



Tecnológico de Monterrey

Escuela de Ingeniería y Ciencias

Campus LATAM

Nombre del Trabajo:

Fase 3b. Ingeniería de datos: enriquecer, agregar, generar features

Curso:

Análisis de Grandes Volúmenes de Datos

Alumnos:

Juan Carlos Alvarado Carricarte, A01793486

Bryan Rodolfo Alvarado Cruz, A01793670

Eduardo Gabriel Arévalo Aguilar, A01793897

Profesor:

Alberto De Obeso Orendain

Fecha de entrega:

12 de Febrero 2023

Introducción

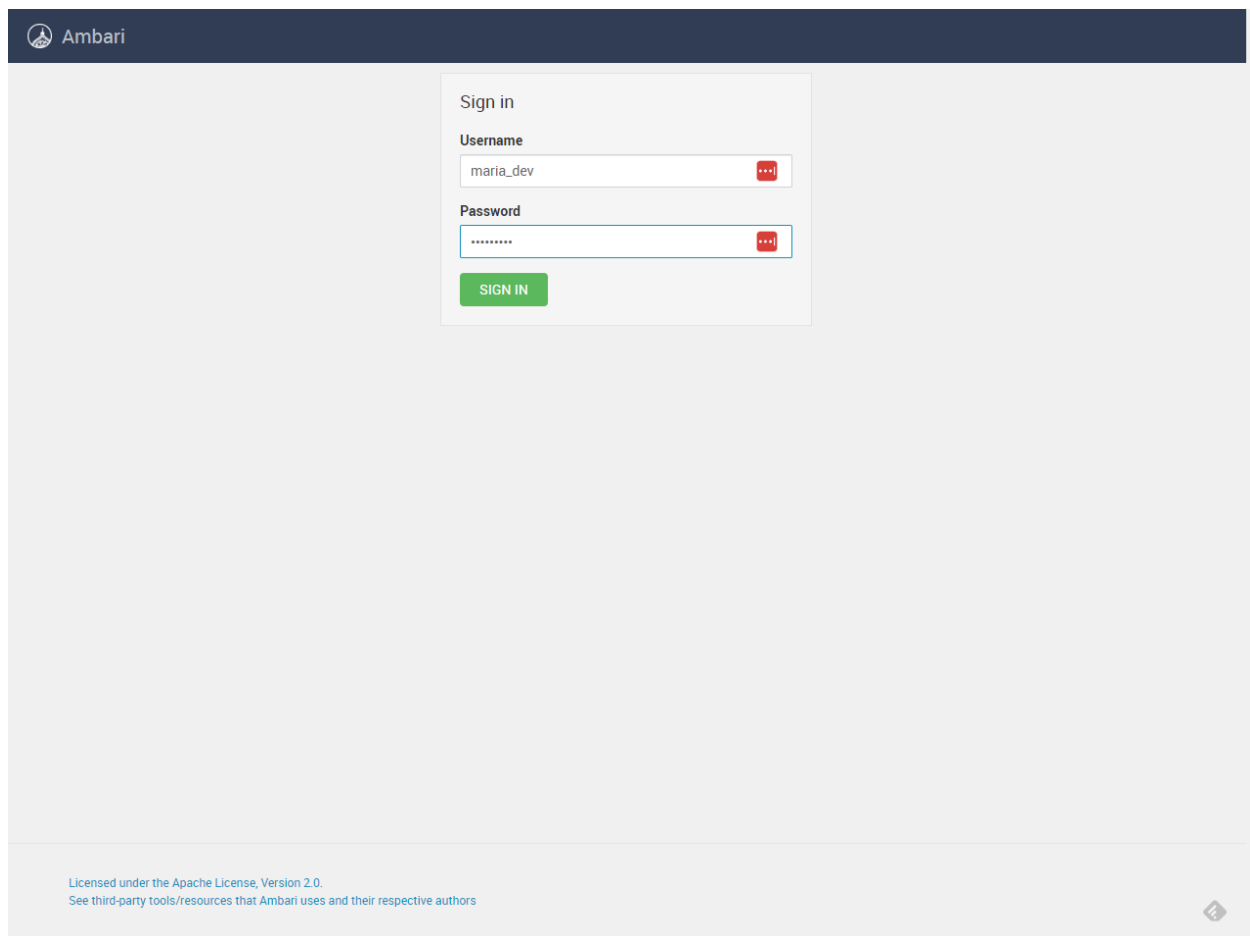
Para este trabajo se tomó como base lo aprendido en la active class 2. Con lo cual buscamos crear un sistema ETL para el procesamiento de los datos que en este caso se toman de la entidad e Jalisco.

Se crean tablas intermedias donde se leen los datos y en cada fila se pone toda la información seoparada con comas, para en un paso posterior poder trasladar a una tabla los datos ya estructurado en una columna por cada tributo de la fila.

Se agregan evidencias sobre multiples consultas al sistema HDFS.

Pasos seguidos:

Iniciar sesión en dashboard Ambari de Hortonworks



The image shows the Ambari web interface for signing in. At the top left, there is a dark blue header with the Ambari logo and the word "Ambari". The main content area is light gray and contains a "Sign in" form. The form has two input fields: "Username" with the value "maria_dev" and "Password" with masked characters "*****". Both fields have a red eye icon to toggle visibility. Below the fields is a green "SIGN IN" button. At the bottom of the page, there is a footer with the text "Licensed under the Apache License, Version 2.0. See third-party tools/resources that Ambari uses and their respective authors" and a small logo on the right.

Ambari

Sign in

Username

maria_dev

Password

SIGN IN

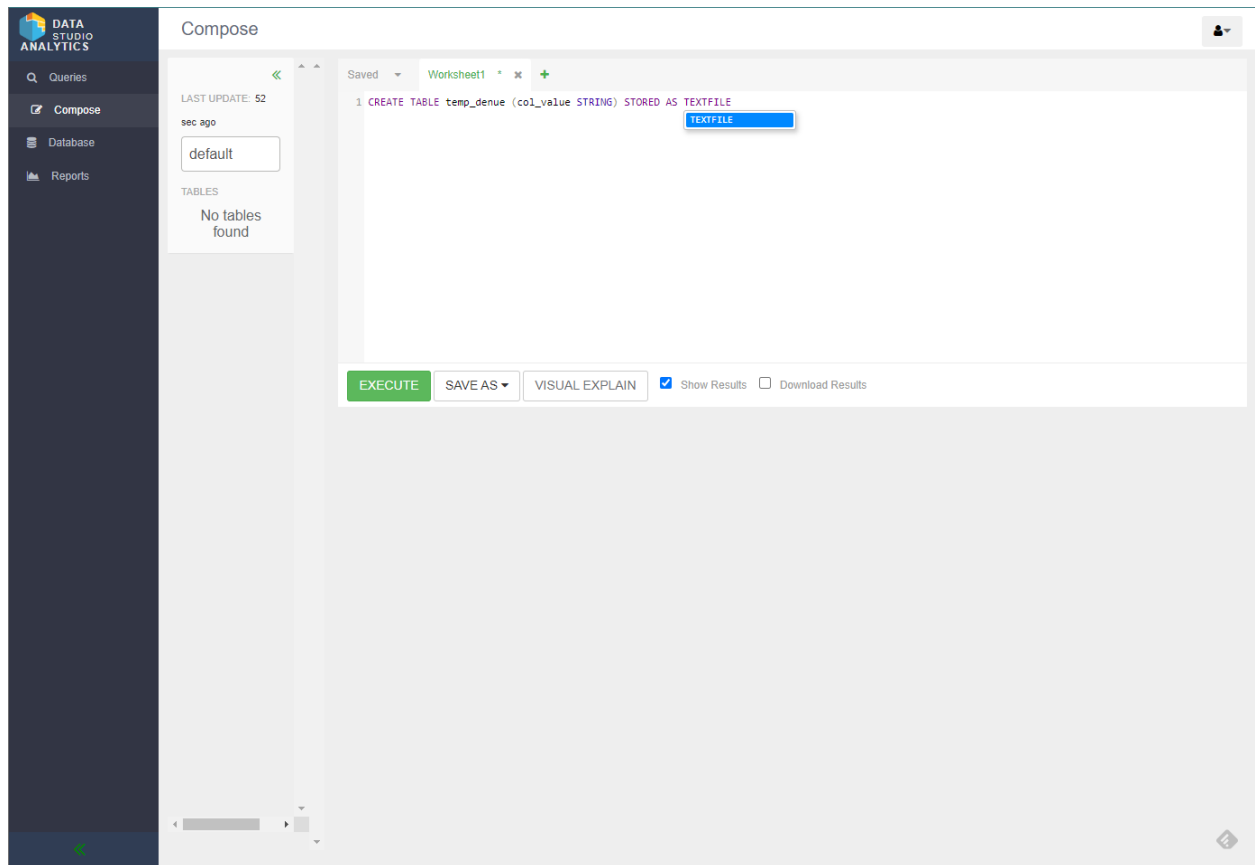
Licensed under the Apache License, Version 2.0.
See third-party tools/resources that Ambari uses and their respective authors

Subir datos de denue ya limpios.

The screenshot displays the Ambari web interface, specifically the 'Files View' section. On the left, a dark sidebar contains a navigation menu with options like Dashboard, Services, and Hosts. The main area shows the file system path '/ > user > maria_dev' with a yellow status bar indicating 'Total: 1 files or folders'. Below this, a table lists the files in the directory. A search bar is located above the table. The table has columns for Name, Size, Last Modified, Owner, Group, Permission, Erasure Coding, and Encrypted. One file is listed: 'denue_inegi_14_clean.csv' with a size of 148.6 MB, last modified on 2023-02-12 at 20:18, owned by 'maria_dev' in the 'hdfs' group, with permissions '-rw-r--r--'. The interface also includes a top navigation bar with a user profile dropdown for 'maria_dev' and a bottom status bar with a green checkmark icon.

Name >	Size >	Last Modified >	Owner >	Group >	Permission	Erasure Coding	Encrypted
denue_inegi_14_clean.csv	148.6 MB	2023-02-12 20:18	maria_dev	hdfs	-rw-r--r--		No

Crear tabla temporal o intermedia.



Verificar carga de taos correcta

DATA STUDIO ANALYTICS

Queries

Compose

Database

Reports

Compose

LAST UPDATE: 12 sec ago

default

TABLES

Search Tables

temp_denue

Saved Worksheet1 Worksheet2 Worksheet3

```
1 SELECT * FROM temp_denue LIMIT 10
```

EXECUTE SAVE AS VISUAL EXPLAIN Show Results Download Results

RESULTS LOG

EXPORT DATA

TEMP_DENUE.COL_VALUE
id,cllee,nom_estab,raz_social,codigo_act,per_ocu,ti more...
8624390,141201125120000220000000000U3,ACUAC ULTORES more...
8838371,140391125110000140000000000U4,ACUAC ULTURA D more...
9233864,140951125120000220000000000U7,ACUIC OLA LA C more...
8341990,140951141190000820000000000U3,ACUIC OLA DE V more...
8908807,140791125120000110000000000U4,ACU COLA EL D more...

Crear tabla donde se guardaran los datos estructurados

DATA STUDIO ANALYTICS

Compose

LAST UPDATE: 16 sec ago

default

TABLES

Search Tables

- denue
- temp_denue

```
1 CREATE TABLE denue(id STRING,clee STRING,nom_estab
2 STRING,rez_social STRING,codigo_ect STRING,per_ocu
3 STRING,tipovial STRING,nom_vial STRING,tipov_e_1
4 STRING,nom_v_e_1 STRING,tipov_e_2 STRING,nom_v_e_2 STRING
5 ,tipov_e_3 STRING ,nom_v_e_3 STRING ,numero_ext STRING
6 ,letra_ext STRING,edificio STRING ,edificio_e STRING
7 ,numero_int STRING ,letra_int STRING ,tipo_asent
8 STRING,nomb_asent STRING ,tipoCenCom STRING ,nom_CenCom
9 STRING ,num_local STRING ,cod_postal STRING,cve_ent
10 STRING,entidad STRING ,cve_mun STRING,municipio STRING,cve_loc
11 STRING,localidad STRING ,ageb STRING,manzana STRING,telefono
12 STRING,correoelec STRING,www STRING,tipouniEco STRING,latitud
13 STRING,longitud STRING,fecha_alta STRING)
```

EXECUTE SAVE AS VISUAL EXPLAIN ☒ Show Results ☐ Download Results

RESULTS LOG

EXPORT DATA

Query completed.

Se pasan los datos de la tabla temporal a la tabla denue.

Queries

Compose

Database

Reports

Queries / hive_20230213045238_fee1be64-5d4d-404b-a1ef-fe8e1e7315d3

RECOMMENDATIONSQUERY DETAILSVISUAL EXPLAINCONFIGSTIMELINEDAG INFO

DOWNLOAD

Recommendations

Table temp_denue is in plain text format. Hive works best with ORC file format. Create a new table using CREATE TABLE new_table (columns) STORED AS ORC; and run INSERT OVERWRITE TABLE new_table SELECT * from temp_denue to convert existing table to ORC file format. Note this creates a copy of your table, for large tables this will be an expensive operation. Consult your database administrator before doing conversion on large tables to avoid running out of space.

Query Details

QUERY ID
hive_20230213045238_fee1be64-5d4d-404b-a1ef-fe8e1e7315d3

USER
hive

STATUS
SUCCESS

START TIME
12 Feb 2023 22:52:38

END TIME
12 Feb 2023 23:04:00

DURATION
11m 21s 461ms

TABLES READ
temp_denue (default)

TABLES WRITTEN
denue (default)

APPLICATION ID
Not Available!

DAG ID
Not Available!

SESSION ID
ffc847a3-3ff2-4d36-bbcc-62cad5b7599d

THREAD ID
HiveServer2-Background-Pool: Thread-159

QUEUE

1 insert overwrite table denue

2 SELECT

3 regexp_extract(col_value, '^?:([*],*)?(1)', 1) id,

4 regexp_extract(col_value, '^?:([*],*)?(2)', 1) clee,

5 regexp_extract(col_value, '^?:([*],*)?(3)', 1) nom_estab,

6 regexp_extract(col_value, '^?:([*],*)?(4)', 1) rat_social,

7 regexp_extract(col_value, '^?:([*],*)?(5)', 1) codigo_act,

8 regexp_extract(col_value, '^?:([*],*)?(6)', 1) per_ocu,

9 regexp_extract(col_value, '^?:([*],*)?(7)', 1) tipo_vial,

10 regexp_extract(col_value, '^?:([*],*)?(8)', 1) nom_vial,

11 regexp_extract(col_value, '^?:([*],*)?(9)', 1) tipo_v_e_1,

12 regexp_extract(col_value, '^?:([*],*)?(10)', 1) nom_v_e_1,

13 regexp_extract(col_value, '^?:([*],*)?(11)', 1) tipo_v_e_2,

14 regexp_extract(col_value, '^?:([*],*)?(12)', 1) nom_v_e_2,

15 regexp_extract(col_value, '^?:([*],*)?(13)', 1) tipo_v_e_3,

16 regexp_extract(col_value, '^?:([*],*)?(14)', 1) nom_v_e_3,

17 regexp_extract(col_value, '^?:([*],*)?(15)', 1) numero_ext,

18 regexp_extract(col_value, '^?:([*],*)?(16)', 1) letra_ext,

19 regexp_extract(col_value, '^?:([*],*)?(17)', 1) edificio,

20 regexp_extract(col_value, '^?:([*],*)?(18)', 1) edificio_e,

21 regexp_extract(col_value, '^?:([*],*)?(19)', 1) numero_int,

22 regexp_extract(col_value, '^?:([*],*)?(20)', 1) letra_int,

23 regexp_extract(col_value, '^?:([*],*)?(21)', 1) tipo_asent,

24 regexp_extract(col_value, '^?:([*],*)?(22)', 1) nomb_asent,

25 regexp_extract(col_value, '^?:([*],*)?(23)', 1) tipoCenCom,

26 regexp_extract(col_value, '^?:([*],*)?(24)', 1) nom_CenCom,

27 regexp_extract(col_value, '^?:([*],*)?(25)', 1) num_local,

28 regexp_extract(col_value, '^?:([*],*)?(26)', 1) cod_postal,

29 regexp_extract(col_value, '^?:([*],*)?(27)', 1) cve_ent,

30 regexp_extract(col_value, '^?:([*],*)?(28)', 1) entidad,

31 regexp_extract(col_value, '^?:([*],*)?(29)', 1) cve_mun,

32 regexp_extract(col_value, '^?:([*],*)?(30)', 1) municipio,

Se verifican los datos cargados correctamente

DATA
STUDIO
ANALYTICS

Queries

Compose

Database

Reports

Compose

LAST UPDATE: 1
sec ago

default

TABLES

Search Tables

» denue

» temp_denue

Saved

Worksheet1

+

1 SELECT * FROM denue WHERE cve_mun = '39' LIMIT 10;

EXECUTE

SAVE AS

VISUAL EXPLAIN

Show Results

Download Results

RESULTS

LOG

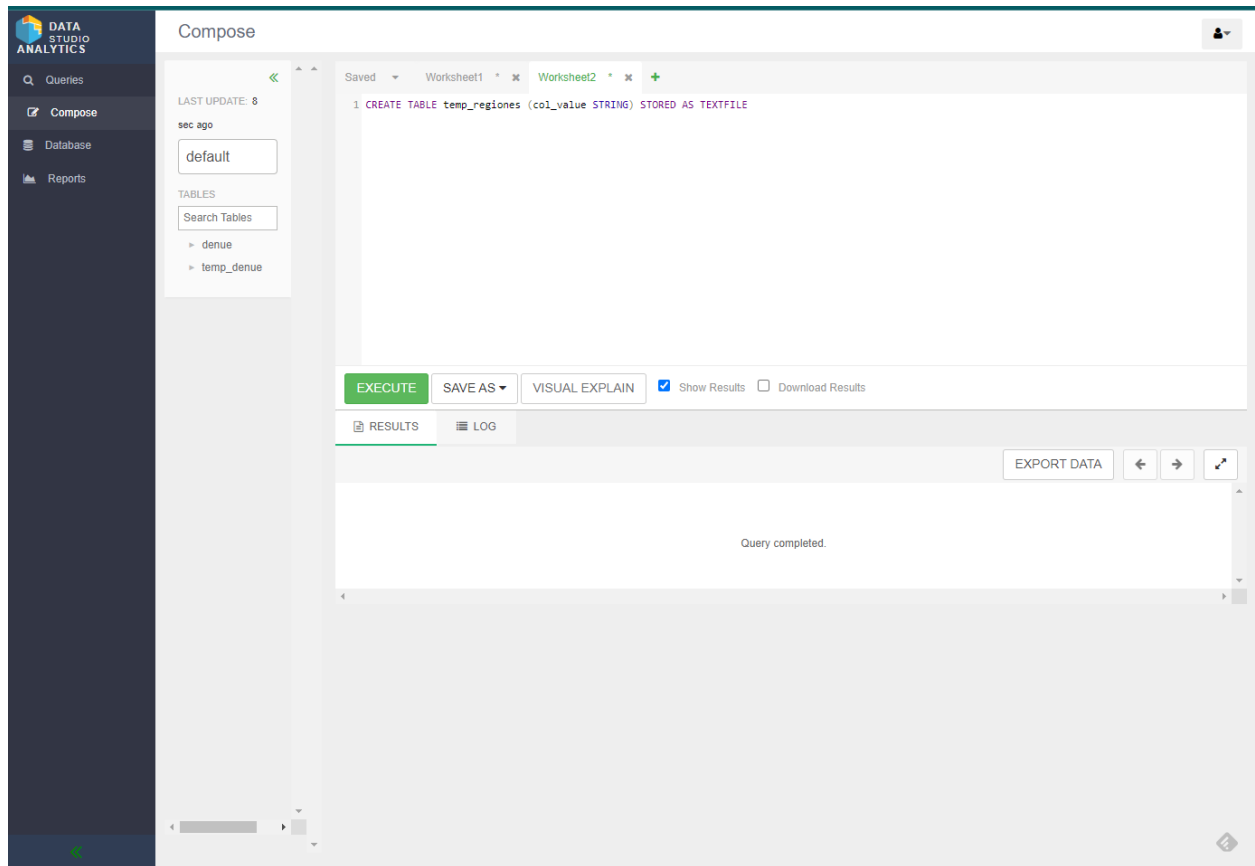
EXPORT DATA

←

→

DENUE.ID	DENUE.CLEE	DENUE.NOM_ESTAB	DENUE.RAZ_SOCIAL	DENUE.CODIGO_ACT	DENUE.PER_OCU	DENUE.TIPO_VIAL	DENUE.NOM_VIAL	DENUE.TIPO_V_E_1	DENU
1962065	140396213 980027410 00000000U 6	CONSULTORIO A CUPUNTURA	JUNG SUB KIM	621398	0 a 5 personas	CALLE	BERNARDO D E BALBUENA	CALLE	COLC
1959234	140396211 130101310 00000000U 3	CONSULTORIO A LARGOLOGO	""	621113	0 a 5 personas	""	JESUS GARC A	""	ADOI MATE
1953972	140396211 110048710 00000000U 6	CONSULTORIO A LOPATA SIN NO MBRE	""	621111	0 a 5 personas	CALLE	GABRIEL RUIZ	CALLE	HACI CARI
1933105	140396212 110159110 00000000U 5	CONSULTORIO A LPHA DENTAL	""	621211	0 a 5 personas	CALLE	BASILIO BADIL LO	CALLE	FEDE
7811357	140396212 110211910	CONSULTORIO A MH DENTAL	""	621211	0 a 5 personas	CALLE	DAVID HINOJO SA	CALLE	LAUR A

Se crea una tabla temporal de regiones



Se suben y se verifican la subida de datos a tabla temporal de regiones

DATA STUDIO ANALYTICS

Queries

Compose

Database

Reports

Compose

LAST UPDATE: 4 sec ago

default

TABLES

Search Tables

» denue

» temp_denue

» temp_regiones

Saved Worksheet1 Worksheet2 Worksheet3

1 LOAD DATA INPATH '/user/maria_dev/regionesjalisco.csv'

2 OVERWRITE INTO TABLE temp_regiones

EXECUTE SAVE AS VISUAL EXPLAIN Show Results Download Results

RESULTS LOG

EXPORT DATA

Query completed.

DATA STUDIO ANALYTICS

Queries

Compose

Database

Reports

Compose

LAST UPDATE: 3 sec ago

default

TABLES

Search Tables

» denue

» temp_denue

» temp_regiones

Saved Worksheet1 Worksheet2 Worksheet3 Worksheet4

1 SELECT * FROM temp_regiones LIMIT 10

EXECUTE

SAVE AS

VISUAL EXPLAIN

Show Results

Download Results

RESULTS

LOG

EXPORT DATA

←

→

↗

TEMP_REGIONES.COL_VALUE

CVE_MUN, REGION

2,centro

29,centro

39,centro

44,centro

45,centro


51,centro

70,centro

71,centro

97,centro

Crear tabla regiones Jalisco

DATA
STUDIO
ANALYTICS

Queries

Compose

Database

Reports

Compose

LAST UPDATE: 4 sec ago

default

TABLES

Search Tables

» denue

» regiones_jalisco

» temp_denue

» temp_regiones

Compose

Saved Worksheet1 Worksheet2 Worksheet3 Worksheet4 Worksheet5

1 CREATE TABLE regiones_jalisco(CVE_MUN STRING,REGION STRING)

EXECUTE SAVE AS VISUAL EXPLAIN Show Results Download Results

RESULTS LOG

EXPORT DATA

Query completed.