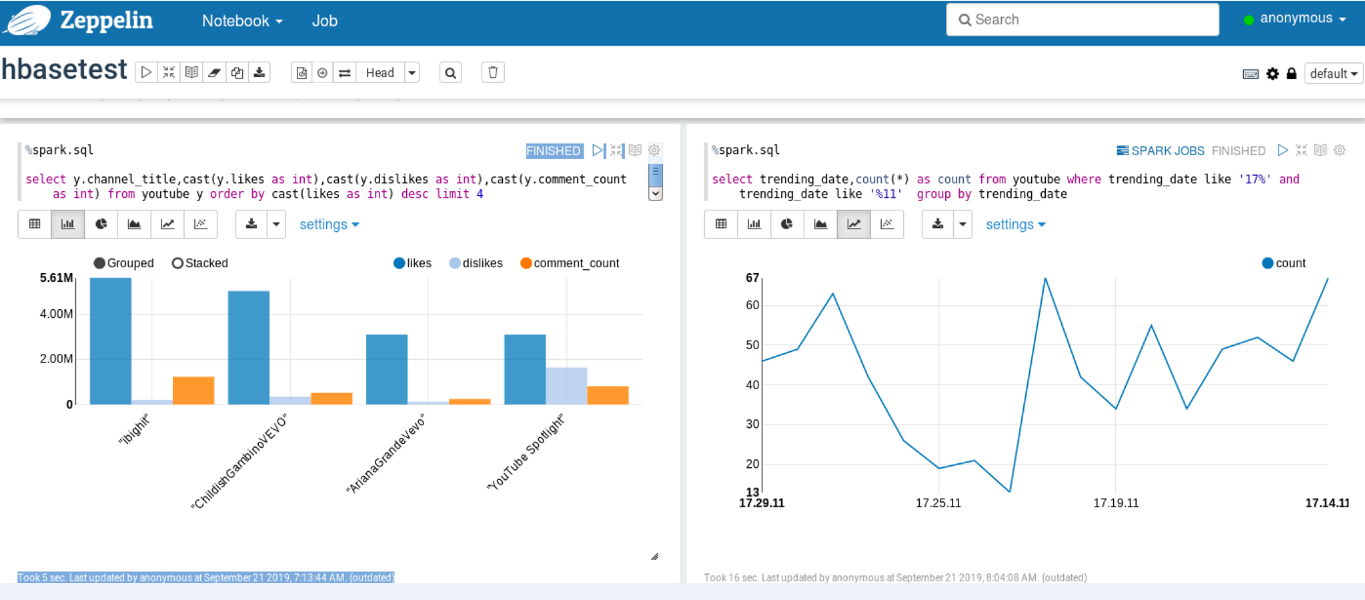
Step 5: Start Zeppelin

Sudo bin/zeppelin-daemon.sh start

Step 6: Open localhost:8080 will display Zeppelin

Step 7: Import HBase data, run the query and observe the result

**Screenshot of Query1 and Query2**



**Query 1**

%spark.sql

select y.channel\_title,cast(y.likes as int),cast(y.dislikes as int),cast(y.comment\_count as int)

from youtube y

order by cast(likes as int) desc

limit 4

**Query 2**

%spark.sql

select trending\_date,count(\*) as count

from youtube

where trending\_date like '17%' and trending\_date like '%11'

by trending\_datePart

**4: Kafka**

**#Kafka installation**Step 1: http://mirrors.sorengard.com/apache/kafka/1.0.1/kafka\_2.12-1.0.1. tgz  
Step 2: tar -xvf kafka\_2.12-1.0.1.tgz  
Step 3: edit .bash\_profile to add 2 line below  
 export KAFKA\_HOME=<location of kafka>  
 export PATH=$PATH:$KAFKA\_HOME/bin  
Step 4: start zookeeper   
 ./bin/zookeeper-server-start.sh config/zookeeper.properties  
Step 5: start kafka  
 ./bin/kafka-server-start.sh config/server.properties  
Step 6: create topic  
 ./bin/kafka-topics.sh --zookeeper localhost:2181 --create --topic --replication- factor 1 -partitions 1 -topic topicname  
Step 7: list topics  
 ./bin/kafka-topics.sh --list --zookeeper --localhost:2181

Step 8: Connect to Zookeeper instance

zookeeper-shell.sh localhost:2181 ls /brokers/topics

Step 9: Remove the topic folder from ZooKeeper

rmr /brokers/topics/yourtopic

Step 10: Start Producer

bin/kafka-console-producer.sh --broker-list localhost:9092 --topic YTVideo

Step 11: Start Consumer

kafka-console-consumer.sh --zookeeper localhost:2181 -topic YTVideo --from-beginning

Testing Impala on Java Demo project:

Created the Java Runnable jar file. This file will able to connect to hive database via JDBC driver according to our project. Just providing the SQL statement after the jar file, you will be able to run the SQL query from this Java file.