



UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

# Trabajo Práctico 1

INF356 - 2025-1 - 200

Computación Distribuida para Big Data

22 de abril de 2025 - v1.0

## Índice

<b>1. Despliegue del cluster</b>	<b>5</b>
1.1. Implementación del tutorial . . . . .	5
1.2. Expansión del cluster . . . . .	7
1.3. Cliente web . . . . .	8
<b>2. Instalación de Apache Hive</b>	<b>9</b>
2.1. Procedimiento . . . . .	9
2.2. Prueba . . . . .	10
<b>3. Exploración del HDFS</b>	<b>12</b>
<b>4. Uso del cluster</b>	<b>13</b>
4.1. Importación . . . . .	13
4.2. Parsing . . . . .	13
4.3. Análisis . . . . .	13

## Índice de figuras

1. Captura de pantalla de la “AWS Management Console - EC2” que muestra las máquinas del cluster creado con el tutorial <b>Se debe ver la consola completa. En las columnas seleccione: Name, Instance ID, Instance state, Instance type, Status check, Public IPv4 address, Private IP Address</b> . . . .	7
2. Captura de pantalla de la “AWS Management Console - EC2” que muestra las máquinas del cluster expandido <b>Se debe ver la consola completa. En las columnas seleccione: Name, Instance ID, Instance state, Instance type, Status check, Public IPv4 address, Private IP Address</b> . . . . .	8

3. Captura de pantalla del cliente web de Hadoop **Mostrar la consola web de Hadoop en la sección de aplicaciones terminadas, en las que se debe ver a lo menos los trabajos de la validación del cluster original y los trabajos que haya usado para validar el funcionamiento de la expansión del cluster** . . . 9
4. Captura de pantalla del cliente web de HDFS **Mostrar la consola web de HDFS mostrando alguno de los archivos que hayan sido agregados al dfs.** 10

## Índice de fragmentos de código

1. Código utilizado en la implementación del tutorial . . . . . 7
2. Código utilizado en la expansión del cluster . . . . . 8
3. Ejemplo de uso de Apache Hive . . . . . 11
4. Resultado esperado para ejemplo de uso de Apache Hive **Algunos de los valores han sido alterados, deben ser corregidos con el resultado de su procedimiento.** . . . . . 12
5. Ejemplo de uso de Apache Hive . . . . . 12
6. Script de reporte de DFS . . . . . 13

## Instrucciones

Este informe corresponde a la plantilla para presentar el reporte de la actividad práctica inicial que debe ser proveído como entregable de la actividad.

Para compilar este informe se requiere del preambulo disponible en <https://github.com/ptoledo-teaching/latex-report-001>. Usted debe clonar este repositorio y colocarlo al mismo nivel que la carpeta **practico-001**.

Todos los textos en rojo a lo largo de la plantilla, junto con esta página de instrucciones, deben ser eliminadas antes de la compilación final.

La evaluación se realizará en base a 3 entregas incrementales con las siguientes fechas:

- **Entrega 1 - 2025/04/21:** Se debe entregar la sección 1 de este informe
- **Entrega 2 - 2025/04/28:** Se debe entregar la sección 2 y 3 de este informe. La evaluación corresponderá en un 80 % a la sección 2 y 3 con un 20 % a la sección 1.
- **Entrega 3 - 2025/05/05:** Se debe entregar la sección 4. La sección 4 corresponderá a un 80 % de la nota de la entrega y un 20 % a las secciones 1, 2 y 3

Los 20 % otorgados a entregas anteriores en las entregas 2 y 3 están pensados para que una mejora de lo entregado anteriormente sea considerado en la calificación final. Recuerde que la nota final del trabajo práctico corresponde al **promedio geométrico** de las entregas 1, 2 y 3.

# 1. Despliegue del cluster

## 1.1. Implementación del tutorial

En esta sección debe explicar como desarrolló el tutorial proveído para implementar un cluster compuesto de una máquina maestra y 4 trabajadores. Debe indicar todos los comandos o scripts no incluidos en el tutorial que haya utilizado para desarrollar esta actividad. Se provee un ejemplo de como incorporar código escrito en bash (puede incluir varios fragmentos independientes, no es necesario que estén todos en el mismo bloque de código).

Todos los snippets de código que sean proveídos deben ser adjuntados al archivo zip de la entrega respetando la numeración que tengan en el informe.

Las imágenes deben ser referenciadas como “Figura 1”, al igual que los snippets de código como “Código 1”

Para la creacion del Cluster en Amazon Web Services, fue necesario seguir los pasos indicados en el tutorial entregado en el Slide T01 - Hadoop Cluster - v1.0 entregado en Aula. El instructivo constaba de los siguientes pasos:

- Crear la cuenta de AWS
- Preparar en entorno del Cluster
- Configurar la maquina Master
- Configurar todos los Workers
- Inicializar el Cluster
- Ejecutar el script de Map Reduce con Hadoop

### Crear la cuenta de AWS

Para la creacion de la cuenta de AWS simplemente segui las indicaciones en el correo de invitacion indicado por el profesor. Posterior a eso, me fui directamente a la seccion de Modules, en donde estaba el Modulo de Launch AWS Academy Learner Lab, que sirve para entrar a la Console Home de AWS.

## Preparar en entorno del Cluster

Una vez en Console Home, creamos una instancia de tipo EC2, para la cual creamos los Key Pairs, el Security Group y la instancia Master.

Para los Key Pairs, creamos una llave de tipo RSA con formato de archivo PEM.

Para el Security Group, creamos tres reglas, una Inbound Rule de tipo SSH con tipo de fuente Anywhere-IPv4 (para entrar por SSH), una Inbound Rule de tipo Custom TCP con rango de puerto 8088 tipo de fuente Anywhere-IPv4 (para la consola web de Hadoop), una Inbound Rule de tipo Custom TCP con rango de puerto 9870 tipo de fuente Anywhere-IPv4 (para la consola web de HDFS) y finalmente, una Inbound Rule de tipo Custom TCP con rango de puerto de 0 a 65535 tipo de fuente Anywhere-IPv4.

Para la instancia Master creamos un nodo con Ubuntu 24.04 de tipo t2.micro y con 16gb de almacenamiento. Esta instancia usara la Key y Security Group recién creados (al igual que las instancias Worker mas adelante). Una vez configurados los parametros, lanzamos la instancia y nos conectamos por SSH.

## Configurar la maquina Master

El primer paso es logicamente actualizar los paquetes del sistema, para esto es necesario ejecutar el siguiente comando:

Una vez hecho esto, instalamos el paquete correspondiente a Java / OpenJDK 11.

Una vez que hayamos resuelto todos los paquetes necesarios, tenemos que crear una llave publica con el siguiente comando:

Esta llave publica la guardaremos en la carpeta de Authorized Keys para SSH, lo cual permitira al Master conectarnos a si mismo por LocalHost.



**Figura 1:** Captura de pantalla de la “AWS Management Console - EC2” que muestra las máquinas del cluster creado con el tutorial *Se debe ver la consola completa. En las columnas seleccione: Name, Instance ID, Instance state, Instance type, Status check, Public IPv4 address, Private IP Address*

## Configurar todos los Workers

## Inicializar el Cluster

## Ejecutar el script de Map Reduce con Hadoop

```
1 for linea in $(cat archivo)
2 do
3     echo $linea
4 done
```

**Código 1:** Código utilizado en la implementación del tutorial

## 1.2. Expansión del cluster

Desarrolle un procedimiento para expandir el tamaño del cluster de 4 a 8 trabajadores. Las máquinas de la ampliación también deben ser de tipo **t2.micro**. Considere que una ampliación de la cantidad de máquinas puede conllevar cambios en algunos de los parámetros de configuración establecidos en los archivos **.xml**. Si es así, indique que parámetros fueron modificados y la razón de la siguiente forma:



**Figura 2:** Captura de pantalla de la “AWS Management Console - EC2” que muestra las máquinas del cluster expandido. Se debe ver la consola completa. En las columnas seleccione: Name, Instance ID, Instance state, Instance type, Status check, Public IPv4 address, Private IP Address

- **aasdf.qwer.zxcv:** Valor cambia de x a y debido a el aumento de la capacidad
- **aasdf.qwer.zxcv:** Valor cambia de x a y debido a el aumento de la capacidad

En esta sección indique el procedimiento desarrollado y agregue cualquier código que haya sido utilizado en caso que este sea diferente o complementario al código utilizado en la subsección anterior. Se provee un ejemplo de como incorporar código escrito en bash (puede incluir varios fragmentos independientes, no es necesario que estén todos en el mismo bloque de código).

```
1 for linea in $(cat archivo)
2 do
3     echo $linea
4 done
```

**Código 2:** Código utilizado en la expansión del cluster

### 1.3. Cliente web

La figura 3 muestra la consola web de Hadoop y la figura 4 la consola web del HDFS.





**Figura 3:** Captura de pantalla del cliente web de Hadoop. Mostrar la consola web de Hadoop en la sección de aplicaciones terminadas, en las que se debe ver a lo menos los trabajos de la validación del cluster original y los trabajos que haya usado para validar el funcionamiento de la expansión del cluster

En esta sección explique el contenido que se está mostrando en cada una de las figuras referenciadas.

## 2. Instalación de Apache Hive

### 2.1. Procedimiento

Desarrolle un procedimiento para instalar **Apache Hive** en su cluster. Utilice la versión 4.0.1. Tenga en consideración:

- Las máquinas t2.micro son muy limitadas para levantar el servicio de Hive. Para esta sección se sugiere subir la máquina maestra a tipo **t3.medium** y los trabajadores a **t3.small**. Para cambiar el tipo de máquina no es necesario volver a desplegarla, basta con detener la máquina, cambiar su tipo desde la consola, y volver a iniciarla.
- Lograr la configuración correcta para Hive es un procedimiento que requiere bastante conocimiento y pruebas. El archivo zip de las instrucciones incluye el archivo **hive-example.xml** con la configuración a utilizar.



**Figura 4:** Captura de pantalla del cliente web de HDFS. **Mostrar la consola web de HDFS mostrando alguno de los archivos que hayan sido agregados al dfs.**

En esta sección indique el procedimiento desarrollado y agregue cualquier código que haya sido utilizado para instalar y probar Apache Hive. Incluya el documento de configuración utilizado y explique cada uno de los parámetros definidos en este.

## 2.2. Prueba

Para probar que la instalación de Hive funciona correctamente, puede utilizar el procedimiento disponible en el fragmento de código 3. Este ejemplo asume que el archivo utilizado para probar el cluster **sw-script-e04.txt** se encuentra disponible en el DFS. El resultado esperado para la prueba indicada se entrega en el fragmento 4

```
1 CREATE DATABASE sw_dialogs;
2
3 USE sw_dialogs;
4
5 CREATE TABLE sw04_dialogs (
6     line      INT,
7     character STRING,
8     dialog    STRING
9 )
10 ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
11 WITH SERDEPROPERTIES (
12     "separatorChar" = " ",
13     "quoteChar" = "\""
14 )
15 STORED AS TEXTFILE
16 TBLPROPERTIES ("skip.header.line.count"="1");
17
18 LOAD DATA INPATH
19 '/data/sw-script-e04.txt'
20 INTO TABLE
21 sw04_dialogs;
22
23 SELECT character, COUNT(*) AS lines
24 FROM sw04_dialogs
25 GROUP BY character
26 ORDER BY lines DESC
27 LIMIT 12;
```

**Código 3:** Ejemplo de uso de Apache Hive

```

1 +-----+-----+
2 | character | lines |
3 +-----+-----+
4 | LUKE      | 524   |
5 | HAN       | 135   |
6 | THREEPIO  | 119   |
7 | BEN       | 82    |
8 | LEIA      | 57    |
9 | VADER     | 41    |
10 | RED LEADER | 73    |
11 | BIGGS     | 34    |
12 | TARKIN    | 28    |
13 | OWEN      | 25    |
14 | TROOPER   | 19    |
15 | WEDGE     | 14    |
16 +-----+-----+
17 12 rows selected (67.678 seconds)

```

**Código 4:** Resultado esperado para ejemplo de uso de Apache Hive. Algunos de los valores han sido alterados, deben ser corregidos con el resultado de su procedimiento.

### 3. Exploración del HDFS

Desarrolle y explique un script que utilizando bash permita obtener la lista de bloques en las que está guardado un archivo en el DFS, incluyendo las direcciones IP privadas de las máquinas que guardan cada copia del bloque. La salida del script se debe ver como lo indicado en el fragmento 5.

```

1 File:          /mydata/myfile.txt
2 Blocks:        2
3 Avg. Replication: 3.0
4
5 0 - blk_1073742511_1694 - 172.31.40.100 172.31.32.111 172.31.35.247
6 1 - blk_1073742511_1695 - 172.31.32.111 172.31.210.111 172.31.35.247

```

**Código 5:** Ejemplo de uso de Apache Hive

Muestre el script desarrollado en el Código 6 y explique cada una de las partes del script.

```
1 for linea in $(cat archivo)
2 do
3     echo $linea
4 done
```

**Código 6:** Script de reporte de DFS

En esta sección indique el procedimiento desarrollado y agregue cualquier código que haya sido utilizado para desarrollar el script. Utilizando la AWS-CLI, descargue en el nodo maestro del cluster el archivo `s3://utfsm-inf356-dataset/vlt_observations_000.csv`<sup>1</sup>

## 4. Uso del cluster

### 4.1. Importación

Desarrolle un procedimiento que permita importar el archivo **vlt\_observations\_000.csv** a Hive desde el DFS respetando las columnas del archivo. Indique el código utilizado.

### 4.2. Parsing

Desarrolle una propuesta para asignar un tipo apropiado a los datos importados y desarrolle un procedimiento que permita dar este formato creando una nueva tabla. Indique el código utilizado.

### 4.3. Análisis

Piense 3 métricas sobre los datos interpretados, describa cada una de estas métricas y provea el código para obtener el resultado. Las métricas deben tener como mínimo 2 elementos de análisis (agrupamiento, contar, promedio, etc.). Ejemplos de métricas posibles son:

- Cantidad de observaciones por cada tipo de observación (agrupa y cuenta)

---

<sup>1</sup>Este archivo es público y está guardado en un bucket S3, por lo que debe utilizar la opción **-no-sign-request**. Este archivo pesa 371.1[MB] y es un archivo **csv** como el mismo formato al descargar [https://archive.eso.org/eso/eso\\_archive\\_main.html](https://archive.eso.org/eso/eso_archive_main.html) con todos los campos. Una vez descargado, coloque el archivo en la carpeta **data** del DFS y muestre el resultado del script desarrollado previamente.

- Ángulo promedio de declination de las observaciones del set para cada instrumento (agrupa y promedia)
- Seeing promedio por hora de observación (agrupa y promedia)

Provea un análisis sobre el desempeño del cluster al realizar estas operaciones. Incluya mediciones como tiempo de cómputo, máquinas usadas, cantidad de trabajos, cantidad de mappers y reducers, etc.