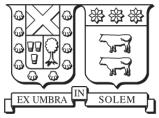


UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

# Trabajo Práctico 1

INF356 - 2025-1 - 200

## Computación Distribuida para Big Data

21 de abril de 2025 - v1.0

Miguel Soto | 204303630-0
miguel.sotod@sansano.usm.com

# Índice



# Índice de figuras

# Índice de fragmentos de codigo

# 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. Despligue 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 proveidos 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”

```
for linea in $(cat archivo)
do
    echo $linea
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 siguueinte forma:

- **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).

**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

```
for linea in $(cat archivo)
do
    echo $linea
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.

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

**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

**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

**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.

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.

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

```
CREATE DATABASE sw_dialogs;

USE sw_dialogs;

CREATE TABLE sw04_dialogs (
  line      INT,
  character STRING,
  dialog    STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.OpenCSVSerde'
WITH SERDEPROPERTIES (
  "separatorChar" = " ",
  "quoteChar" = "\""
)
STORED AS TEXTFILE
TBLPROPERTIES ("skip.header.line.count"="1");

LOAD DATA INPATH
'/data/sw-script-e04.txt'
INTO TABLE
sw04_dialogs;

SELECT character, COUNT(*) AS lines
FROM sw04_dialogs
GROUP BY character
ORDER BY lines DESC
LIMIT 12;
```

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

```
+-------------+--------+
|  character  | lines  |
+-------------+--------+
| LUKE        | 524    |
| HAN         | 135    |
| THREEPIO    | 119    |
| BEN         | 82     |
| LEIA        | 57     |
| VADER       | 41     |
| RED LEADER  | 73     |
| BIGGS       | 34     |
| TARKIN      | 28     |
| OWEN        | 25     |
| TROOPER     | 19     |
| WEDGE       | 14     |
+-------------+--------+
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.

```
File:             /mydata/myfile.txt
Blocks:           2
Avg. Replication: 3.0

0 - blk_1073742511_1694 - 172.31.40.100 172.31.32.111 172.31.35.247
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.

```
for linea in $(cat archivo)
do
    echo $linea
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`[1]

# 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)

[1]Este archivo es público y está guardado en un bucket S3, por lo que debe utilizar la opcion **–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` 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.