CentraleSupélec

2024-2025

Bases de données pour les données massives

Spark et Parquet

The purpose of this set of exercises is to work with Spark and explore the Spark UI with attention for the DAG visualization.
Remember that every SparkContext launches a Web UI:

- by default on port 4040
- that displays useful information about the application

1. A list of scheduler stages and tasks
2. A summary of RDD sizes and memory usage
3. Environmental information.
4. Information about the running executors

1 Environment

1. Start Spark
2. open a jupiter notebook

1.1 Spark UI

In a second web browser open `http://<driver-node>:4040` (probably `http://localhost:4040/jobs/` for your local configuration).

2 Explore the interface

Look at the interface and try to find, with the help of the slides of the last lesson, the main components:

- Jobs Tab
- Stages Tab
- Storage Tab
- Environment Tab
- Executors Tab
- SQL Tab
- Structured Streaming Tab
- Streaming (DStreams) Tab
- JDBC/ODBC Server Tab

Exercise 1

Is there any Tab of the list that is missing in your configuration? List the tabs that you cannot find-visualize

Exercise 2

Look at the executor tab. How many cores do you see? How can you take advantage of this information in order to better configure and run your Spark application?

3 Practical exercise 1

In this exercise you will be asked to check the execution of one simple application that counts how many words composed by an even number of characters, and how many words composed by a odd number of characters are present in the provided `books.json` text file.

Exercise 3

Write a program that performs this count according to a "silly procedure", for example not using a map-reduce algorithm.
Analyze the DAG of the operators, describe what happens and provide a description of how data is accessed.

Exercise 4

Write a program that performs this count according to a a map-reduce algorithm. Analyze the DAG of the operators, describe what happens and provide a description of how data is accessed.

Exercise 5

Compare the two executions focusing on data access.

Remember the parameters that you can specify in the `SparkConf`[1]

Exercise 6

Look at all the configurations and how the executors work. Is your Spark configuration the best for your machine? Comment and if necessary change it.

4 Practical exercise 2

In this exercise you will be asked to check the execution of a little more complicated application.

Exercise 7

Write a program that counts how many times each word appears in the file `books.json`. Look at the DAG and at the UI interface. What can you say?

Exercise 8

Write a program that finds the most used word in the file:

1. using the sort function of spark
2. using a sort function that you implement

Can you remark any difference in the analysis of data access and performances of the two programs?

5 Practical exercise lost

For this exercise we had not studied the theory as deeply as necessary but you can still have an intuition about something. Think about a set of documents that must be stored in a distributed environment. Think about the fact that the structure of this documents is almost homogeneous and that you have two possibilities:

[1] `https://spark.apache.org/docs/latest/configuration.html`

- store them as they are in a distributed file system that can be accessed by Spark
- store them in a distributed database that can be queried using Spark

Exercise 9

How can a database (think about parquet) help the performances for the access of certain fields of a set of documents sharing the same pattern?

6 BONUS - HOW PARQUET CAN HELP YOU TO MANAGE BIG COMPUTATIONS AND BIG FILES

For running this series of exercises, we are going to use a dataset coming from [2]
As stated in the description of the dataset: "The TripAdvisor dataset includes 1,083,397 restaurants with attributes such as location data, average rating, number of reviews, open hours, cuisine types, awards, etc.
The dataset combines the restaurants from the main European countries".

6.1 THE DATASET

The dataset is in a .csv file, and among the columns, you can find:

- `Restaurant_link`the link of the restaurant in TripAdvisor
- `Restaurant_name` The name of the restaurant
- `Original_location` The location of the restaurnt
- `Country` The country
- `...`

6.2 SPARK AND PANDAS.

For this set of exercises you must import data in Spark. After this first import you can pass any dataset to Pandas for the data analysis. At the end of each exercise (when the question is pertinent) you must return (reconvert) the dataframe in Spark.
Each time you do this conversion you must comment about this. Example:

```
creating a Spark dataframe

 df = ...
```

[2] `hhttps://www.kaggle.com/datasets/stefanoleone992/tripadvisor-european-restaurants`

```
using Pandas and creating a Pandas dataframe

 dfp = df ...

back to Spark

 dfs = ...

df = pa.read_csv("tripadvisor_european_restaurants.csv")
```

Exercise 10

Try now to use parquet for reading data and find the country with the most number of restaurant

Exercise 11

And the most common cuisine.