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, hospitalied 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.