NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTAMENTO DE
INFORMÁTICA

DEPARTAMENT OF
COMPUTER SCIENCE

Big Data Processing Systems — 2021/22

Project Assignment

November 17, 2021

Abstract

This project assignment is to be executed in groups of three students. All students must contribute to the final outcome somehow (although they may have different tasks). Please notice that in the last page of the report you are expected to list the contributions of each group member, together with it's relevance to the final result in percentage (the percentages given to the group members must add up to 100%).

1 The Data Set

Hazardous air pollutants, also known as toxic air pollutants or air toxics, are those pollutants that are known or suspected to cause cancer or other serious health effects, such as reproductive effects or birth defects, or adverse environmental effects. The USA Environmental Protection Agency (EPA) tracks 187 air pollutants. See https://www.epa.gov/haps/ for more information.

1.1 The Main Data Set

The Data Set is a 2.6 GB CSV file (`epa_hap_daily_summary.csv`) that contains data for every monitor (sampled parameter) in the Environmental Protection Agency (EPA) database for each day (available from Kaggle[1]). Each entry contains a daily summary record that is:

1. The aggregation of all sub-daily measurements taken at the monitor; or
2. A single sample value, if the monitor takes a single, daily sample (e.g., there is only one sample with a 24-hour duration). In this case, the mean and max daily sample will have the same value.

For development purposes, you may use the smaller version of the Data File with only 117 MB (`epa_hap_daily_summary-small.csv`)! See below for the links.

1.1.1 The Main Data Set Structure

The main Data Set contains 29 columns with the following contents:

1. **State Code**: The Federal Information Processing Standards (FIPS) code of the state in which the monitor resides.

[1] https://www.kaggle.com/epa/hazardous-air-pollutants

Campus de Caparica
2829-516 CAPARICA | Portugal
T: +351 212 948 536

2. **County Code**: The FIPS code of the county in which the monitor resides.
3. **Site Num**: A unique number within the county identifying the site.
4. **Parameter Code**: The AQS code corresponding to the parameter measured by the monitor.
5. **POC**: This is the "Parameter Occurrence Code" used to distinguish different instruments that measure the same parameter at the same site.
6. **Latitude**: The monitoring site's angular distance north of the equator measured in decimal degrees.
7. **Longitude**: The monitoring site's angular distance east of the prime meridian measured in decimal degrees.
8. **Datum**: The Datum associated with the Latitude and Longitude measures.
9. **Parameter Name**: The name or description assigned in AQS to the parameter measured by the monitor. Parameters may be pollutants or non-pollutants.
10. **Sample Duration**: The length of time that air passes through the monitoring device before it is analyzed (measured). So, it represents an averaging period in the atmosphere (for example, a 24-hour sample duration draws ambient air over a collection filter for 24 straight hours). For continuous monitors, it can represent an averaging time of many samples (for example, a 1-hour value may be the average of four one-minute samples collected during each quarter of the hour).
11. **Pollutant Standard**: A description of the ambient air quality standard rules used to aggregate statistics.
12. **Date Local**: The calendar date for the summary. All daily summaries are for the local standard day (midnight to midnight) at the monitor.
13. **Units of Measure**: The unit of measure for the parameter.
14. **Event Type**: Indicates whether data measured during exceptional events are included in the summary. A wildfire is an example of an exceptional event; it is something that affects air quality, but the local agency has no control over. *None* means no events occurred. *Included* means events occurred and the data from them is included in the summary. *Excluded* means that events occurred but data form them is excluded from the summary. *Concurred Events Excluded* means that events occurred but only EPA concurred exclusions are removed from the summary. If an event occurred for the parameter in question, the data will have multiple records for each monitor.
15. **Observation Count**: The number of observations (samples) taken during the day.
16. **Observation Percent**: The percent representing the number of observations taken with respect to the number scheduled to be taken during the day. This is only calculated for monitors where measurements are required (e.g., only certain parameters).
17. **Arithmetic Mean**: The average (arithmetic mean) value for the day.

Campus de Caparica
2829-516 CAPARICA | Portugal
T: +351 212 948 536

18. **1st Max Value**: The highest value for the day.
19. **1st Max Hour**: The hour (on a 24-hour clock) when the highest value for the day (the previous field) was taken.
20. **AQI**: The Air Quality Index for the day for the pollutant, if applicable.
21. **Method Code**: An internal system code indicating the method (processes, equipment, and protocols) used in gathering and measuring the sample. The method name is in the next column.
22. **Method Name**: A short description of the processes, equipment, and protocols used in gathering and measuring the sample.
23. **Local Site Name**: The name of the site (if any) given by the State, local, or tribal air pollution control agency that operates it.
24. **Address**: The approximate street address of the monitoring site.
25. **State Name**: The name of the state where the monitoring site is located.
26. **County Name**: The name of the county where the monitoring site is located.
27. **City Name**: The name of the city where the monitoring site is located. This represents the legal incorporated boundaries of cities and not urban areas.
28. **CBSA Name**: The name of the core bases statistical area (metropolitan area) where the monitoring site is located.
29. **Date of Last Change**: The date the last time any numeric values in this record were updated in the AQS data system.

1.1.2 Downloading the Data Set

The three Data Sets (the main data set, the reduced data set, the secondary data set) are available as a 579 MB ZIP file from

`https://tinyurl.com/drnfupyb`

1.2 The Secondary Data Set

There is an additional auxiliary CSV file (`usa_states.csv`) that contains an approximation of the limits (latitude and longitude) of all the USA states, e.g.,

State	Name	MinLat	MaxLat	MinLon	MaxLon
AK	Alaska	52.5964	71.5232	-169.9146	-129.9930
AL	Alabama	30.1463	35.0041	-88.4743	-84.8927
⋮	⋮	⋮	⋮	⋮	⋮

This data set will be necessary for answering some of the questions.

Campus de Caparica
2829-516 CAPARICA | Portugal
T: +351 212 948 536

2 Project Assignment

You are asked to use the technologies we studied in this course, namely,

1. Map-Reduce;
2. Spark (*plain Spark*);
3. SparkDF (*Spark with Dataframes*);
4. SparkSql (*Spark with SQL*); and
5. Hive;

to prepare a set of indexes that will help answering the following questions:

Q.1) Which states have more/less monitors? (*Rank states*!)

Q.2) Which counties have the best/worst air quality? (*Rank counties considering pollutants' level!*)

Q.3) Which states have the best/worst air quality in each year? (*Rank states per year considering pollutants' level!*)

Q.4) For each state, what is the average distance (in km) of the monitors in that state to the state center? For simplicity, assume that 1 degree of latitude or logitude equals to 111 km. (Monitor dispersion per state!)

Q.5) How many sensors there are per quadrant (NW, NE, SE, SW) in each state? To answer this question, you should approximate each state's area to a rectangle as defined in the file "`usa_satates.csv`", and divide that area in 4 quadrants (NW, NE, SE, SW). (*Count monitors per sate qudrant!*)

2.1 Experimental Work

Please create a Jupyter Notebook for each technology, where the questions are answered in the same order as listed above. Each Jupyter file should be named as:

`<Gnn>-<TECHNOLOGY>.ipynb`

where `<Gnn>` is the group numer, e.g., `G05`, and `<TECHNOLOGY>` is one of: `mapreduce`, `spark`, `sparkDF`, `sparkSQL`, `hive`.

Start by developing your solution in your personal computer with the small data set. Once you believe your solutions are ok, then you may try them in the Department's Cluster. Additional instructions on how to access and use the Department's Cluster will be given later.

Campus de Caparica
2829-516 CAPARICA | Portugal
T: +351 212 948 536

2.2 Project Report

Prepare a project report in PDF format with name "Gnn-report.pdf" (where "Gnn" is the group number), with the look and feel of a research paper, with a maximum of 4 pages. If necessary a 5th page may be added **only for** "*acknowledgments*", "*description of the individual contributions*", and "*references*". It is mandatory to follow the IEEE template for Computer Society Journals using either Word or LaTeX. Please remember to show, for each exercise, a small sample of the produced index and the corresponding execution time.

To get the IEEE template site at `https://goo.gl/Xtjdh4` and

Select Publication Type → Transactions, Journals and Letters;

Select Publication → IEEE Transactions on Parallel and Distributed Systems;

Select Article Type → Original Research;

Select Format → LaTeX or Word;

Download Template → Download Template;

Remember the Bloom's Pyramid... the report should have a stronger emphasis on the upper layers of the pyramid (*evaluate*, *analyze*, *create*) rather than on the lower layers (*remember*, *understand*, *apply*). This means that, although *How did you do your work?* is important to me, *What you learned with your work (and how did you learned it)?* is much more important!

Please remember that in the last page of the report you are expected to have three sections:

1. List all the documentation relevant to your work in the "Bibliography/References". The entries listed in this section must be cited somewhere in the main text.

2. List the contributions of each group member, together with it's relevance to the final result in percentage (the percentages given to the group members must add up to 100%).

3. Acknowledgments to other colleagues (*non-group members*) that somehow were relevant to your work. Please be sure you identify the colleagues clearly (if possible include their student number) and how they helped you.

2.3 Submission Instructions

1. Create a folder named `Gnn_AAAAA_BBBBB_CCCCC`, where
   Gnn → group number, e.g., G04
   AAAAA → student 1 number, e.g., 45454
   BBBBB → student 2 number, e.g., 54321
   CCCCC → student 3 number, e.g., 56321
   (*the numbers AAAAA, BBBBB, and CCCCC, must be in increasing order*).

2. Copy your Jupiter Notebooks to the folder;

3. Copy your project report to the folder;

4. Zip the folder (and its contents) into a file named "`Gnn_AAAAA_BBBBB_CCCCC.zip`".

Campus de Caparica
2829-516 CAPARICA | Portugal
T: +351 212 948 536

5. Submit your ZIP file in the form at the address:

   https://forms.gle/2zSjCJb4oh4hGsLV9

   no later than Monday, December 20, 2021 @ 23:59.

3 Document Versioning

1.0	2021-11-17	Initial version.

Campus de Caparica
2829-516 CAPARICA | Portugal
T: +351 212 948 536