Downloaded File from

<https://www.kaggle.com/datasnaek/youtube-new>

<https://gist.github.com/dgp/1b24bf2961521bd75d6c>

Import data into HDFS:

hduser@naren:~$ hadoop fs -ls /

Found 4 items

drwxr-xr-x - hduser supergroup 0 2018-07-18 20:21 /WordCountTutorial

drwxr-xr-x - hduser supergroup 0 2018-07-20 09:42 /mydata

drwx------ - hduser supergroup 0 2018-07-19 23:39 /tmp

drwxr-xr-x - hduser supergroup 0 2018-07-19 23:42 /user

hduser@naren:~$ hadoop fs -mkdir /Youtube

hduser@naren:~$ hadoop fs -ls /

Found 5 items

drwxr-xr-x - hduser supergroup 0 2018-07-18 20:21 /WordCountTutorial

drwxr-xr-x - hduser supergroup 0 2018-08-07 17:56 /Youtube

drwxr-xr-x - hduser supergroup 0 2018-07-20 09:42 /mydata

drwx------ - hduser supergroup 0 2018-07-19 23:39 /tmp

drwxr-xr-x - hduser supergroup 0 2018-07-19 23:42 /user

hduser@naren:~$ hadoop fs -put /home/hduser/Downloads/Youtube/GBvideos.csv /Youtube

Found 5 items

drwxr-xr-x - hduser supergroup 0 2018-07-18 20:21 /WordCountTutorial

drwxr-xr-x - hduser supergroup 0 2018-08-07 18:01 /Youtube

drwxr-xr-x - hduser supergroup 0 2018-07-20 09:42 /mydata

drwx------ - hduser supergroup 0 2018-07-19 23:39 /tmp

drwxr-xr-x - hduser supergroup 0 2018-07-19 23:42 /user

hduser@naren:~$ hadoop fs -ls /Youtube

Found 1 items

-rw-r--r-- 1 hduser supergroup 53213441 2018-08-07 18:01 /Youtube/GBvideos.csv

hduser@naren:~$ hadoop fs -put /home/hduser/Downloads/Youtube/DEvideos.csv /Youtube

hduser@naren:~$ hadoop fs -put /home/hduser/Downloads/Youtube/USvideos.csv /Youtube

hduser@naren:~$ hadoop fs -put /home/hduser/Downloads/Youtube/CAvideos.csv /Youtube

hduser@naren:~$ hadoop fs -put /home/hduser/Downloads/Youtube/FRvideos.csv /Youtube

hduser@naren:~$ hadoop fs -ls /Youtube

Found 5 items

-rw-r--r-- 1 hduser supergroup 64067991 2018-08-07 18:02 /Youtube/CAvideos.csv

-rw-r--r-- 1 hduser supergroup 63040138 2018-08-07 18:02 /Youtube/DEvideos.csv

-rw-r--r-- 1 hduser supergroup 51424708 2018-08-07 18:02 /Youtube/FRvideos.csv

-rw-r--r-- 1 hduser supergroup 53213441 2018-08-07 18:01 /Youtube/GBvideos.csv

-rw-r--r-- 1 hduser supergroup 62756152 2018-08-07 18:02 /Youtube/Usvideos.csv

Architecture used can be explained by using diagram given below: -

Output Stored

Analysis

Programming, Querying, Visualization

DATA

SOURCES

STAGE

Results

Evaluation

Operational Database

HDFS

MySQL

Mongo DB

Cassandra

HBase

Spark

Output

RDD



Local

Sqoop

MySQL

Output

Pig

Hive

ETL

ETL



|  |
| --- |
| Video\_id |
| trending\_date |
| title |
| channel\_title |
| category\_id |
| publish\_time |
| tags |
| views |
| likes |
| dislikes |
| comment\_count |
| thumbnail\_link |
| comments\_disabled |
| ratings\_disabled |
| video\_error\_or\_removed |
| description |

|  |
| --- |
| Category\_Id |
| Category\_name |