Ministry of Higher Education and Scientific Research
University of Eloued
Department of Computer Science
Master 2 I2A – Big Data
2025/2026

TP08: Simulating MapReduce Jobs on Google Colab

Objective:

In this session, we will simulate MapReduce programs directly on Google Colab, without installing Hadoop or Spark. The goal is to help you understand the map → group → reduce pattern that underlies distributed data processing systems.
You will experiment with three examples:
1. Word Count (introductory example)
2. Sales per Region Analysis (applied example)
3. Web Log Analysis (exercise to complete)

We will use Google Colab for all exercises. Each example has its own Colab notebook. You can open them directly using the provided links.

Example 1 – Word Count

Description:
This is the classic MapReduce example. You will count how many times each word appears in a text file.
Notebook link:
https://colab.research.google.com/drive/1tXy3WPq7M9Hw_6wB4CDVkbdgEP1KSC3K?usp=sharing

Concepts illustrated:
- Mapping: splitting sentences into words and emitting (word, 1) pairs
- Reducing: summing up counts for each word
- Viewing the output as key-value pairs

Example 2 – Sales per Region Analysis

Description:
You are given a CSV file of product sales. Each line contains:
Product, Quantity, UnitPrice, Region
Your goal is to calculate the total revenue per region using MapReduce logic.
Notebook link: https://colab.research.google.com/drive/1XoGr_NYFQ6HsvzUlRr9N-0nhJqWMvZY6?usp=sharing
Concepts illustrated:

- Parsing structured CSV-like data
- Calculating totals using key-value grouping
- Simulating distributed aggregation in Python

Exercise : Log Analysis

Problem Statement:
You are given a log file of HTTP requests. Each line contains:
Date, Time, IP, Method, URL, Status, ResponseSize

Your task is to use the MapReduce logic to compute:

- The total number of requests for each HTTP status code (e.g., 200, 404, 500).

We have prepared a Colab template with partially written code: your job is to fill in the missing parts.
Notebook link: https://colab.research.google.com/drive/14dJiyQQ1OGGSUD9enzZin4ml_Fnnflnh?usp=sharing

Dataset Preview:

Example entries from the weblogs.txt file:
2025-10-10,12:01:32,192.168.1.2,GET,/index.html,200,1024
2025-10-10,12:01:33,192.168.1.3,GET,/products.html,200,850
2025-10-10,12:01:35,192.168.1.4,GET,/contact.html,404,512
...

Your Task:

Complete the missing code in the *mapper()* function to **extract the status code** and **output a pair (status, 1)** for each line. Then run the full notebook to display the number of requests per status code.

Expected Output

HTTP 200: 8 requests
HTTP 403: 2 requests
HTTP 404: 4 requests
HTTP 500: 3 requests

Bonus Exploration

After completing the base task, try the following:
1. Count requests per URL instead of per status code.
2. Compute total response size per status code.
3. Filter out successful responses (status 200) and analyze only errors.