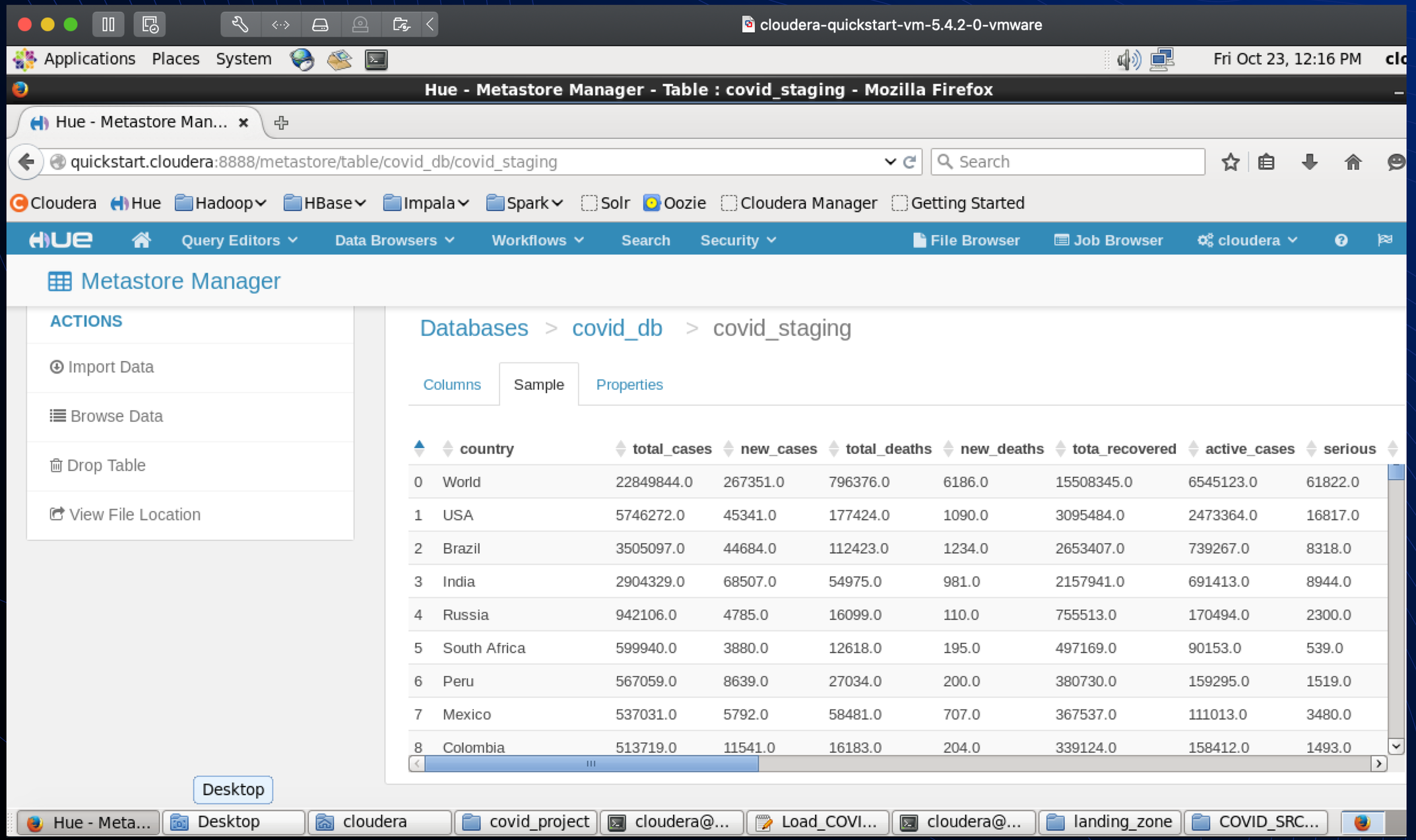
Omar Hazem Mohamed — 2020

SPRINTS GRADUATION PROJECT

LOADING DATA AND HDFS

- *Get the covid19.csv file into virtual machine*
  - *I use macbook so i used drag and drop on my VMware fusion Virtual machine to copy files from my native OS to virtual machine*
- *Create directory "/home/cloudera/covid_project/landing_zone/COVID_SRC_LZ" on vm and add covid19.csv file on it*
- *Add the dataset from "COVID_SRC_LZ" to HDFS directory "/user/cloudera/ds/COVID_HDFS_LZ" by using the script* "Load_COVID_TO_HDFS.sh"

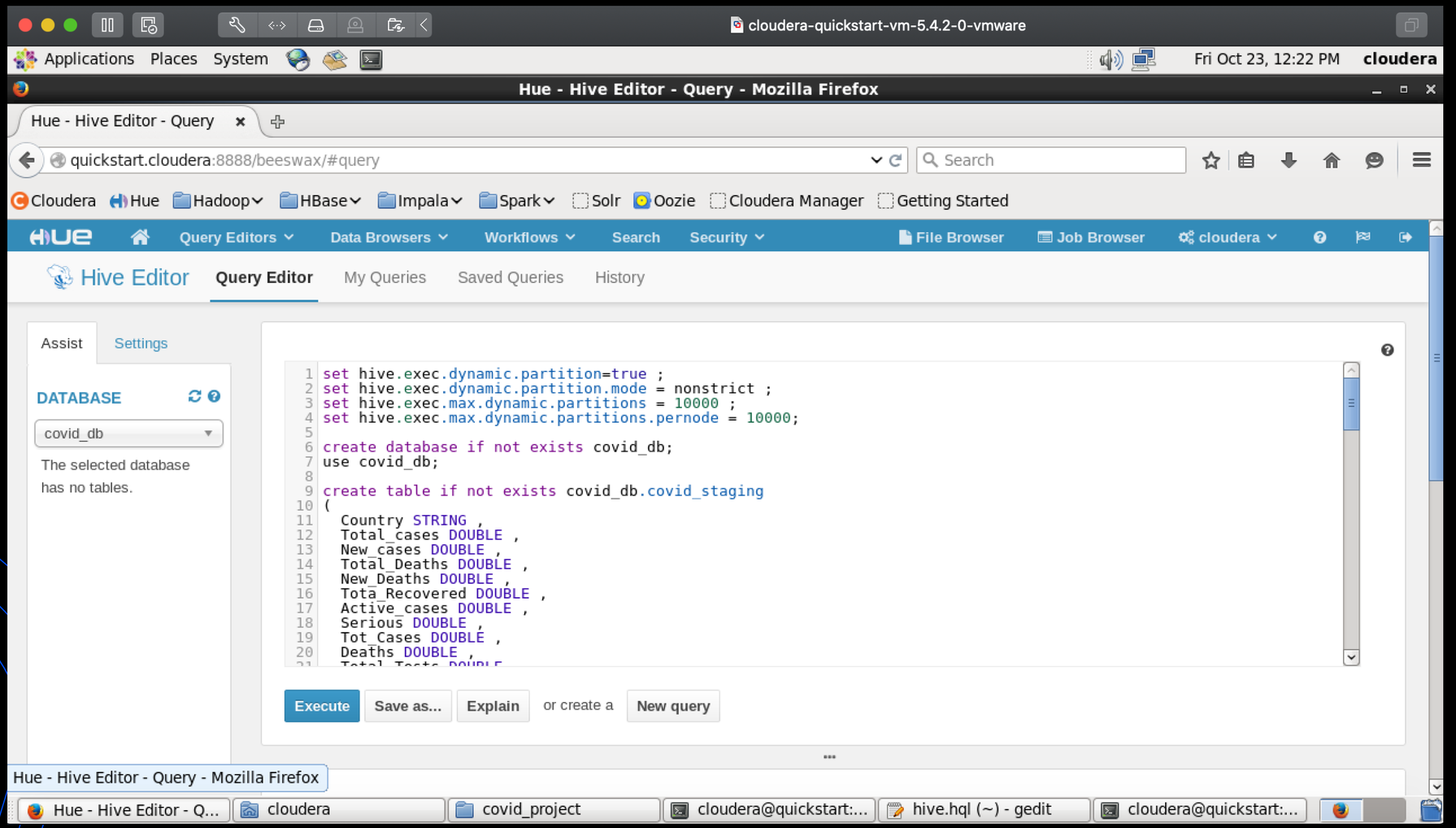


	country	total_cases	new_cases	total_deaths	new_deaths	tota_recovered	active_cases	serious
0	World	22849844.0	267351.0	796376.0	6186.0	15508345.0	6545123.0	61822.0
1	USA	5746272.0	45341.0	177424.0	1090.0	3095484.0	2473364.0	16817.0
2	Brazil	3505097.0	44684.0	112423.0	1234.0	2653407.0	739267.0	8318.0
3	India	2904329.0	68507.0	54975.0	981.0	2157941.0	691413.0	8944.0
4	Russia	942106.0	4785.0	16099.0	110.0	755513.0	170494.0	2300.0
5	South Africa	599940.0	3880.0	12618.0	195.0	497169.0	90153.0	539.0
6	Peru	567059.0	8639.0	27034.0	200.0	380730.0	159295.0	1519.0
7	Mexico	537031.0	5792.0	58481.0	707.0	367537.0	111013.0	3480.0
8	Colombia	513719.0	11541.0	16183.0	204.0	339124.0	158412.0	1493.0

HIVE

- *Create database (covid_db) and different schema for each Hive loading stage*
  - *Hive staging table for pointing to dataset location to select data from*
  - *ORC table is partitioned by Cofinal reportuntry*
  - *final report output table to visualize in further steps*
- *I used hive.hql script to execute the above commands*
- *Note :*
  - *the provided .hql file was not running so I modified it becaues it had some typos and some missing commands*
  - *I needed to rewrite the whole commands all over again in texteditor in cloudera vm as they don't work if I copied them directly to vm*

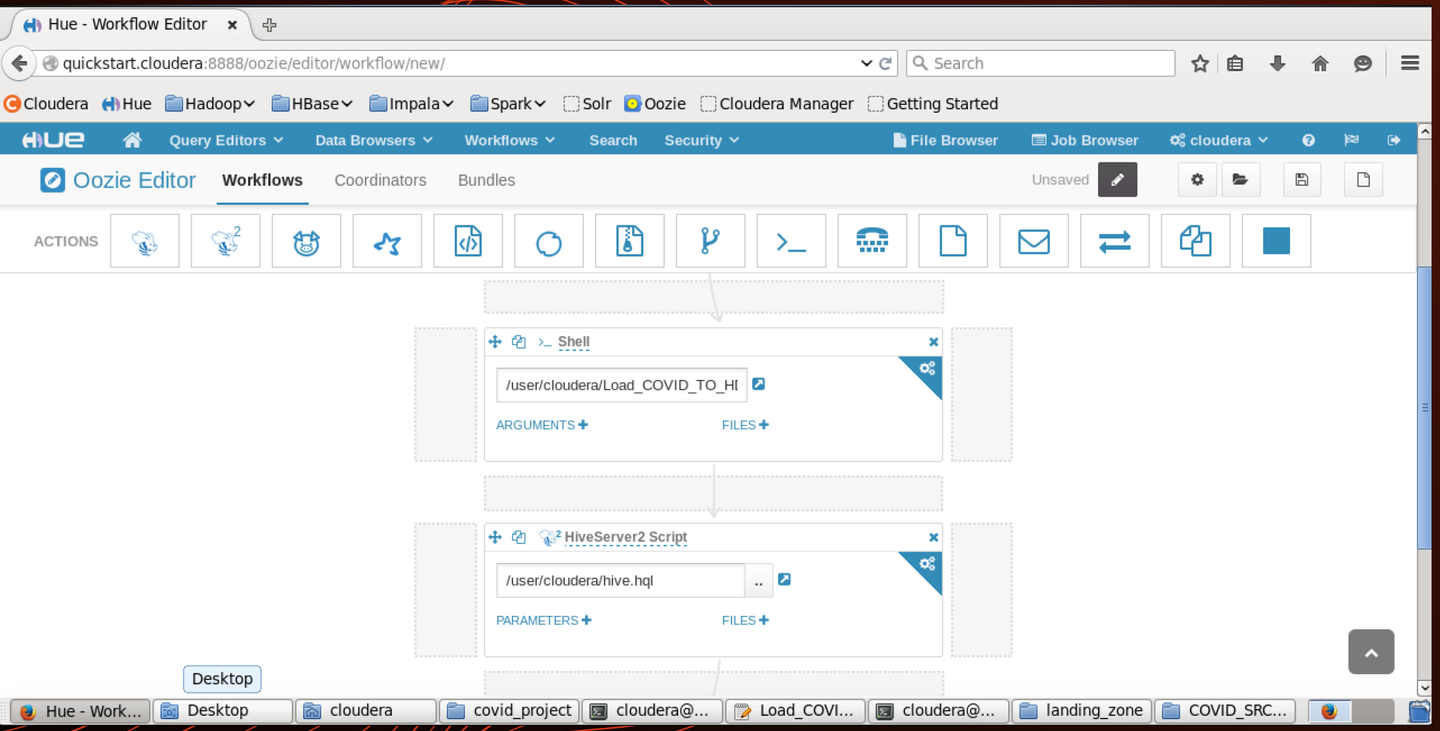
HIVE

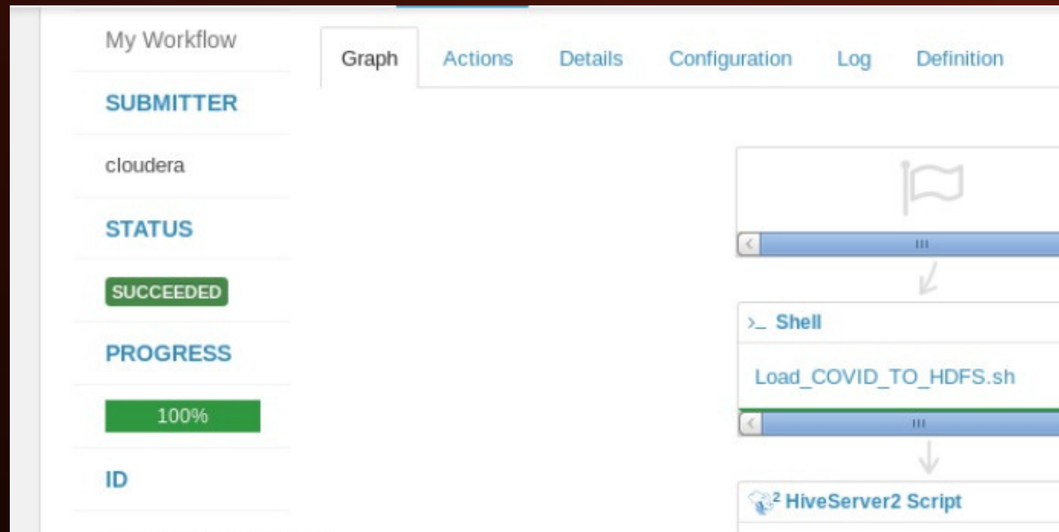

cloudera-quickstart-vm-5.4.2-0-vmware
Applications Places System
Fri Oct 23, 12:22 PM cloudera
Hue - Hive Editor - Query - Mozilla Firefox
Hue - Hive Editor - Query
quickstart.cloudera:8888/beeswax/#query
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloudera
Hive Editor Query Editor My Queries Saved Queries History
Assist Settings
DATABASE
covid_db
The selected database has no tables.
set hive.exec.dynamic.partition=true ;
set hive.exec.dynamic.partition.mode = nonstrict ;
set hive.exec.max.dynamic.partitions = 10000 ;
set hive.exec.max.dynamic.partitions.pernode = 10000;
create database if not exists covid_db;
use covid_db;
create table if not exists covid_db.covid_staging
(
Country STRING ,
Total_cases DOUBLE ,
New_cases DOUBLE ,
Total_Deaths DOUBLE ,
New_Deaths DOUBLE ,
Tota_Recovered DOUBLE ,
Active_cases DOUBLE ,
Serious DOUBLE ,
Tot_Cases DOUBLE ,
Deaths DOUBLE ,
Execute Save as... Explain or create a New query
Hue - Hive Editor - Query - Mozilla Firefox
Hue - Hive Editor - Q...
cloudera
covid_project
cloudera@quickstart:...
hive.hql (~) - gedit
cloudera@quickstart:...

OOZIE WORKFLOW

- I used oozie workflow from Hue to execute the terminal .sh command first "Load_COVID_TO_HDFS.sh" and then run the .hql hive script created from above step after "hive.hql" to get out the output files
- Note :
  - I ran the following script to get single csv ouptut file to use on Microsoft power Bi
  - (echo "TOP_DEATH,RANK_D,DEATHS, TOP_TEST,RANK_T,TESTS" > output.csv
    cat 000000_0_copy_2 >> output.csv
    cat 000000_0_copy_3 >> output.csv)

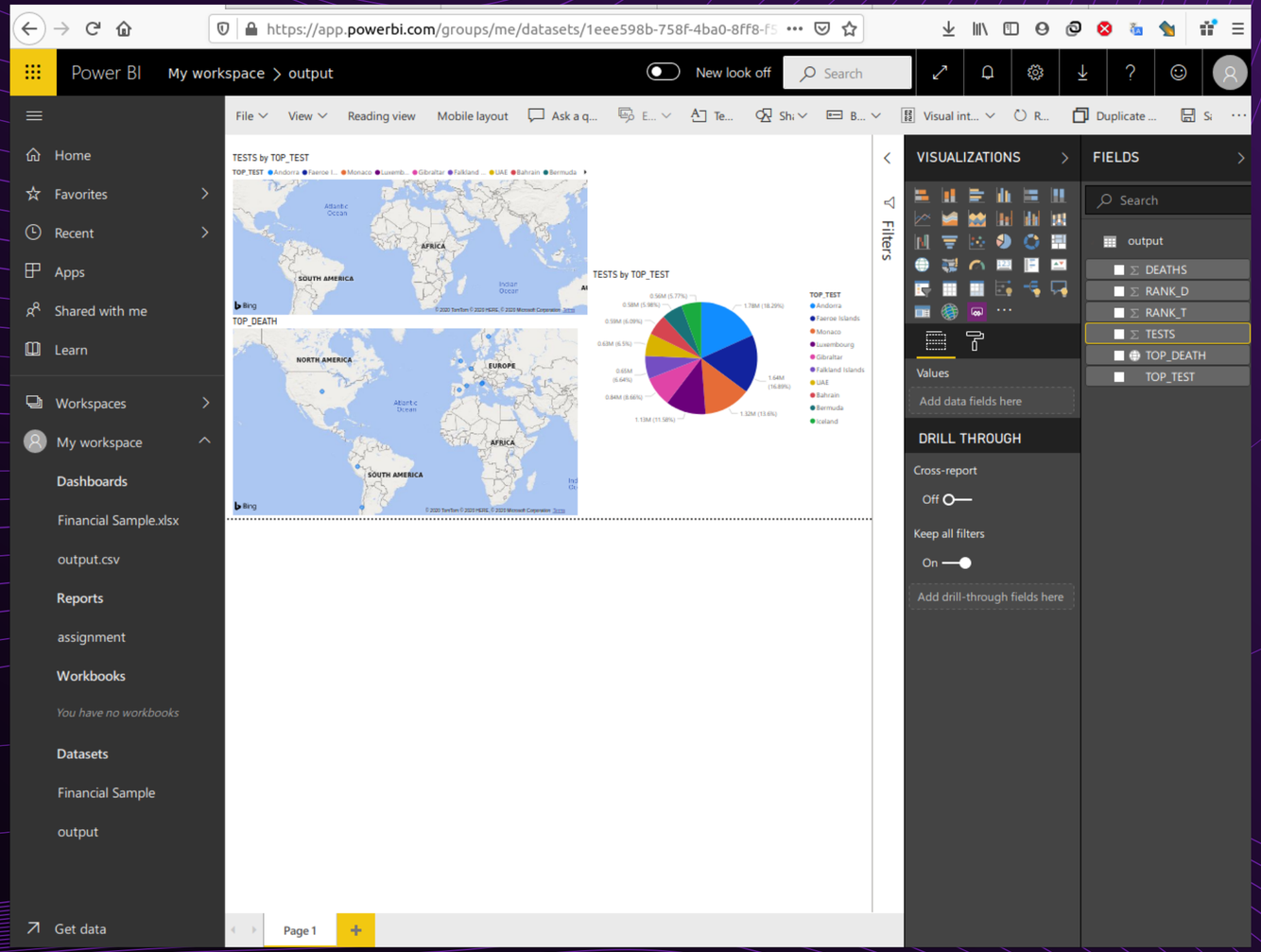
OOZIE WORKFLOW





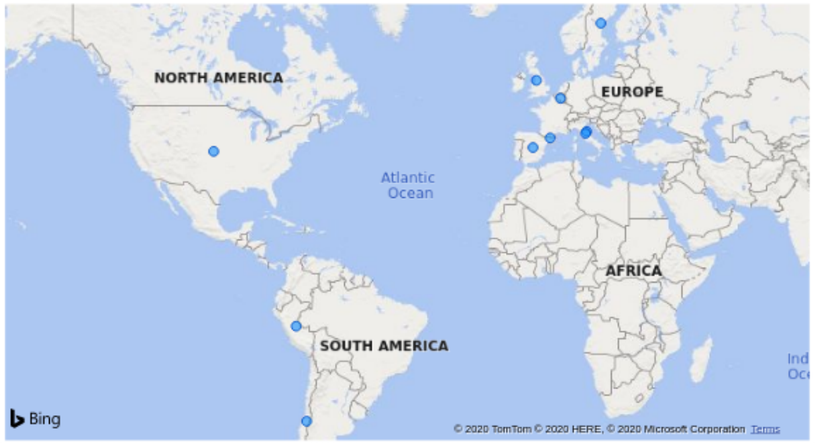
VISUALIZATION

- *Using Microsoft PowerBi to visualize the results extracted from output.csv*

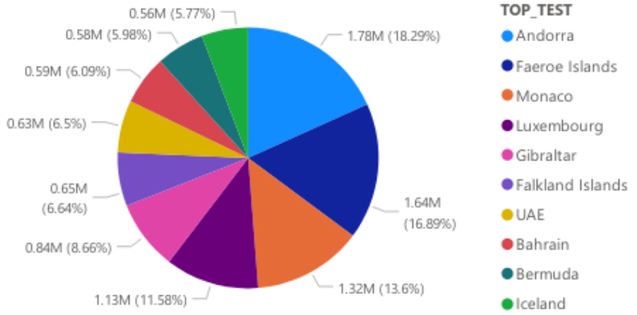


TESTS by TOP_TEST

TOP_DEATH



TESTS by TOP_TEST



Filters

VISUALIZATIONS

Values

Add data fields here

DRILL THROUGH

Cross-report

Off

Keep all filters

On

Add drill-through fields here

Page 1

06