**Project Report :**

1. **Strategy of computation of new\_brazil\_covid19.csv** :

* brazil\_covid19.csv have date,region,state,cases,deaths columns while brazil\_covid19\_cities.csv have date,state,name,code,cases,deaths columns
* Select only date,state,cases,deaths columns from brazil\_covid19\_cities.csv and group based on date and state to find aggregate of cases and deaths
* brazil\_covid19\_cities.csv do not have region field. Select distinct state and region from brazil\_covid19.csv and join with brazil\_covid19\_cities.csv(which is now aggregated based on date and state) based on state to get region column.
* Save above dataframe to new\_brazil\_covid19.csv. Now this dataset will have same structure as brazil\_covid19.csv.

1. **strategy of comparison of new\_brazil\_covid19.csv and brazil\_covid19.csv** : **by assuming following -> Source: brazil\_covid19.csv , Destination: new\_brazil\_covid19.csv**

* load above 2 datasets and add 1 addition column “date\_state” which is a concatination of date and state. Say we have dataset as readSource and readDestination
* To get source count and destination count, use below
  + readSource.count and readDestination.count
* To get number of rows present in source but not in destination, use below
  + do left join readSource and readDestination based on date\_state and filter date\_state from readDestination df as null i.e. readSource.join(readDestination,readSource("date\_state") === readDestination("date\_state"),"left").filter(readDestination("date\_state").isNull)
* To get number of rows present in destination but not in source , use below
  + do left join readDestination and readSource based on date\_state and filter date\_state from readSource df as null i.e. readDestination.join(readSource,readSource("date\_state") === readDestination("date\_state"),"left").filter(readSource("date\_state").isNull)
* To get number of rows identical in source and destination - sameCasesAndDeaths
  + Do inner join between readDestination and readSource based on date\_state,cases,deaths
* To get number of rows with different values in source and destination
  + Do inner join between readDestination and readSource based on date\_state
  + subtract its output with above output(sameCasesAndDeaths)

1. **report-diff data explanation:**

Report have 6 fields:

* **source\_count** – count of rows in brazil\_covid19.csv
* **destination\_count** – count of rows in new\_brazil\_covid19.csv
* **data\_in\_source\_but\_not\_in\_destination** – count of brazil\_covid19.csv records which is not available in new\_brazil\_covid19.csv
* **data\_in\_destination\_but\_not\_in\_source** – count of new\_brazil\_covid19.csv records which is not available in brazil\_covid19.csv (it’s 0 because brazil\_covid19.csv is suprset of newbrazil\_covid19.csv)
* **data\_same\_in\_source\_destination** – records in both data set which have same number of cases and deaths.
* **data\_different\_in\_source\_destination** - records in both data have different number of cases or deaths.

**<Architecture diagram on next page below>**

**4. Architechture:**

Diagram, schematic

Description automatically generated