TUNIS BUSINESS SCHOOL
UNIVERSITY OF TUNIS

Report

IT-300
Business Intelligence and Database Management Systems

Business Intelligence Research

Global Salaries in AI, ML, Data Science

Authors:

Montaha DRIDI Firas ELLOUMI

Fatma BELLAJ Eya SALEM

Submitted to:
Prof. Ameni AZZOUZ

1 . Introduction

This project focuses on the analysis of global salary data in the fields of Artificial Intelligence (AI), Machine Learning (ML), and Data Science. The goal is to provide open salary data for individuals at all career stages, including beginners, seasoned professionals, hiring managers, recruiters, and those considering career transitions.

Additionally, the project aims to uncover and understand various factors influencing these jobs, considering their prominence as trend-setting careers in today's workforce.

By offering comprehensive insights into both salary trends and influential factors, the initiative seeks to empower stakeholders to make informed decisions, foster better hiring practices, and support individuals navigating these dynamic and sought-after professions.

2 . Implementation

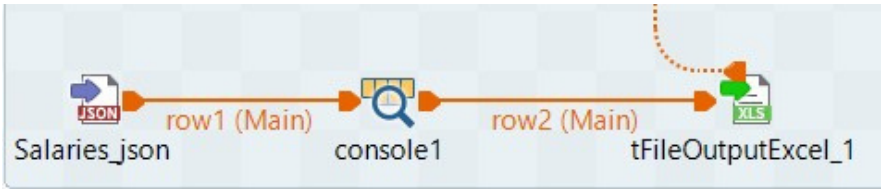
2.1 Data Gathering

We extracted Global Salaries in AI, ML, Data Science from KAGGLE , This is a link to the dataset:
https://www.kaggle.com/datasets/aijobs/global-salaries-in-ai-ml-data-science?fbclid=IwAR26__4Lz_IDkynCweNfyVS5EtAu7n20mnQrAW61yPh-c7xF44d9d2I3sdA&select=salaries.csv

2.2 Data Preparation

During this phase, we employed the "Talend" software for the Extract, Transform, and Load (ETL) processes. This involved extracting data, loading it into the system, and performing necessary transformations to ensure its suitability for analysis.

At first, we converted the (JSON) data into Excel format ,We changed the structure of the data so that it could be easily manipulated and managed.



Attached here is a part the code for the conversion process

```
/**
 * [tLogRow_1 main ] start
 */

currentComponent = "tLogRow_1";

////////////////////////

String[] row_tLogRow_1 = new String[11];

if (row1.work_year != null) { //
    row_tLogRow_1[0] = String.valueOf(row1.work_year);

} //

if (row1.experience_level != null) { //
    row_tLogRow_1[1] = String.valueOf(row1.experience_level);

} //

if (row1.employment_type != null) { //
    row_tLogRow_1[2] = String.valueOf(row1.employment_type);
```

```
if (row2.work_year != null) {
//modif start

    columnIndex_tFileOutputExcel_1 = 0;

    jxl.write.WritableCell cell_0_tFileOutputExcel_1 = new jxl.write.Number(
            columnIndex_tFileOutputExcel_1,
            startRowNum_tFileOutputExcel_1 + nb_line_tFileOutputExcel_1,

//modif end
            row2.work_year);
//modif start
    // If we keep the cell format from the existing cell in sheet

//modif ends

    writableSheet_tFileOutputExcel_1.addCell(cell_0_tFileOutputExcel_1);
    int currentWith_0_tFileOutputExcel_1 = String
            .valueOf(((jxl.write.Number) cell_0_tFileOutputExcel_1).getValue()).trim().length();
    currentWith_0_tFileOutputExcel_1 = currentWith_0_tFileOutputExcel_1 > 10 ? 10
            : currentWith_0_tFileOutputExcel_1;
    fitWidth_tFileOutputExcel_1[0] = fitWidth_tFileOutputExcel_1[0] > currentWith_0_tFileOutputExcel_1
            ? fitWidth_tFileOutputExcel_1[0]
            : currentWith_0_tFileOutputExcel_1 + 2;
}

if (row2.experience_level != null) {
//modif start

    columnIndex_tFileOutputExcel_1 = 1;
```

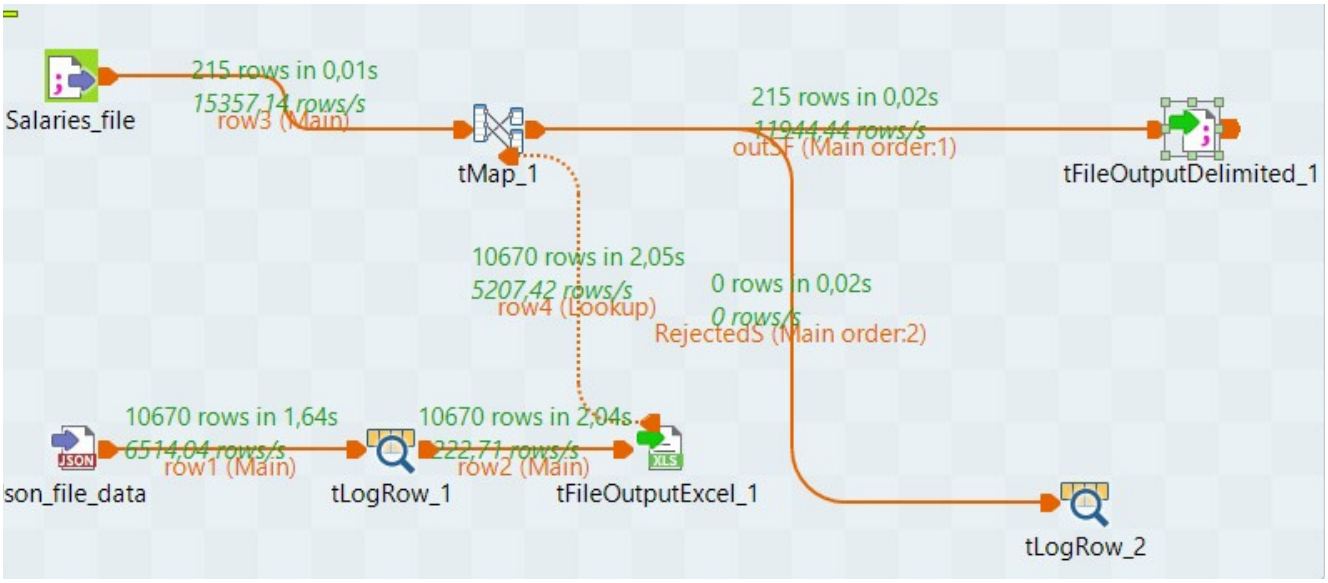
we limited the number of rows to 215 , keeping the data related to the year 2024 in order to make so that the data could be easily manipulated and managed.

The output is:

- **a normalized table encompassing all 2024-related data, excluding the "salary" and "salary currency" columns. This omission was made because the dataset features salaries in diverse currencies, and our objective is to standardize them uniformly to the common currency, USD.**
- **We've introduced a "category" column to facilitate the comparison of individual salaries with the average salary. This column allows for a quick assessment of whether a specific salary is above or below the established average of 122,738 USD in 2024.**

- **We employ the "tLogRow" tool to display rejected data on the console, providing visibility into information that did not meet the defined criteria**

An in-depth video will be provided during the project defense, in which we will explain the various steps taken.



2.3 Data Storage

2.3.1 Storage

For data storage, we employed the use of MySQL workbench

- **work_year: Represents the year which the salary was paid**
- **experience_level: Represents the employee's experience level in the job, indicated as: EN (Entry-level), MI (Mid-level), SE (Senior-level), or EX (Executive-level).**
- **employment_type: Represents the type of employment for the role, such as: FT (Full-time)**
- **job_title: Represents the specific job title held by the employee.**
- **salary_in_usd: Represents The annual salary paid to the employee, expressed in US dollars.**
- **employee_residence: Represents the country where the employee primarily resided during the work year**
- **remote_ratio: Represents the overall percentage of work done remotely, indicated as: 0 (no remote work), or 100 (fully remote).**
- **company_location: Represents the country where the company's main office is located**
- **company_size: Represents the approximate size of the company according to the number of employees work there**
- **Category:Represents whether the salary is above or below average for the given job title**

2.3.2 Fact/Dimensions

Salary Fact Table :

The central table, containing the core salary data and linking to other dimensions.

Dimensions included in this data are:

Time Dimension:

Stores information about time, specifically the years for which salary data is available,derived from: work_year

Company Dimension:

Stores information about companies, including their size and location,derived from:
company_id
company_size
company_location

Employee Dimension:

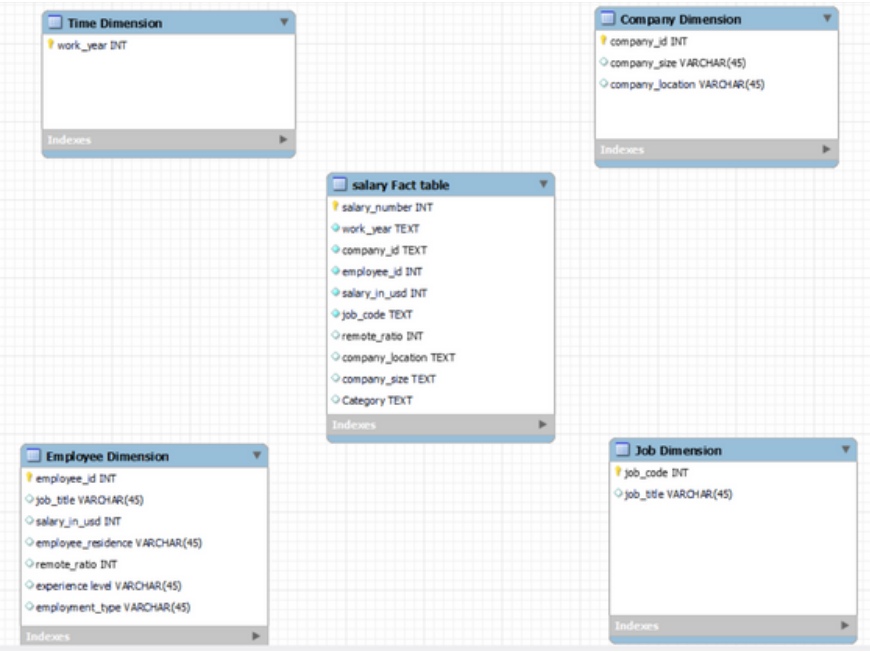
Stores information about individual employees. derived from :
salary_in_usd
employee_residence
remote_ratio
experience_level
employment_type

Job Dimension:

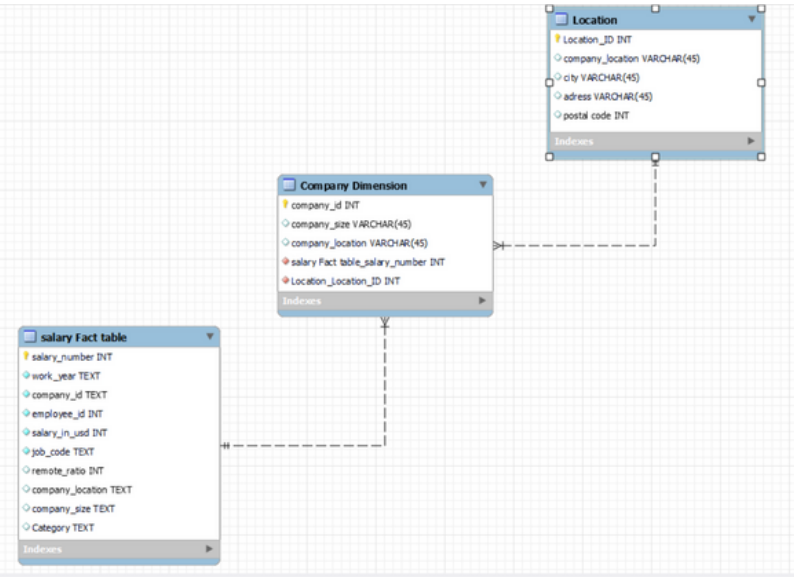
Stores information about job titles. derived from :
job_title

2.3.3 construction of schemas:

star schema



existence of relationship of type one-to-many between salary Fact table and its 4 dimension tables



company dimension table normalized into a sub category location which contains all informations of the company location

2.4 OLAP analysis:

we used OLAP tools to explore salary data from different angles, helping to uncover patterns and trends in employees ,such as :

Calculation of average salary for each work year and experience level combination

```
#average salary for each work year and experience level combination
SELECT work_year, experience_level, AVG(salary_in_usd) AS average_salary
FROM Fout
GROUP BY work_year, experience_level;
```

work_year	experience_level	average_salary
2024	MI	124396.0135
2024	EN	94907.2083
2024	SE	159901.7368
2024	EX	177825.0000

Count of employees by job title and company location:

```
#count of employees by job title and company location
SELECT job_title, company_location, COUNT(*) AS employee_count
FROM Fout
GROUP BY job_title, company_location;
```

job_title	company_location	employee_cour
Data Engineer	US	46
Data Analyst	US	34
Business Intelligence Developer	US	6
BI Developer	US	4
Business Intelligence Analyst	US	4

Calculation of maximum salary for each employment type and job title

```
#Maximum salary for each employment type and job title
SELECT employment_type, job_title, MAX(salary_in_usd) AS highest_salary
FROM Fout
GROUP BY employment_type, job_title;
```

employment_type	job_title	highest_salary
FT	Data Engineer	308000
FT	Data Analyst	220000
FT	Business Intelligence Developer	144138
FT	BI Developer	120000
FT	Business Intelligence Analyst	132000

Count of jobs with different remote ratios grouped by company location

```
#Count of jobs with different remote ratios, grouped by company location
SELECT company_location, remote_ratio, COUNT(*) AS job_count
FROM Fout
GROUP BY company_location, remote_ratio;
```

company_location	remote_ratio	job_count
US	0	141
US	100	58
CA	0	4
GB	100	6
PT	0	2

Average salary for remote vs. non-remote jobs, grouped by category:

```
erage salary for remote vs. non-remote jobs, grouped by categor
ECT category, remote_ratio, AVG(salary_in_usd) AS average_salar
M Fout
RE remote_ratio IN ('100% Remote', 'Onsite')
UP BY category, remote_ratio;
```

category	remote_ratio	average_salary
Above AVG	0	184437.7849
Below AVG	0	88491.4643
Below AVG	100	83622.6765
Above AVG	100	167963.0938

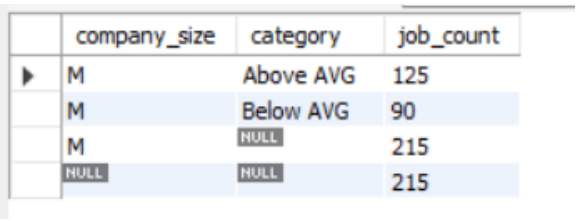
Employee count by company size and employee residence

```
#Employee count by company size and employee residence:
SELECT company_size, employee_residence, COUNT(*) AS employee_count
FROM Fout
GROUP BY company_size, employee_residence;
```

company_size	employee_residence	employee_count
M	US	199
M	CA	4
M	GB	8
M	PT	2
M	IE	2

Retrieve the total job count, grouped by company size and category, with hierarchical summarization using ROLLUP:

```
#Hierarchical Grouping with ROLLUP:
SELECT company_size, category, COUNT(*) as job_count
FROM Fout
GROUP BY company_size, category WITH ROLLUP;
```



company_size	category	job_count
M	Above AVG	125
M	Below AVG	90
M	NULL	215
NULL	NULL	215

2.4 Data Visualization

For the Data visualization, we used many metrics to understand the data and get insights from it.

- **Average salary for each work year and experience level combination**
- **Number of employees per job title and company location combination**
- **Maximum salary for each employment type and job title combination**
- **Number of jobs with different remote ratios for each company location**
- **Average salary for remote vs. non-remote jobs for each category**
- **Number of employees categorized by company size and employee residence combination**
- **Maximum salary**
- **Minimum salary**
- **Average salary**
- **Number and percentage of employees per work year**

we wanted to bring to your attention that there are additional files included in the folders attached with this report. These supplementary materials aim to provide a more comprehensive understanding of the topic discussed in the report.