KLE Society’s

KLE Technological University



**Data Integration and Cloud Services Course Project Report**

**On**

**LinkedIn Database**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

***Submitted By***

|  |  |
| --- | --- |
| **NAME** | **USN** |
| Aditi Khyadad | 01FE21BCS072 |
| Vandana | 01FE21BCS125 |
| Sakshi T | 01FE21BCS265 |
| Soumya Neelakari | 01FE21BCS156 |

**Faculty In charges**

**Neha T**

School of Computer Science & Engineering,

Vidyanagar, Hubli – 580031 (India).

Academic year 2023-24

|  |  |
| --- | --- |
| **Table of Content** |  |
| **Chapters** | **Page**  **No** |
| **1. Introduction**  1.1 Preamble  1.2 Problem Definition  1.3Objectives | **3-4**  **3**  **3**  **4** |
| **2. ER diagram** | **4** |
| **3. Data set description** | **5-9** |
| **4. Transformations** | **9-21** |
| **5. Conclusion** | **22** |

**1. Introduction:**

**1.1 Preamble:**

Designing a database for a complex system like LinkedIn requires careful consideration of various entities and their relationships. The goal is to ensure efficient data storage, retrieval, and management while maintaining data integrity and supporting various functionalities like user profiles, connections, posts, comments, likes, shares, groups, and more. Given the diverse set of features that LinkedIn offers, such as professional networking, job searching, skill endorsements, and content sharing, the database design must be robust and scalable to handle large volumes of data and high transaction rates. It should accommodate the dynamic nature of user interactions and relationships, providing seamless integration and data flow across different modules. Moreover, security and privacy are paramount, necessitating strict access controls and data protection measures to safeguard user information. The design should also consider future growth and adaptability, allowing for easy incorporation of new features and functionalities. Effective indexing, query optimization, and efficient ETL (Extract, Transform, Load) processes are essential to enhance performance and provide real-time analytics and insights. Overall, the database design should be user-centric, ensuring a smooth and responsive experience for millions of LinkedIn users worldwide.

**1.2 Problem Definition :**

Design a relational database schema for a LinkedIn-like system. The database should support the following features:

1. **User Management**: Storing user information and their profiles.
2. **Connections**: Managing user connections with status indicators (pending, accepted).
3. **Education and Experience**: Recording users' educational background and work experiences.
4. **Skills**: Tracking skills users possess.
5. **Content Creation**: Enabling users to create posts, comments, likes, and shares.
6. **Groups**: Allowing users to create and join groups with different statuses (pending, accepted, blocked).

**1.3 Objectives:**

1. Utilize Informatica PowerCenter to extract, transform, and load (ETL) detailed user information and profiles.
2. Support profile customization with summaries, headlines, and industry information using Informatica transformations.
3. Implement primary keys, foreign keys, and unique constraints within Informatica to maintain data integrity.
4. Leverage Informatica's optimization techniques to analyze query patterns and enhance database performance.
5. Use Informatica transformations to provide functionality for viewing and managing user connections.
6. Maintain comprehensive records of user professional experiences, including companies, roles, and durations, through Informatica mappings.
7. Enable Informatica workflows to create mechanisms for liking and sharing posts.
8. Facilitate the creation and management of user connections, including connection status and dates, through Informatica workflows.

**2. ER diagram:**

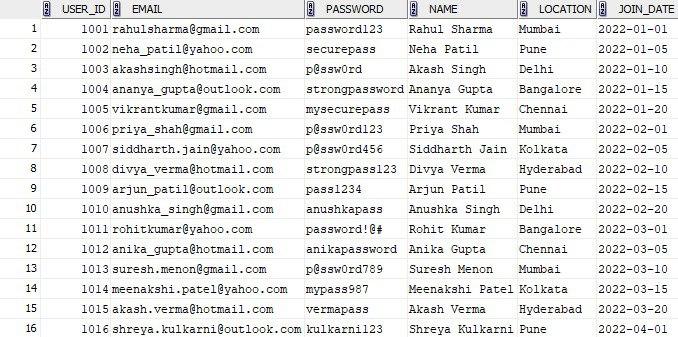
****

**Fig 2.1 ER Diagram**

1. **One-to-one** relationship between Users and Profiles, where one user can have only one profile, but one profile can belong to only one user.
2. **One-to-many** relationship between Users and Connections, where one user can have multiple connections, but one connection can belong to only one user.
3. **One-to-many** relationship between Users and Education, where one user can have multiple education, but one education can belong to only one user.
4. **One-to-many** relationship between Users and Experience, where one user can have multiple experiences, but one experience can belong to only one user.
5. **One-to-many** relationship between Users and Skills, where one user can have multiple skills, but one skill can belong to only one user.
6. **One-to-many** relationship between Users and Posts, where one user can have multiple posts, but one post can belong to only one user.
7. **One-to-many** relationship between Posts and Comments, where one post can have multiple comments, but one comment can belong to only one post.
8. **One-to-many** relationship between Posts and Likes, where one post can have multiple likes, but one like can belong to only one post.
9. **One-to-many** relationship between Posts and Shares, where one post can have multiple shares, but one share can belong to only one post.
10. **One-to-many** relationship between Users and Groups, where one user can have multiple groups, but one group can belong to only one user.
11. **Many-to-many** relationship between Users and Group\_members, where one user can join multiple groups and one group can have multiple members.

**3. Data set description:**

1. **Users:** This table stores basic information about users such as their name, email, password, and registration date.

****

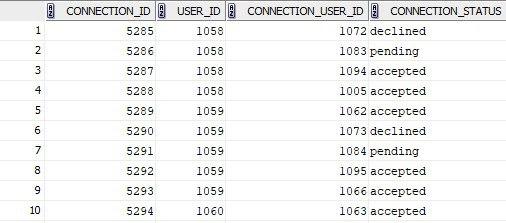
**Fig 3.1 :Users table**

1. **Profiles:** Each user would have a unique profile, which would contain information such as their current job title, industry, location, and summary. The profile table also includes a foreign key linking it to the user who owns it.

****

**Fig.3.2 :Profiles table**

1. **Connections:** This table stores information about the relationships between users, including the user ID of the person making the connection and the user ID of the person they are connecting with. This table also has a status column to indicate whether the connection is pending, accepted, or blocked.

****

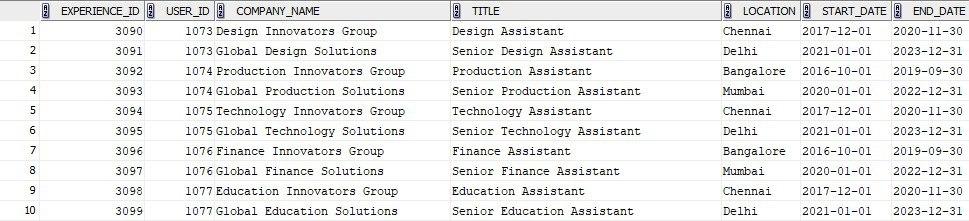
**Fig.3.3:Connection table**

1. **Education:** This table stores information about the user’s education like school name, degree, field of study, and start and end date. This table also includes a foreign key linking it to the user who owns it.

****

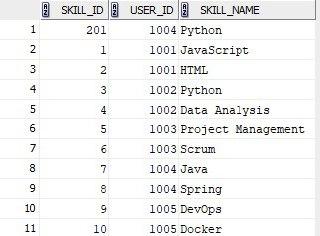
**Fig.3.4 :Education table**

1. **Experience:** This table stores information about the user’s work experience, company, name, job title, location, and start and end date. This table would also include a foreign key linking it to the user who owns it.

****

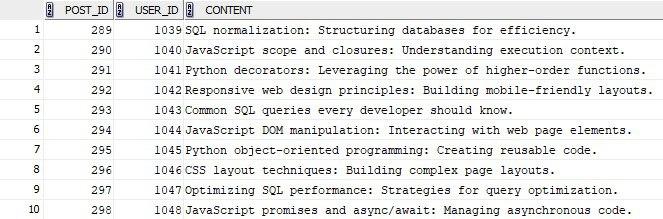
**Fig.3.5:Experience table**

1. **Skills:** This table stores information about the user’s skills. This table would also include a foreign key linking it to the user who owns it.

****

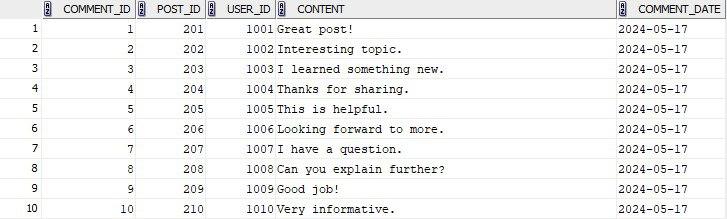
**Fig.3.6: Skills table**

1. **Posts:** This table stores information about posts made by users, including the user ID of the person who made the post, the post’s content, and the date it was posted.

****

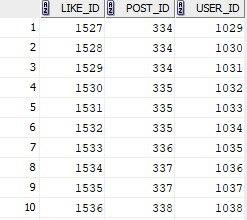
**Fig.3.7:Posts table**

1. **Comments:** This table stores information about comments made on posts, including the user ID of the person who made the comment, the comment’s content, and the date it was posted. It also includes a foreign key linking it to the post it is commenting on.

****

**Fig.3.8: Comments table**

1. **Likes:** This table stores information about likes on posts, including the user ID of the person who made the like and the date it was posted. It also includes a foreign key linking it to the post it is liking on.

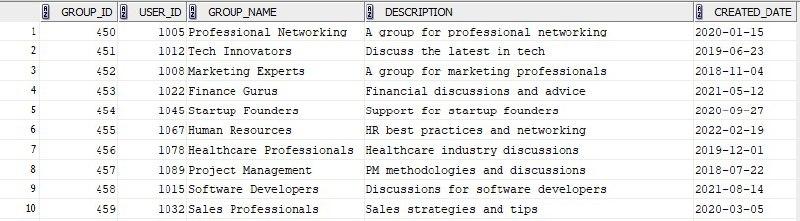
**Fig.3.9: Likes table**

1. **Shares:** This table stores information about shares on posts, including the user ID of the person who made the share and the date it was posted. It also includes a foreign key linking it to the post it is sharing.

****

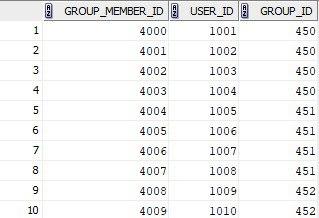
**Fig.3.10: Shares table**

1. **Groups:** This table stores information about groups created by users, including the group name, description, and the user ID of the person who created the group.

****

**Fig.3.11: Groups table**

1. **Group\_members:** This table stores information about the relationship between groups and users, including the user ID of the person and the group ID they are joining. It also has a status column to indicate whether the request is pending, accepted, or blocked

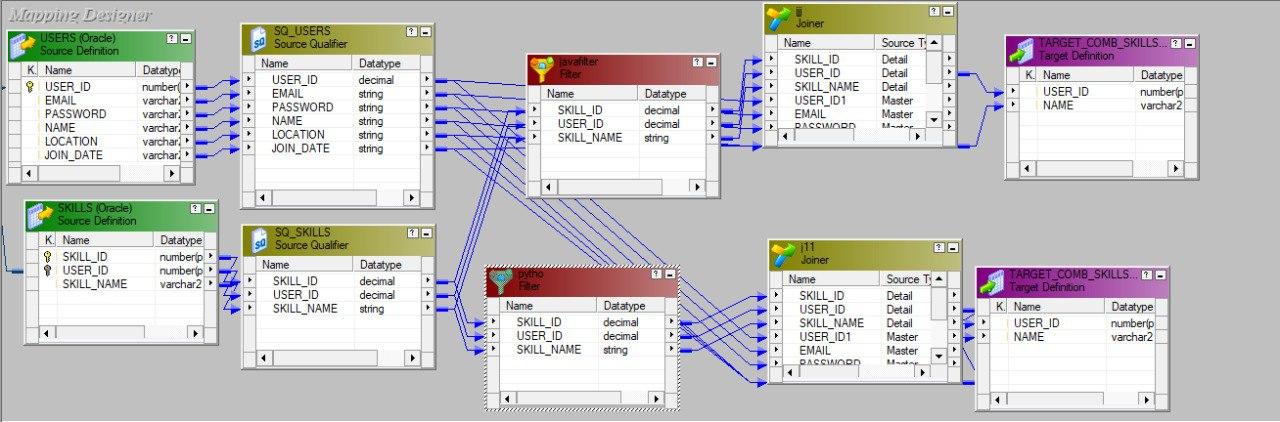


**Fig.3.12: Groups\_members table**

**4. Transformations:**

**4.1 Users Who Have a Combination of Specific Skills**

A project manager is tasked with forming a team of individuals proficient in both Java and Python. To ensure the selected members meet these specific criteria, he must identify users who possess both skill sets.

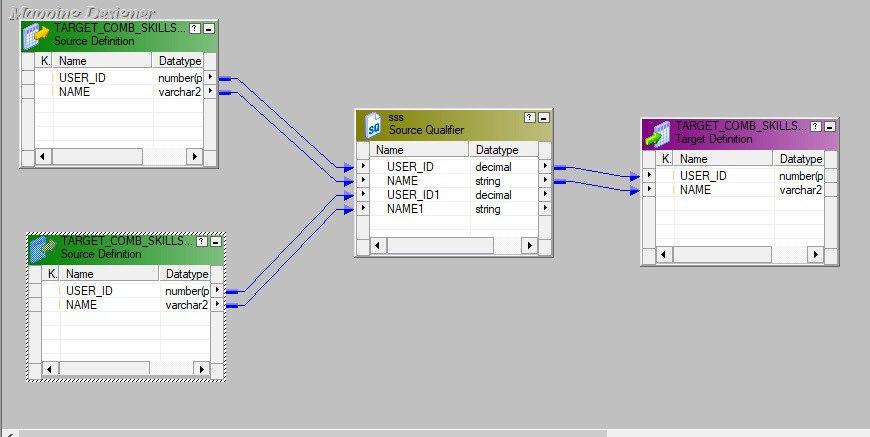
****

**Fig 4.1 Users Who Have a Combination of Specific Skills**

In this transformation, data is extracted from the USERS and SKILLS tables. The Source Qualifier transformations SQ\_USERS and SQ\_SKILLS prepare the data for further processing. The SQ\_USERS source qualifier extracts user information such as USER\_ID, EMAIL, PASSWORD, NAME, LOCATION, and JOIN\_DATE. The SQ\_SKILLS source qualifier extracts skill information including SKILL\_ID, USER\_ID, and SKILL\_NAME.

Two Filter transformations are applied: java filter and python. The java filter transformation filters data to pass through only the rows where the SKILL\_NAME matches "Java". The python transformation filters data to pass through only the rows where the SKILL\_NAME matches "Python".

Next, Joiner transformations (Joiner and J11) are used to join the filtered data from the USERS and SKILLS tables based on the USER\_ID. The Joiner transformation combines data where the SKILL\_NAME is "Java", while the J11 transformation combines data where the SKILL\_NAME is "Python".

****

**Fig 4.2 Users Who Have a Combination of Specific Skills part2**

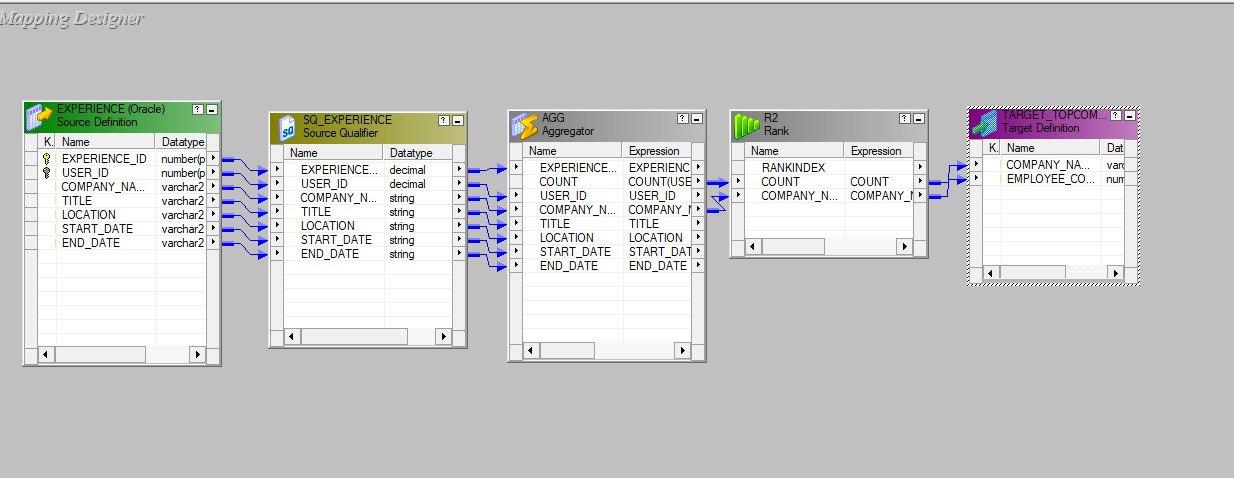
Finally, the relevant data (USER\_ID and NAME) from the joined results is loaded into the TARGET\_COMB\_SKILLS target table, ensuring that only the relevant user information with their respective skills is displayed. This process ensures that the target table TARGET\_COMB\_SKILLS contains a list of users and their skills, filtered for specific skills like Java and Python.

Output:



**Fig 4.3 Users Who Have a Combination of Specific Skills Output**

**4.2 Top Industries by User Count**

By identifying the industries with the highest user count on LinkedIn, the agency can allocate their marketing resources more effectively, focusing on sectors with the largest potential audience.  ****

**Fig 4.4 Top Industries by User Count**

In this transformation, data is extracted from the EXPERIENCE table. The Source Qualifier transformation, SQ\_EXPERIENCE, prepares the data for further processing. The Aggregator transformation, AGG, groups the data by COMPANY\_NAME and counts the number of users (USER\_ID) associated with each company using the expression COUNT(USER\_ID). The Rank transformation, R2, then ranks these companies based on the user count in descending order, utilizing the expression COUNT(USER\_ID) to assign a RANKINDEX to each company. Finally, the relevant data (COMPANY\_NAME and EMPLOYEE\_COUNT) is loaded into the TARGET\_TOPCOMPANIES target table, ensuring that the top companies by user count are identified and displayed.

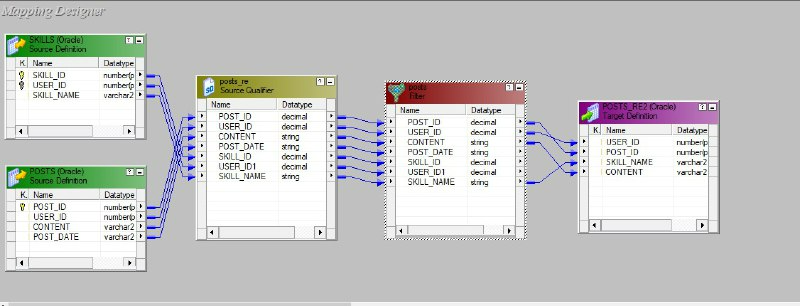
Output:



**Fig 4.5 Top Industries by User Count Output**

**4.3 Recommend posts to users based on their skills.**

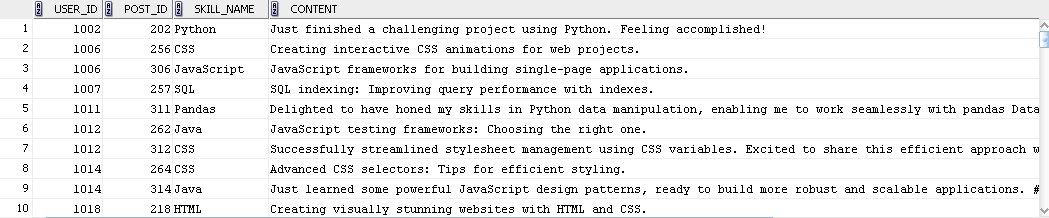
When companies post job openings seeking skills like CSS, Java, and Python, LinkedIn's algorithm prioritizes these posts in the user's feed, ensuring they don't miss relevant career opportunities. Additionally, invitations to webinars, workshops, and conferences related to their skills are highlighted, providing opportunities for networking and staying updated with industry trends.

****

**Fig 4.6 Recommend posts to users based on their skills.**

In this transformation, data is extracted from the SKILLS and POSTS tables. The Source Qualifier transformation joins these tables based on USER\_ID and prepares the combined data for further processing. The Filter transformation applies a condition using the expression IIF(INSTR(POSTS.CONTENT, SKILLS.SKILL\_NAME) > 0, TRUE, FALSE) to ensure that only posts containing the mentioned skills, such as CSS, are passed through. If a post's content includes "CSS" and the user has CSS listed as a skill, the relevant data (USER\_ID, SKILL\_NAME, SKILL\_ID, and CONTENT) is then loaded into the POSTS\_RES target table, ensuring that only relevant posts are displayed for each user's skills.

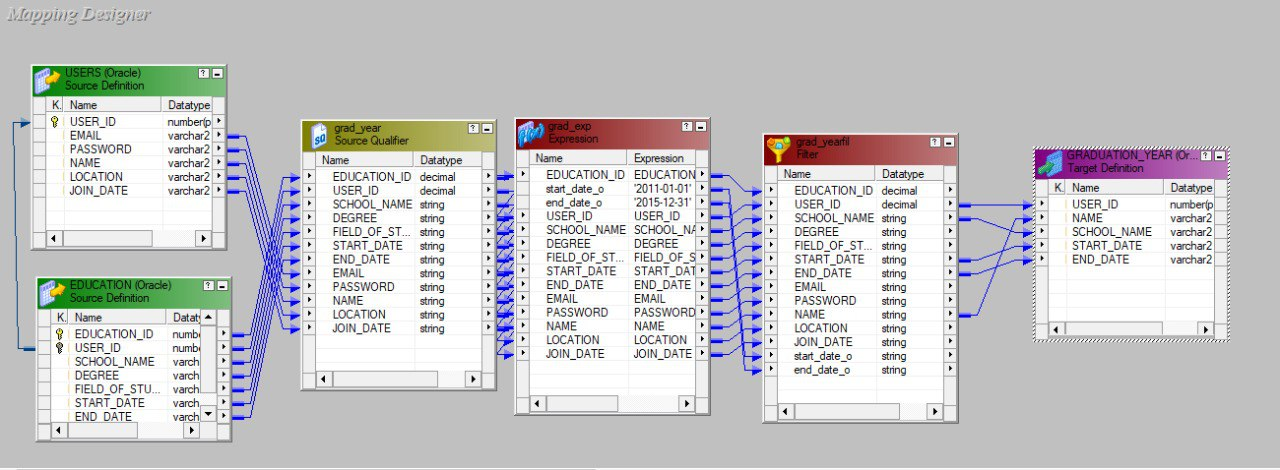
**Output:**

****

**Fig 4.7 Recommend posts to users based on their skills output**

**4.4 Users Who Graduated Within a Specific Year Range**

The organization uses the graduation year filter to create a targeted list of potential mentors. By identifying alumni who graduated between 2011 and 2015, they can focus their recruitment efforts on individuals who are likely to have relevant recent experiences and the willingness to give back to the community. They aim to leverage LinkedIn’s recommendation system to identify and reach out to these individuals.



**Fig 4.8 Users Who Graduated Within a Specific Year Range**

This data flow transformation extracts data from the USERS and EDUCATION tables. The Source Qualifier transformation joins these tables based on USER\_ID to combine user and education details. An Expression transformation calculates and filters the graduation year, with an expression such as end\_date\_o >= '2011-01-01' AND end\_date\_o <= '2015-12-31', ensuring that only users who graduated between 2011 and 2015 are included. A Filter transformation then applies this condition to pass through only the relevant records. Finally, the filtered data is loaded into the target table, containing the USER\_ID, NAME, SCHOOL\_NAME, START\_DATE, and END\_DATE.

**Output**:

****

**Fig 4.9 Users Who Graduated Within a Specific Year Range Output**

**4.5 Users Who Have Experience in Two or More Different Companies**

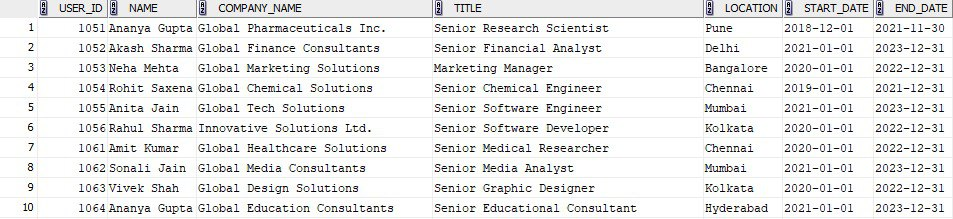
A tech company is looking to hire a senior project manager with a diverse background in managing projects across different industries. The company wants to leverage LinkedIn's recommendation system to identify candidates who have worked in two or more different companies, as this experience can bring a wealth of knowledge and adaptability to their team.

****

**Fig 4.10 Users Who Have Experience in Two or More Different Companies**

This data flow transformation extracts data from the USERS and EXPERIENCE tables. The Source Qualifier transformation joins these tables based on USER\_ID to combine user and work experience details. An Aggregator transformation calculates the total number of companies each user has worked at by counting distinct COMPANY\_NAME values. A Filter transformation applies a condition to select only users who have worked at three or more companies, with an expression such as company\_count >= 2. Finally, the filtered data, is loaded into the target table.

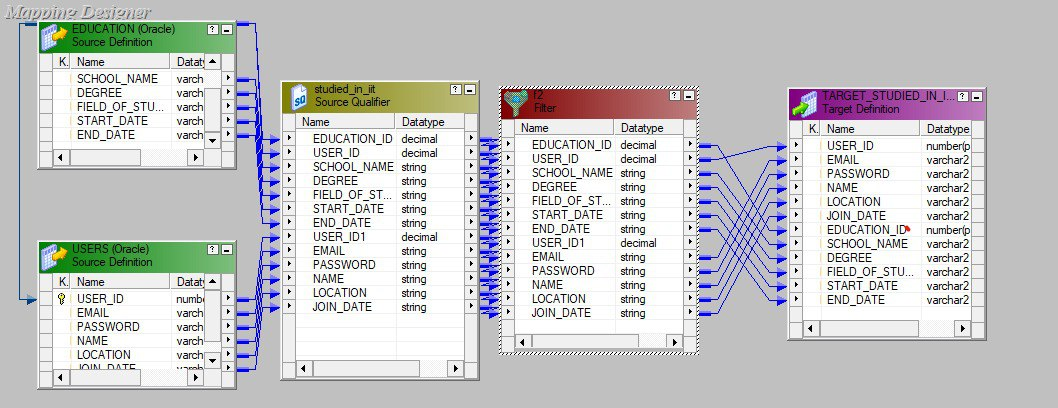
**Output:**

****

**Fig 4.11 Users Who Have Experience in Two or More Different Companies Output**

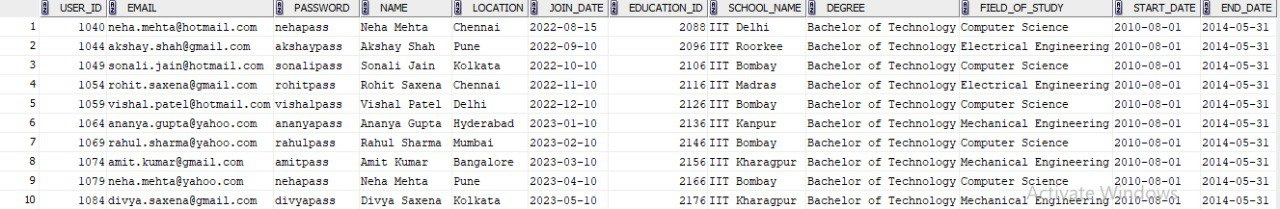
**4.6 Users Studied in IIT or NIT**

Graduates from IIT and NIT are sought after for their robust technical skills and innovative capabilities, grounded in rigorous educational training. Focusing recruitment on these candidates enables the startup to cultivate a team poised to drive substantial contributions to technical projects and foster a culture of innovation within the company.

 **Fig 4.12 Users Studied in IIT or NIT**

The Filter transformation in Informatica enables the startup to specifically identify and recruit users who have studied at prestigious institutions like IIT and NIT. By configuring the Filter transformation, the startup can efficiently narrow down its candidate pool to those with relevant educational backgrounds, ensuring that the team is composed of individuals who bring strong technical prowess and innovative thinking to the table. This strategic approach not only enhances the startup's technical capabilities but also positions it to tackle complex challenges with confidence and competence.

**Output:**

****

**Fig 4.13 Users Studied in IIT or NIT Output**

**4.7 Evaluating Candidates Based on Average Tenure at Companies**

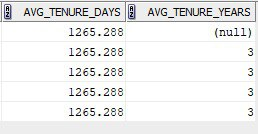
A multinational corporation seeks a senior manager prioritizing stability and loyalty in leadership. Using LinkedIn's average tenure data, they assess candidates for long-term commitment, aiming to ensure continuity in managing corporate initiatives and projects. This strategy enhances strategic advantages within their leadership team.



**Fig 4.14 Evaluating Candidates Based on Average Tenure at Companies**

The Source Qualifier, Expression, and Aggregator transformations in Informatica play crucial roles in this evaluation process. The Source Qualifier transformation extracts relevant data from LinkedIn, while the Expression transformation computes the average tenure at companies for each candidate based on their employment history. Subsequently, the Aggregator transformation aggregates this data to provide insights into candidates' career stability metrics. By leveraging these transformations, the multinational corporation can make informed decisions, selecting senior managers who not only possess requisite skills but also demonstrate a history of commitment and longevity in their professional careers.

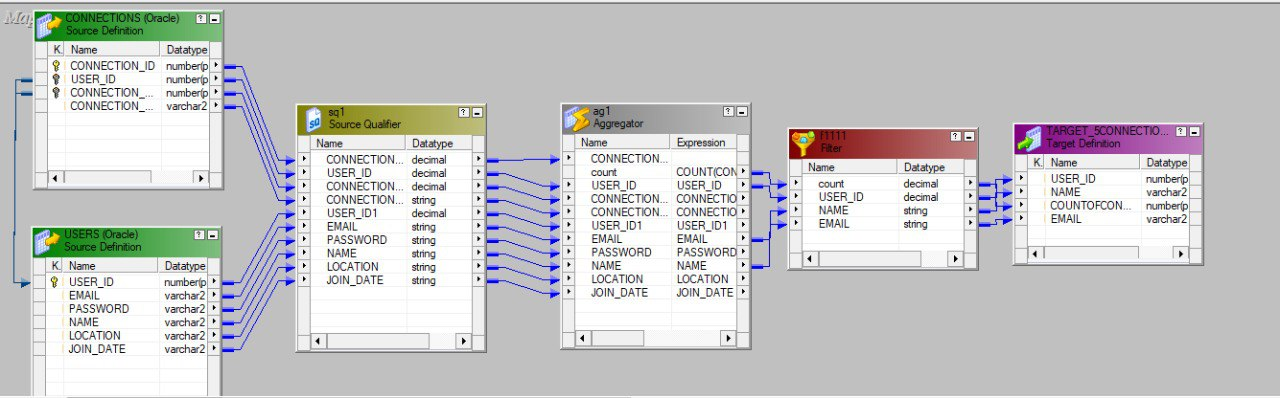
**Output:**

****

**Fig 4.15 Evaluating Candidates Based on Average Tenure at Companies Output**

**4.8 Find Users Who Have More Than 2 Connections**

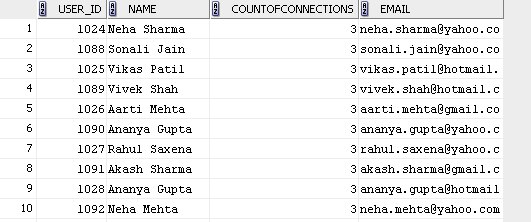
Finding users with more than 2 connections helps LinkedIn recommend active users for better networking and show their content more widely. It also helps LinkedIn offer special features and ads to these users, increasing engagement and revenue.

****

**Fig 4.16 Find Users Who Have More Than 2 Connections**

Extracting essential data from the USERS and CONNECTIONS tables. This includes USER\_ID from USERS and CONNECTION\_ID, SOURCE\_USER\_ID, and TARGET\_USER\_ID from CONNECTIONS to establish connections between users.After extraction, the data is aggregated by USER\_ID to count the number of connections each user has, providing an overview of their network size. A filter is then applied to isolate users who have more than 2 connections, identifying active users likely to engage extensively on the platform. The final results, comprising USER\_ID and their connection counts, are stored in the ACTIVE\_USERS table. This dataset serves as a basis for LinkedIn to enhance user engagement through targeted recommendations, content promotion, and advertising strategies, thereby boosting overall platform activity and revenue generation.

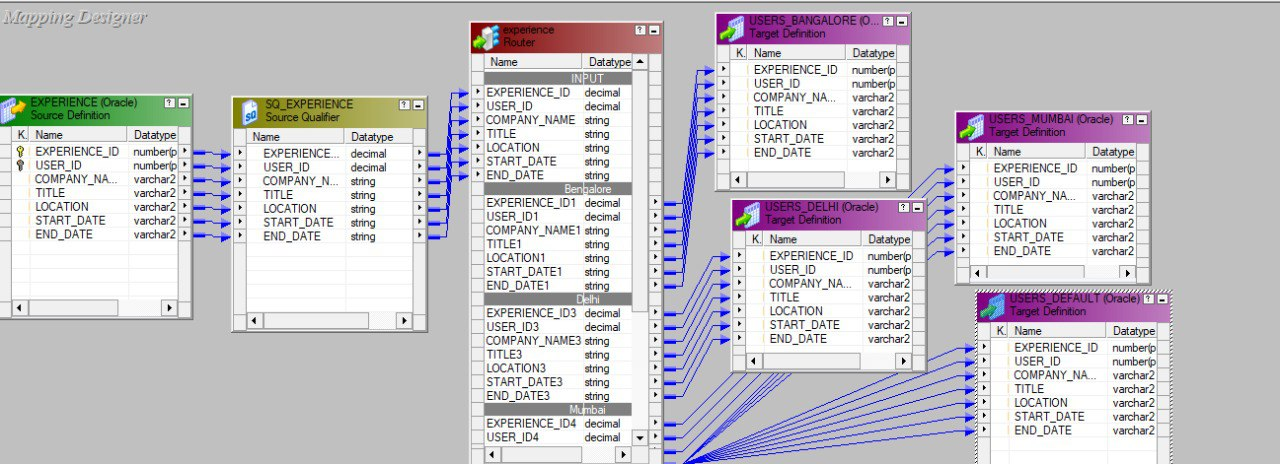
**Output:**



**Fig 4.17 Find Users Who Have More Than 2 Connections Output**

**4.9 Group User Experiences by Specific Locations Using Router Transformation**

This is useful for LinkedIn-like applications to manage user experiences based on their locations.

****

**Fig 4.18 Group User Experiences by Specific Locations Using Router Transformation**

The provided image illustrates an ETL process where data from the EXPERIENCE table is extracted, transformed, and loaded into multiple target tables based on user location. The process begins with extracting fields such as EXPERIENCE\_ID, USER\_ID, COMPANY\_NAME, TITLE, LOCATION, START\_DATE, and END\_DATE from the source table. The data is then routed through a Source Qualifier and a Router Transformation to segregate it into groups according to location: Bangalore, Delhi, Mumbai, and other locations. Each group is subsequently loaded into respective target tables (USERS\_BANGALORE, USERS\_DELHI, USERS\_MUMBAI, and USERS\_DEFAULT). This segregation helps in managing data efficiently and enables location-specific analytics and user engagement strategies. Ultimately, this structured approach aids in delivering targeted content and advertisements, thereby enhancing platform functionality and potentially increasing revenue.

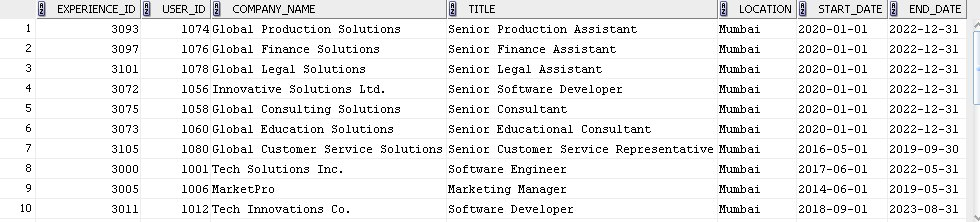
**Output:**

who are in Bangalore:



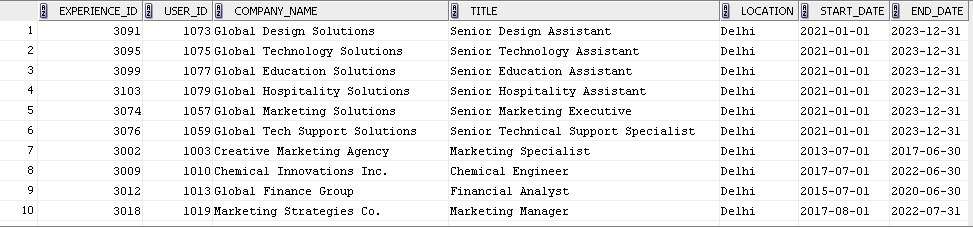
**Fig 4.19 Group User Experiences by Specific Locations Using Router Transformation Ouput1**

who are in Mumbai :



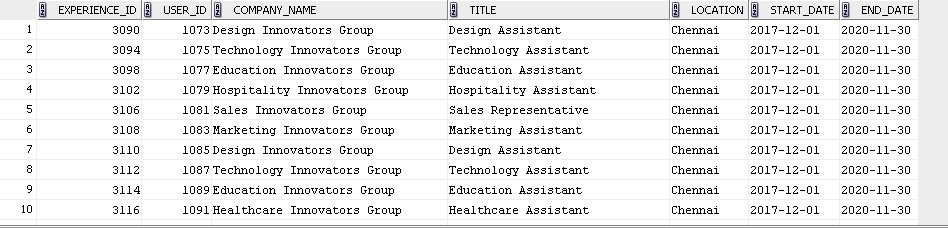
**Fig 4.20 Group User Experiences by Specific Locations Using Router Transformation Ouput2**

who are in Delhi:



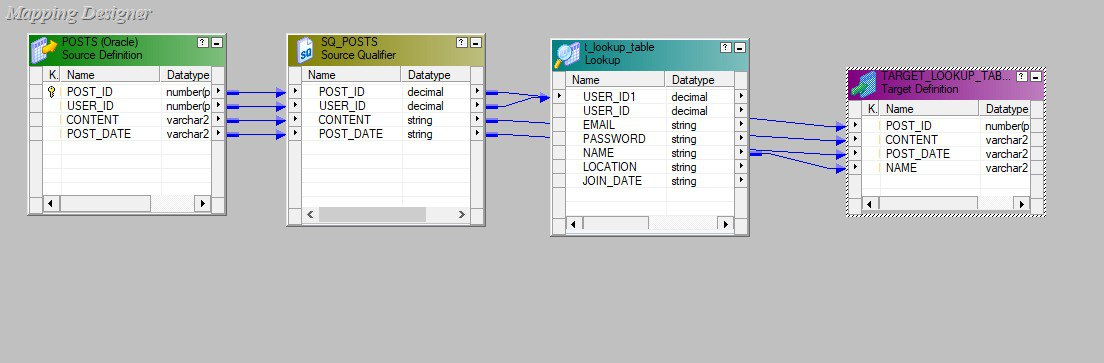
**Fig 4.21 Group User Experiences by Specific Locations Using Router Transformation Output3**

who are in default:



**Fig 4.22 Group User Experiences by Specific Locations Using Router Transformation Output4**

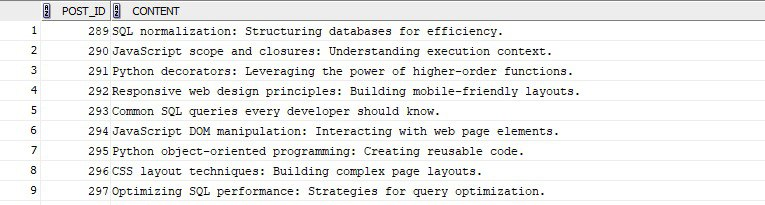
**4.10 Retrieve User Names for Posts Using Lookup Table**

To identify users who have posted content, the startup will need to match user IDs from the Posts table with those in the Users table. This process allows for a clear understanding of user engagement and activity. By linking user IDs to their corresponding usernames, the startup can create a target table that provides comprehensive insights into who is posting content and how active they are on the platform. 

**Fig 4.23 Retrieve User Names for Posts Using Lookup Table**

The above figure demonstrates how the Lookup transformation in Informatica facilitates the task of matching user IDs from the Posts table with usernames from the Users table. By configuring the Lookup transformation, the startup can efficiently link each user ID in the Posts table to its corresponding username in the Users table. This integration enables the creation of a comprehensive target table that not only shows which users are posting content but also provides insights into their activity levels and engagement on the platform.

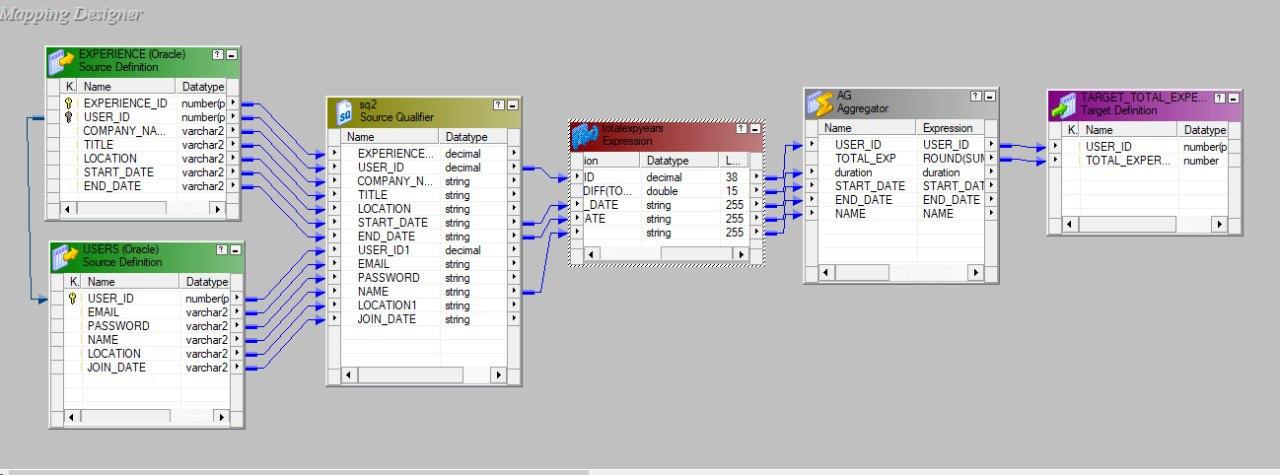
**Output:**



**Fig 4.24 Retrieve User Names for Posts Using Lookup Table Output**

**4.11** **Total experience of a user in all the companies.**

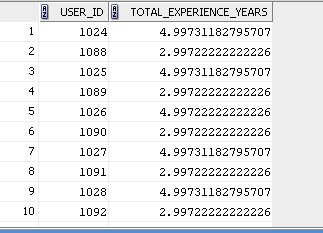
The consulting firm posts a job listing for a senior consultant, specifying the requirement of no of years of total experience. They use LinkedIn’s advanced targeting options to ensure the listing reaches the right audience



**Fig 4.25 Total experience of a user in all the companies.**

These tables are combined in a Source Qualifier transformation (SQ2), which joins the data on USER\_ID and selects the necessary fields. An Expression transformation (totalexpyears) is then used to calculate the duration of each work experience by computing the difference between START\_DATE and END\_DATE using condition DATEDIFF('yyyy', TO\_DATE(END\_DATE, 'yyyy-mm-dd'), TO\_DATE(START\_DATE, 'yyyy-mm-dd')).This calculated duration is passed to an Aggregator transformation (AG), which aggregates the data by USER\_ID to sum the total experience duration for each user, rounded to the nearest whole number. The final result, comprising USER\_ID and TOTAL\_EXPERIENCE, is loaded into the target table TARGET\_TOTAL\_EXPERIENCE.

**Output:**



**Fig 4.26 Total experience of a user in all the company’s output**

**5. Conclusion**

The insights derived from our LinkedIn-like system data provide invaluable guidance for strategic decision-making and operational improvements. By analyzing user interactions and behaviors within the platform, we gain a deeper understanding of user preferences and engagement patterns. This knowledge allows us to optimize features and functionalities to better meet user needs, thereby enhancing overall user satisfaction and retention.

Additionally, leveraging data analytics enables us to identify trends and emerging patterns in user activity. This proactive approach not only helps in anticipating user demands but also facilitates timely adjustments and innovations in our platform's offerings. Understanding these insights empowers us to stay agile in a competitive market, continually evolving our services to remain relevant and compelling to users.

Moreover, business intelligence insights inform our content strategy, guiding us in delivering relevant and engaging content that resonates with our audience. By tailoring our content to user interests and preferences, we foster deeper engagement and encourage active participation within our community. In essence, harnessing the power of data-driven insights enables us to optimize operations, drive growth, and maintain a strong competitive edge in the professional networking landscape.