University of Tunis El Manar
Faculty of Sciences of Tunis

Data Warehouse Project

Salaries in Jobs Around the World

Prepared by:

Makrem Ben Zekri
Yassine Aljene

Supervisor:

Rihab Mersni

Date:
December 6, 2024

Contents

1 Introduction

The job market, particularly in the IT sector, plays a crucial role in the global economy, with varying trends in salaries, employment rates, and required skills across different regions. This project, “Data Warehouse Project: Salaries in Jobs Around the World”, aims to analyze these trends globally, with a focus on the Tunisian job market.

The objectives include collecting data from multiple sources, designing a data warehouse for structured analysis, and creating an interactive Power BI dashboard to visualize key metrics such as salary comparisons, skill distributions, and employment trends.

This report outlines the methodology used in the project, from data collection and ETL processes to the creation of the dashboard, and presents the findings and recommendations for future work.

2 Methodology

2.1 Data Collection

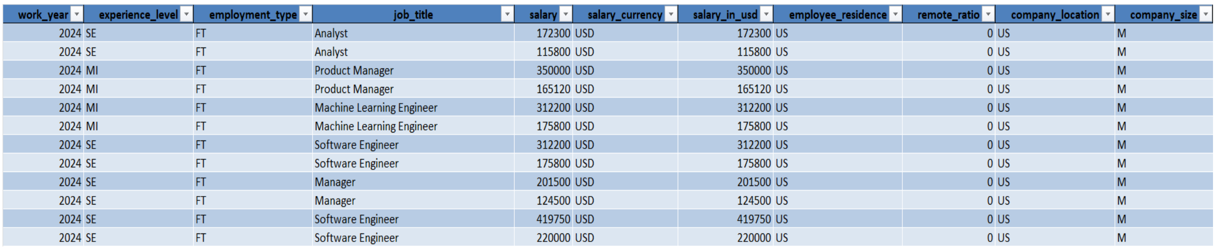
To ensure a comprehensive and accurate analysis of IT job trends, salaries, and required skills, data was gathered from multiple reliable sources. The datasets collected include:

2.1.1 Jobs Salary Dataset:

Sourced from `AIJobs.net`[1], this dataset provides salary data for various IT roles across different regions. The dataset contains the following key columns:

- `work_year`: Year of the salary data.
- `experience_level`: The experience level of the employee (e.g., junior, mid, senior).
- `employment_type`: Type of employment (e.g., full-time, part-time, contract).
- `job_title`: The job title of the employee.
- `salary`: The salary in the specified currency.
- `salary_currency`: The currency of the salary.
- `salary_in_usd`: The salary converted to USD for easier comparison.
- `employee_residence`: Location of the employee.
- `remote_ratio`: The percentage of remote work in the role.
- `company_location`: The location of the company.
- `company_size`: Size of the company (e.g., small, medium, large).

Figure 1 illustrates a sample of the salaries dataset.



work_year	experience_level	employment_type	job_title	salary	salary_currency	salary_in_usd	employee_residence	remote_ratio	company_location	company_size
2024	SE	FT	Analyst	172300	USD	172300	US	0	US	M
2024	SE	FT	Analyst	115800	USD	115800	US	0	US	M
2024	MI	FT	Product Manager	350000	USD	350000	US	0	US	M
2024	MI	FT	Product Manager	165120	USD	165120	US	0	US	M
2024	MI	FT	Machine Learning Engineer	312200	USD	312200	US	0	US	M
2024	MI	FT	Machine Learning Engineer	175800	USD	175800	US	0	US	M
2024	SE	FT	Software Engineer	312200	USD	312200	US	0	US	M
2024	SE	FT	Software Engineer	175800	USD	175800	US	0	US	M
2024	SE	FT	Manager	201500	USD	201500	US	0	US	M
2024	SE	FT	Manager	124500	USD	124500	US	0	US	M
2024	SE	FT	Software Engineer	419750	USD	419750	US	0	US	M
2024	SE	FT	Software Engineer	220000	USD	220000	US	0	US	M

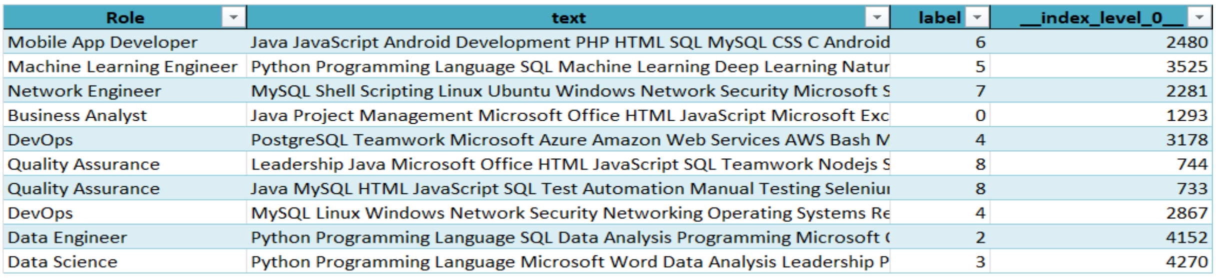
Figure 1: Salaries Dataset

2.1.2 Roles Based on Skills Dataset:

Obtained from **Hugging Face**'s Datasets Repository[2], this dataset maps IT job roles to the skills required for each position. It was used to assess skill distribution across different IT roles and regions. The dataset contains the following key columns:

- `Role`: The IT job role.
- `text`: A list of skills associated with the role, providing insight into the required skill set for each position.
- `label`: The label associated with the job role .
- `__index_level_0__`: An index level used for internal purposes.

Figure 2 illustrates a sample of the roles based on skills dataset.

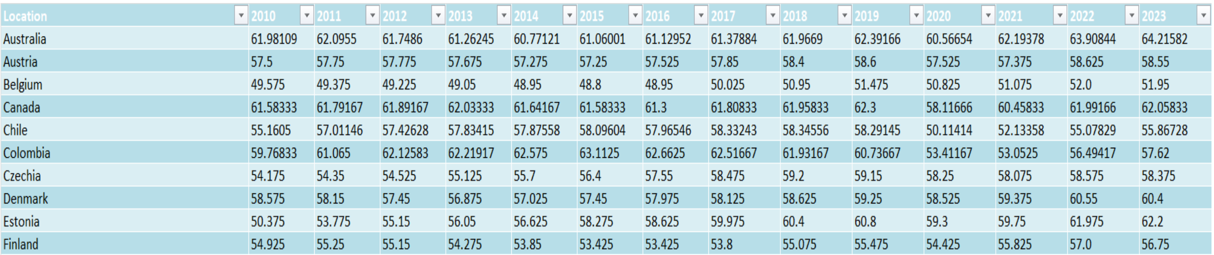


Role	text	label	__index_level_0__
Mobile App Developer	Java JavaScript Android Development PHP HTML SQL MySQL CSS C Android	6	2480
Machine Learning Engineer	Python Programming Language SQL Machine Learning Deep Learning Natur	5	3525
Network Engineer	MySQL Shell Scripting Linux Ubuntu Windows Network Security Microsoft S	7	2281
Business Analyst	Java Project Management Microsoft Office HTML JavaScript Microsoft Exc	0	1293
DevOps	PostgreSQL Teamwork Microsoft Azure Amazon Web Services AWS Bash M	4	3178
Quality Assurance	Leadership Java Microsoft Office HTML JavaScript SQL Teamwork Nodejs S	8	744
Quality Assurance	Java MySQL HTML JavaScript SQL Test Automation Manual Testing Seleniu	8	733
DevOps	MySQL Linux Windows Network Security Networking Operating Systems Re	4	2867
Data Engineer	Python Programming Language SQL Data Analysis Programming Microsoft (	2	4152
Data Science	Python Programming Language Microsoft Word Data Analysis Leadership P	3	4270

Figure 2: Roles Based on Skills Dataset

2.1.3 Employment Rate by Region:

Data from the **OECD** Employment and Labour Force Statistics provided valuable insights into global employment trends and workforce participation, enhancing the understanding of the labor market dynamics in relation to IT jobs.[3]
Figure 3 shows a sample of the OECD Employment and Labour Force dataset.



Location	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Australia	61.98109	62.0955	61.7486	61.26245	60.77121	61.06001	61.12952	61.37884	61.9669	62.39166	60.56654	62.19378	63.90844	64.21582
Austria	57.5	57.75	57.775	57.675	57.275	57.25	57.525	57.85	58.4	58.6	57.525	57.375	58.625	58.55
Belgium	49.575	49.375	49.225	49.05	48.95	48.8	48.95	50.025	50.95	51.475	50.825	51.075	52.0	51.95
Canada	61.58333	61.79167	61.89167	62.03333	61.64167	61.58333	61.3	61.80833	61.95833	62.3	58.11666	60.45833	61.99166	62.05833
Chile	55.1605	57.01146	57.42628	57.83415	57.87558	58.09604	57.96546	58.33243	58.34556	58.29145	50.11414	52.13358	55.07829	55.86728
Colombia	59.76833	61.065	62.12583	62.21917	62.575	63.1125	62.6625	62.51667	61.93167	60.73667	53.41167	53.0525	56.49417	57.62
Czechia	54.175	54.35	54.525	55.125	55.7	56.4	57.55	58.475	59.2	59.15	58.25	58.075	58.575	58.375
Denmark	58.575	58.15	57.45	56.875	57.025	57.45	57.975	58.125	58.625	59.25	58.525	59.375	60.55	60.4
Estonia	50.375	53.775	55.15	56.05	56.625	58.275	58.625	59.975	60.4	60.8	59.3	59.75	61.975	62.2
Finland	54.925	55.25	55.15	54.275	53.85	53.425	53.425	53.8	55.075	55.475	54.425	55.825	57.0	56.75

Figure 3: Employment rate by country by year

2.1.4 Country Codes and Currency Exchange Rates:

To enrich the dataset and ensure compatibility between regions, additional datasets were downloaded that included country codes and currency exchange rates. This allowed for more accurate regional comparisons and salary adjustments based on currency differences.

2.1.5 Job Title Mapping Dataset:

The Job Title Mapping.xlsx was obtained to match job titles with corresponding roles, enabling more precise mapping between different job descriptions and facilitating deeper analysis of job trends across regions.

2.2 Conceptual Design

The conceptual design of the data warehouse is structured around a star schema to facilitate efficient analysis of job market data. This schema integrates multiple dimensions and fact tables to provide a comprehensive view of salary trends, job roles, employment rates, and related attributes. The primary components of the design are as follows:

2.2.1 Fact Tables

- **Fact_Salary:** This table serves as the central fact table, capturing salary-related data across various dimensions. Key attributes include:
  - `salary`: The original salary amount.
  - `salary_amount_usd`: The salary converted into USD for standardization.
  - Foreign keys linking to dimensions such as `Dim_Job`, `Dim_Company`, `Dim_Employee`, `Dim_Currency`, and `Dim_Time`.
- **Fact_Role_Skills:** This fact table maps roles to associated skills, allowing for skill distribution analysis. It contains:
  - `role_id`: A reference to `Dim_Role`.
  - `skill`: The specific skill associated with the role.
- **Fact_Employment_Rate:** This table tracks employment rates by country and time, providing insights into workforce trends. Attributes include:
  - `employment_rate`: The employment rate percentage.

  - Foreign keys linking to `Dim_Location` and `Dim_Time`.

2.2.2 Dimension Tables

- **Dim_Job:** Captures job-related metadata, such as:
  - `job_title`: The title of the job.
  - `role_id`: A reference to `Dim_Role` for mapping job roles.
- **Dim_Role:** Provides a high-level categorization of job roles:
  - `role_id`: The unique identifier for each role.
  - `role_name`: The name of the role.
- **Dim_Employee:** Stores employee-specific details, such as:
  - `employment_type`: Type of employment (e.g., full-time, part-time).
  - `experience_level`: The employee's experience level.
- **Dim_Company:** Includes information about companies:
  - `company_size`: A classification of the company's size (e.g., small, medium, large).
  - `location_id`: Links to the `Dim_Location` table for geographical context.
- **Dim_Location:** Captures geographic information:
  - `country_name`: The name of the country.
  - `country_code`: The corresponding ISO country code.
- **Dim_Currency:** Tracks currency information for salary standardization:
  - `currency_name`: The name of the currency.
  - `exchange_rate`: The exchange rate for conversion into USD.
- **Dim_Time:** Provides temporal granularity for analysis:
  - `work_year`: The year associated with the data point.

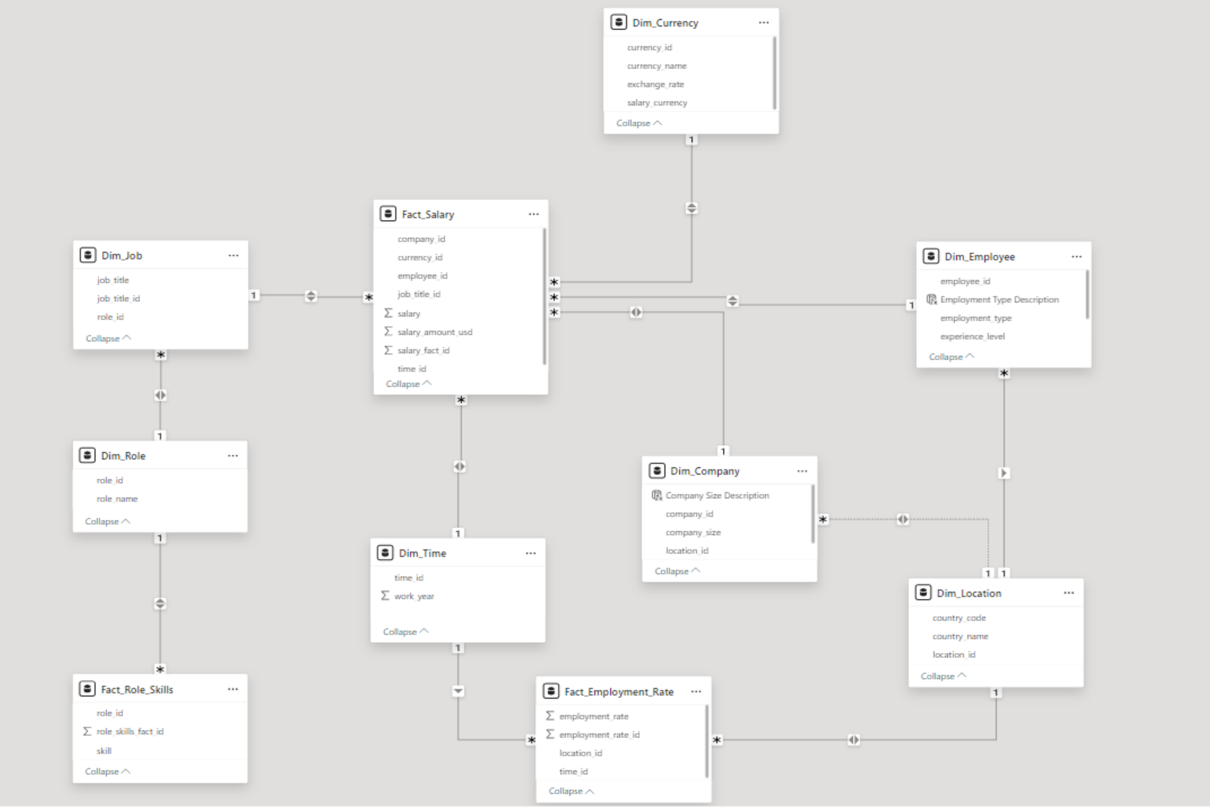


Figure 4: Conceptual Design of the Data Warehouse

This conceptual design ensures scalability, consistency, and ease of querying, supporting diverse analytical use cases such as salary comparisons, skill gap analysis, and regional employment trends. The star schema structure minimizes redundancy and enhances performance for dashboard visualizations and business intelligence tools.

2.3 Data Warehouse Implementation

The data warehouse implementation was carried out using Microsoft SQL Server Management Studio (SSMS) and SQL Server Integration Services (SSIS). The star schema design from the conceptual phase was translated into physical tables and relationships.

2.3.1 Tables Creation

This section details the creation of these tables and their relationships.
The SQL code below shows the implementation of these tables:

```
-- Table 1: Dim_Time
CREATE TABLE Dim_Time (
    time_id INT PRIMARY KEY IDENTITY(1,1),
    work_year INT NOT NULL
);

-- Table 2: Dim_Currency
```

```
CREATE TABLE Dim_Currency (
    currency_id INT PRIMARY KEY IDENTITY(1,1),
    salary_currency VARCHAR(10) NOT NULL,
    exchange_rate DECIMAL(18,6) NOT NULL,
    currency_name VARCHAR(50) NOT NULL
);

-- Table 3: Dim_Location
CREATE TABLE Dim_Location (
    location_id INT PRIMARY KEY IDENTITY(1,1),
    country_code VARCHAR(100) NOT NULL,
country_name VARCHAR(255) NOT NULL
);

-- Table 4: Dim_Employee (References Dim_Location)
CREATE TABLE Dim_Employee (
    employee_id INT PRIMARY KEY IDENTITY(1,1),
    location_id INT NOT NULL,
    experience_level VARCHAR(20) NOT NULL,
    employment_type VARCHAR(20) NOT NULL,
    FOREIGN KEY (location_id) REFERENCES Dim_Location(location_id)
);

-- Table 5: Dim_Company (References Dim_Location)
CREATE TABLE Dim_Company (
    company_id INT PRIMARY KEY IDENTITY(1,1),
    location_id INT NOT NULL,
    company_size VARCHAR(20),
    remote_ratio INT,
    FOREIGN KEY (location_id) REFERENCES Dim_Location(location_id)
);

-- Table 6: Dim_Role
CREATE TABLE Dim_Role (
    role_id INT PRIMARY KEY IDENTITY(1,1),
    role_name VARCHAR(100) NOT NULL
);

-- Table 7: Dim_Job (References Dim_Role)
CREATE TABLE Dim_Job (
    job_title_id INT PRIMARY KEY IDENTITY(1,1),
    role_id INT NOT NULL,
    job_title VARCHAR(100) NOT NULL,
    FOREIGN KEY (role_id) REFERENCES Dim_Role(role_id)
);
```

```
-- Table 8: Fact_Employment_Rate (References Dim_Time and Dim_Location)
CREATE TABLE Fact_Employment_Rate (
    employment_rate_id INT PRIMARY KEY IDENTITY(1,1),
    time_id INT NOT NULL,
    location_id INT NOT NULL,
    employment_rate VARCHAR(100),
    FOREIGN KEY (time_id) REFERENCES Dim_Time(time_id),
    FOREIGN KEY (location_id) REFERENCES Dim_Location(location_id)
);

-- Table 9: Fact_Role_Skills (References Dim_Role)
CREATE TABLE Fact_Role_Skills (
    role_skills_fact_id INT PRIMARY KEY IDENTITY(1,1),
    role_id INT NOT NULL,
    skill VARCHAR(100) NOT NULL,
    FOREIGN KEY (role_id) REFERENCES Dim_Role(role_id)
);

-- Table 10: Fact_Salary (References Dim_Time, Dim_Currency, Dim_Job, Dim_Employee, Di
CREATE TABLE Fact_Salary (
    salary_fact_id INT PRIMARY KEY IDENTITY(1,1),
    time_id INT NOT NULL,
    currency_id INT NOT NULL,
    job_title_id INT NOT NULL,
    employee_id INT NOT NULL,
    company_id INT NOT NULL,
    salary DECIMAL(15, 2),
    salary_amount_usd VARCHAR(50),
    FOREIGN KEY (time_id) REFERENCES Dim_Time(time_id),
    FOREIGN KEY (currency_id) REFERENCES Dim_Currency(currency_id),
    FOREIGN KEY (job_title_id) REFERENCES Dim_Job(job_title_id),
    FOREIGN KEY (employee_id) REFERENCES Dim_Employee(employee_id),
    FOREIGN KEY (company_id) REFERENCES Dim_Company(company_id)
);
```

This implementation ensures that the star schema is accurately represented in the database, with well-defined relationships between fact and dimension tables. The use of primary keys and foreign keys enforces data integrity, while the identity columns simplify the generation of unique identifiers.

2.3.2 ETL Process in SSIS

The Extract, Transform, and Load (ETL) process was implemented using SQL Server Integration Services (SSIS) to efficiently load and transform data into the data warehouse. The SSIS package was structured with modular and reusable components to enhance maintainability and scalability.

Loading Data Using Excel Connection Manager

The data was first loaded into the SSIS model using the Excel Connection Manager. Each dataset was processed in separate **Data Flow Tasks** to maximize parallelism and improve performance. This design allowed faster execution by leveraging concurrent processing for multiple datasets.

Loading Data into Staging Tables

Before transforming the data, staging tables were created to temporarily hold raw data. An **Execute SQL Task** was included at the beginning of the SSIS workflow to truncate the staging tables. This ensured that they were emptied before each run, avoiding any duplication or contamination from previous executions.

2.3.3 Data Transformation

The transformation phase involved a series of steps to clean, reshape, and prepare the data for loading into the data warehouse.

- **Data Cleaning and Normalization:**
    - **Unpivoting the Employment Rate Table:** The Employment Rate dataset was unpivoted to reduce the number of columns and standardize the structure. The result was a staging table with three columns (`work_year`, `country`, and `employment_rate`), making it easier to execute queries, especially for `GROUP BY` operations.
- **Deduplication Using Lookup Transformations:** For each data loading task, a **Lookup Transformation** was used to ensure that only new records were inserted into the destination tables. This process checked for existing entries and prevented data duplication, maintaining the integrity of the warehouse.

Enhancing the Skills Dataset

While most transformations were conducted within SSIS, the enhancement of the skills dataset required a more flexible approach. Attempts to implement a solution in SSIS using C# scripting were unsuccessful, so the task was completed in Python.

- **Custom Skill Mapper:** A Python script was developed to iterate through each list of skills associated with job titles. Relevant skills were identified and extracted using predefined mapping logic.
- **Unpivoting Skills:** The extracted skills were unpivoted alongside their corresponding job roles to create a clean, relational structure. This transformation ensured that each skill was represented as an individual row linked to its respective role, simplifying future analyses.

The use of Python for this specific task ensured better control over the data manipulation process and enabled the successful preparation of the skills dataset for loading into the data warehouse.

2.3.4 Data Loading into Dimensional and Fact Tables

The process of loading data into the dimensional and fact tables was carried out systematically to ensure coherence and data integrity. Separate **Data Flow Tasks** were created for each dimension table, followed by the fact tables, maintaining the correct order to respect the relationships defined in the star schema.

To address inconsistencies in data types between staging tables and destination tables, **Derived Columns** transformations were used. These transformations allowed us to perform necessary conversions and adjustments, ensuring compatibility with the schema design.

Lookup Transformations played a dual role in this process. Firstly, they were employed to populate foreign key columns in the fact tables by referencing the corresponding values from related dimension tables, ensuring accurate relationships across the schema. Secondly, they were utilized for **deduplication**, preventing duplicate entries by verifying whether the incoming data already existed in the target tables. This ensured that only unique and new records were inserted, maintaining the integrity of the data warehouse.

Figure 5 illustrates the Data Flow Tasks in SSIS.

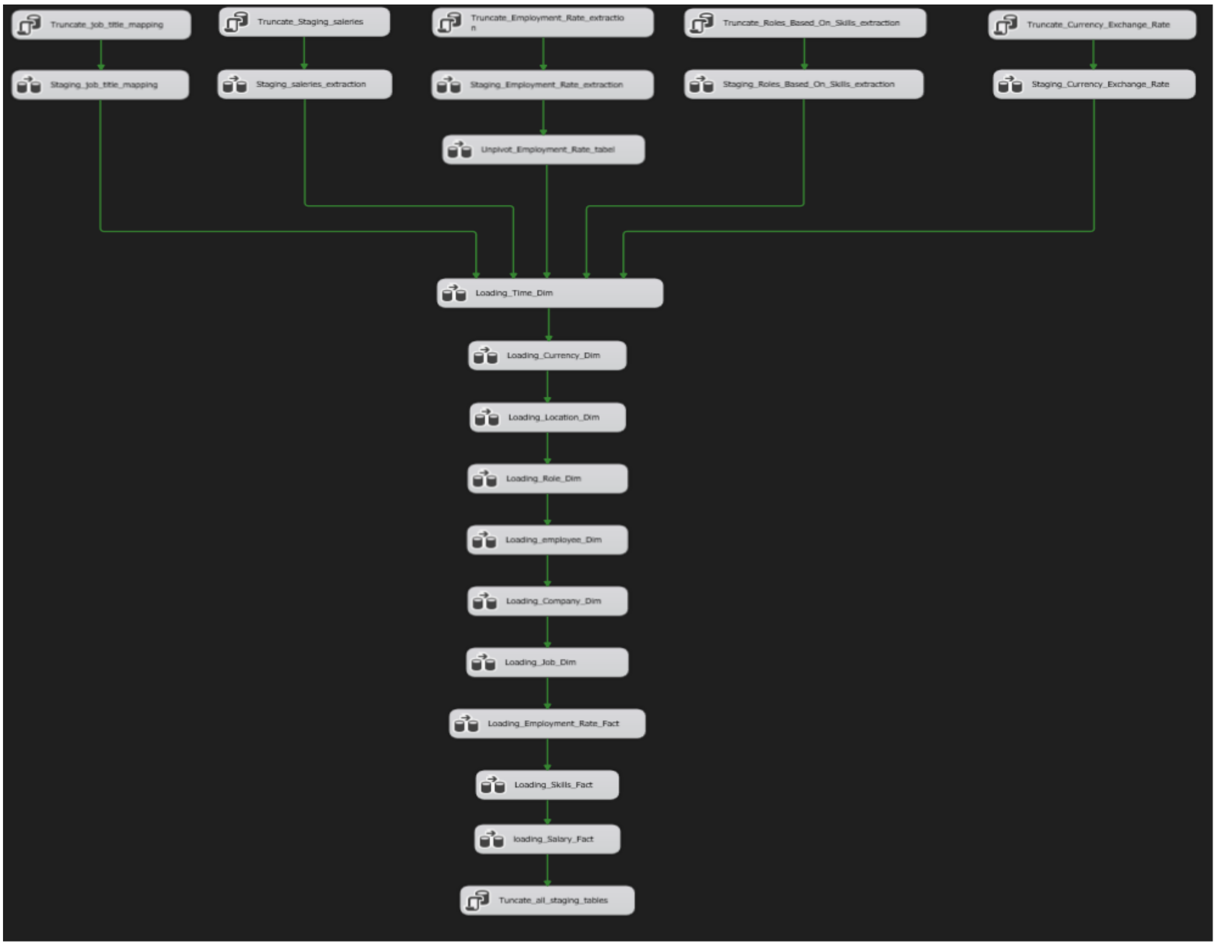

Truncate_job_title_mapping
Truncate_Staging_saleries
Truncate_Employment_Rate_extraction
Truncate_Roles_Based_On_Skills_extraction
Truncate_Currency_Exchange_Rate
Staging_job_title_mapping
Staging_saleries_extraction
Staging_Employment_Rate_extraction
Staging_Roles_Based_On_Skills_extraction
Staging_Currency_Exchange_Rate
Unpivot_Employment_Rate_tabel
Loading_Time_Dim
Loading_Currency_Dim
Loading_Location_Dim
Loading_Role_Dim
Loading_employee_Dim
Loading_Company_Dim
Loading_Job_Dim
Loading_Employment_Rate_Fact
Loading_Skills_Fact
loading_Salary_Fact
Tuncate_all_staging_tables

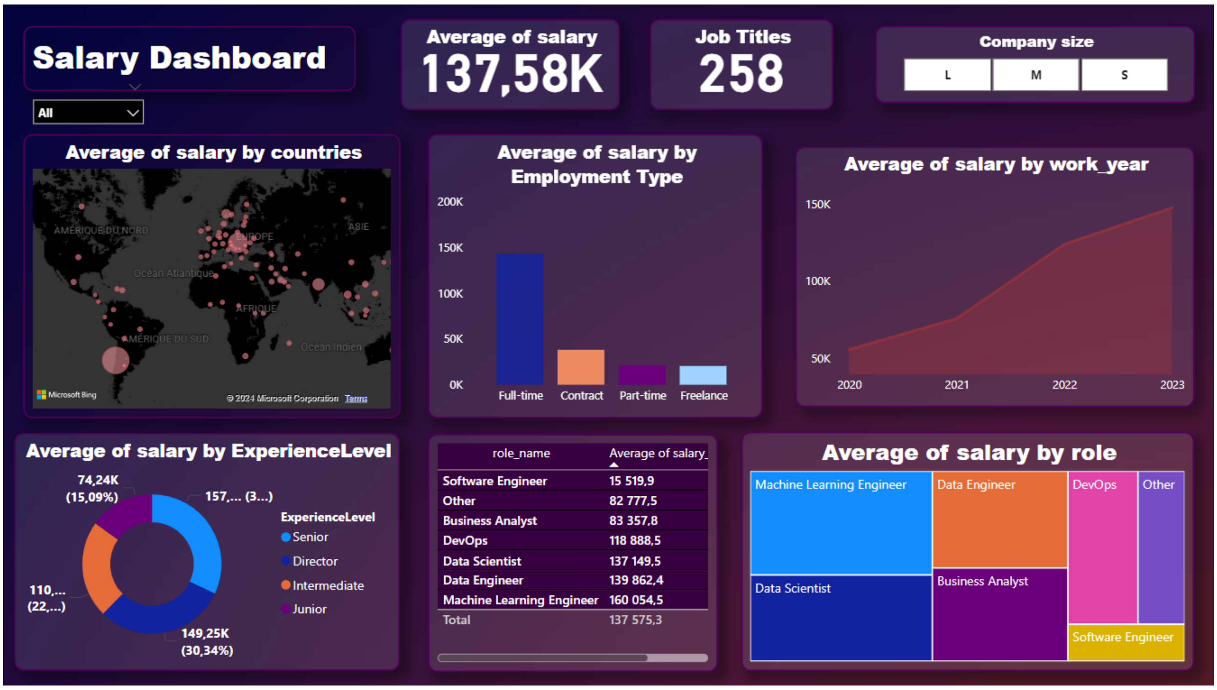
Figure 5: SSIS Project

2.4 Dashboard Creation

The data visualization component of this project was implemented using **Power BI** to create interactive and visually appealing dashboards that provide comprehensive insights into IT market trends and job-related data. Four primary dashboards were designed, each focusing on a specific aspect of the analysis:

2.4.1 IT Market Overview

- Displays the **average salary by countries** using a world map to highlight geographic disparities.
- Includes insights on the **average salary by employment type** (full-time, contract, and part-time) and **experience level** (e.g., junior, mid-level, senior).
- Tracks salary progression over time using the **average salary by work year** graph, alongside a breakdown of salaries by **job roles**.



role_name	Average of salary_
Software Engineer	15 519,9
Other	82 777,5
Business Analyst	83 357,8
DevOps	118 888,5
Data Scientist	137 149,5
Data Engineer	139 862,4
Machine Learning Engineer	160 054,5
Total	137 575,3

Figure 6: IT Market Overview Dashboard

2.4.2 Global Comparison

- Analyzes **remote work trends** by comparing salary averages based on remote ratio and company size.
- Highlights the **top 10 countries** with the highest average salaries.
- Explores **disparities across roles**, showing how average salaries vary among different IT positions globally.

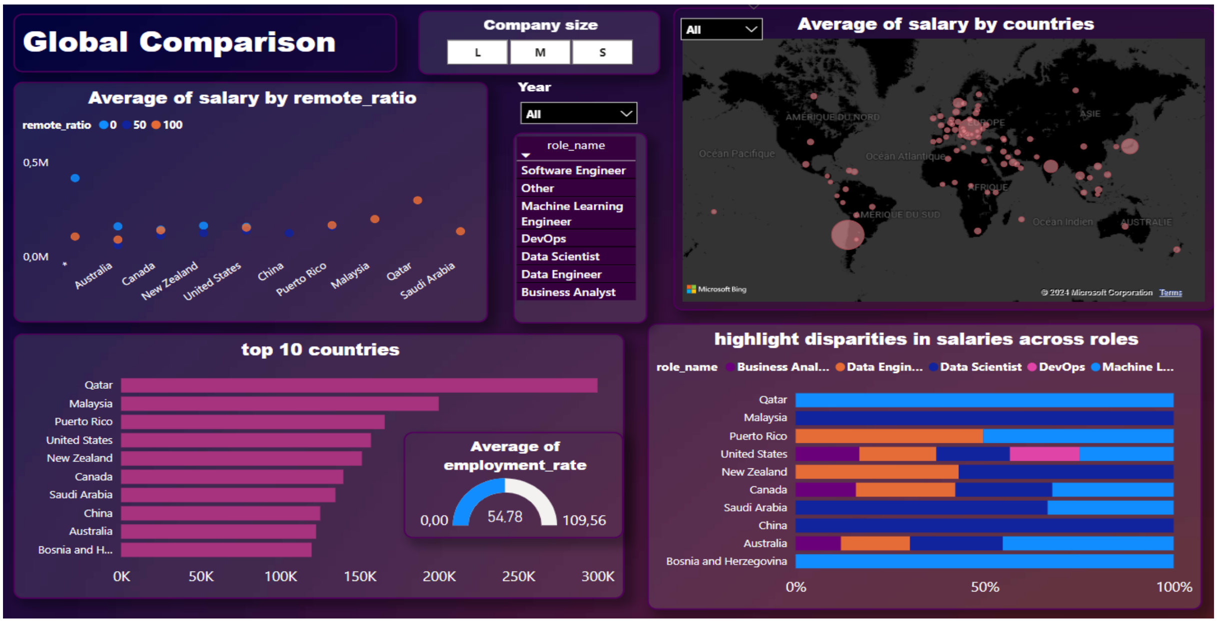
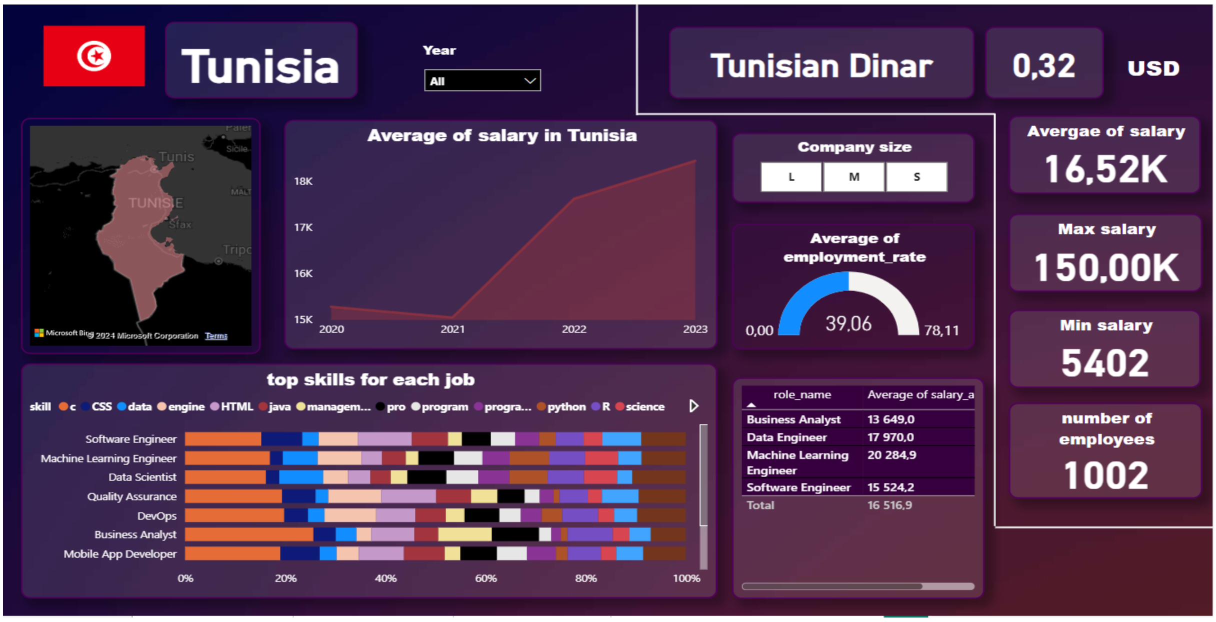


Figure 7: Global Comparison Dashboard

2.4.3 Tunisian IT Market

- Highlights Tunisia-specific **salary trends**, illustrating how average salaries have evolved over the years.
- Features a **stacked bar chart** for top jobs like software engineer and data scientist, showcasing the distribution of key skills in demand for each role.
- Includes key metrics such as **employment rates** and **salary conversions into Tunisian dinars**, enabling localized and contextualized analysis.



role_name	Average of salary_a
Business Analyst	13 649,0
Data Engineer	17 970,0
Machine Learning Engineer	20 284,9
Software Engineer	15 524,2
Total	16 516,9

Figure 8: Tunisian IT Market Dashboard

2.4.4 Skills Insights and Employment Trends

- Displays the **most sought-after skills** in IT roles globally, with a bar chart ranking technical skills like programming languages and frameworks.
- Monitors **employment rate trends over time** for leading IT markets such as the United States, Canada, and European countries.
- Analyzes **remote work adoption by region**, showing remote work trends across years and the impact of the global shift toward flexible work models.

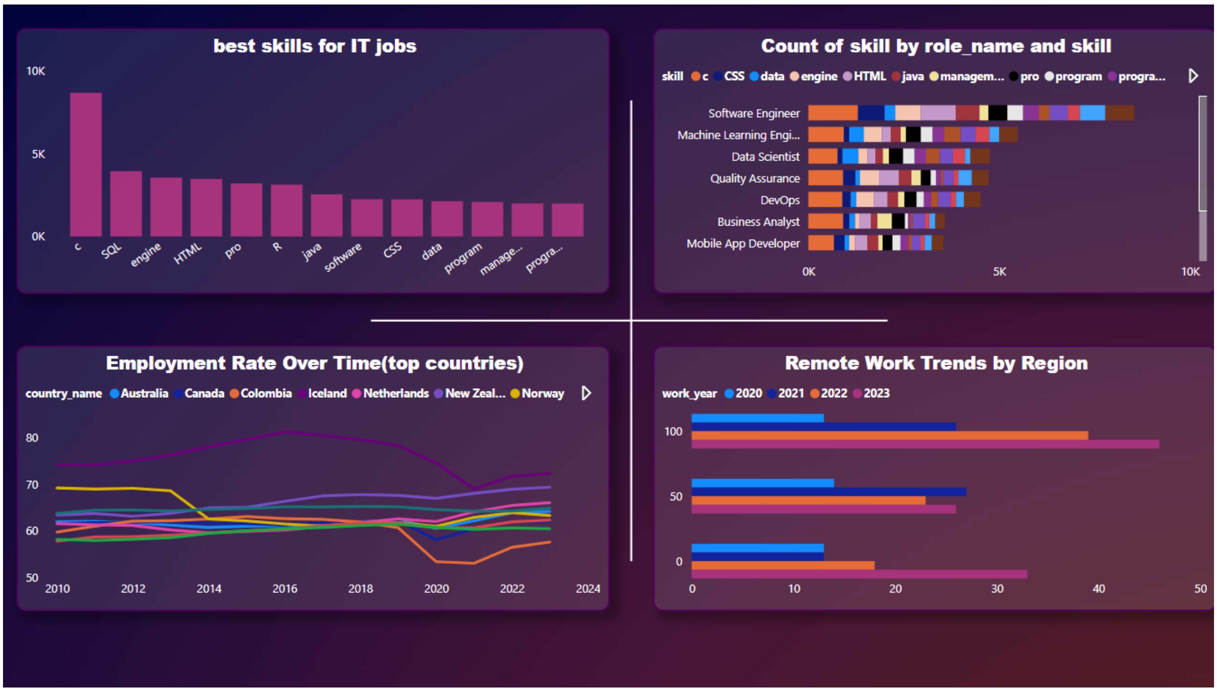


Figure 9: Skills Insights and Employment Trends Dashboard

These dashboards provide **interactive filters**, enabling users to drill down into specific metrics such as job roles, company size, and employment types. The combination of maps, charts, and heatmaps ensures clear and actionable insights into IT market trends worldwide.

3 Results

The implementation of the project led to the creation of dynamic, interactive dashboards that deliver actionable insights into global and localized IT market trends. These dashboards effectively transform complex datasets into intuitive visualizations, enabling stakeholders such as job seekers, employers, and policymakers to make informed decisions. By simplifying intricate data and fostering data-driven decision-making, the dashboards empower users to explore and customize insights according to their specific interests. Their interactive design ensures versatility, making them indispensable tools for comprehensively understanding the IT labor market.

References

[1] AIJobs.net. *AI Jobs Salary Dataset.* Accessed: 2024-12-05. 2024. URL: https://aijobs.net/salaries/download/.

[2] Fazni. *Roles Based on Skills Dataset.* https://huggingface.co/datasets/fazni/roles-based-on-skills. Accessed: 2024-12-05. 2024.

[3] OECD. *OECD Employment Statistics: Employment to Population Ratio.* https://data-explorer.oecd.org/vis?lc=en&tm=DF_IALFS_INDIC&pg=0&snb=1&vw=tb&df[ds]=dsDisseminateFinalDMZ&df[id]=DSD_LFS%40DF_IALFS_INDIC&df[ag]=OECD.SDD.TPS&df[vs]=&pd=2010%2C2023&dq=.EMP_WAP...Y._T.Y_GE15..A&to[TIME_PERIOD]=false. Accessed: 2024-12-05. 2023.