# Trabajo Práctico Especial 2

**Álvarez Escalante, Tomás** (60127)
tomalvarez@itba.edu.ar

**Ferreiro, Lucas Agustín** (61595)
lferreiro@itba.edu.ar

**Gómez Kiss, Román** (61003)
romgomez@itba.edu.ar

01/11/2023

# Alquiler de Bicicletas

Grupo 5

72.42 Programación de Objetos Distribuidos

Instituto Tecnológico de Buenos Aires

# 1. Diseño de los componentes del MapReduce

## 1.1. Query 1

Para la query 1, decidimos que el mapper emita el par $key =< startStation, endStation >$ , $value = 1$, para luego en el reducer realizar la sumatoria y así obtener la cantidad total de viajes entre estaciones.

## 1.2. Query 2

Para la query 2, decidimos que el mapper emita el par $key = startStation, value =< 1, haversineDistance >$, para luego en el reducer realizar obtener el total de viajes y realizar la sumatoria de todos las distancias y así obtener el promedio de distancia aproximada para cada estación de inicio.

Se emite un 1 que simboliza la cantidad de viajes para esa distancia total, para poder utilizar el Combiner asociado para las sumas parciales, sin hacer aún la división.

## 1.3. Query 3

Para la query 3, decidimos que el mapper emita el par $key =< startStation, endStation >$ , $value =< startDate, tripDuration >$, para luego en el reducer obtener el viaje con mayor duración (y en caso de coincidencia quedarnos con el del día más "temprano").

## 1.4. Query 4

Para la query 4, se hace uso de dos operaciones MapReduce.

En el primero de ellos el mapper emite dos pares key/value:

1. $key =<$ startStation, startDate $>, value = -1$
2. $key =<$ endStation, endDate $>, value = 1$

De esta forma, en el reducer obtenemos la afluencia neta de una estación para cierto día, donde el -1 representa viajes salientes, y el 1 viajes entrantes.

En el segundo mapper se emite el par $key =$ stationId, $value =$ afluenciaNeta de un cierto día. En el reducer utilizamos 3 contadores que representaran la afluencia positiva, neutra y negativa según la afluencia neta emitida, y se devolverá una List<Long> con los 3 valores totales.

# 2. Análisis de tiempos para cada query

El análisis se realiza con un dataset de 200.000 líneas y hasta 4 nodos. Los parámetros utilizados para las queries respectivas fueron:

- Query 2: N=15
- Query 4: startDate=01/05/2021 y endDate=31/05/2021

Tabla 1: Tiempo de carga

| Query | 1 Nodo [ms] | 2 Nodos [ms] | 3 Nodos [ms] | 4 Nodos [ms] |
|---|---|---|---|---|
| Query 1 | 1015 | 2015 | 4975 | 3031 |
| Query 2 | 998 | 2921 | 2996 | 7013 |
| Query 3 | 997 | 2024 | 3027 | 2056 |
| Query 4 | 990 | 2021 | 4929 | 3033 |

## 2.1. Sin combiner

Tabla 2: Tiempo del MapReduce sin combiner

| Query | 1 Nodo [ms] | 2 Nodos [ms] | 3 Nodos [ms] | 4 Nodos [ms] |
|---|---|---|---|---|
| Query 1 | 30932 | 16980 | 80062 | 58029 |
| Query 2 | 932 | 533 | 2021 | 30926 |
| Query 3 | 29983 | 19983 | 124002 | 89929 |
| Query 4 | 345 | 577 | 2001 | 1979 |

## 2.2. Con combiner

Tabla 3: Tiempo del MapReduce con combiner

| Query | 1 Nodo [ms] | 2 Nodos [ms] | 3 Nodos [ms] | 4 Nodos [ms] |
|---|---|---|---|---|
| Query 1 | 33011 | 16928 | 21033 | 49026 |
| Query 2 | 45 | 60 | 1947 | 3932 |
| Query 3 | 6961 | 5042 | 36030 | 155983 |
| Query 4 | 1968 | 3907 | 46971 | 52984 |

# 3. Potenciales puntos de mejora y/o expansión

Al finalizar el trabajo, y analizar los tiempos de ejecución de las tareas, pensamos que un posible punto de mejora podría ser modificar la carga según la query a realizar. En este momento se cargan los datos siempre de la misma manera, subiendo toda la información de los viajes. Se podría hacer que aquellos que no debieran estar, como los de estaciones que no se encuentran en el stations.csv, o viajes de no miembros, se descartaran al momento de cargar los datos, en vez de dentro del mapper. De esta forma, el mapper tendría menos lógica, y se podrían procesar más líneas del archivo.