EFREI

APPLICATION OF BIG DATA

Documentation

Louis CAILLAREC – Paul JOUVANCEAU – Noa ANDRE – Mazen CHOUCHARA

01/12/2024

TABLE OF CONTENTS

1. INTRODUCTION

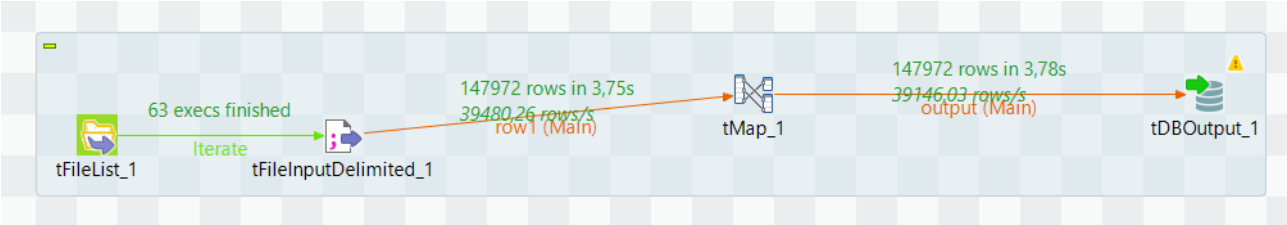
1.1 DATA VISUALIZATION PROJECT

We decided to work on the data visualization project because this is the subject we are the most comfortable with. This subject is about loading data inside a relational database and then making a dashboard to picture relevant information. We have a folder with 62 csv files named after users on Lastfm website. These files contain lines of artists, albums, songs and timestamps. These are songs listened to by users at a time. Thanks to this data, we have the opportunity to see which artists are the most heard for example and the goal is to picture these elements, these facts into a dashboard.

2. CHOICE OF SOFTWARES

2.1 ETL SOFTWARE

First part was ETL, we discovered for some of us what is an ETL which stands for Extract, Transform and Load. We decided to go with Talaxie Open Studio which is a free ETL and the new version of Talend that we heard about before. We did some tests from extracting file by file to realize that we can also extract from an entire folder which is better. We learnt some tips as it's usually better to work from small files even if they are big than having a lot of small ones. We saw it in the process with a faster loading into the database.



2.2 DATABASE AND DASHBOARD

For the database, we decided to be simple and use MySQL Workbench that we used plenty of time for past projects. We understood that it wasn't mandatory to have a relational database, but it is a convention when we are working with more complex data. We didn't have much to do with it except create the table, but we still tried some changes needed to make sure nullable values were possible.

For the Dashboard we chose PowerBI as some of us had already worked with it, so it was obvious, and we didn't have trouble using it except for the timestamp that made us reload the entire process with a new data type for the timestamp that we put as a varchar first and then switch to date type to use it correctly on the dashboard.

3. DIFFICULTIES AND CHALLENGES

3.1 OBSTACLES OF THE PROJECT

We had some troubles at the beginning of the project. Indeed, we chose to work on this project because we all like manipulation of data. Also some of us, already use it during their internship so it was a benefice to reuse it on the project. Unfortunately, none of us used ETL before. We've heard about ETL but never test them. We had the first session entirely to discover the data provided and determine what will be the software used for the project. We started to test the first one and we were a bit lost on how to use it. It was obvious that often when a solution is free to use as Talaxie, the user interface won't be friendly, so we had to identify which elements were useful here. We also had no idea of what we had to do with other than: "We need to load the data inside the database."

After learning the basics of Talaxie, we succeeded to load the data from a folder into the database but we still needed to obtain the username which is the name of each csv file. It took us a lot of time during the second session but we also succeeded it trying and making research online about that.

We had finally some minor details to resolve by making the dashboard with creating the metrics for a week. But as previously, after some tests, we succeeded.

3.2 USAGE OF AI

During the project, we used AI to help ourselves at the beginning as said previously. Indeed, none of us ever used ETL software. So, we asked for help to know which elements could be useful to us to create the job.

4. CONCLUSION

4.1 LEARNINGS

With this project, we had the opportunity to discover new tools. For some of us it was PowerBI, but all of us discovered ETL software and how to use one of them as Talaxie Open Studio. This was also the first project in school where we came from having the data to present a dashboard. It was interesting to see how it was possible to do it with basic data. We had the opportunity to talk about how things could be with more data, how we could optimize it.