Loading data, getting list of countries/continents

--load data

hadoop fs -mkdir -p CA675-Assignment-2

hadoop fs -put ~/Desktop/Covid\_CSVs/csv CA675-Assignment-2

hadoop fs -ls CA675-Assignment-2/csv

pig

CaseCSV = LOAD 'CA675-Assignment-2/csv/coronavirus.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (date:chararray, province:chararray, country:chararray, lat:chararray, long:chararray, type:chararray, cases:int);

VaccineCSV = LOAD 'CA675-Assignment-2/csv/covid19\_vaccine.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (country:chararray, date:chararray, doses:int, people\_partially\_vaccinated:int, people\_fully\_vaccinated:int, report\_date\_string:chararray, uid:int, province\_state:chararray, iso2:chararray, iso3:chararray, code3:int, fips:chararray, lat:chararray, long:chararray, combined\_key:chararray, population:int, continent\_name:chararray, continent\_code:chararray);

--Get list of distinct countries for vaccine/cases data

Case\_countries = FOREACH CaseCSV generate country;

Vaccine\_Countries = FOREACH VaccineCSV generate country;

Case\_Vaccine\_Countries = union Case\_countries, Vaccine\_Countries;

Case\_Vaccine\_Distinct\_Countries = Distinct Case\_Vaccine\_Countries;

STORE Case\_Vaccine\_Distinct\_Countries INTO 'CA675-Assignment-2/csv/Case\_Vaccine\_Distinct\_Countries' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',' ,'NO\_MULTILINE' ,'UNIX', 'SKIP\_OUTPUT\_HEADER');

--add continents manually in spreadsheet

hadoop fs -put ~/Desktop/Covid\_CSVs/csv/Distinct\_Countries\_Continents.csv CA675-Assignment-2/csv

Countries\_continents = Load 'CA675-Assignment-2/csv/Distinct\_Countries\_Continents.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (country:chararray, continent:chararray);

Deaths, recovered, confirmed cases by Continent

CaseCSV = LOAD 'CA675-Assignment-2/csv/coronavirus.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (date:chararray, province:chararray, country:chararray, lat:chararray, long:chararray, type:chararray, cases:int);

--Format date, remove unnecessary columns

Cases\_pruned= FOREACH CaseCSV generate country, cases, type, ToDate(date, 'MM/dd/yyyy') as (Formatted\_Date:DateTime);

--Filter by case type, we will later join to have all 3 case type data dor each country/date in one relation

Confirmed\_filtered = Filter Cases\_pruned BY type == 'confirmed';

Deaths\_filtered= Filter Cases\_pruned BY type == 'death';

Recovered\_filtered= Filter Cases\_pruned BY type == 'recovered';

--label fields with case type

Country\_date\_confirmed = FOREACH Confirmed\_filtered generate country, Formatted\_Date as date, cases as confirmed;

Country\_date\_deaths = FOREACH Deaths\_filtered generate country, Formatted\_Date as date, cases as deaths;

Country\_date\_recovered = FOREACH Recovered\_filtered generate country, Formatted\_Date as date, cases as recovered;

--join deaths and confirmed cases, label fields clearly

Deaths\_confirmed\_join = JOIN Country\_date\_confirmed BY (country, date), Country\_date\_deaths by (country, date);

Deaths\_confirmed = FOREACH Deaths\_confirmed\_join Generate Country\_date\_confirmed::country AS country,

Country\_date\_confirmed::date AS date,

Country\_date\_confirmed::confirmed AS confirmed,

Country\_date\_deaths::deaths AS deaths;

--Join deaths&confirmed cases with recovered cases, label fields

--this will give us schema (country, date, confirmed\_cases, deaths, recovered)

Deaths\_confirmed\_recovered\_join = JOIN Deaths\_confirmed BY (country, date), Country\_date\_recovered BY (country, date);

Deaths\_confirmed\_recovered\_country = FOREACH Deaths\_confirmed\_recovered\_join GENERATE Deaths\_confirmed::country AS country,

Deaths\_confirmed::date AS date,

Deaths\_confirmed::confirmed AS confirmed\_cases,

Deaths\_confirmed::deaths AS deaths,

Country\_date\_recovered::recovered AS recovered;

--Join with Countries\_continents data, clarify fields

Countries\_continents = Load 'CA675-Assignment-2/csv/Distinct\_Countries\_Continents.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (country:chararray, continent:chararray);

Deaths\_confirmed\_recovered\_country\_continent\_join = Join Deaths\_confirmed\_recovered\_country BY country, Countries\_continents BY country;

Deaths\_confirmed\_recovered\_country\_continent = FOREACH Deaths\_confirmed\_recovered\_country\_continent\_join GENERATE

Deaths\_confirmed\_recovered\_country::country AS country,

Deaths\_confirmed\_recovered\_country::date AS date,

Deaths\_confirmed\_recovered\_country::confirmed\_cases AS confirmed\_cases,

Deaths\_confirmed\_recovered\_country::deaths AS deaths,

Deaths\_confirmed\_recovered\_country::recovered AS recovered,

Countries\_continents::continent AS continent;

Cases\_grouped\_continent = GROUP Deaths\_confirmed\_recovered\_country\_continent BY (continent, date);

--Generate counts of each case type for each continent, date pair

Cases\_count\_continent\_flattened= FOREACH Cases\_grouped\_continent GENERATE FLATTEN(group),

SUM (Deaths\_confirmed\_recovered\_country\_continent.confirmed\_cases) AS confirmed\_cases,

SUM (Deaths\_confirmed\_recovered\_country\_continent.deaths) AS deaths,

SUM (Deaths\_confirmed\_recovered\_country\_continent.recovered) AS recovered;

Cases\_count\_continent = FOREACH Cases\_count\_continent\_flattened GENERATE

group::continent AS continent, group::date AS date, confirmed\_cases, deaths, recovered;

STORE Cases\_count\_continent INTO 'CA675-Assignment-2/csv/Cases\_count\_continent5' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',' ,'NO\_MULTILINE' ,'UNIX', 'WRITE\_OUTPUT\_HEADER');

Partially and fully vaccinated by Continent, date

VaccineCSV = LOAD 'CA675-Assignment-2/csv/covid19\_vaccine.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (country:chararray, date:chararray, doses:int, people\_partially\_vaccinated:int, people\_fully\_vaccinated:int, report\_date\_string:chararray, uid:int, province\_state:chararray, iso2:chararray, iso3:chararray, code3:int, fips:chararray, lat:chararray, long:chararray, combined\_key:chararray, population:int, continent\_name:chararray, continent\_code:chararray);

Countries\_continents = Load 'CA675-Assignment-2/csv/Distinct\_Countries\_Continents.csv' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (country:chararray, continent:chararray);

--generate required fields

Vaccine\_Pruned = FOREACH VaccineCSV generate country, doses, people\_partially\_vaccinated, people\_fully\_vaccinated, population, ToDate(date, 'yyyy-MM-dd') as (date:DateTime);

Vaccine\_continent\_join = JOIN Vaccine\_Pruned BY country, Countries\_continents BY country;

Vaccines\_grouped\_continent = GROUP Vaccine\_continent\_join BY (continent, date);

Vaccine\_count\_continent\_flattened = FOREACH Vaccines\_grouped\_continent GENERATE FLATTEN(group),

SUM (Vaccine\_continent\_join.people\_partially\_vaccinated) AS partially\_vaccinated,

SUM (Vaccine\_continent\_join.people\_fully\_vaccinated) AS fully\_vaccinated,

SUM (Vaccine\_continent\_join.population) AS population;

Vaccine\_count\_continent = FOREACH Vaccine\_count\_continent\_flattened GENERATE group::Countries\_continents::continent AS continent,

group::Vaccine\_Pruned::date AS date,

partially\_vaccinated, fully\_vaccinated, population;

Store Vaccine\_count\_continent INTO 'CA675-Assignment-2/csv/Vaccine\_count\_continent2' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',' ,'NO\_MULTILINE' ,'UNIX', 'WRITE\_OUTPUT\_HEADER');

Vaccine and cases date joined

Cases = LOAD 'CA675-Assignment-2/csv/Cases\_count\_continent5' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (continent:chararray, date:DateTime, confirmed\_cases:int, deaths:int, recovered:int);

Vaccines = LOAD 'CA675-Assignment-2/csv/Vaccine\_count\_continent2' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',', 'NO\_MULTILINE', 'WINDOWS', 'SKIP\_INPUT\_HEADER') AS (continent:chararray, date:DateTime, partially\_vaccinated:int, fully\_vaccinated:int, population:int);

Cases\_vaccines\_join = Join Cases by (date, continent)LEFT OUTER, Vaccines by (date, continent);

Cases\_Vaccines = FOREACH Cases\_vaccines\_join GENERATE Cases::continent as continent, Cases::date as date, Cases::confirmed\_cases as confirmed\_cases, Cases::deaths as deaths, Cases::recovered as recovered, Vaccines::partially\_vaccinated as partially\_vaccinated, Vaccines::fully\_vaccinated as fully\_vaccinated, Vaccines::population as population;

Store Cases\_Vaccines INTO 'CA675-Assignment-2/csv/Cases\_vaccines\_by\_date\_continent1' USING org.apache.pig.piggybank.storage.CSVExcelStorage(',' ,'NO\_MULTILINE' ,'UNIX', 'WRITE\_OUTPUT\_HEADER');