

Instituto Tecnológico de Costa Rica
Escuela de Ingeniería en Computación
Bases de datos ii

Grupo 40

Profesor Erick Hernandez

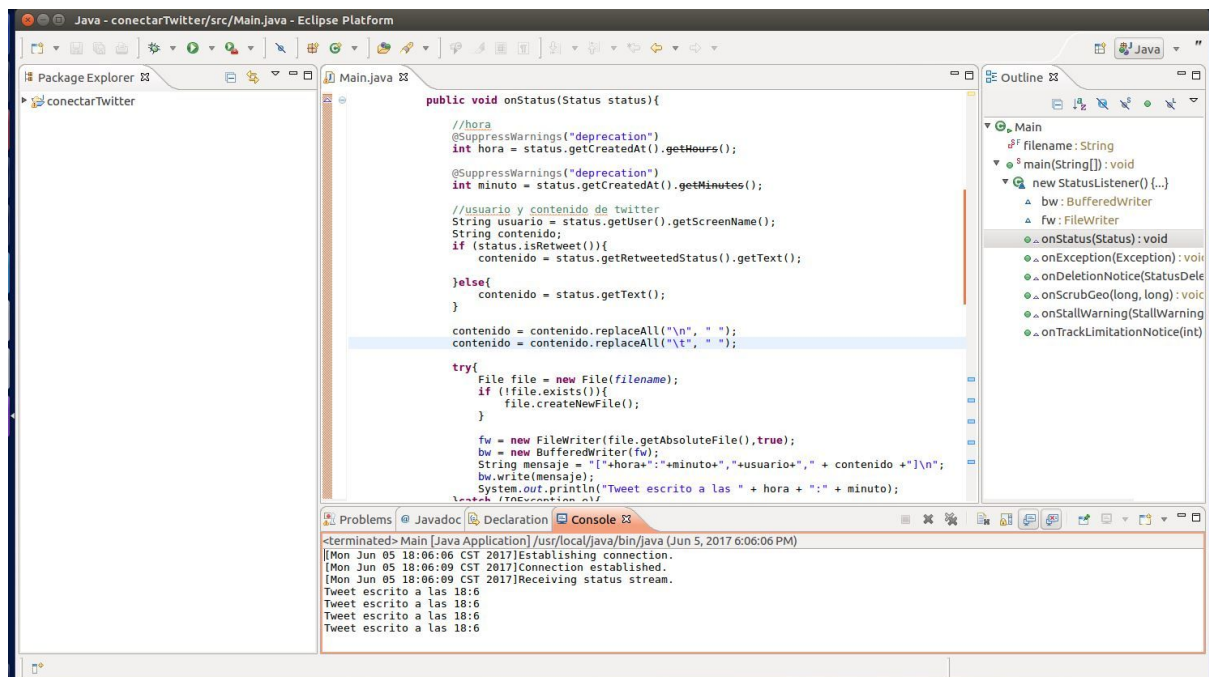
Estudiantes Jose Pablo Murillo, Carlos Villalobos, Esteban Santamaría

Manual de usuario proyecto 2: análisis de tweets

El siguiente manual de usuario se realizó en una máquina virtual corriendo Ubuntu 16.04 y Hadoop 2.7.

Obtención de datos de twitter

1. Abrir la aplicación en Eclipse.
2. Correr la aplicación presionando el botón verde en la barra superior de Eclipse.
3. La aplicación va a ir corriendo y registrando datos en “datos.txt”, que se encuentra en el mismo folder que el de la aplicación JAVA. La aplicación puede correr hasta que el usuario decide detenerla.



Ejecución del mapreduce usando Python

Para esta demostración, se guardaron los archivos.py y se copio el archivo producto de correr la aplicación JAVA “datos.txt” en el folder “twitter”. Por conveniencia se puso el archivo en el Desktop del usuario “jose”.

1. Ingrese como el usuario que fue definido para usar hadoop. En nuestro caso, se llama "hduser".
2. Inserte el comando ssh localhost para acceder al localhost de la máquina.

```
hduser@jose-VirtualBox: ~
jose@jose-VirtualBox:~$ su hduser
Password:
hduser@jose-VirtualBox:/home/jose$ ssh localhost
hduser@localhost's password:
Welcome to Ubuntu 16.04.2 LTS (GNU/Linux 4.8.0-52-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

96 packages can be updated.
0 updates are security updates.

Last login: Mon Jun  5 15:41:44 2017 from 127.0.0.1
hduser@jose-VirtualBox:~$
```

3. Ingrese los comandos start-dfs.sh y start-yarn.sh para inicializar los servicios para el servidor.

```
twitter
hduser@jose-VirtualBox: ~
hduser@localhost's password:
Welcome to Ubuntu 16.04.2 LTS (GNU/Linux 4.8.0-52-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:    https://landscape.canonical.com
 * Support:       https://ubuntu.com/advantage

96 packages can be updated.
0 updates are security updates.

Last login: Mon Jun  5 15:41:44 2017 from 127.0.0.1
hduser@jose-VirtualBox:~$ start-dfs.sh
Starting namenodes on [localhost]
hduser@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hduser-namenode-jose-VirtualBox.out
hduser@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hduser-datanode-jose-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
hduser@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hduser-secondarynamenode-jose-VirtualBox.out
hduser@jose-VirtualBox:~$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hduser-resourcemanager-jose-VirtualBox.out
hduser@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hduser-nodemanager-jose-VirtualBox.out
hduser@jose-VirtualBox:~$
```

4. Dirijase a la carpeta "twitter" donde se encuentran los archivos.py

```
hduser@jose-VirtualBox:~$ cd ..
hduser@jose-VirtualBox:/home$ cd jose/Desktop/twitter
hduser@jose-VirtualBox:/home/jose/Desktop/twitter$ hadoop fs -ls analisis
hduser@jose-VirtualBox:/home/jose/Desktop/twitter$ hadoop fs -put datos.txt analisis
```

5. Use los comandos “hadoop fs -mkdir analisis y hadoop fs -put datos.txt analisis para crear la carpeta donde se va a guardar el archivo “datos.txt”.

NOTA: En caso de que le muestre un error con “ConnectionRefused” o “no such file or directory” puede ver la secuencia de pasos a seguir en el archivo “comandos”, incluido en los archivos de la aplicación.

6. Corra el siguiente comando:

```
“$HADOOP_PREFIX/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -input analisis/datos.txt -output output -mapper mapperTweets.py -reducer reducerTweets.py
```

En el comando anterior se especifica lo siguiente:

- input: archivo de datos que se va a usar para el map reduce.
- output: donde se guardan lo que retorne el reducer, en este caso nada.
- mapper: el archivo mapper que va a mapear los datos
- reducer: el archivo reducer que va a analizar los datos pasados por el

mapper.

Cuando presione enter, se va a comenzar el job como se muestra a continuación.

```
hduser@jose-VirtualBox: /home/jose/Desktop/twitter$ hadoop jar $HADOOP_PREFIX/share/hadoop/tools/lib/hadoop-streaming-2.7.3.jar -input analisis/datos.txt -output output -mapper mapperTweets.py -reducer reducerTweets.py
17/06/05 16:11:23 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
17/06/05 16:11:23 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
17/06/05 16:11:23 INFO jvm.JvmMetrics: Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
17/06/05 16:11:24 INFO mapred.FileInputFormat: Total input paths to process : 1
17/06/05 16:11:24 INFO mapreduce.JobSubmitter: number of splits:1
17/06/05 16:11:24 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local275502567_0001
17/06/05 16:11:25 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
17/06/05 16:11:25 INFO mapreduce.Job: Running job: job_local275502567_0001
17/06/05 16:11:25 INFO mapred.LocalJobRunner: OutputCommitter set in config null
17/06/05 16:11:25 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
17/06/05 16:11:25 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
17/06/05 16:11:25 INFO mapred.LocalJobRunner: Waiting for map tasks
17/06/05 16:11:25 INFO mapred.LocalJobRunner: Starting task: attempt_local275502567_0001_m_000000_0
17/06/05 16:11:25 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1
17/06/05 16:11:25 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
17/06/05 16:11:25 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/user/hduser/analisis/datos.txt:0+371723
17/06/05 16:11:25 INFO mapred.MapTask: numReduceTasks: 1
17/06/05 16:11:25 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
17/06/05 16:11:25 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
```

7. Una vez finalizado el job, se puede usar mySQL Workbench y examinar los resultados en las tablas.

The screenshot shows the MySQL Workbench interface. The query editor at the top contains the SQL statement: `SELECT * FROM twitter_data.tema;`. The query has been executed, and the results are displayed in the 'Result Grid' pane. The results show 7 rows of data from the 'tema' table. The columns are: #, idTema, cantUsuarios, and cantTweets. The data rows are: 1. #TRAFICOCR (5 users, 5 tweets), 2. #MAKEAMERICAGREATAGAIN (44 users, 44 tweets), 3. #2030NOW (1 user, 1 tweet), 4. #WOMEN (209 users, 209 tweets), 5. #PURAVIDA (2 users, 2 tweets), 6. #COSTARICA (39 users, 39 tweets), and 7. #TRUMPRUSSIA (1086 users, 1086 tweets). The 'Object Info' pane on the left shows the schema for the 'tema' table, with columns: idTema (varchar(25)), cantUsuarios (int(11)), and cantTweets (int(11)). The 'Action Output' pane at the bottom shows the execution details: 'SELECT * FROM twitter_data.tema LIMIT 0, 1000' returned 7 row(s).

| # | idTema | cantUsuarios | cantTweets |
|---|------------------------|--------------|------------|
| 1 | #TRAFICOCR | 5 | 5 |
| 2 | #MAKEAMERICAGREATAGAIN | 44 | 44 |
| 3 | #2030NOW | 1 | 1 |
| 4 | #WOMEN | 209 | 209 |
| 5 | #PURAVIDA | 2 | 2 |
| 6 | #COSTARICA | 39 | 39 |
| 7 | #TRUMPRUSSIA | 1086 | 1086 |

8. Para visualizar los datos de manera gráfica, diríjase al navegador web e ingrese a la dirección “localhost”.



Se muestra una página donde se pueden realizar las siguientes consultas de datos para los temas que usaron para el desarrollo de la aplicación. Las acciones que se pueden hacer son las siguientes:

- Cantidad de usuarios que mencionaron el tema.
- Cantidad de tweets que fueron capturados por tema.
- El top 10 de palabras mencionadas por tema con su respectivo total de apariencias.
- La distribución de tiempo (en horas) dado un tema.
- Qué otros temas se mencionan para cierto tema y en cuantos tweets paso eso.

9. El usuario selecciona la accion y selecciona los temas sobre los que quiere ver los datos.

10. Los datos son visualizados en un cuadro debajo de las opciones de los temas.

