Anexo Proyecto
INF-464 Computación Distribuida para Big Data
Segundo Semestre 2021

Apache Spark:
AWS EMR vs Local Host

Héctor Labraña Rojas hector.labrana.13@sansano.usm.cl

1. **Enlace Video: https://youtu.be/jd-or8408i0**
2. **Enlace Código Fuente y Documentos: https://github.com/hlabrana/ProyectoCDBD**
3. **Enlace Dataset: http://files.pushshift.io/reddit/comments/**

I. Acceso a código fuente:

El archivo SparkJob.py resume el código fuente necesario para procesar el dataset en estudio, este se puede encontrar en los siguientes lugares:

1. Github: https://github.com/hlabrana/ProyectoCDBD/blob/main/code/SparkJob.py
2. En las máquinas virtuales proporcionadas SPVM: ~/CDBD/SparkJob.py

II. Resultados de experimentos:

Se incluye una captura de pantalla de cada resultado obtenido para cada entorno en ejecución (3 clústers EMR + 3 máquinas virtuales SPVM)

1. https://github.com/hlabrana/ProyectoCDBD/blob/main/img/Log-EMR01.png
2. https://github.com/hlabrana/ProyectoCDBD/blob/main/img/Log-EMR02.png
3. https://github.com/hlabrana/ProyectoCDBD/blob/main/img/Log-EMR03.png
4. https://github.com/hlabrana/ProyectoCDBD/blob/main/img/Log-SPVM01.png
5. https://github.com/hlabrana/ProyectoCDBD/blob/main/img/Log-SPVM02.png
6. https://github.com/hlabrana/ProyectoCDBD/blob/main/img/Log-SPVM03.png

III. Instalación de Software:

Para cada máquina fue necesario ejecutar los siguientes comandos:

1. pip install pyspark==3.1.1 (En EMR viene incorporado)
2. pip install pandas
3. pip install matplotlib
4. sudo apt install default-jre
5. pip install sparkmeasure
6. Mover el jar de sparkmeasure a la carpeta ~/local/lib/pyspark/jars

Para ingresar a las máquinas virtuales se debe seguir los siguientes pasos:

1. Ingresar desde una terminal con cuenta informática (DI)
2. Desde una terminal, ejecutar el siguiente comando:
    a. > ssh dockeruser@<IPSPVM>
3. Ingresar la contraseña
4. Ir a la carpeta /CDBD
5. Ejecutar SparkJob.py con el comando:

    a. > PYTHONSTARTUP=SparkJob.py pyspark

6. Esperar los resultados por consola
7. Ver gráfica generada en el mismo directorio con nombre plotResults.png

IV. Recomendaciones:

1. Al detener la EC2 correspondiente del clúster EMR este termina abruptamente sin posibilidad de restaurar.
2. Se recomienda revisar la tabla de ruteo del grupo de seguridad del clúster EMR para asegurar conexión tráfico inbound / outbound para SSH.
3. Ejecutar SparkJob.py con el siguiente comando:

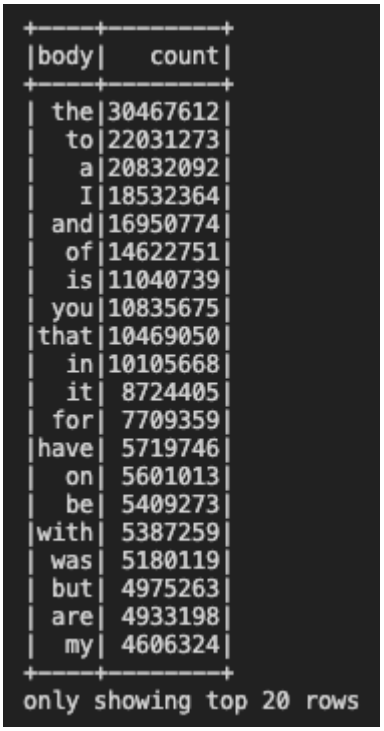
    a. > PYTHONSTARTUP=SparkJob.py pyspark

4. La versión de pyspark recomendada es 3.1.1
5. Al parecer los clústers de EMR con versión 6.3.1 tienen problemas al ejecutar el kerner de un bloc de notas (jupyter notebook) se recomienda conectarse por SSH

V. Gráficos Adicionales:

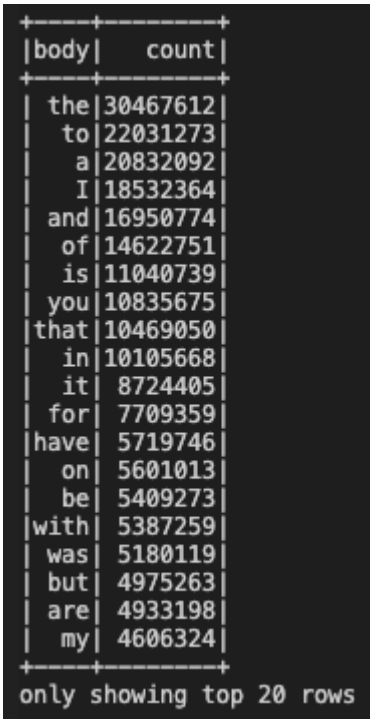
I. Dataframe final con 15 palabras más frecuentes

SPVM01



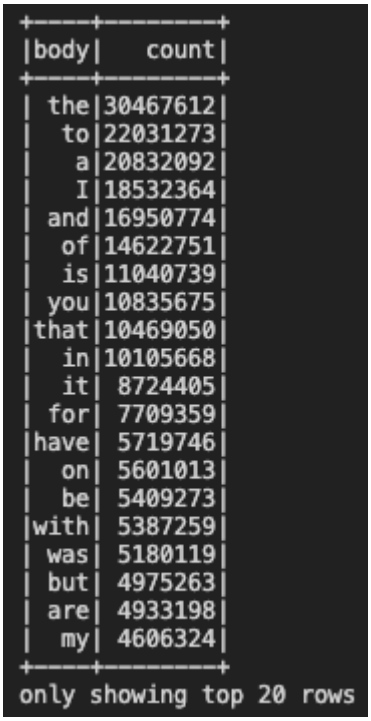
```
+----+--------+
|body|   count|
+----+--------+
| the|30467612|
|  to|22031273|
|   a|20832092|
|   I|18532364|
| and|16950774|
|  of|14622751|
|  is|11040739|
| you|10835675|
|that|10469050|
|  in|10105668|
|  it| 8724405|
| for| 7709359|
|have| 5719746|
|  on| 5601013|
|  be| 5409273|
|with| 5387259|
| was| 5180119|
| but| 4975263|
| are| 4933198|
|  my| 4606324|
+----+--------+
only showing top 20 rows
```

SPVM02



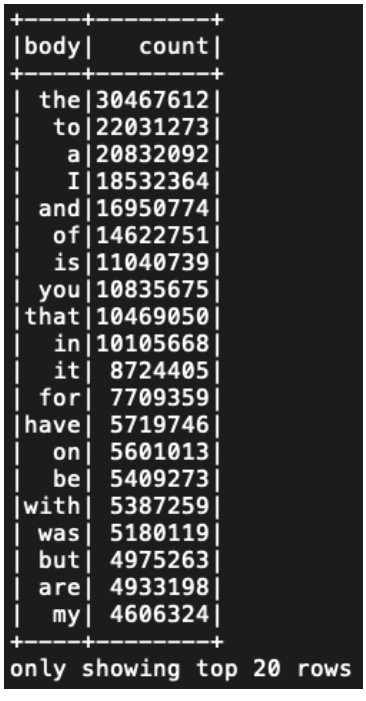
```
+----+--------+
|body|   count|
+----+--------+
| the|30467612|
|  to|22031273|
|   a|20832092|
|   I|18532364|
| and|16950774|
|  of|14622751|
|  is|11040739|
| you|10835675|
|that|10469050|
|  in|10105668|
|  it| 8724405|
| for| 7709359|
|have| 5719746|
|  on| 5601013|
|  be| 5409273|
|with| 5387259|
| was| 5180119|
| but| 4975263|
| are| 4933198|
|  my| 4606324|
+----+--------+
only showing top 20 rows
```

SPVM03



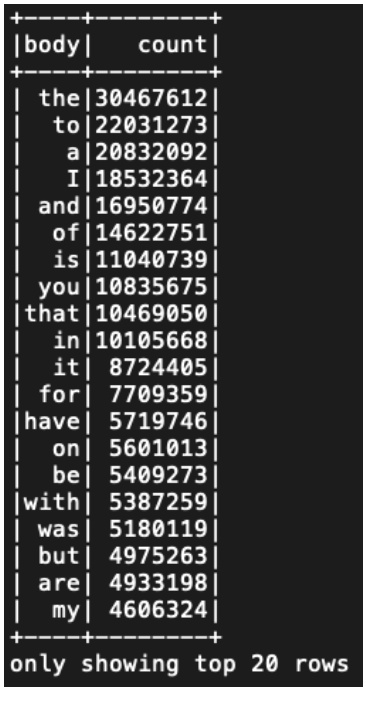
```
+----+--------+
|body|   count|
+----+--------+
| the|30467612|
|  to|22031273|
|   a|20832092|
|   I|18532364|
| and|16950774|
|  of|14622751|
|  is|11040739|
| you|10835675|
|that|10469050|
|  in|10105668|
|  it| 8724405|
| for| 7709359|
|have| 5719746|
|  on| 5601013|
|  be| 5409273|
|with| 5387259|
| was| 5180119|
| but| 4975263|
| are| 4933198|
|  my| 4606324|
+----+--------+
only showing top 20 rows
```

EMR01



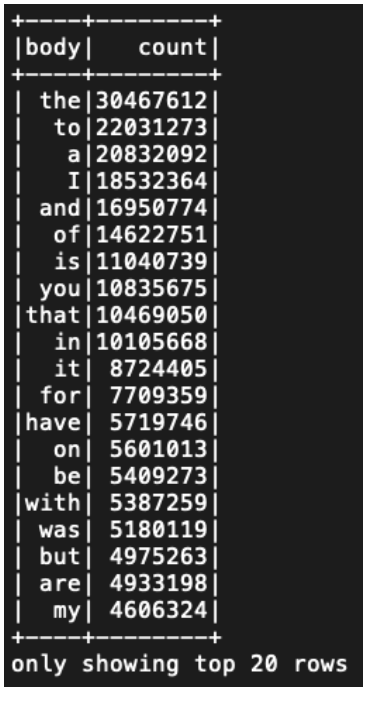
```
+----+--------+
|body|   count|
+----+--------+
| the|30467612|
|  to|22031273|
|   a|20832092|
|   I|18532364|
| and|16950774|
|  of|14622751|
|  is|11040739|
| you|10835675|
|that|10469050|
|  in|10105668|
|  it| 8724405|
| for| 7709359|
|have| 5719746|
|  on| 5601013|
|  be| 5409273|
|with| 5387259|
| was| 5180119|
| but| 4975263|
| are| 4933198|
|  my| 4606324|
+----+--------+
only showing top 20 rows
```

EMR02



```
+----+--------+
|body|   count|
+----+--------+
| the|30467612|
|  to|22031273|
|   a|20832092|
|   I|18532364|
| and|16950774|
|  of|14622751|
|  is|11040739|
| you|10835675|
|that|10469050|
|  in|10105668|
|  it| 8724405|
| for| 7709359|
|have| 5719746|
|  on| 5601013|
|  be| 5409273|
|with| 5387259|
| was| 5180119|
| but| 4975263|
| are| 4933198|
|  my| 4606324|
+----+--------+
only showing top 20 rows
```

EMR03



```
+----+--------+
|body|   count|
+----+--------+
| the|30467612|
|  to|22031273|
|   a|20832092|
|   I|18532364|
| and|16950774|
|  of|14622751|
|  is|11040739|
| you|10835675|
|that|10469050|
|  in|10105668|
|  it| 8724405|
| for| 7709359|
|have| 5719746|
|  on| 5601013|
|  be| 5409273|
|with| 5387259|
| was| 5180119|
| but| 4975263|
| are| 4933198|
|  my| 4606324|
+----+--------+
only showing top 20 rows
```

II. **Fórmula error porcentual**

$$\varepsilon r = \left| \frac{X_i - X_v}{X_v} \right| 100\%$$

Dónde:

$\Delta x = Cociente\ de\ error\ Absoluto$

$X_v = Valor\ real\ de\ la\ Magnitud$

VI. **Referencias:**

1. **Reddit comments:** http://files.pushshift.io/reddit/comments/
2. **Pyspark Documentation:** https://spark.apache.org/docs/3.1.1/api/python/index.html