

UNIVERSIDAD EAFIT

Tópicos Especiales en Telemática

Laboratorio 3-3

Juan Manuel Gómez - Ingeniería de Sistemas

(jmgomezp@eafit.edu.co)

Profesor: Edwin Nelson Montoya

Medellín, 7 de noviembre de 2024

## Data Processing usando PySpark en JupyterHub

En este caso se hizo una exploración de los datos usando PySpark en JupyterHub, leyendo los datos de un archivo *csv* almacenado en un Bucket de S3. Además, Una vez hecha la exploración, se exportaron los datos nuevamente a S3/

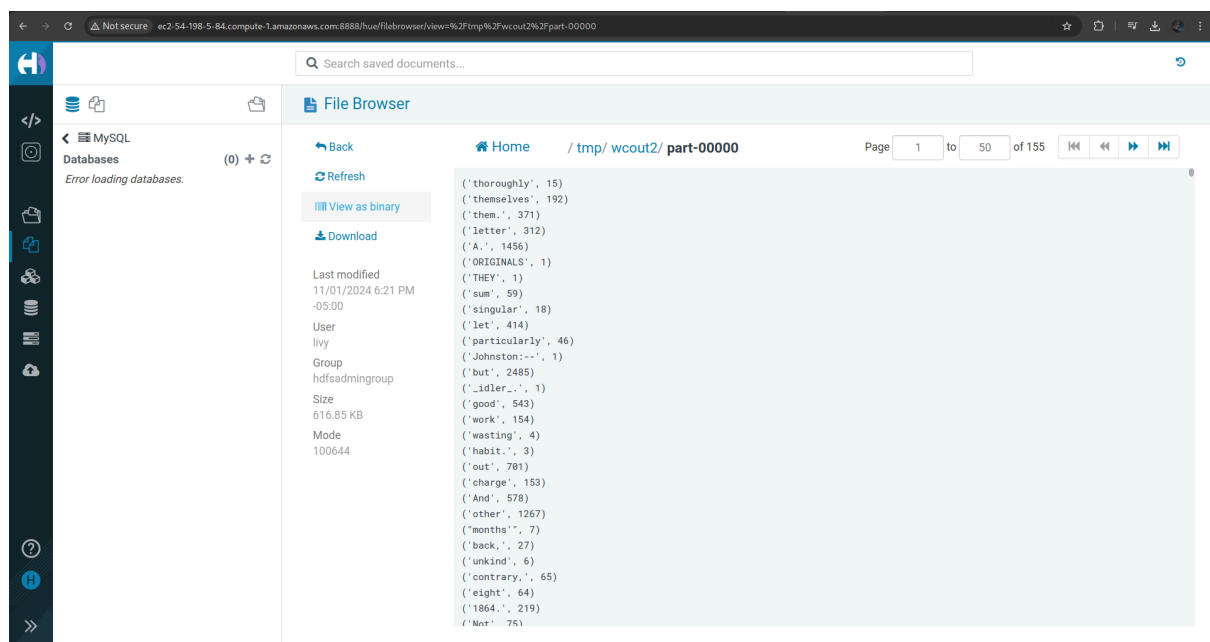
Dentro del repositorio de GitHub se encuentra el Notebook (Data\_processing\_using\_PySpark\_jupyterhub.ipynb) correspondiente a esta actividad.

## Word Count

En este apartado se creó un Notebook desde la aplicación JupyterHub que está corriendo en un cluster EMR. En este Notebook se hizo un Word Count del dataset gutenber-small. Donde el output de este programa fue consultado desde Hue.

Dentro del repositorio de GitHub se encuentra el Notebook (wordcount.ipynb) con los resultados de este laboratorio.

A continuación, se muestra el resultado en Hue:



The screenshot shows the Hue web interface. On the left is a sidebar with navigation icons. The main area is titled 'File Browser' and shows a file named 'part-00000' in the directory '/ tmp/ wcout2/'. The file is 616.85 KB and was last modified on 11/01/2024 at 6:21 PM. The file content is displayed as a list of word counts, such as ('thoroughly', 15), ('themselves', 192), ('them', 371), ('letter', 312), ('A.', 1456), ('ORIGINALS', 1), ('THEY', 1), ('sum', 59), ('singular', 18), ('let', 414), ('particularly', 46), ('Johnston:--', 1), ('but', 2485), ('\_idler\_', 1), ('good', 543), ('work', 154), ('wasting', 4), ('habit.', 3), ('out', 781), ('charge', 153), ('And', 578), ('other', 1267), ('months', 7), ('back', 27), ('unkind', 6), ('contrary.', 65), ('eight', 64), ('1864.', 219), and ('Not', 75).

Not secure ec2-54-198-5-84.compute-1.amazonaws.com:8888/hue/

File browser view=13a%3A%2F%2Fbig-data-topicos%2Fwcout2%2Fpart-00000

☆ ↻ ⌂

Search saved documents...

↻

MySQL

Databases

Error loading databases.

(0) + ↻

File Browser

Back Home

Refresh

View as binary

Download

Last modified  
11/01/2024 6:22 PM  
-05:00

User

Group

Size  
616.85 KB

Mode  
100666

s3a://big-data-topicos/ wcout2/ part-00000

```
(`thoroughly`, 15)
(`themselves`, 192)
(`them`, 371)
(`letter`, 312)
(`A`, 1456)
(`ORIGINALS`, 1)
(`THEY`, 1)
(`sum`, 59)
(`singular`, 18)
(`let`, 414)
(`particularly`, 46)
(`Johnston:--`, 1)
(`but`, 2485)
(`_idler_`, 1)
(`good`, 543)
(`work`, 154)
(`wasting`, 4)
(`habit`, 3)
(`out`, 781)
(`charge`, 153)
(`And`, 578)
(`other`, 1267)
(`months`, 7)
(`back`, 27)
(`unkind`, 6)
(`contrary`, 65)
(`eight`, 64)
```