**Maharishi University of Management**

**CS532 – Big Data Technology**

**Instructor Prof. Mrudula Mukadam**

**Project report**

**Spark Streaming With Kafka**

**Group name: Penguin**

**Members:**

**Hoang Duy Vu – 986170**

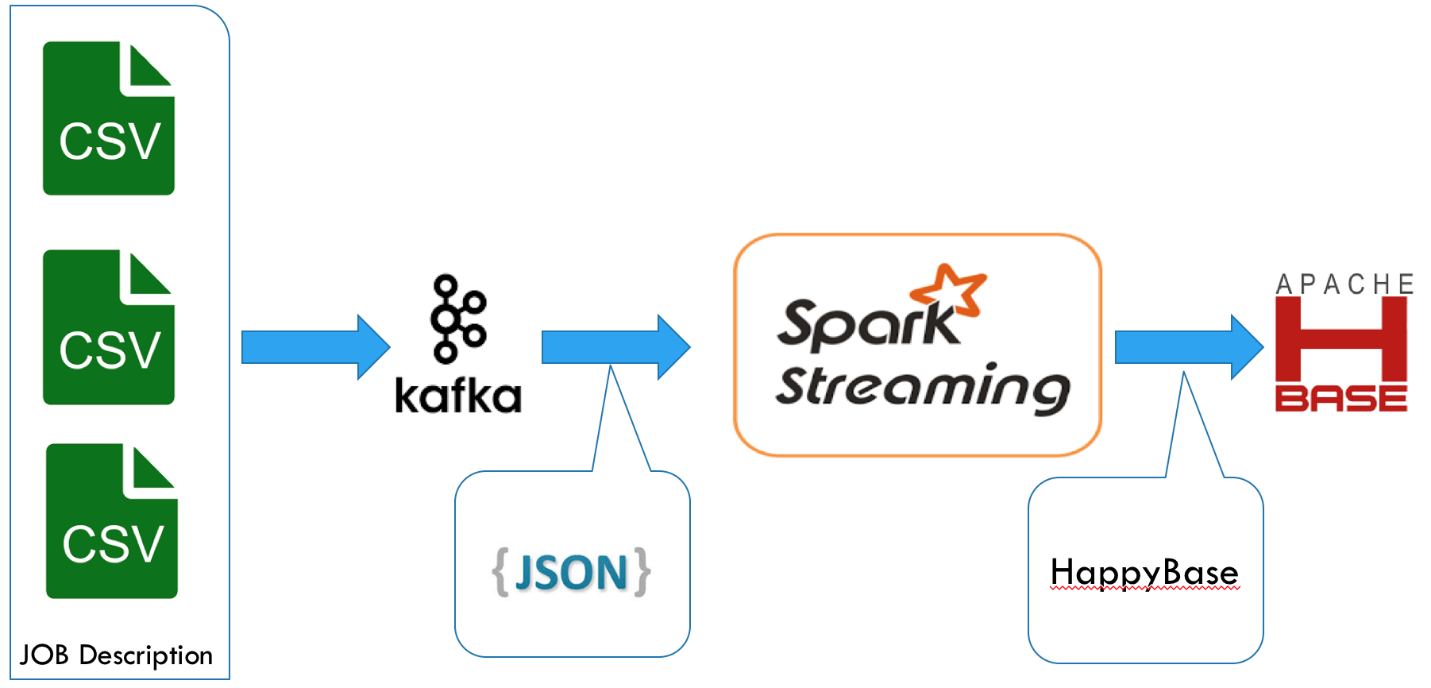
**Nguyen Trung Vo – 986241**

Mar – 19 - 2018

**Overview**

Project is using Kafka, Spark Streaming, HBase, Hive, Spark SQL, Jupyter for analysis Jobs requirement data set.

Data flow streaming required for question 1, 4



Data query and visualization flow for question 2, 3



**Technologies are used**

Cloudera Express 5.13.0

Java 1.8.0\_121

Anaconda 4.4 (Python 2.7)

Spark 2.2

Kafka

**Step by Step**

**Prepare environment**

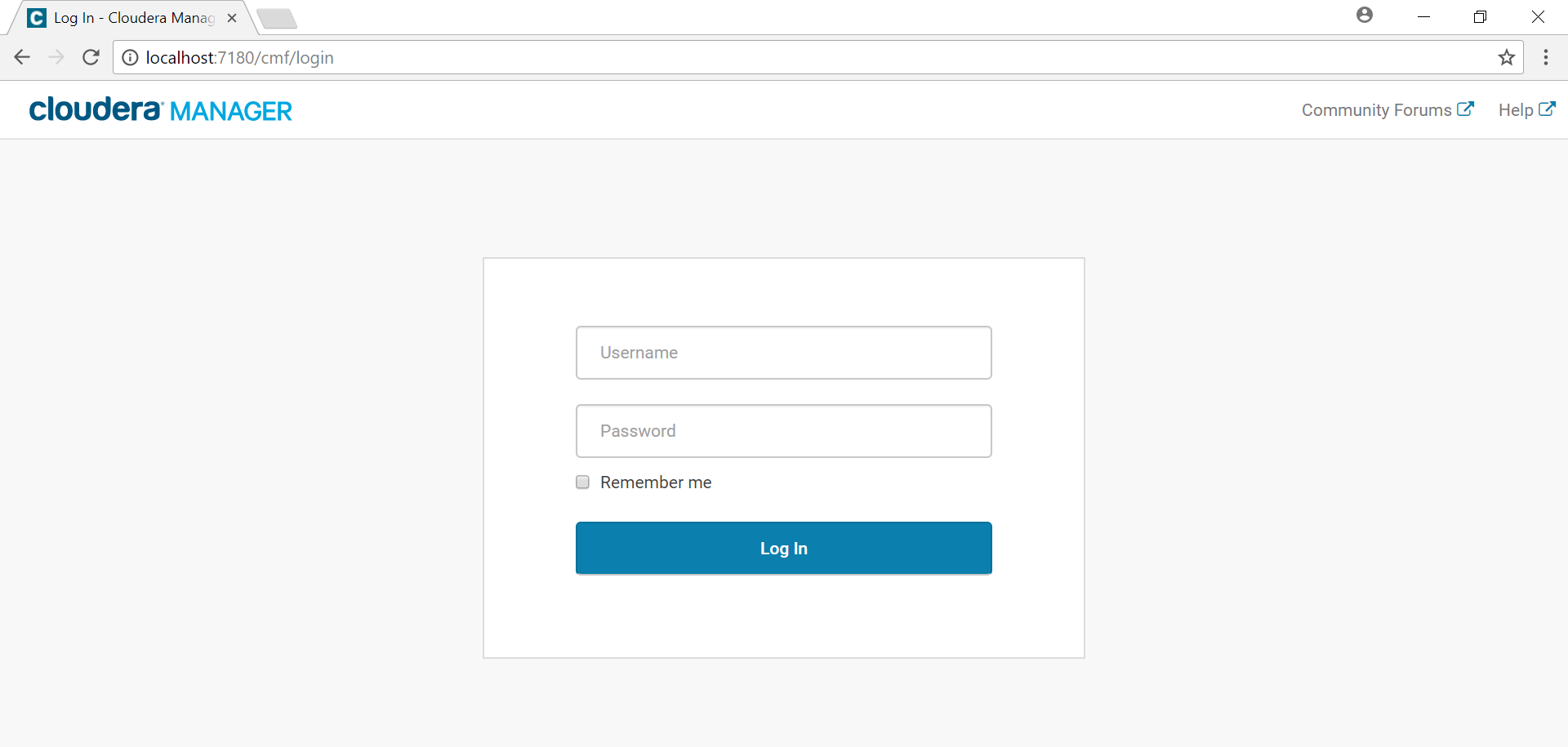
Data set, script files put on location /home/cloudera/

(job\_skills.csv, Visualization.ipynb, SparkStreaming.py, KafkaProducer.py)

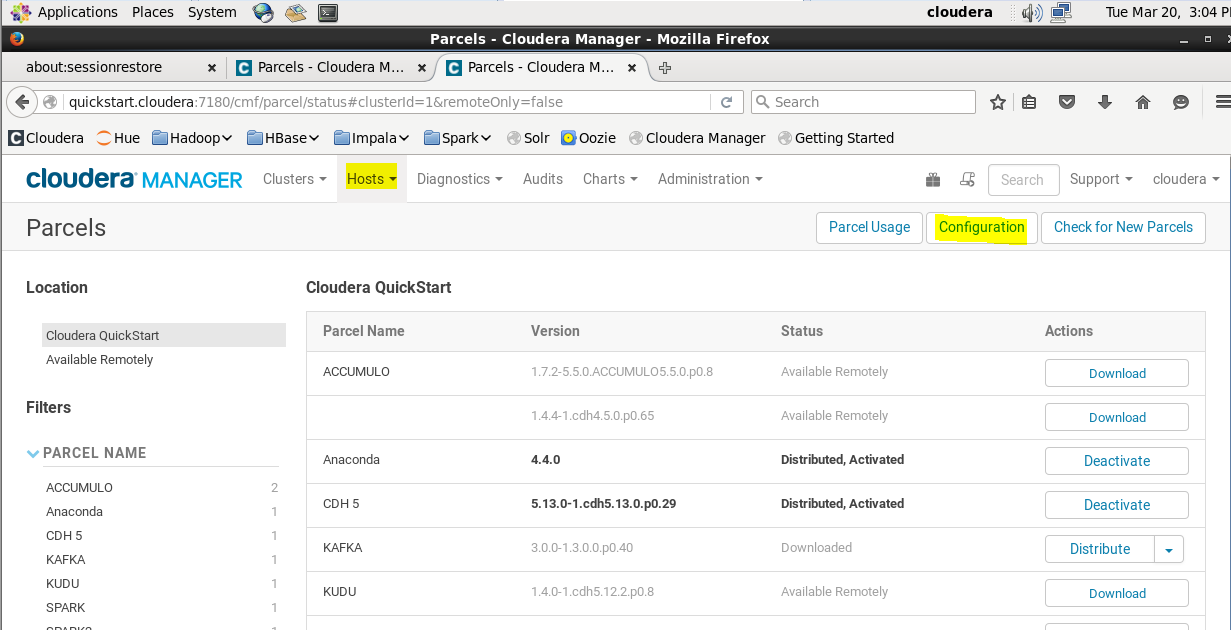
Need to **sudo** **chmod 777** **/tmp/hive** for query data on Jupyter Notebook

**Upgrade CDH Manager**

Login to Cloudera manager

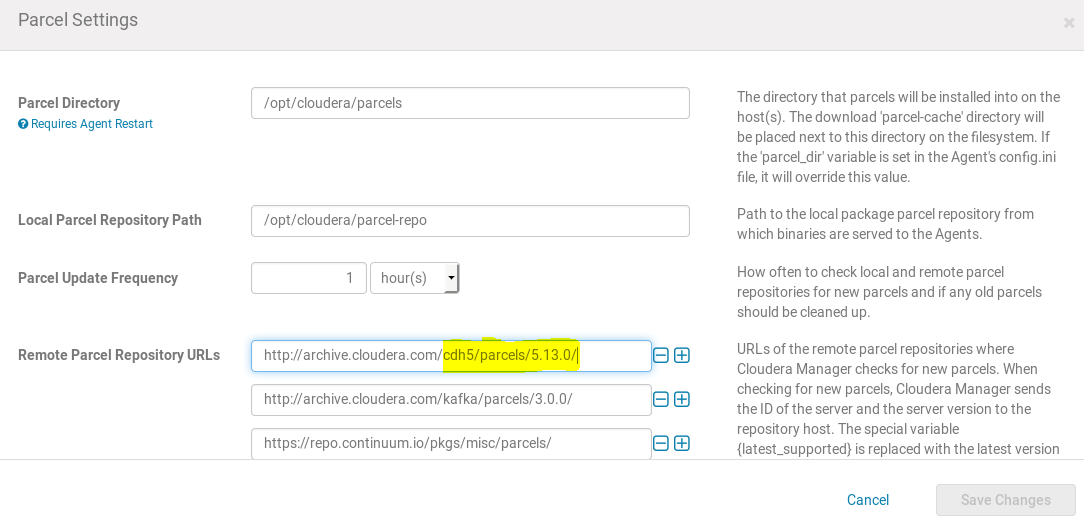


Click to Host, then click to Configuration



Add the repository CDH

<http://archive.cloudera.com/cdh5/parcels/5.13.0/>



Save change, back to the previous screen, click download then Distribute, then Activate

Restart server if needs.

**Upgrade Spark**

Same process above for upgrade Spark 2 by add remote parcel repositories URL below

<http://archive.cloudera.com/spark2/parcels/2.2.0.cloudera2/>

**Install Anaconda**

Same process above for install Anaconda in CDH by add remote parcel repositories URL below

<https://repo.continuum.io/pkgs/misc/parcels/>

**Install Kafka server**

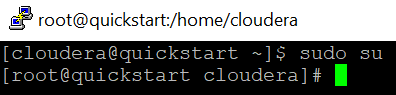
$ sudo wget <http://mirrors.sorengard.com/apache/kafka/1.0.1/kafka_2.12-1.0.1.tgz>

$ sudo tar -xzvf kafka\_2.12-1.0.1.tgz

$ sudo mv kafka\_2.12-1.0.1 /usr/local/kafka

**Install other modules requirement**

Change to root account before run below commands



$ sudo su

# pip install --upgrade pip

# pip install happlybase

# pip install pyspark

# pip install kafka

**Create table on HBase**

$ hbase shell

hbase(main):001:0> create 'jobs', 'info'

**Create external table for Hive**

$ hive

hive >

create external table job\_python\_hbase(rowid String,company String,title String,category String,location String,responsibility String,minimum String,prefer String)

STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' with serdeproperties("hbase.columns.mapping"=":key,info:company,info:title,info:category,info:location,info:responsibility,info:minimum,info:prefer")

tblproperties("hbase.table.name"="jobs");

**Ready to run**

1. Start Kafka server

$ cd /usr/local/kafka

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

1. Submit job for streaming

$ spark2-submit --master local SparkStreaming.py

1. Start Producer

$ python KafkaProducer.py

1. Start Jupyter notebook for data queries and visualization.

$ jupyter notebook

Open book Visualization.ipynb

