



*Licence Professionnelle*

**Systèmes d'Information et BIG DATA**

***Submitted by***

**Hamza Elhaiki**

**Abdelhay Amani**

**Hicham Ben Hsine**

# Linkedin Workforce Analytics

Data warehouse

***Under the supervision of Professor Youssef Lefdaoui.***

Academic year 2023/2024

# Contents

# Introduction

In today's dynamic business landscape, the effective utilization of data has become imperative for informed decision-making. This project focuses on the realm of workforce analytics, where the convergence of data science methodologies and advanced technologies unravels profound insights within the complex tapestry of professional engagements.

The project aims to delve into LinkedIn data, specifically within Moroccan organizational frameworks, to extract multifaceted insights. By harnessing data warehousing, advanced analytics, and visualization techniques, the goal is to illuminate critical aspects of employee profiles, job trends, skill landscapes, and temporal patterns.

The primary objective is to uncover actionable insights that inform strategic workforce decisions, enhance talent management strategies, and elucidate correlations between skill acquisition, professional engagements, and geographical nuances. The project will meticulously prepare data, employ dimensional modeling techniques, and utilize Tableau visualizations to present a comprehensive narrative.

Special acknowledgment and gratitude are extended to Professor Youssef Lefdaoui, whose expertise and guidance in the field of data warehousing significantly enriched this project.

# Chapter 1

# Project Overview

## 1.1 Objectives and Goals

The primary objective of this project is to conduct comprehensive workforce analytics utilizing LinkedIn data. The specific goals include:

- Analyzing and understanding trends in employment and skill acquisition.
- Creating detailed insights into job distributions and geographical employment hotspots.
- Extracting valuable insights to assist in strategic decision-making for recruitment and talent management.

## 1.2 Scope and Context

The project scope encompasses:

- Utilizing LinkedIn data provided by the professor of the data warehouse course.
- Performing data cleaning, transformation, and integration using Talend.
- Developing a dimensional model based on Kimball methodology for better data representation.
- Employing Tableau for visualization to provide a user-friendly interface for data exploration and analysis.

The context of this project revolves around leveraging LinkedIn's rich dataset to derive meaningful insights into workforce patterns and industry trends, aiming to facilitate informed decision-making in talent acquisition and management.

# Chapter 2

# Data Preparation

## 2.1 Sources of Data

The project relies on datasets provided by our esteemed professor specializing in data warehousing. These datasets encompass a diverse range of variables and parameters crucial for workforce analytics within Moroccan organizational contexts.

### 2.1.1 Challenges Faced

Accessing and organizing the dataset posed certain challenges. One of the prominent hurdles encountered was the structuring of the data due to its extensive nature. The dataset encompassed various facets of information, including:

- **Personal Information:** Full names, first names, last names, titles.
- **Location and Organization Details:** Detailed information regarding multiple organizations, their titles, start and end dates, descriptions, and locations.
- **Educational Background:** Multiple educational entries with degrees, fields of study, and timeframes.
- **Skills and Interests:** Comprehensive listings of skills, follower counts, relationship details, and interests.

### 2.1.2 Data Structure Complexity

The dataset's intricate structure, particularly the multiplicity of organizations, educational details, and diverse skill sets, demanded meticulous handling. Organizing these varied aspects into a coherent structure for analysis and modeling posed a significant challenge.

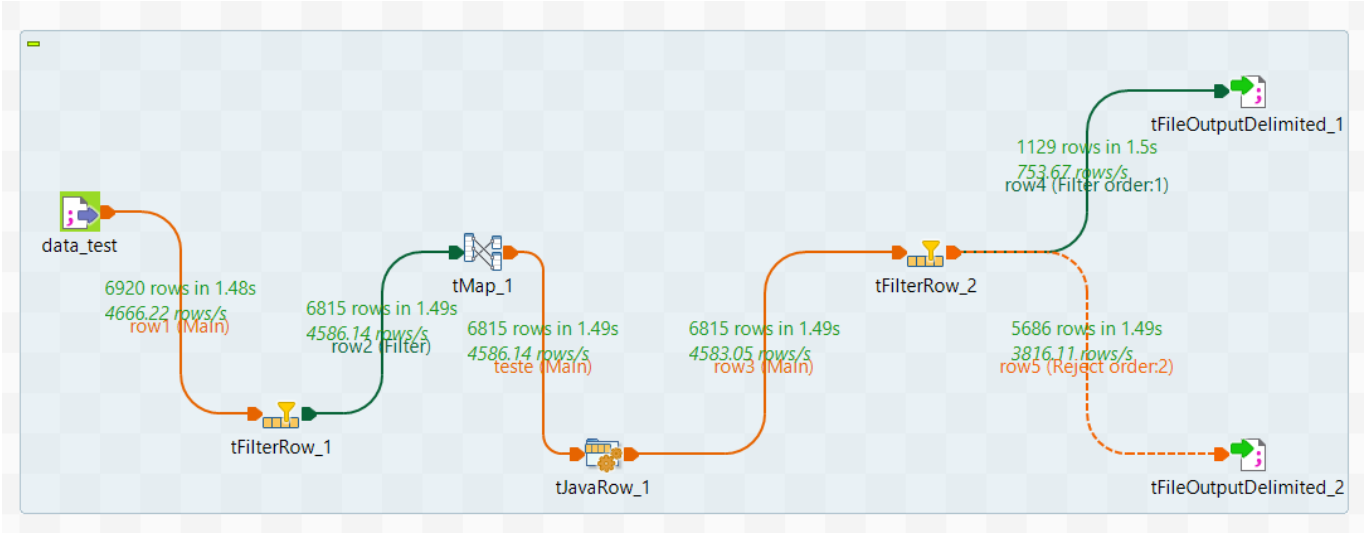
### 2.1.3 Key Parameters

Among the numerous parameters, the focus lay on establishing relationships between individuals and their affiliations, understanding educational qualifications, discerning skill proficiencies, and gauging interests and follower dynamics within professional networks. These dimensions formed the crux of our analytical exploration and insights generation for workforce analytics on LinkedIn data.

## 2.2 Data Cleaning and Transformation

The phase of Data Cleaning and Transformation constitutes a pivotal stage in this project, focused on refining raw data obtained from LinkedIn's employment-related datasets. This segment involves a meticulous process aimed at ensuring the accuracy, consistency, and reliability of the data for subsequent analysis.

Utilizing Talend as the primary tool, this phase encompasses comprehensive data integration, rectification of inconsistencies, and addressing missing or erroneous data points. Components such as `TFilterRow`, `TJavaRow`, and `TMap` are employed to filter the dataset, eliminate invalid entries, and transform complex data structures into standardized formats ,the process involves various aspects, including the removal of missing values, rectification of invalid formats in job titles, locations, skills, education, and interests. Additionally, tools such as `TFileExcelInput` and `TFileOutputDelimited` facilitate the extraction and storage of cleansed data into a PostgreSQL database, ensuring that subsequent analysis is based on accurate and well-prepared datasets.




`TFilterRow 1` - Removing Invalid Employee Titles: Purpose: Eliminating invalid or incomplete employee titles from the dataset. Criteria: Removing entries containing "–", "!", or empty strings within the employee title column. Function: This filter ensures that only valid and complete employee titles are retained for further processing.

`TFilterRow 2` - Eliminating Specific Date from TJavarow Output: Purpose: Removing specific date provided by the `TJavaRow` component due to invalid or empty values. Functionality: Filtering out rows with specific date that are considered invalid or empty. These entries are identified and removed from the dataset to maintain data integrity.

`TMap`: We use for splitting data into: `TFileOutputDelimited_1`: Saving data from `tfilter` that satisfy the condition. `TFileOutputDelimited_2`: Saving data from `tfilter` that don't satisfy the condition, reject.

`TJavaRow`: Date Conversion: Converting unstructured date strings into a valid date format using Java code within a `TJavaRow` component.

**Functionality:**

*Input Data Handling:* The component takes an input date value (`input_row.date2`) along with other related fields (`Full_name`, `Location`, etc.).

*Mapping Month Abbreviations:* Utilizes a mapping of month abbreviations (in French) to their full names using a predefined map (`monthMap`).
*Parsing and Formatting:* Splits the input date string into parts (month abbreviation and year). Constructs a full date string in the format of "1 February 2012" based on the mapped month full name and year. Attempts to parse this constructed full date string using a default date format ("dd/MM/yyyy").
*Error Handling:* Handles exceptions that might arise during the parsing process, providing a default date ("10/10/1929") if needed.
*Output:* Generated Date: The script creates a `java.util.Date` object (`outputDate`) representing the converted date. Assigned Output: This converted date (`output_row.date2`) is assigned to an output column for further processing in your Talend job.

```
String inputDate = input_row.date2;
output_row.Full_name = input_row.Full_name;
output_row.Location = input_row.Location;
output_row.date = input_row.date;

java.util.Date outputDate = null;

// Map to store month abbreviations and their full names
java.util.Map<String, String> monthMap = new java.util.HashMap<>();
monthMap.put("janv.", "January");
monthMap.put("févr.", "February");
monthMap.put("mars", "March");
monthMap.put("avr.", "April");
monthMap.put("mai", "May");
monthMap.put("juin", "June");
monthMap.put("juil.", "July");
monthMap.put("août", "August");
monthMap.put("sept.", "September");
monthMap.put("oct.", "October");
monthMap.put("nov.", "November");
monthMap.put("déc.", "December");

// Default date format
java.text.SimpleDateFormat defaultDateFormat = new java.text.SimpleDateFormat("dd/MM/yyyy",
java.util.Locale.ENGLISH);

// Split the input date by space and get the month abbreviation
String[] parts = inputDate.split(" ");
String monthAbbr = parts[0];

// Check if the month abbreviation exists in the map
if (monthMap.containsKey(monthAbbr)) {
    String monthFullName = monthMap.get(monthAbbr);
    String year = parts[1];

    // Construct the full date string in the format of "1 February 2012"

    String fullDateString = "1 " + monthFullName + " " + year;

    try {
        // Parse the constructed full date string
        outputDate = defaultDateFormat.parse(defaultDateFormat.format(new
java.text.SimpleDateFormat("d MMMM yyyy", java.util.Locale.ENGLISH).parse(fullDateString)));
    } catch (java.text.ParseException e) {
        e.printStackTrace();
        // Handle the parse exception if needed
    }
} else {
    // Set a default date
    String defaultDateString = "10/10/1929";
    try {
        outputDate = defaultDateFormat.parse(defaultDateString);
    } catch (java.text.ParseException e) {
        e.printStackTrace();
        // Handle the parse exception if needed
    }
}

output_row.date2 = outputDate;
```

# Chapter 3

# Dimensional Modeling

## 3.1 Star Schema Architecture

The Star Schema architecture, a fundamental concept derived from the Kimball Dimensional Modeling methodology, forms the backbone of the data warehouse design in this project. It comprises a centralized fact table surrounded by dimension tables, resembling a star-shaped structure.

### 3.1.1 Dimension Tables

The dimension tables represent the core entities or business aspects being analyzed and include:
**dim_location**: Captures details related to different geographical locations, providing insights into the regional distribution of employees or organizations.
**dim_time**: Encompasses time-related attributes such as start and end dates of job roles, facilitating temporal analysis and trend identification.
**dim_employee**: Contains details about individual employees, including personal information, job titles, and other relevant attributes.
**dim_job**: Stores information about job roles or positions, aiding in job-related analysis and understanding the organizational structure.
**dim_company**: Represents data pertaining to different companies or organizations, providing a context for employment and organizational affiliation.

### 3.1.2 Fact Table

The fact table serves as the centerpiece and contains foreign keys linking to the dimension tables along with various measures and metrics. In this project, the fact table includes attributes such as skills, education details, interests, and other relevant dimensions, allowing comprehensive analysis and querying.

**Fact Table Attributes**: The fact table combines multiple dimensions and measures to facilitate analytical queries and insights generation. It includes foreign keys referencing dimension tables alongside skill-related, education-related, and interest-related attributes.

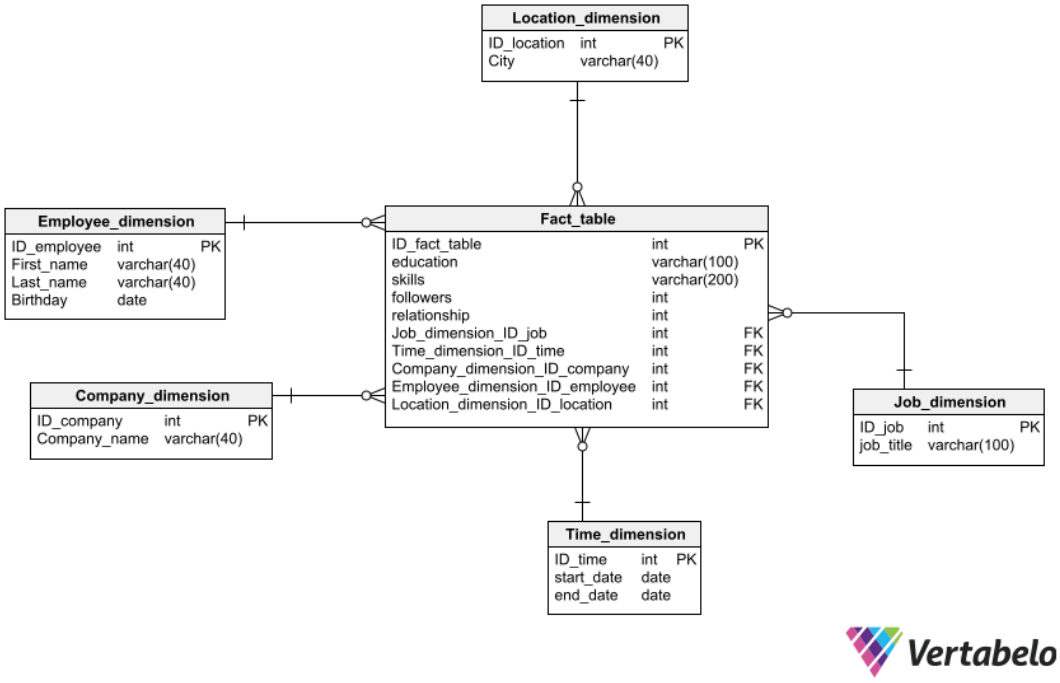
**Granularity and Aggregation**: The fact table operates at a granular level, capturing detailed information about employees' skills, educational qualifications, and interests. However, due to data limitations, some aggregations might be necessary, especially concerning time and location attributes.

### 3.1.3 Star Schema Design Benefits

The Star Schema architecture offers several advantages:

- **Simplified Queries:** Facilitates straightforward and efficient query execution by organizing data into clear dimensions and measures.
- **Enhanced Performance:** Improves performance due to reduced join operations, enabling faster data retrieval and analysis.
- **Ease of Understanding:** Provides a user-friendly design, enhancing the comprehension of data relationships and hierarchies.

The Star Schema architecture, with its inherent simplicity and performance benefits, underpins the data warehouse design, enabling effective analysis and extraction of actionable insights.

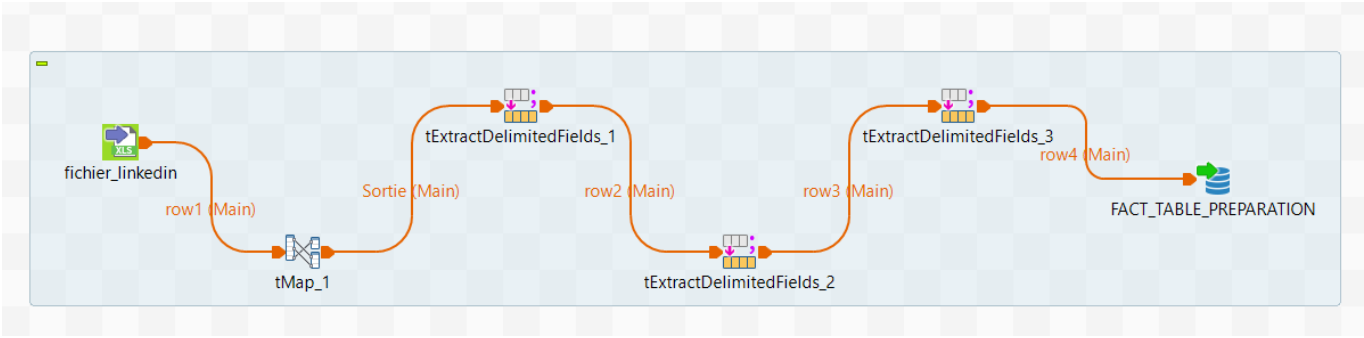


# Chapter 4

# Implementation of Data warehouse

## 4.1 Data preparation Job:

`TFileExcelInput` was the primary component used to read Excel files containing essential workforce data during the data preparation phase. Following this initial ingestion, the `TMap` component played a pivotal role in executing various transformations.




**Extracting City Names from Complex Phrases:**
Using `TMap`, intricate location strings such as "Préfecture de Rabat, Morocco" were dissected to isolate specific city names like "Rabat." This process significantly enhanced data clarity and facilitated more insightful analysis.

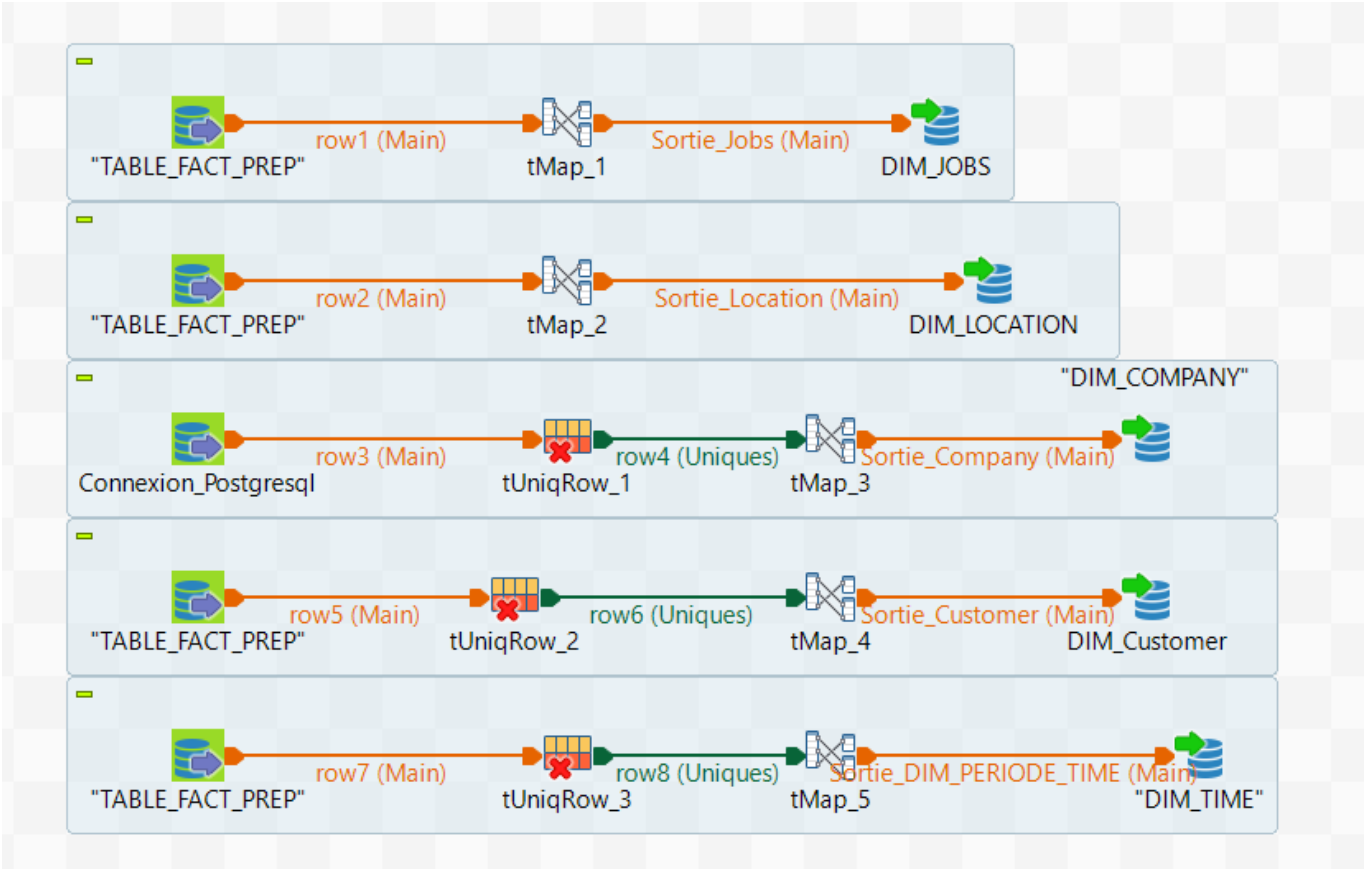
**String Handling and Cleaning:**
Within `TMap`, the data underwent refinement using StringHandling functions. Numeric characters, colons, and quotes within the 'Skills' field were removed utilizing regular expressions. This cleaning ensured standardized and uniform information across the dataset.

**Text Parsing with `TExtractDelimitedFields`:**
The use of `TExtractDelimitedFields` enabled the parsing of delimited fields such as 'Skills' and 'Interests.' This extraction process allowed for the segregation of individual skill sets and interests from strings containing multiple items separated by specific delimiters (',' and ';').

**Output to PostgreSQL with `TDBOutput`:**
The meticulously prepared, refined, and transformed data was loaded into a PostgreSQL database

using `TDBOutput`. This action ensured the data's compatibility and accessibility for subsequent analysis and modeling procedures.

## 4.2 Create dimensions Jobs:




**Dimension Table Creation - dim_jobs:**
*Data Retrieval from Postgres:*
Utilized the `TDBInput` component to extract distinct job titles from the 'TABLE_FACT_PREP' in the PostgreSQL database. The SQL query was structured to retrieve unique job titles where the 'Title' field was not null.
*Transformation in `TMap`:*
Employed the `TMap` component to define the transformation logic. Attributes such as job titles obtained from the `TDBInput` formed the basis of this transformation. Added primary keys or any necessary transformations within `TMap` to enrich the dataset for dimension table creation.
*Output to PostgreSQL:*
The transformed and structured job titles data was then written into a new table, namely 'dim_jobs,' within the PostgreSQL database using `TDBOutput`. This process aimed to extract unique job titles, assign primary keys or relevant attributes, and organize this information into the 'dim_jobs' table for use as a dimension in the overall analytics framework.

**Dimension Table Creation - dim_Location:**

*Data Extraction from Postgres:*
Leveraged the `TDBInput` component to retrieve distinct location values from the 'TABLE_FACT_PREP' table in the PostgreSQL database. A SQL query was employed to select unique location entries while excluding specific values like 'Maroc' or 'MA'.
*Transformation in `TMap`:*
The obtained locations underwent transformations in the `TMap` component, which could include tasks like assigning primary keys or restructuring the data if required.
*Output to PostgreSQL:*
The refined location data was written into a new table named 'dim_Location' within the PostgreSQL database using the `TDBOutput` component.

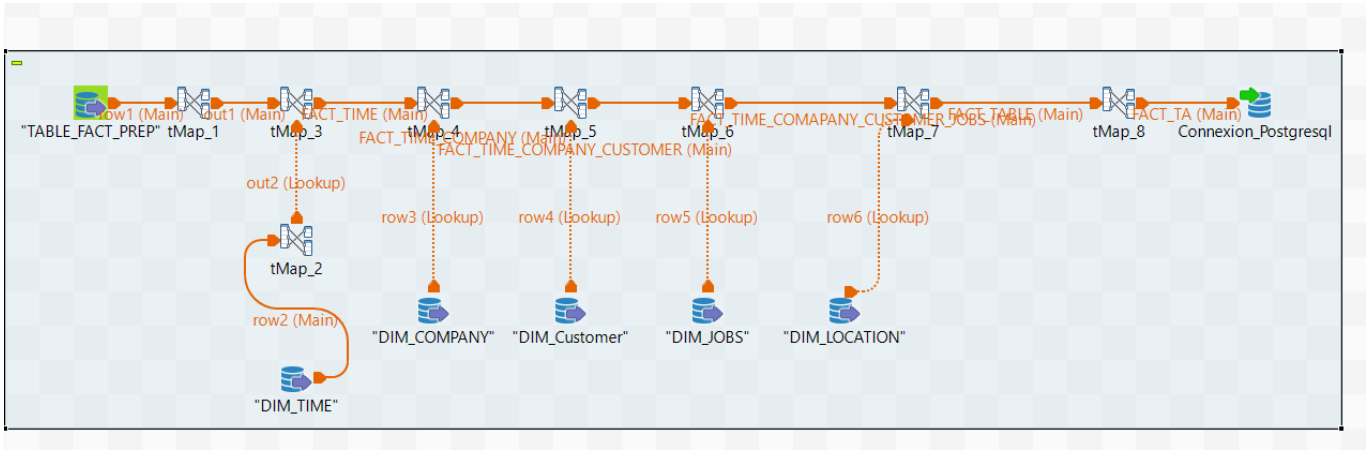
**Dimension Table Creation - dim_company:**
*Data Extraction from Postgres:*
Utilized `TDBInput` to retrieve distinct organization/company names from various columns ('Organization_1', 'Organization_2',..., 'Organization_7') in the 'TABLE_FACT_PREP' table within the PostgreSQL database. A SQL query was used to gather unique organization names across these columns.
*Unique Row Filtering:*
Likely used the `TUniqRow` component to ensure only unique company names were considered for the dimension table.
*Transformation in `TMap`:*
Within the `TMap` component, added operations to assign primary keys or perform any necessary transformations on the retrieved organization names.
*Output to PostgreSQL:*
The refined list of unique organization names was saved into a new table named 'dim_company' within the PostgreSQL database using the `TDBOutput` component.

**Dimension Table Creation - dim_employee:**
*Data Extraction and Transformation:*
Utilized the `TDBInput` component to fetch specific attributes ('Full_name', 'First_name', 'Last_name', 'Birthday') from the 'TABLE_FACT_PREP' in PostgreSQL. Incorporated a SQL query to extract these columns from the source table.
*Unique Rows Handling:*
Employed the `tUniquerow` component on the 'Full_name' column to ensure uniqueness within the employee names.
*Mapping and Assignment in `TMap`:*
Routed the selected attributes through a `TMap` component to potentially add a primary key or perform any necessary transformations.
*Output to PostgreSQL:*
Stored the transformed and enriched data into the 'dim_employee' table within the PostgreSQL database using the `TDBOutput` component.

**Dimension Table Creation - dim_time:**
*Data Extraction from Postgres:*
Utilized `TDBInput` to fetch 'Organization_Start_1' and 'Organization_End_1' values from the 'TABLE_FACT_PREP' table in PostgreSQL. The SQL query retrieved non-empty date entries from the specified columns.

*Unique Dates Identification:*
Employed the `tUniquerow` component to gather distinct start and end dates to ensure uniqueness in the time dimension.
*Transformation in `TMap`:*
Integrated the retrieved dates into a `TMap` component where modifications or augmentations such as formatting or the addition of primary keys could be applied.
*Output to PostgreSQL:*
Stored the processed time-related data into the 'dim_time' table within the PostgreSQL database using the `TDBOutput` component.

## 4.3 Create Fact table Job:

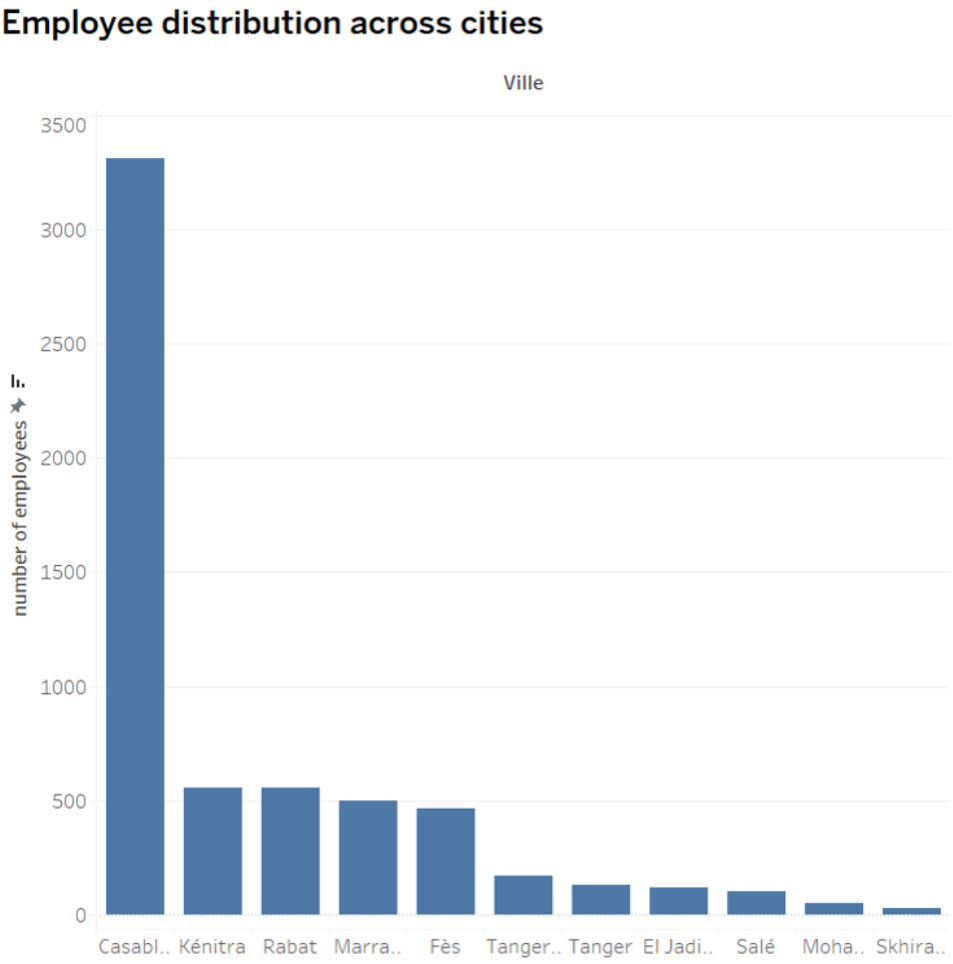



**Fact Table Creation:**
*Data Retrieval from Postgres:*
Employed `TDBInput` for each dimension table (`dim_employee`, `dim_location`, `dim_time`, `dim_job`, `dim_company`) to fetch necessary attributes from the respective tables.
*Joining Dimensions in `TMap`:*
Integrated `TMap` components to perform the necessary joins between the retrieved dimensions using keys such as employee ID, location ID, time ID, job ID, and company ID. Merged the related attributes from each dimension to form the complete job-related fact data.
*Fact Table Creation:*
Constructed the fact table 'fact_jobs' in the `TMap` by combining the relevant attributes from the joined dimensions.
*Storing the Fact Data:*
Utilized `TDBOutput` to save the finalized job-related fact data into the 'fact_jobs' table within the PostgreSQL database.

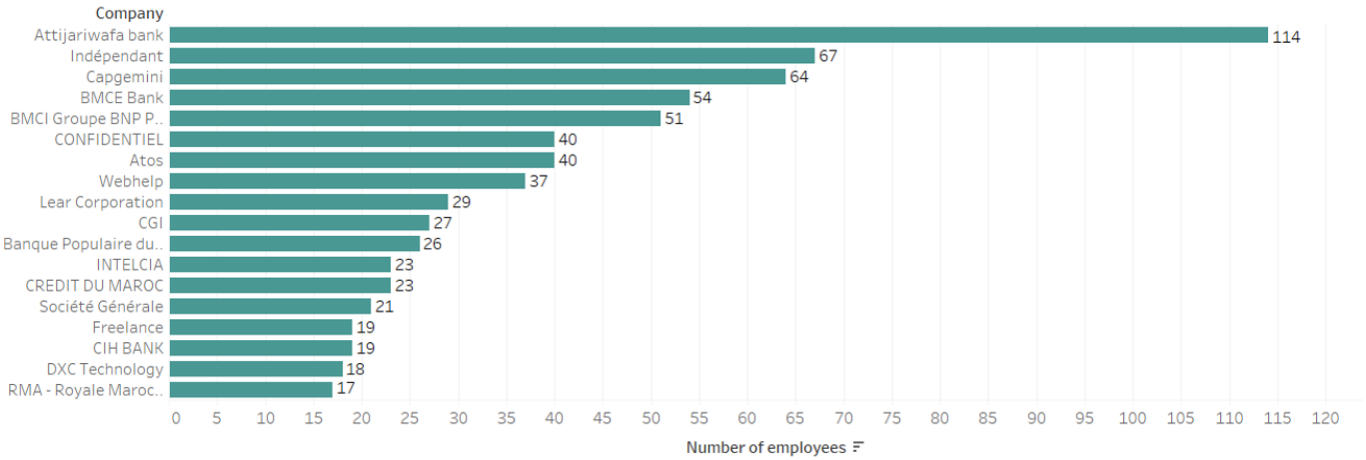
# Chapter 5

# Tableau Visualizations

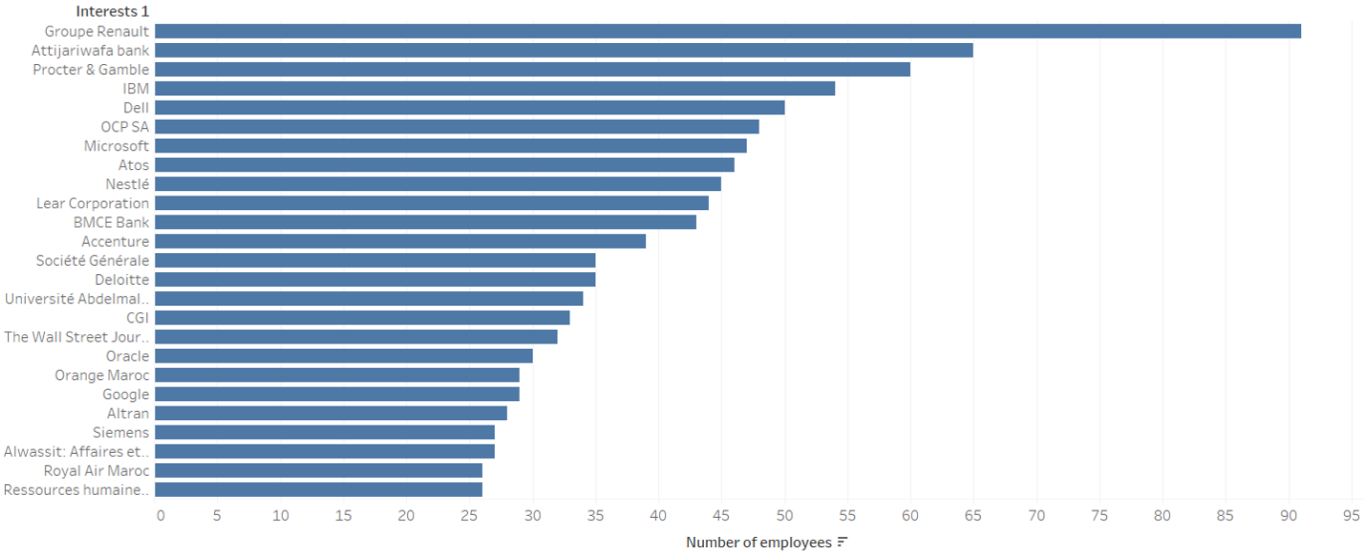
Utilizing Tableau software, we aimed to present insightful visual representations derived from the cleaned and structured data. This phase involved the creation of interactive and informative dashboards and visualizations, enabling stakeholders to grasp critical workforce analytics insights effectively.



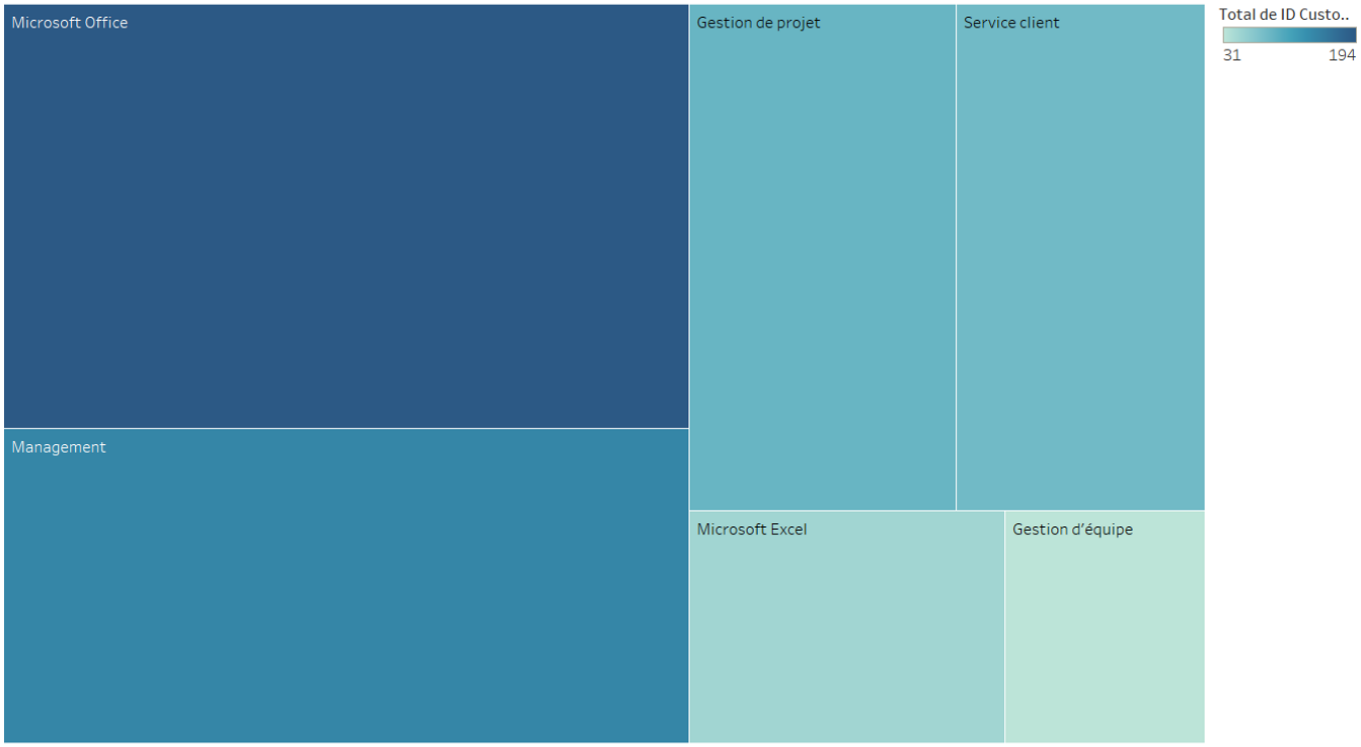
## Employee Distribution Across Top Companies in Morocco



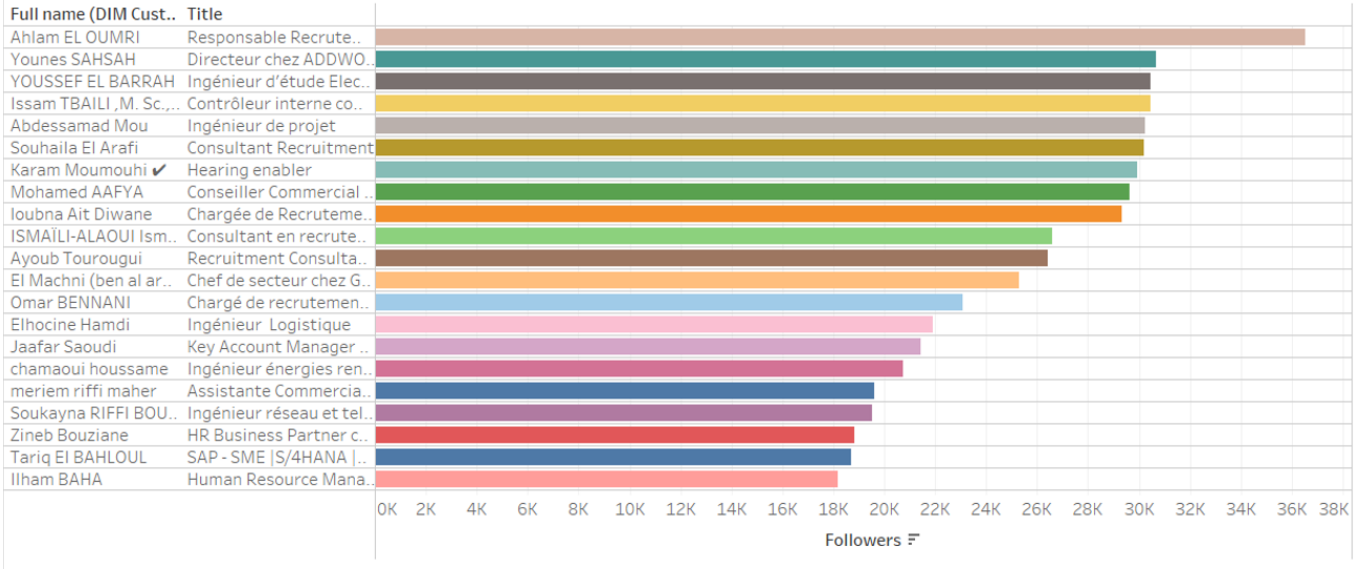

## Employee Interests Landscape



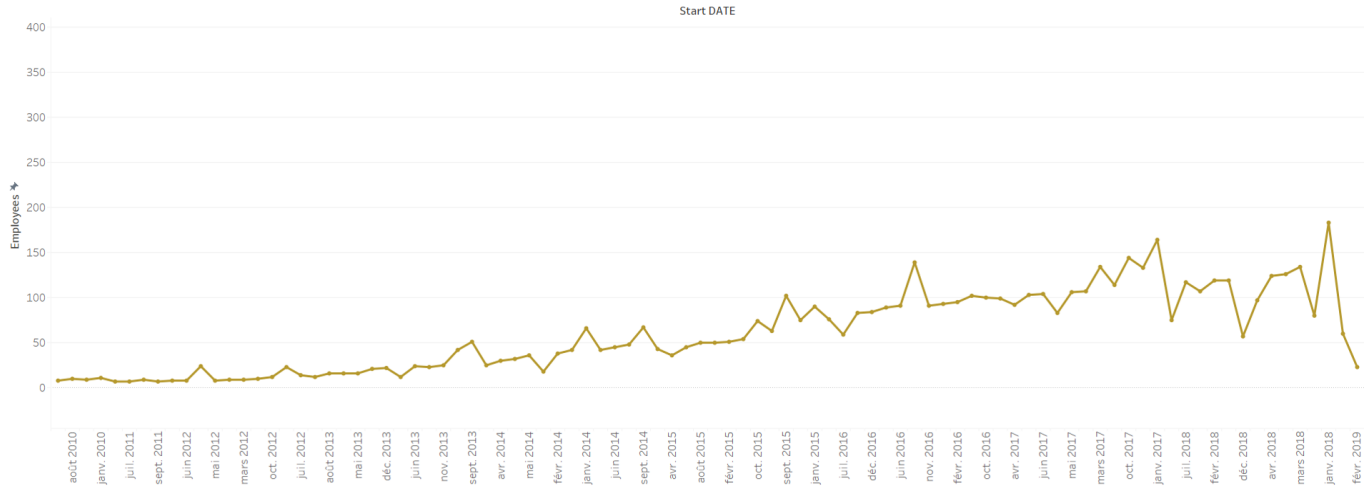
## Skills distribution among employees




## Top LinkedIn Profiles with followers References in Titles



**Annual Employee Job Start Count (2009 - 2019)**



# Conclusion

Our project navigated LinkedIn data through data warehousing, ETL processes, dimensional modeling, and Tableau visualization, tailored for Moroccan organizational contexts. The challenges in data access and structuring highlighted the complexity of real-world data. Cleaning and structuring this information was crucial and showcased the resilience of our methodologies.

Talend's ETL processes—TFilterRow, TJavaRow, and TMap—played a vital role in data integrity and transformation. Dimensional modeling created enriched tables that formed the backbone of our analytics. These dimensions, tied through a fact table, provided comprehensive insights into workforce dynamics.

Tableau visualizations presented these insights intuitively, offering deeper understanding of job trends, geographic distributions, and skill landscapes. These visualizations translated complex data into actionable narratives, empowering informed decision-making.

In conclusion, our project underscores the power of data-driven insights in workforce analytics. From meticulous data prep to insightful visualizations, we've unveiled valuable patterns for informed decision-making in the evolving world of work.