MÁSTER

INFRAESTRUCTURAS COMPUTACIONALES PARA EL PROCESAMIENTO DE DATOS MASIVOS

PRÁCTICA DEL MÓDULO 1: ECOSISTEMA DE PROCESAMIENTO PARALELO PARA DATOS MASIVOS: APACHE HADOOP

Versión 1.0

Dr. Agustín C. Caminero Herráez — Dr. Rafael Pastor Vargas

MÁSTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS

Contenido

Introducción

En este documento se presenta el Trabajo Práctico (TP) del módulo 1 de la asignatura "INFRAESTRUCTURAS COMPUTACIONALES PARA EL PROCESAMIENTO DE DATOS MASIVOS", del "MÁSTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS" de la UNED.

Este trabajo se realiza de forma individual. En las siguientes secciones se exponen los diferentes ejercicios que es necesario implementar para este trabajo.

Se proponen dos ejercicios diferentes. El primer ejercicio consiste en la implementación de un trabajo MapReduce con Python, y se valora sobre 8 puntos. El segundo ejercicio consiste en la implementación de un trabajo utilizando Hive y se valora sobre 2 puntos.

La forma de evaluar el trabajo se hará en base a lo siguiente:

- **Jupyter notebook** con el código realizado. Se debe incluir no solamente el código sino también las explicaciones necesarias, imágenes, … de forma que el notebook sea autocontenido. **Al principio del notebook se debe indicar el nombre del/la estudiante**. **Los notebooks tienen que estar ejecutados (deben mostrar la salida de la ejecución del código realizado), y se deben incluir las órdenes necesarias para demostrar el correcto funcionamiento de los desarrollos.** Por ejemplo, para demostrar que una tabla se ha creado y cargado de datos correctamente, habría que ejecutar un select que devuelva algunos de los contenidos de la tabla. Los notebooks no deben contener órdenes que ralenticen su carga de forma innecesaria, por ejemplo no se deben utilizar consultas select * from tabla para tablas con más de una decena de filas, habría que utilizar la cláusula limit.
- De forma optativa, se puede incluir una **memoria explicativa**.

Se valorará positivamente se incluya un apartado final donde se explique la opinión del/la estudiante sobre este trabajo, puntos fuertes y/o débiles, recomendaciones para el futuro, así como una valoración general de este módulo.

Este material se deberá incluir en un fichero comprimido y enviado a través del curso virtual dentro de los plazos establecidos para su entrega. El nombre de dicho fichero comprimido deberá tener la estructura *MR-ApellidosNombre.zip*, donde *Apellidos* y *Nombre* deben sustituirse por los valores correspondientes para el/la estudiante que realiza el envío.

A continuación se detallan los ejercicios a completar.

Ejercicio 1: MapReduce.

Este ejercicio consta de 3 partes, que se detallan en este apartado. Para cada uno de las partes es necesario presentar lo siguiente:

- El diseño del programa MapReduce. Este diseño debe contener al menos respuestas a las siguientes preguntas:
    - ¿Cuántos pasos MapReduce son necesarios?
    - ¿Qué hace cada función de cada paso? (No es necesario código ejecutable, una descripción en texto o en pseudocódigo es suficiente).
    - ¿Qué datos se pasan de una función a la siguiente?
- Implementación de este diseño. Este código debe ejecutarse en el entorno de desarrollo propuesto para el tema de MapReduce.

El detalle de cada apartado se encuentra a continuación.

Ejercicio 1.1: Contador de clientes valorados por países.

Partiendo de los ficheros de datos de países y desarrolla el código que devuelva cuántos clientes con valoración "bueno" hay en cada país. En concreto, la salida del programa MapReduce debe ser un fichero con el contenido que se muestra en la Figura 1 . Este ejercicio tiene una puntuación de hasta 3 puntos sobre 10.

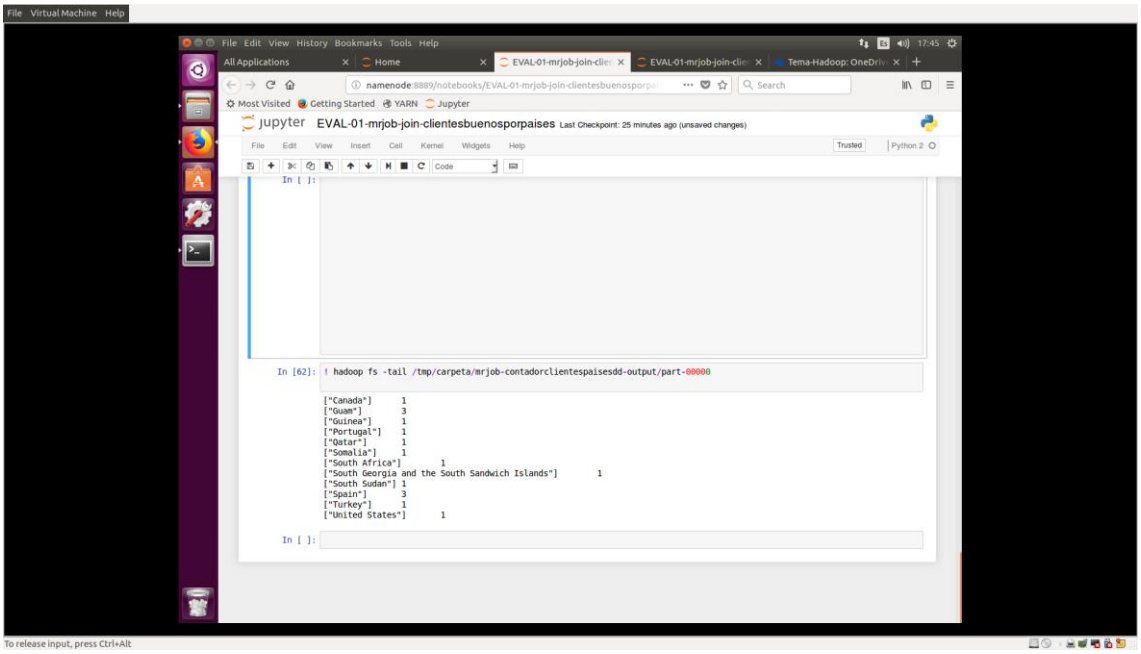

Figura 1. Fichero de salida del contador de buenos clientes.

Ejercicio 1.2: País con mejores clientes.

Partiendo del código implementado en el ejercicio anterior, extiéndelo para que devuelva el país en el que hay más clientes valorados como "bueno". En el caso de que haya más de un país con el mismo número de clientes buenos empatados en el primer lugar, se devolverá solamente uno de ellos. El resultado de este ejercicio se muestra en la Figura 2. Este ejercicio tiene una puntuación de hasta 3 puntos sobre 10.

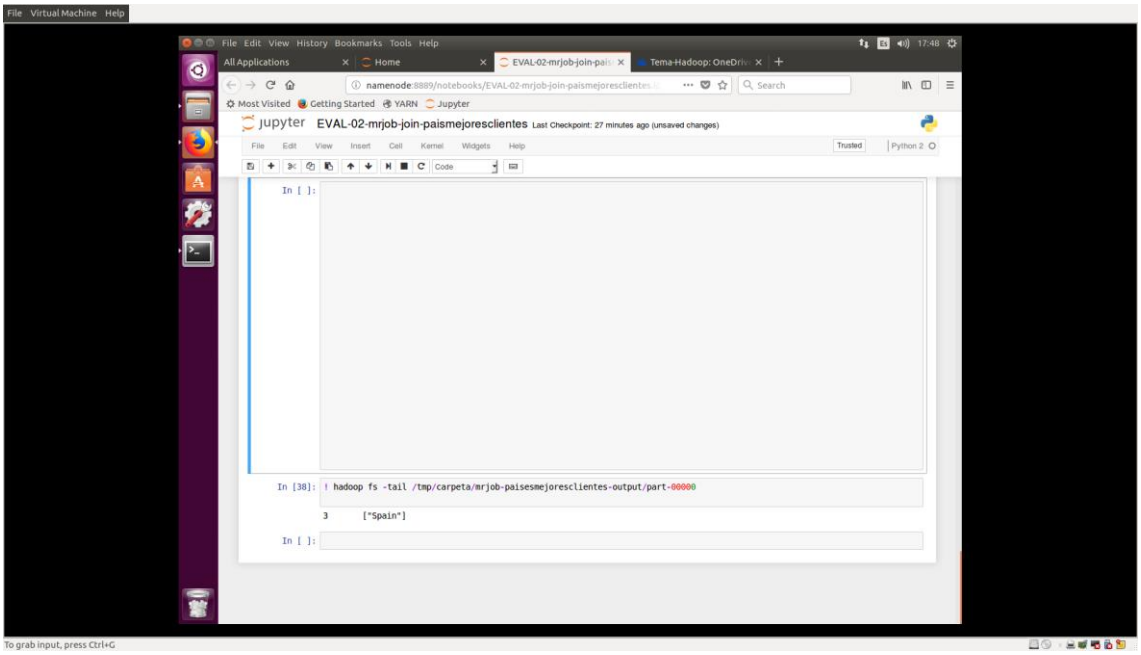

Figura 2. Fichero de salida con los mejores clientes.

Ejercicio 1.3: Mejorando el país con mejores clientes.

Partiendo del código implementado para el ejercicio anterior, mejóralo para que, en el caso de que haya más de un país empatado con el mayor número de buenos clientes, se devuelvan todos esos países. Utilizando los ficheros de datos sobre países y clientes vistos anteriormente, la salida de este programa MapReduce debería ser la que se muestra en la Figura 3. Este ejercicio tiene una puntuación de hasta 4 puntos sobre 10.

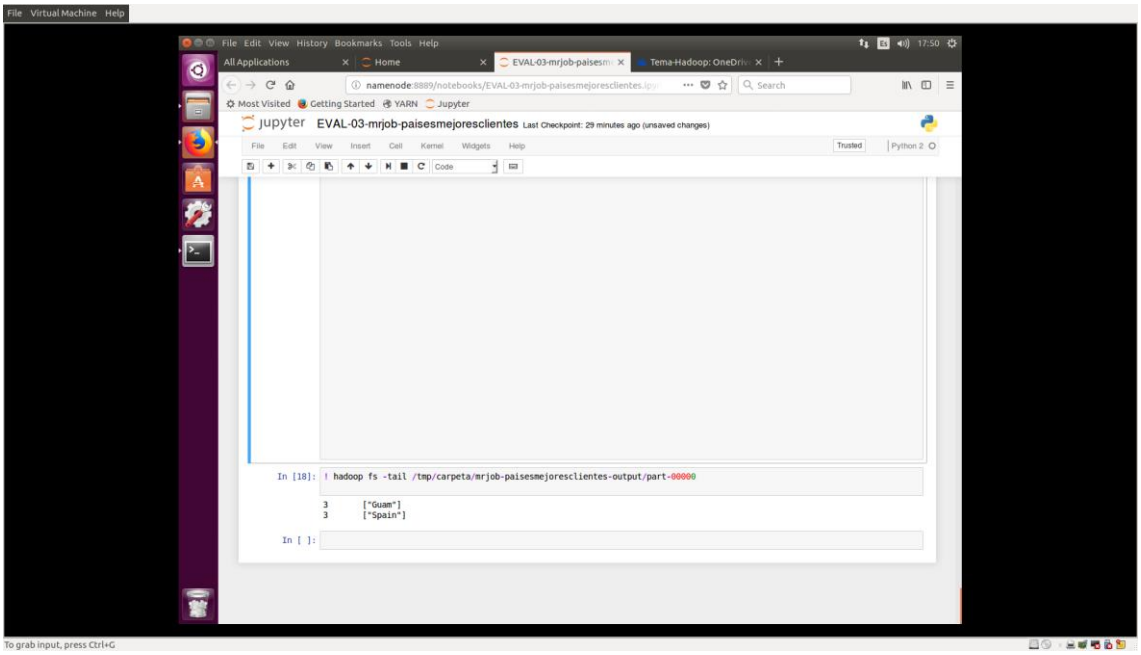

Figura 3. Fichero de salida con los mejores clientes, mejorado.

Ejercicio 2: Hive

En este trabajo se van a utilizar ficheros (**disponibles en el curso virtual**) que contienen datos del Gobierno de Argentina (https://datos.gob.ar/), con información sobre todos los ámbitos del país: Ciencia y tecnología, Educación, Energía, Medio Ambiente, etc.

Como la licencia es abierta y se permiten todo tipo de modificaciones, se han limpiado previamente los datasets para conseguir una mejor comprensión de estos, eliminar valores erróneos, etc.

El dataset a utilizar es:

- "Convocatorias 2020": Se trata de un dataset del gobierno argentino con el listado de las convocatorias abiertas durante el año 2020 a través de la Plataforma COMPR.AR. Podemos encontrar la información en (https://www.datos.gob.ar/dataset/jgm-sistema-contrataciones-electronicas/archivo/jgm_4.19), de la fuente de datos abiertos argentinos (https://datos.gob.ar/).
  Contiene un total de 13549 registros con las convocatorias a los procedimientos de las convocatorias abiertas durante el año 2020. Los campos que encontramos son:
  - Número Procedimiento: Un identificador único para cada procedimiento de convocatoria.
  - Nro SAF: Número del Servicio Administrativo Financiero.
  - Descripcion SAF: Descripción del Servicio Administrativo Financiero.
  - Nro UOC: Número de la Unidad Operativa de Contrataciones.
  - Descripcion UOC: Descripción de la Unidad Operativa de Contrataciones.
  - Tipo de Procedimiento: Tipo de proceso de contratación (ej., licitación pública, contratación directa).
  - Modalidad: Modalidad específica del procedimiento dentro del tipo de procedimiento.
  - Apartado Directa: Indica el apartado para la contratación directa.
  - Ejercicio: Año fiscal en el que se realiza la convocatoria.
  - Fecha de Publicación: Fecha en la que se publicó la convocatoria.
  - Fecha de Apertura: Fecha en la que se abrirán las ofertas o se realizará la adjudicación.
  - Etapa: Indica si la etapa es única o múltiple.
  - Alcance: Extensión o ámbito geográfico de la convocatoria (nacional o internacional).
  - Nombre del Procedimiento: Nombre específico del procedimiento.

- Objeto del Procedimiento: Descripción del propósito del procedimiento.
- Monto Estimado: Valor estimado del contrato o adquisición.
- Tipo de Operación: Clasificación de la operación.

Partiendo de estos datos, desarrolla las órdenes de HiveQL que implementen las siguientes tareas.

1. (1 puntos) Crea las tablas necesarias para almacenar los datos. Pueden ser internas o externas en función de los datos que se desee. La decisión de interna o externa debe estar razonada.

2. (1 puntos) Importa los datos en las tablas creadas.

3. (4 puntos) Crea las consultas de Hive necesarias para responder las siguientes cuestiones:
   - (0.5 puntos) Devuelve los registros cuyo tipo de procedimiento de contratación sea contratación directa.
   - (0.5 puntos) Devuelve los registros cuyo tipo de procedimiento de contratación sea contratación directa y cuyo ejercicio fiscal sea anterior a 2020.
   - (1 punto) ¿Cuántos tipos de SAF tenemos para los registros de las convocatorias? Hay que contar el número de valores únicos de la columna que contiene las descripciones del Servicio Administrativo Financiero (Descripcion SAF).
   - (1 punto) ¿Cuál es el máximo presupuesto estimado de estas convocatorias? Utilizaremos el campo monto estimado de la tabla convocatorias.
   - (1 punto) ¿Cuál es el monto estimado medio que se emplea en las convocatorias agrupado por su alcance (nacional e internacional)?

4. (4 puntos) Selecciona un dataset formado por uno o más ficheros, impórtalo en las tablas de Hive necesarias e implementa al menos dos consultas sobre dicho dataset. Para realizar este ejercicio, es necesario seguir los siguientes pasos:
   - Selecciona una fuente de datos públicamente disponible en Internet. Puede ser un dataset o alguna otra fuente de datos como por ejemplo una red social. Si es una fuente de datos, deberás utilizar alguna de las herramientas de inyección de datos y serdes vistos en la asignatura para capturar datos para realizar este ejercicio. Si se decide utilizar un dataset, en los siguientes enlaces se pueden encontrar algunos de libre acceso (existen más repositorios de datasets):
     - https://www.kaggle.com/datasets
     - https://data.worldbank.org/

  - https://datasetsearch.research.google.com/
  - https://console.cloud.google.com/marketplace/browse?filter=solution-type:dataset&_ga=2.67976598.645382453.1601640482-195064274.1601640482
  - https://datos.gob.ar/
  - https://datos.gob.es/es

- Una vez seleccionada la fuente de datos, el primer paso será publicar un mensaje en el foro del módulo indicando el enlace a dicha fuente de datos. Además, revisa si hay otro/a estudiante que haya publicado previamente dicha fuente de datos, en cuyo caso deberás elegir otra distinta. El objetivo es evitar que dos o más estudiantes utilicen los mismos datos en sus trabajos.

- Utilizando la fuente de datos seleccionada, importa los datos en Hive e implementa al menos dos consultas sobre dichos datos.

- Esta parte se evaluará de la siguiente forma:
  - (1 puntos) Selección de dataset, creación de las tablas e importación de los datos.
  - (3 puntos) Creación de consultas.

- Se valorará positivamente la complejidad de los datos utilizados así como la de las consultas implementadas. En el caso de que no sea necesario utilizar herramientas de inyección de datos y serdes, se podrá conseguir la máxima puntuación en este apartado si las consultas tienen complejidad suficiente.