

Data ingestion

Chris Xu qx428

Data Source

1. Economy by region

Data Size: 10 KB

Site:

<https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-economy.md>

has 5 columns:

```
location_key,  
gdp_usd,  
gdp_per_capita_usd,  
human_capital_index
```

2. COVID-19 Hospitalizations

Data Size: 63.4 MB

Site:

<https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-hospitalizations.md>

has 11 columns:

```
date,  
location_key,  
new_hospitalized_patients,  
cumulative_hospitalized_patients,  
current_hospitalized_patients,  
new_intensive_care_patients,  
cumulative_intensive_care_patients,  
current_intensive_care_patients,  
new_ventilator_patients,  
cumulative_ventilator_patients,  
current_ventilator_patients
```

Shell Command

```
javac -classpath `hadoop classpath` *.java
jar cvf CovidRegion.jar *.class
hadoop jar CovidRegion.jar CovidRegion project/hospitalizations.csv
project/economy.csv project/output1 project/output2
```

Process

Two jobs are done in my code. All keys are based on location. And only country data and US states data are used.

First job: cleaned up useless features and reordered the dataset by location key

Only the hospitalization dataset was used in this job. In the mapper, I only kept new and cumulative, hospitalized and intensive care patients numbers. Date and location key were set as keys. In the reducer, it simply output each row.

If we need to visualize the patient number changes over the time, this output dataset can be very helpful.

Second job: combined economy and hospitalization dataset by location key, and summed up daily increases of patients.

In the mapper, two files are mapped with the location key as keys. Values from Hospitalization have a prefix “CASES” and values from economy have “ECON”. In reducers, the number of patients from “new patient” column were summed up, and combined with economy data to one row.

This dataset give a clear relationship between hospitalization of covid and economies in different regions.

```
qx428_nyu_edu@nyu-dataproc-m1:~/project$ mv ../*.java .
qx428_nyu_edu@nyu-dataproc-m1:~/project$ hadoop fs -rm -r project/output*
Deleted project/output2
qx428_nyu_edu@nyu-dataproc-m1:~/project$ javac -classpath `hadoop classpath` *.java
qx428_nyu_edu@nyu-dataproc-m1:~/project$ jar cvf CovidRegion.jar *.class
added manifest
adding: CovidRegion.class(in = 1874) (out= 1037) (deflated 44%)
adding: CovidRegionDateMapper.class(in = 2072) (out= 876) (deflated 57%)
adding: CovidRegionDateReducer.class(in = 1473) (out= 606) (deflated 58%)
adding: CovidRegionOnlyMapper.class(in = 2424) (out= 1055) (deflated 56%)
adding: CovidRegionOnlyReducer.class(in = 2117) (out= 1229) (deflated 43%)
qx428_nyu_edu@nyu-dataproc-m1:~/project$ hadoop jar CovidRegion.jar CovidRegion project/hospitalizations.csv project/economy.csv project/output1 project/output2
2024-04-20 02:30:25,777 INFO client.RMProxy: Connecting to ResourceManager at nyu-dataproc-m/192.168.1.77:8032
2024-04-20 02:30:25,971 INFO client.AHSProxy: Connecting to Application History server at nyu-dataproc-m/192.168.1.77:10200
2024-04-20 02:30:26,150 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-04-20 02:30:26,165 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/qx428_nyu_edu/.staging/job_1704906963891_53649
2024-04-20 02:30:26,674 INFO mapreduce.JobSubmitter: number of splits: 1
2024-04-20 02:30:26,801 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1704906963891_53649
2024-04-20 02:30:26,804 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-20 02:30:27,042 INFO conf.Configuration: resource-types.xml not found
2024-04-20 02:30:27,042 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-04-20 02:30:27,466 INFO impl.YarnClientImpl: Submitted application application_1704906963891_53649
2024-04-20 02:30:27,505 INFO mapreduce.Job: The url to track the job: http://nyu-dataproc-m:8088/proxy/application_1704906963891_53649/
2024-04-20 02:30:27,506 INFO mapreduce.Job: Running job: job_1704906963891_53649
2024-04-20 02:30:36,658 INFO mapreduce.Job: Job job_1704906963891_53649 running in uber mode : false
2024-04-20 02:30:36,660 INFO mapreduce.Job: map 0% reduce 0%
2024-04-20 02:30:44,858 INFO mapreduce.Job: map 100% reduce 0%
2024-04-20 02:30:49,889 INFO mapreduce.Job: map 100% reduce 100%
2024-04-20 02:30:51,910 INFO mapreduce.Job: Job job_1704906963891_53649 completed successfully
2024-04-20 02:30:52,006 INFO mapreduce.Job: Counters: 54
File System Counters
```