Cloud Computing: Servicios y aplicaciones

Curso 2019-2020

Práctica entregable 2: Procesamiento *Big Data* en *Cloud Computing*

Descripción

Cloud computing es la plataforma natural para el procesamiento de datos con características genuinas de "Big Data". Las dos plataformas más utilizadas para ello son Hadoop y Spark. El objetivo de esta práctica es ayudar al alumno a alcanzar conocimientos y desarrollar habilidades básicas en el uso de estas dos plataformas.

De forma más concreta, los objetivos específicos que el alumno debe alcanzar con esta práctica son:

- Implementar programas que usen técnicas de procesamiento de datos masivos con Hadoop.
- Implementar programas para el procesamiento de datos masivos en Spark con Python.

Estos objetivos deben alcanzarse a través de la realización de las tareas que se detallan a continuación.

Tareas

1. Implementación de algoritmos MapReduce en Hadoop

Usando la plataforma Hadoop, debe implementarse (en el lenguaje de programación Java) un programa basado en el paradigma MapReduce que calcule la función estadística "desviación estándar".

La formula para calcular la desviación estándar es la siguiente:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

donde *n* es el número de elementos totales del conjunto de datos, x_i es cada dato de la muestra y $\bar{x}$ es el media aritmética del conjunto de datos.

Para verificar el funcionamiento correcto de la implementación realizada, se aplicará al cálculo de la desviación estándar de un conjunto de datos. El conjunto a emplear está ubicado en:

- Volumen HDFS del servidor hadoop.ugr.es,

- La ruta dentro del volumen HDFS es:
  /user/mp2019/ECDB-2012.training

Como resultado de esta tarea, el alumno deberá entregar:

a) El código fuente, organizado en tres módulos:

- `DevSTDMapper.java` ; que contiene la fase Map
- `DevSTDReducer.java` ; que contiene la fase Reduce
- `DevSTD.java` ; contiene el main para ejecutar la aplicación.

b) El valor de la desviación estándar de las nueve primeras columnas del conjunto de datos.

2. Implementación de algoritmos de clasificación en Spark con Python

La segunda parte de la práctica consiste en emplear distintos algoritmos para resolver un problema de clasificación. Se utilizará métodos implementados en la biblioteca MLlib de la plataforma Spark, invocados desde programas en Python (o R). Una vez creados se comparará el rendimiento de los distintos clasificadores para identificar cuál sea el más adecuado para el problema en cuestión. El estudio empírico debe incorporar al menos tres de los métodos incluidos en la biblioteca.

El problema a considerar está definido por un conjunto de datos, sobre el que ya se ha realizado una partición en partes de entrenamiento y prueba. Los resultados de rendimiento a comparara serán los obtenidos en el conjunto de prueba.

- Los conjuntos de datos están el volumen HDFS de hadoop.ugr.es.
- La ruta dentro de HDFS es:
  - Train: /user/mp2019/ECDB-2012.training
  - Test: /user/mp2019/ECDB-2012.test

Para acometer esta tarea, el alumno debe:

- Procesar el fichero de datos para crear uno nuevo que conste de:
  - Las 5 primeras columnas y la variable de clase (la última columna). En total tendrá 6 columnas.
  - Se llamará /user/*tuusuario*/ECDB-2012.small.training.
- Equilibrar el fichero resultante de datos para que tenga el mismo número de registros de las clases 0 y 1.
- Aplicar al menos tres clasificadores de la MLlib al conjunto de entrenamiento nuevo creado.
- Aplicar el modelo creado al conjunto de entrenamiento:
  - /user/mparra/ECDB-2012.test.
- Obtener los resultados clasificación para cada uno de los modelos utilizando al menos dos variaciones en los hiperparámetros de los algoritmos de clasificación.

- Crear una tabla con los resultados para los tres algoritmos y las variantes de parámetros aplicadas, para conocer la efectividad de la clasificación aplicada.
- Identificar el algoritmo que ha obtenido los mejores resultados.

Para esta parte será necesario entregar las aplicaciones en código Python (o R) que realizan el procesamiento de los datos y la ejecución de los clasificadores.

Documentación

La documentación constará de un documento breve con un máximo de tres páginas, donde se incluya:

- Nombre y DNI del alumno.
- Parte 1: Hadoop. Explicación de cómo se ha realizado (sin incluir código fuente) y responder a las preguntas de la sección.
- Parte 2: Spark+Python. Explicación de cómo se ha realizado, los resultados obtenidos y responder a las preguntas de la sección.
- Conclusiones
- Apartados extras aportados por los alumnos.

Evaluación

El trabajo realizado para esta práctica se evaluará con hasta 2 puntos sobre el total de puntos de la parte práctica de la asignatura. En esta evaluación se incluirá tanto el software (código fuente en java y python) como la documentación entregada (a través de la plataforma prado.ugr.es).

Plazo de entrega: 4 de junio de 2019 (hasta las 23:55).