


MÁSTER



INFRAESTRUCTURAS COMPUTACIONALES PARA EL
PROCESAMIENTO DE DATOS MASIVOS

TRABAJO PRÁCTICO MÓDULO 4
SERVICIOS GESTIONADOS EN LA NUBE PARA EL
PROCESAMIENTO DE DATOS MASIVOS



2021-2022

Dr. Rafael Pastor Vargas
Dr. Agustín C. Caminero Herráez — Dr. Salvador Ros Muñoz

MÁSTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS

Contenido

1	Introducción	1
2	Opción 1: Implementación del Benchmark para Spark en AWS.	2
3	Opción 2: Implementación del benchmark para Spark en GCP.....	3

1 Introducción

En este documento se presenta el Trabajo Práctico (TP) del módulo 4 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 usando dos proveedores de servicio en la nube, de los cuáles el estudiante deberá seleccionar uno: Amazon Web Services (AWS) y Google Cloud Platform (GCP). En ambos casos se debe implementar un *benchmark* que permita la evaluación de un despliegue basado en Spark, en términos de velocidad, rendimiento y utilización de recursos.

Existen multitud de *benchmarks* que permiten la evaluación de *frameworks* de *big data*. Estos *benchmarks* se basan en la utilización con mayor o menor exhaustividad de diversas configuraciones hardware y software con el objetivo de extraer datos y realizar análisis sobre las implicaciones derivadas de su rendimiento.

El estudiante deberá seleccionar el *benchmark* que se utilizará en este trabajo, y describir las características del mismo: tipos de tareas que se ejecutan (*workloads*), configuraciones y recursos necesarios, etc. A continuación, se proponen algunos *benchmarks* que se pueden utilizar, aunque el estudiante podrá seleccionar cualquier otro argumentando su utilización:

<https://github.com/Intel-bigdata/HiBench/blob/master/docs/run-sparkbench.md>

<https://github.com/databricks/spark-perf>

<https://github.com/ehiggs/spark-terasort>

<https://github.com/CODAIT/spark-bench>

<https://github.com/BBVA/spark-benchmarks/blob/master/docs/TestDFSIO.md> (Solo version Spark 2.1.x)

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

TP4: SERVICIOS GESTIONADOS EN LA NUBE PARA EL PROCESAMIENTO DE DATOS MASIVOS

- Scripts de ejecución (Python o Notebooks) con el código realizado, listos para ser ejecutado en cada uno de los proveedores de servicio. Se debe incluir no solamente el código sino también las explicaciones necesarias, imágenes, ... de forma que el script/notebook sea autocontenido. Al principio de los notebooks/scripts se debe indicar el nombre del/de la estudiante.
- Se debe incluir una memoria explicativa indicando la metodología de desarrollo empleada y recursos/servicios usados en cada plataforma. Es decir, cual es la relación entre servicios usados (almacenamiento y procesamiento) así como las conexiones realizadas entre servicios para implementar la solución asociada al benchmark. Se debe añadir también un apartado indicando cómo se ha desarrollado la solución concreta en cada plataforma, indicando los pasos seguidos.
- Estas memorias podrán incrementar la nota completa de la práctica ejercicio si aporta contenido significativo adicional (referencias, documentos adicionales, reflexiones, etc.)

Se valorará positivamente que los notebooks/scripts o la memoria contenga un apartado final donde se explique la opinión del/la estudiante sobre cada ejercicio del 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 *TP4-ApellidosNombre.zip*, donde Apellidos y Nombre deben sustituirse por los valores correspondientes para el/la estudiante que realiza el envío (evitar el uso de acentos o símbolos).

A continuación, se detallan los proveedores de servicios en la nube entre los cuáles se tendrá que elegir el utilizado para la implementación del *benchmark* elegido: AWS y GCP.

2 Opción 1: Implementación del Benchmark para Spark en AWS.

Amazon Web Services (AWS) es una plataforma que proporciona una amplia gama de servicios de infraestructura para su utilización profesional, como potencia de cómputo, almacenamiento, redes o bases de datos. **Amazon Elastic Compute Cloud (EC2)** es la parte central de su plataforma basada en el alquiler de computadoras virtuales (denominadas instancias) en las cuáles el cliente puede

TP4: SERVICIOS GESTIONADOS EN LA NUBE PARA EL PROCESAMIENTO DE DATOS MASIVOS

ejecutar sus aplicaciones. Estas instancias presentan multitud de configuraciones hardware y software, por lo que el usuario puede seleccionar aquellas que mejor se adecuen a sus intereses particulares.

Para este trabajo práctico en concreto y la implementación del *benchmark*, se recomienda el uso en AWS de los siguientes servicios:

- **Amazon Simple Storage Service (S3)** es un servicio de almacenamiento diseñado para facilitar a los desarrolladores recursos de computación escalables, basados en una interfaz de servicios Web. Esta interfaz simple permite el almacenamiento y recuperación de cualquier cantidad de datos en cualquier momento. Sus características de administración permiten la organización de los datos y la configuración de controles de acceso personalizados. Gracias a S3, cualquier desarrollador puede utilizar una infraestructura de almacenamiento de datos estándar, altamente escalable, fiable, segura y rápida. De esta forma, se facilitan enormemente tareas de recuperación, copia de seguridad o utilización de grandes conjuntos de datos para sistemas big data. En S3, los datos se almacenan en contenedores denominados buckets, a los que se les proporcionan diversos permisos de acceso, características de localización (región geográfica donde se almacenarán), así como logs de acceso.
- **Amazon Elastic MapReduce (EMR)**, por su parte, es la plataforma de big data nativa en la nube, que permite el procesamiento eficiente de grandes volúmenes de datos, con una gran escalabilidad y rentabilidad. Diversas herramientas de código abierto como Apache Spark, Apache Hive o Apache HBase se integran en EMR. En combinación con la escalabilidad proporcionada por EC2 y S3, se consigue la creación, mantenimiento y utilización a gran escala de clústeres por un coste mucho menor en relación a infraestructuras locales tradicionales. Además, la utilización a demanda, escalabilidad automática, alta disponibilidad y duración prolongada de dichos clústeres permite la ejecución de casos de uso único.

3 Opción 2: Implementación del benchmark para Spark en GCP.

Los servicios de computación en la nube ofrecidos por Google se engloban dentro de **Google Cloud Platform (GCP)**, su plataforma que reúne soluciones de acceso, almacenamiento y gestión de datos, así como todo tipo de tecnologías para el procesamiento de dichos datos. La plataforma dispone de recursos físicos distribuidos por todo el mundo que se ponen a disposición de los usuarios en forma

TP4: SERVICIOS GESTIONADOS EN LA NUBE PARA EL PROCESAMIENTO DE DATOS MASIVOS

de servicios. La interconexión de estos servicios permite la implementación de infraestructuras y casos de uso específicos.

Para este trabajo práctico en concreto y la implementación del *benchmark*, se recomienda el uso en GCP de los siguientes servicios:

- El servicio de ejecución de clústeres de Apache Spark y Apache Hadoop de ofrecido en GCP se denomina **Dataproc**. Este servicio permite acelerar de forma sencilla y rentable todo tipo de procesos de computación en la nube. El servicio está integrado con el resto de los servicios de la plataforma. Su escalabilidad permite modificar el tamaño de los clústeres en cualquier momento, lo que permite una mejora sustancial en lo referente al tiempo que se emplea en supervisar la infraestructura utilizada. Al igual que EMR, se proporciona un ecosistema completo de código abierto, incluyendo herramientas, librerías y documentación de Apache Spark y Hadoop. Dataproc ofrece igualmente gestión y escalado automático de clústeres, herramientas de desarrollo, configuración y gestión, alta flexibilidad y disponibilidad, gestión de versiones y herramientas de seguridad, entre otras características.
- **Google Cloud Storage** es el servicio de almacenamiento en la nube ofrecido por Google. Ofrece un almacenamiento seguro, con alta durabilidad y escalabilidad, así como una API unificada para integrar todo tipo de almacenamiento, optimizando la gestión del ciclo de vida de los objetos y el acceso instantáneo a los mismos.