**Guidelines to fulfill the report template:**

1. Use **precise** and **technical** vocabulary as far as needed.
2. Be as **simple** and **direct** as possible.   
   *Number of pages does not matter, contents are more important than quantity! Nevertheless, a report of about 10 pages (not including the table of contents) is expected.*
3. The report is for you and for your customer (partner or ESIVL) as well.  
      
   **For you:** to improve your skill in project management *(if this is the first closure report you write, this won’t be the last!).*  
   **For the customer**: to pass the baton to another one, in order to have the project continued or developed if necessary, or to have a better understanding of the global project process you had.
4. This report is to be completed in English.
5. In case you have any questions related to how to fulfill some sections, please ask them to EGPE team.
6. The Project Closure Report is completed by extracting the data already produced throughout the project: the project file, the initial planning documents, the progress reports.  
   The added value consists in the aggregation of the data within a single document, the analysis of the overall differences.

**Project Title: Dataprep Starter**

***Brief description:***

***Preparing data can be a long and complicated task, specifically if done manually. Moreover, the manipulation of bigger datasets can drastically increase the difficulty of the process and decrease the quality of the results. That’s why our team, at the demand of the company Sopra Steria, and supervised by Mr. Mickael PATRON, is commissioned to automatize the data preparation of large datasets, using a programming language that can use Spark, a powerful tool for this task.***

**Team Number: 147**

**Project Partner: Sopra Steria**

**Project Team: Data Blender Co-pilot (DBC)**

|  |  |
| --- | --- |
| **Team Member** | **Project Role (if defined)** |
| Yassine CHENIK | Data Analyst & Referent |
| Emine BOUCHIBA | Software Developer |
| Nicolas BERLIOZ | Quality Assurance Analyst |
| Loick CUER | Technical Writer |

*The Project Closure Report contains key descriptive information about the project.*

*As the* ***last document written*** *on the project it* ***analyzes the outcome*** *of the project and the* ***process*** *by which that outcome was produced.*

|  |
| --- |
| *Its purpose is twofold: to ensure that the closure activities are carried out properly and to facilitate the transfer of experience (or the transfer of experience if the project is not completed or if a follow-up to the project is envisaged) to the customer’s organization.*  Table des matières  [1 Objectives Set versus Results Achieved 2](#_Toc93367879)  [1.1 Project Initial Objectives 2](#_Toc93367880)  [1.2 Results Achieved 2](#_Toc93367881)  [1.3 List of Deliverables 3](#_Toc93367882)  [2 Methodological review 3](#_Toc93367883)  [3 Risk Management 3](#_Toc93367884)  [4 Constraints 3](#_Toc93367885)  [5 Technical Review 3](#_Toc93367886)  [6 Engineering standards 3](#_Toc93367887)  [7 “Post-project” tasks 4](#_Toc93367888)  [8 Project Completion Recommendations 4](#_Toc93367889)  [9 Annexes to the closure report. 5](#_Toc93367890) |

# Objectives Set versus Results Achieved

## Project Initial Objectives

The initial objective of the project was to use the Java programming language and a tool called Spark to automatize the process of Data Preparation.  
Data preparation is the process of cleaning, transforming, and organizing data to make it ready for analysis. The goal of data preparation is to ensure that the data is accurate, complete, and consistent so that it can be effectively analyzed. This task is time-consuming and, if done manually, prone to error and inconsistency. It is even more complicated when we work with large datasets.

We defined four different steps for this project:

* Data Ingestion: The load of multiple sources with different vectors and frequencies
* Data Fusion: The integration of these multiple data sources
* Data Cleaning: An important step that consists of detecting and correcting the corrupt or inaccurate records
* Data Delivery: The easiest but most important step, that comes down to make the data available

However, before manipulating our datasets, we had to prepare our workspaces and our tools.

Preparing our workspace and setting up Spark took some time, and our team members are more used to Python than Java. With our referent, Mr. Patron, we decided to use a Python Notebook rather than a Java environment, to gain some time and efficiency.

## Results Achieved

* Objective: The objective of the project was to create a Python algorithm that could clean and fuse two datasets that share the same primary key.
* Results: The project was successful in achieving its objectives. The team was able to create a Python algorithm that cleaned and fused the datasets. Additionally, Spark was utilized to manage the large amount of data, resulting in a more efficient process.
* Positive Deviations: There may have been some positive deviations from the initial expectations, such as increased speed or accuracy of the algorithm or identifying additional insights or trends in the data during the fusion process.
* Negative Deviations: There may have been unforeseen challenges in the data that made the fusion process more difficult than anticipated. Additionally, limitations in the resources available (such as computing power or time constraints) may have made it difficult to achieve certain objectives.
* Overall, the project was successful in achieving its objectives and utilizing Spark to manage large amounts of data helped to make the process more efficient. It's important to consider any positive or negative deviations from initial expectations and to maintain open communication with partners throughout the project.

## List of Deliverables

Here is a list of the deliverables of the project:

* Closure Report:
* Video: A two-minute video presenting and promoting the project
* Visual: A short document representing our project
* Summary: A half a page document that summarize the main aspects of the project with a French and an English version
* One Pager:

# Methodological review

From the outset of our project, we recognized the importance of establishing a clear and effective communication strategy to ensure that we could work together with efficiency and purpose. To this end, we decided to schedule regular Teams calls every two weeks, in which we could review our progress and plan our next steps.

We adopted a two-step approach to our work. Firstly, we dedicated time to preparing our workspace, ensuring that we had all the necessary tools and resources at our disposal. This included setting up our development environments and testing our code to ensure that we could work as smoothly as possible.

Once our workspace was prepared, we moved on to the second step of our approach, which involved developing the project itself. Here, we drew inspiration from the Agile methodology, which allowed us to work on the project iteratively and incrementally, reviewing and revising our work with each meeting.

This approach proved to be highly effective, as it allowed us to stay focused and on track throughout the project. By breaking our work down into manageable steps and regularly reviewing our progress, we were able to ensure that we were working towards our goals with purpose and clarity. Overall, our experience taught us the importance of clear communication and effective project management when working on complex projects.

# Risk Management

Data quality issues: The data being ingested or merged contained errors and missing values which could lead to incorrect results or poor data quality. (Solution: We created a data cleaning algorithm)

Technical issues: Technical issues such as system performance (we had only 8Go of RAM) bottlenecks caused delays or disruptions in data processing or delivery. (Solution: we used Mr. Patron’s computer with 32Go of RAM)

Resource constraints: Inadequate resources, such as insufficient budget (Spark is not a free solution when we want to use it for huge computing units), could impede progress and limit the project's success. (Solution: Mr. Patron helped us to install a former version of Spark in our computers in order to execute the code with a huge amount of data)

# Constraints

The professionalism of our referent from Sopra Steria allowed us to point a few constraints from the very beginning of the project. The constraints were the following:

* The Spark tool’s free version is limited: Spark is a powerful computing system designed for large volumes of data. In order to use it to its fullest capacity, we needed to pay. Thus, we had to find a free alternative, that our referent, Mr. Patron, provided us with.
* Huge needs in RAM: Since Spark is used to manipulate large datasets, it requires a specific hardware. The minimum amount of RAM required is 16GB, which not all of our team possesses. However, Mr. Patron could provide us with a computer that fulfills this condition.

# Technical Review

As we embarked on our project using Spark, we quickly realized that the installation process was not a straightforward one. We had to install a multitude of tools and packages via our Terminal and Docker, and the process was not without its complications. Each team member encountered different difficulties at different stages of the installation, which further complicated matters, as we all had to work on our own computers.

Despite our best efforts, these complications had a snowball effect, gradually leading to a delay in our expected deadlines. We knew we had to find a way to make up for the lost time if we wanted to meet our goals. After careful consideration, we decided to switch from a Java environment to a Python environment, as we were all more familiar with it.

The decision to switch to Python proved to be a wise one. We found that we were able to work more efficiently in this environment and were able to catch up on the work we had missed. By leveraging our knowledge of Python, we were able to work faster and more effectively, allowing us to meet the expected deadlines for the rest of the project. The experience taught us the importance of being adaptable and flexible when working on complex projects, and we were proud of the progress we were able to make despite the initial setbacks.

# Engineering standards

Engineering Standards that were used I during our project development:

* ISO/IEC 25010:2011
* ISO/IEC 12207:2008
* IEEE 830-1998
* IEEE 1016-1998
* ANSI/IEEE 1471-2000
* Data Management Body of Knowledge (DMBOK)

# “Post-project” tasks

Post-project tasks:

1. Optimizing the code to make it more efficient (i.e: Increasing the performance of the fusion and reducing the execution time).
2. Setting up a Human-Machine environment to make it much easier to use. (Because we decided to let the code in a notebook format to make it much easier to understand)

# Project Completion Recommendations

* Best Practices:
  + Applying our own cleaning methods:
    - Between 0% and 10% of missing values, we decided to just drop the row.
    - Between 10% and 20% of missing values, we decided to pick a random value from the same column in our dataset.
    - Over 20% of missing values, we decided to pick the value we the most occurrence.
  + Using PySpark in order to test our code in Google Colab, and so, being able to execute it using Spark with Docker.
* Improvements:
  + Optimizing the code to make it more efficient (i.e: Increasing the performance of the fusion and reducing the execution time).
  + Setting up a Human-Machine environment to make it much easier to use. (Because we decided to let the code in a notebook format to make it much easier to understand)
* Peripheral action to be taken:
  + Increasing the group performance by splitting it into two sub-groups, one focusing in installing and Launching Spark and the other one focusing in implementing the Python code, in order to gain time.

# Annexes to the closure report.

[PI² - Annexes](https://drive.google.com/drive/folders/10q2dndyveVgqgsnsBSqO4IXohf-VXYuh?usp=share_link)