Práctica 2 – Parte 3: Inyección de datos

viu Universidad Internacional de Valencia

09.05.23

Tipos de inyección de datos

- Streaming:
  - Spark
  - Apache Storm
  - Kafka
  - etc
- Se verá en asignaturas posteriores en detalle.

- Batch
  - Datasets
  - Databases
  - CSV
- Múltiples herramientas que se pueden usar.
- Veremos algunas de ellas para Cassandra y trabajaremos con un programa Python

- sstableloader
  - Permite la inyección de datos en el formato sstable a Cassandra.
  - Desventaja: El formato tiene que ser sstable, que es el que usa Cassandra para almacenar la información. Hay que usar una api JAVA llamada **CQLSSTableWriter** para convertir la información.
- dsbulk
  - Permite la inyección de datos en en Cassandra en diferentes formatos como CSV o JSON.
  - Permite conexión remota a ficheros de datos
  - Compatible con librerías de java para Cassandra

- Otra alternativa es la de hacer la migración a través de un script propio hecho en un lenguaje como Python.
- Utilizaremos un csv con datos de casos de COVID-19 en el estado de California en el que se detalle diferente información como positivos, sospechosos de COVID, hospitalizados, camas en UCI libres y ocupadas, etc.
- En Cassandra creamos dos tablas, una en la que la partition key sea el condado y la clustering key la fecha y otra table con los roles intercambiados.

```
CREATE KEYSPACE hospitals WITH durable_writes = true AND replication = {
'class' : 'SimpleStrategy', 'replication_factor' : 1};
CREATE TABLE hospitals.countypk (
county text,todays_date date,all_hospital_beds
int,hospitalized_covid_confirmed_patients int,hospitalized_covid_patients int,
hospitalized_suspected_covid_patients int,icu_available_beds
int,icu_covid_confirmed_patients int,icu_suspected_covid_patients int,
PRIMARY KEY (county, todays_date)
) WITH CLUSTERING ORDER BY ( todays_date ASC );
CREATE TABLE hospitals.datepk (
todays_date date,county text,all_hospital_beds
int,hospitalized_covid_confirmed_patients int,hospitalized_covid_patients int,
hospitalized_suspected_covid_patients int,icu_available_beds
int,icu_covid_confirmed_patients int,icu_suspected_covid_patients int,
PRIMARY KEY (todays_date, county)
) WITH CLUSTERING ORDER BY ( county ASC );
```

Ejemplo de inserción

```
def insertData ():
    with open ('covid19hospitalbycounty.csv', newline='') as csvhospitals:
        spamreader = csv.DictReader(csvhospitals,)
        insertStatement = session.prepare(
            "INSERT INTO countypk (county,todays_date,hospitalized_covid_confirmed_patients,hospitalized_suspected_covid_patients,"
            "hospitalized_covid_patients,all_hospital_beds,icu_covid_confirmed_patients,icu_suspected_covid_patients,icu_available_beds) VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?
        futures = []
        for row in spamreader:
            dataCovid = hospitalInfo(row['county'], row['todays_date'], row['hospitalized_covid_confirmed_patients'], row['hospitalized_suspected_covid_patients'],
                                     row['hospitalized_covid_patients'], row['all_hospital_beds'], row['icu_covid_confirmed_patients'], row['icu_suspected_covid_patient

            futures.append(session.execute_async(insertStatement, [dataCovid.county,dataCovid.todays_date,
                                                                    dataCovid.hospitalized_covid_confirmed_patients,
                                                                    dataCovid.hospitalized_suspected_covid_patients,
                                                                    dataCovid.hospitalized_covid_patients,dataCovid.all_hospital_beds,
                                                                    dataCovid.icu_covid_confirmed_patients,dataCovid.icu_suspected_covid_patients,
                                                                    dataCovid.icu_available_beds]))
        # Garantiza que el programa no finalice hasta que se ejecuten todas las inserciones
        for f in futures:
            f.result()  # bloquea hasta que se ejecuta la inserción
```

- Crea una función que consulte según un county los valores obtenidos en cada día que se ha registrado
- Crea una función que consulte según un día (`todays_date`) los valores de cada county

- Para realizar una importación o exportación de datos desde dsbulk se debe ejecutar un comando con los siguientes parámetros:

```
dsbulk ( load | unload | count ) [options] (( -k | --keyspace ) keyspace_name
( -t | --table ) table_name) | ( --schema.query string ) [ help | --help ]
```

- Puedes ver más ayuda en:

https://docs.datastax.com/en/dsbulk/docs/reference/dsbulkCmd.html

dsbulk

- Para descargar e instalar dsbulk:
  https://docs.datastax.com/en/dsbulk/doc/dsbulk/install/dsbulkInstall.html

- Ejemplo para cargar datos:
  https://docs.datastax.com/en/dsbulk/doc/dsbulk/reference/dsbulkLoad.html

- **Ejemplos para exportar datos**
  https://docs.datastax.com/en/dsbulk/doc/dsbulk/reference/dsbulkUnload.html

1. Cree la siguiente tabla en el keyspace newzealand:

```
CREATE TABLE keyspace1.census(
        age text,
        area text,
        ethnic text,
        sex text,
        year text,
        count text,
        PRIMARY KEY (age, area, ethnic, sex, year)
)
```

2. Descargar el siguiente enlace (contiene un csv comprimido de casi 900 megas)
https://www3.stats.govt.nz/2018census/Age-sex-by-ethnic-group-grouped-total-responses-census-usually-resident-population-counts-2006-2013-2018-Censuses-RC-TA-SA2-DHB.zip?_ga=2.191883050.1886570868.1667915962-816741319.1667915961
3. Descomprima el fichero, y localice el fichero **Data8277.csv**

4. Ejecute el siguiente comando en la carpeta donde se encuentre ***dsbulk***

./dsbulk load -url ~/Downloads/Age-sex-by-ethnic-group-grouped-total-responses-census-usually-resident-population-counts-2006-2013-2018-Censuses-RC-TA-SA2-DHB/Data8277.csv -k newzealand -t census5 --schema.mapping "Year = year, Age = age, Ethnic=ethnic ,Area=area , Sex = sex, count=count"

El valor **~/Downloads/Age-sex-by-ethnic-group-grouped-total-responses-census-usually-resident-population-counts-2006-2013-2018-Censuses-RC-TA-SA2-DHB/Data8277.csv** debe ser la ubicación del fichero **Data8277.csv.**

Nota: Al copiar y pegar, tenga especial cuidado con símbolos como las comillas.