**EVALUACIÓN**

**PROGRAMACIÓN CON MAPREDUICE**

**2024-1**

**CRITERIOS DE EVALUACIÓN:**

* Comprensión de los conceptos básicos de MapReduce y HDFS
* Habilidad para diseñar algoritmos MapReduce eficientes
* Habilidad para implementar algoritmos MapReduce en Hadoop

Los criterios para autoevaluación, coevaluación y heteroevaluación, se define a partir de la rúbrica anexa.

**Construcción del DataSet (individual): 10 pts**

Seleccionar, cada estudiante del grupo, 15 libros del proyecto web <https://www.gutenberg.org/>

General un listado con el link de descarga del libro en el formato Plain Text UTF-8

Realizar un script (Shell o Python) para descarga los 15 libros (con ***wget*** *o* ***curl***) al interior de la carpeta llamada **input**. Se puntúa si el script es genérico y hace la descarga de cualquier listado.

* *Entrega 1: Script individual con el nombre* ***1****\_****download\_book.sh.***

**Análisis de palabras comunes (grupal) 20 pts**

Emplear MapReduce para identificar las 20 palabras más recurrentes, con más de 10 caracteres de longitud, que aparezcan en al menos el 80 % de los libros del conjunto de datos analizado

* *Entrega 2: Script de MapReduce, con el nombre* ***2\_Wordcount.py****. se debe explicar dentro del script las líneas fundamentales de código.*
* *Entrega 3: Archivo de salida con los resultados, con el nombre* ***3\_out.txt***

**Análisis de tweets (grupal) 20pts**

Con el DataSet [tweets2016\_olympic\_rio.test](https://uredu-my.sharepoint.com/:u:/g/personal/ferney_beltran_urosario_edu_co/EYoc_dtRAQxOvRoAKwoL3hgBtlGlQ5RtF9KNNDT8YqzZqA?e=jY4LVL), colección de mensajes de Twitter recopilados durante los Juegos Olímpicos de Río 2016 de la API responder:

* ¿Cuál es el **minuto** más popular respecto a la actividad de Twitter y cuantos tweets hay de ese minuto? Se recomienda usar la librería **time** con los métodos **gmtime** y **strftime**
* ¿Cuál es el promedio de la longitud de texto de cada tweet?

Entregables:

* Archivos de texto plano con el resultado para cada pregunta en los archivos **4a\_out.txt y 4b\_out.txt**
* Script de mapreduce utilizado para cada pregunta en los archivos **4a\_tweetcount.py y 4b\_tweet average.**

Recuerdes Los datos se almacenan en formato CSV, y cada línea contiene los siguientes campos:

* 1. **epoch**\_**time**: Marca de tiempo UNIX del tweet en milisegundos desde el 01-01-1970.
  2. **tweetId**: ID único del tweet
  3. **tweet**: contenido del tweet que incluye el #hashtags
  4. **device**: información adicional meta-datos, donde se incluye el dispositivo usado

**MapReduce con DataSet del curso (grupal): 50 pts**

La base de datos "routes.csv" proviene de la base de datos de rutas de vuelo disponible en Kaggle, específicamente del conjunto de datos "Flight Route Database". Este conjunto de datos contiene información detallada sobre las rutas de vuelo a nivel mundial, incluyendo el código de la aerolínea, el ID de la aerolínea, aeropuertos de origen y destino, ID de aeropuertos de origen y destino, información sobre compartición de código (codeshare), cantidad de paradas y el equipo (tipo de avión) utilizado en cada ruta. Link (https://www.kaggle.com/datasets/open-flights/flight-route-database?resource=download)

**Descripción Detallada de la Base de Datos**

* **airline**: Código IATA o ICAO de la aerolínea que opera la ruta.
* **airline ID**: Identificador único de la aerolínea, según la base de datos de OpenFlights.
* **source airport**: Código IATA o ICAO del aeropuerto de origen.
* **source airport id**: Identificador único del aeropuerto de origen, según la base de datos de OpenFlights.
* **destination airport**: Código IATA o ICAO del aeropuerto de destino.
* **destination airport id**: Identificador único del aeropuerto de destino, según la base de datos de OpenFlights.
* **codeshare**: Indica si la ruta es operada bajo un acuerdo de compartición de código (codeshare). Los valores pueden estar ausentes si no aplica.
* **stops**: Número de paradas de la ruta. Una ruta directa tiene 0 paradas.
* **equipment**: Código(s) del equipo (tipo de avión) utilizado en la ruta.

**Problemática a resolver**

Objetivo: Analizar la conectividad y accesibilidad entre diferentes regiones del mundo a través de la red de rutas aéreas. Se busca identificar hubs aéreos globales, evaluar la accesibilidad de regiones remotas y analizar la densidad de la red de rutas aéreas entre continentes para ello resuelva con mapreduice las siguientes preguntas:

1. **Cuántas rutas opera cada aerolínea?**

El resultado debe ser algo de este estilo:

Aerolínea 2B: 42 rutas.

Aerolínea 2G: 21 rutas.

Aerolínea 2I: 18 rutas.

Aerolínea 2J: 22 rutas.

Aerolínea 2K: 12 rutas.

1. ¿Cuáles son las 5 rutas más transitada con conexión directa?

El resultado debe ser parecido a:

ORD-ATL: 20 conexiones directas.

ATL-ORD: 19 conexiones directas.

ORD-MSY: 13 conexiones directas.

HKT-BKK: 13 conexiones directas.

HKG-BKK: 12 conexiones directas.

1. ¿Cuáles son los 10 aeropuerto con más conexiones directas? El resultado puede ser algo de este estilo

ATL: 915 destinos directos.

ORD: 558 destinos directos.

PEK: 535 destinos directos.

LHR: 527 destinos directos.

CDG: 524 destinos directos.

FRA: 497 destinos directos.

LAX: 492 destinos directos.

DFW: 469 destinos directos.

JFK: 456 destinos directos.

AMS: 453 destinos directos.

Entregables:

* Archivos de texto plano con el resultado para cada pregunta en los archivos **5a\_out.txt, 5b\_out.txt y 5c\_out.txt**
* Script de mapreduce utilizado para cada pregunta en los archivos **5a\_airline.py, 5b\_route.py y 5b\_airport.py.**