UNIVERSIDAD
EL BOSQUE

Ingeniería de Sistemas
Big Data Analytics
Taller 1

Enlace de entrega: aquí
Fecha máxima de entrega: **Septiembre 16, 2021 11:59 PM**

El ecosistema de Hadoop hace referencia a un conjunto de tecnologías o herramientas que permiten implementar proyectos de Big Data, dado que están construidas para solucionar tareas de procesamiento distribuido y escalable de grandes volúmenes de datos. En su aspecto más básico, Hadoop está principalmente compuesto por Hadoop common, (*Hadoop Distributed File System*), YARN (*Yet Another Resource Negotiator*), MapReduce, componentes que se explorarán en este taller. A través de los años han aparecido otros componentes que han ganado gran relevancia y que amplían el espectro de casos de uso que pueden ser implementados dentro del ecosistema. Elementos como Spark, para el procesamiento distribuido in-memory, también será objeto de experimentación en este taller.

Parte 1

Hadoop se encuentra optimizado para ser ejecutado en sistemas operativos basados en el kernel Linux y requiere de la JVM (*Java Virtual Machine*) para su ejecución.

1. Si no dispone de un sistema operativo Linux, descargue e instale VirtualBox o VMWare.
2. Cree una máquina virtual e instale sobre esta un sistema operativo basado en Linux como Ubuntu.
3. Siga los pasos dispuestos en esta guía para desplegar Hadoop en modo pseudo-distribuido. También puede apoyarse de la guía oficial que encuentra en la página de Apache.

Parte 2

MapReduce es el componente de Hadoop que se encarga de procesar grandes volúmenes de datos de manera distribuida y escalable.

1. Una vez instalado Hadoop, en la guía oficial de Apache, sección “Execution”, pasos 4 al 7, encontrará las instrucciones para correr un programa de ejemplo. Ejecútelo, intente analizar los logs que se muestran en consola así como los archivos generados en la carpeta output. También puede ser útil analizar el código fuente del ejemplo. Responda:
    - ¿Qué resultados generó el programa y cuales son los pasos MapReduce que implementa?
2. En el mismo jar de ejemplos encontrará otros algoritmos con los que puede experimentar. Cargue al HDFS cualquier documento en texto plano (un poema, un libro, una canción, etc.) y ejecute el programa WordCount. Realice el mismo análisis anterior. Responda:
    - ¿Qué resultados generó el programa y cuales son los pasos MapReduce que implementa?

Parte 3

En los últimos años, Spark ha ganado una importante popularidad respecto a Hadoop/MapReduce para procesamiento distribuido de datos. La clave de Spark es su procesamiento en memoria. También tiene la ventaja de que puede ser programado en otros lenguajes más compactos como Scala, Python y R.

1. En una nueva máquina Linux, deseablemente, siga las instrucciones dispuestas en esta guía para instalar Spark en Ubuntu.
2. Revise el código de WordCount dispuesto como ejemplo en la web oficial de Spark y entiéndalo para que realice los cálculos de conteo de palabras sin tener en cuenta mayúsculas/minúsculas y signos de puntuación.
3. Ejecute los códigos original y extendido en PySpark y analice los resultados.
4. (BONO) Investigue una forma de realizar el conteo de palabras categorizando por sustantivos, adjetivos, verbos, preposiciones, etc.

Parte 4

Como con cualquier otra herramienta, desarrollar un programa de computador desde la interfaz de línea de comandos es una labor tediosa. Por lo general, los desarrolladores recurren a entornos de desarrollo (IDE) que ofrecen gran variedad de apoyos a la codificación en términos de evaluación de sintaxis, *debugging*, conexión con recursos externos, entre muchos otros.

1. Dentro de la máquina virtual, en la carpeta de su preferencia, clone el repositorio localizado en este link. Este repositorio contiene 2 scripts que serán trabajados más adelante así como una carpeta data con un archivo TXT. Sobre esta misma carpeta descargue y descomprima el archivo ubicado aquí.
2. Dentro del mundo de Python y el análisis de datos, Jupyter Notebooks es una de las herramientas de desarrollo más ampliamente usadas. Descargue e instale Anaconda en la misma máquina virtual en donde instaló Spark.
3. Este paquete de software, entre muchas otras cosas, contiene el entorno de desarrollo Jupyter Notebook. Para iniciarlo, desde una terminal y ubicado sobre la misma carpeta en la que clono el repositorio previamente, digite el comando: *jupyter lab --ip=0.0.0.0*.
4. Desde la misma máquina o desde otra máquina virtual o física abra un navegador web y diríjase a: http://<puerto>:8888/lab/. Si todo va bien hasta este punto, desde la página web que se abre debería poder visualizar la estructura de archivos del repositorio clonado.
5. Ejecute los scripts *spark-basics.ipynb* y *spark-data-analysis.ipynb*. Analice cada archivo y detalle los elementos más importantes de cada uno.

Forma de entrega

Cree un repositorio de GitHub y en un archivo Readme reporte con screenshots, más las descripciones que considere, todos los pasos solicitados anteriormente. También agregue los scripts IPYNB abordados en el último punto con soporte de su correcta ejecución.