A) High-level data staging plan

- Data Acquisition : The dataset is about the data scientist jobs salary. So we get the dataset from Kaggle with the following link : https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor?select=glassdoor_jobs.csv

- Data Cleaning : The cleaning of the dataset consisted first of removing the irrelevant column for the analysis. These columns were `"Unnamed: 0","Headquarters","Sector","age",”Company size”,"Salary Estimate","JobDescription","hourly","employer_provided","company_txt","job_state","Competitors","same_state","python_yn","R_yn","spark","aws","excel","seniority","desc_len","num_comp"`.
  After that, we removed the “Founded” and the “Revenue” columns because they had a lot of null values. In this stage we decide to keep the rating column although they have some null values columns because they are relevant for the analysis.
  We removed also the rows with inconsistent values in “Industry”,”Rating” and “Size” columns

- Data transformation : In the original form of the dataset, the “job_simp” column had a lot of null value. We noticed that we can solve this issue by analyzing the first column and using this information to fill the missing rows. We filled the missing rows with the label “Research”.
  We also transform the “location” column data. Indeed, they had the form “City,State”, we then split them into two different columns. The first column was the “City” column and the second one was the “State”. We transformed the “Size” columns data into range values.

- Data storage : We stored the staging data in MySql data database.

DATA QUALITY ISSUE

During the data staging we faced certain issues. The first one was the amount of rows, we only got 742 rows which made our data cleaning more delicate because we didn’t want to remove anymore rows. But the advantage of this dataset was the large number of columns. So we took advantage of that.

In “job_simp” columns we got more than 100 missing values. So to avoid removing them, we used the first column to make some correspondence and complete the missing rows.

Also in the “Industry”,”Size” and “Rating” columns we gave us some inconsistent values. To solve this problem, we checked the total number of inconsistency rows and the number was very low, less than 10. So we decided to remove them.