Refugee Management System

Team Members: Karthikeyan Thorali, Pavan Soma, Srinivas Pusapati

Abstract:

Our project mainly concentrates on merging the different department databases to a **data warehouse**. We have three databases namely refugee(all refugee details),Job_consultancy(all job consultancy details) and funding (details of sponsors) databases. We choose Star schema to built a Data Warehouse based on the data and Analysis Requirements. Then we came up with the **Talend** tool for ETL process. In this process we get the data from the source databases and after performing required transformation data was loaded into the target database(Data Warehouse). After loading the data into data warehouse , based on the problem statements we performed analysis on the data and we used **Tableau** to visualize the analyzed data.

Introduction:

Relationship of records across databases has become an important tool for combining records that belong to the same entity across different data sources. When unique identifiers are not available or deterministic record linkage is not possible, probabilistic record linkage may be used to create additional matches based on probability scores that pair records belonging to the same individual. The success of record linkages is dependent on the quality of the individual data sources and identifiers as well as the accuracy of the record linkage process, which often involves manual review. The goal is therefore to reduce the number of mismatches and unlinked records and in turn to reduce the potential for systematic biases.

In this project we are dealing with three databases namely

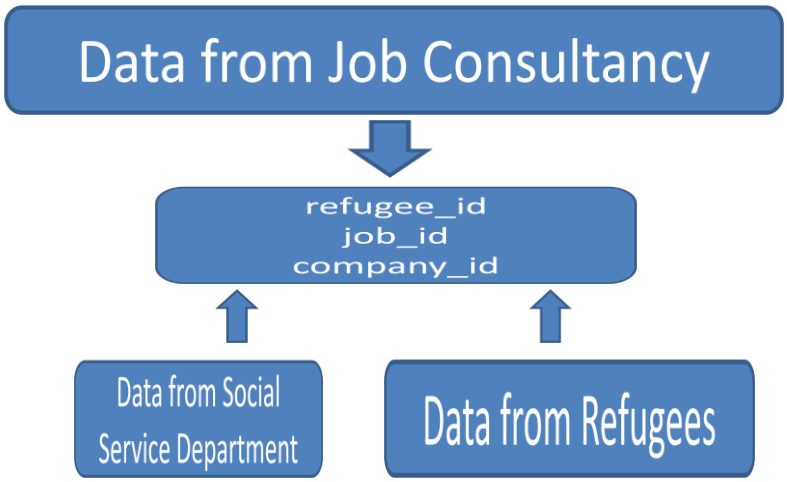
i)Refugee Database: This Database consist of refugee details like refugee personal details, educational background, locality and the physical status of refugee.

ii) Job_Consultancy Database: This database consist of all consultancy details like address, type of consultancy, type of jobs provided by the consultancies and the requirements of the jobs.

iii) Funding Database: This database consist of sponsor personal details, sponsors address, the amount they are willing to fund for refugee's etc.

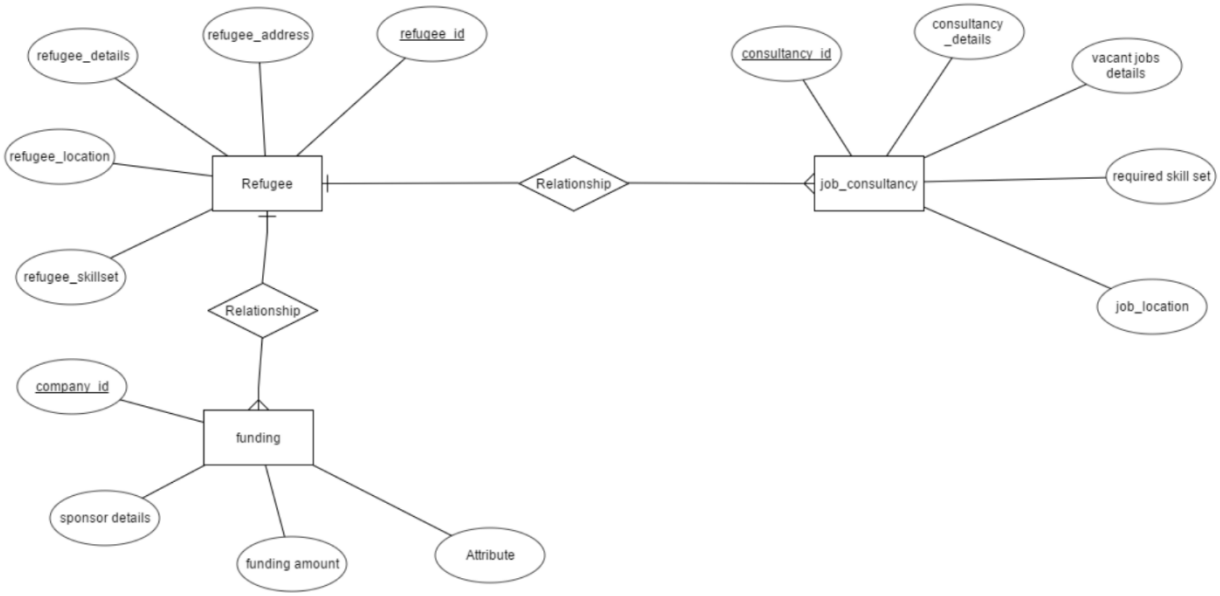
Data Flow :

The primary key refugee_id is the common field in three databases to keep track of the relation among records among the tables. A Refugee has one to many relationship with the job consultancy table which indicates that a refugee can be put in one or more jobs. This makes sure that the refugee is guaranteed to be using his time and saving more money. Thus, it also helps in funding him less since the refugee keeps him paid with his work.



Technical Description and Experiment:

Entity Relationship Diagram:



In this entity refugee has different attributes and it has one to many relationship with the job_consultancy entity and funding entity. So once after getting the refugee data the data was sent to both job consultancy entity and funding entity. Based on the refugee details the job and funding was given to refugee.

Data Warehouse Schema

We used **star schema** in building the data warehouse. Below are the reasons for choosing star schema. The main advantages of star schema are

Query Performance

Queries run faster against a star schema database than an OLTP system because the star schema has smaller number tables and plain join paths. In this, dimensions tables are linked through the central fact table. Dimensions are linked with each other through one join path intersecting the fact table.

Efficient Navigation through Data

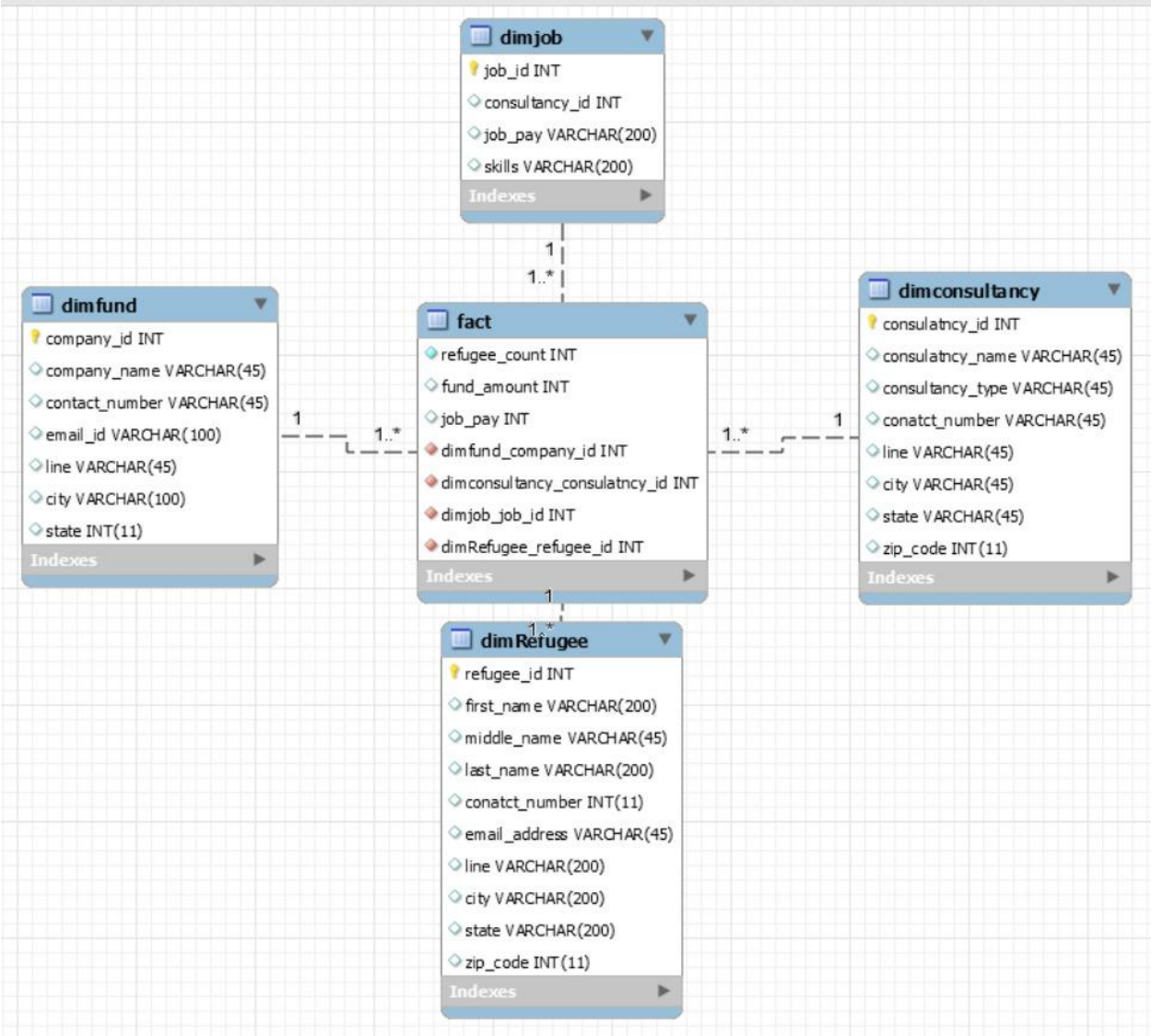
In this dimensions tables are joined through fact tables. These joins are important because they stand for elementary relationships of real business processes. You can surf a single dimension table in order to select attribute values to create an efficient query.

Built-in Referential Integrity

A star schema is planned to enforce referential integrity of loaded data. Referential integrity is imposed by the use of primary and foreign keys. Primary keys in dimension tables become foreign keys in fact tables to link each record across dimension and fact tables.

Considering the query performance and efficiency in navigating through data, Star schema has been used to build the Data Warehousing. In our Data Warehouse, we have a one fact table and four dimension tables. Each of these four dimension tables are (dimconsultancy, dimfund, dimjob, dimrefugee) are connected to the single fact table. Our fact table consist of all the count values in our databases and the foreign keys.

In the data warehouse each table is related to particular entity. Let's consider the below star schema of Data warehouse in which it has four dimension tables and one fact table. The refugee dimension table 'dimrefugee' contains all details of refugee, funding dimension table 'dimfund' contains all details of sponsors, 'dimconsultancy' table consist of all details job consultancies and the 'dimjob' dimension table consist of all details regarding the jobs provided by consultancies. All this dimension tables are connected to a single fact table through one to many relationship as shown in the above figure. Now using the UML diagram data warehouse was created.



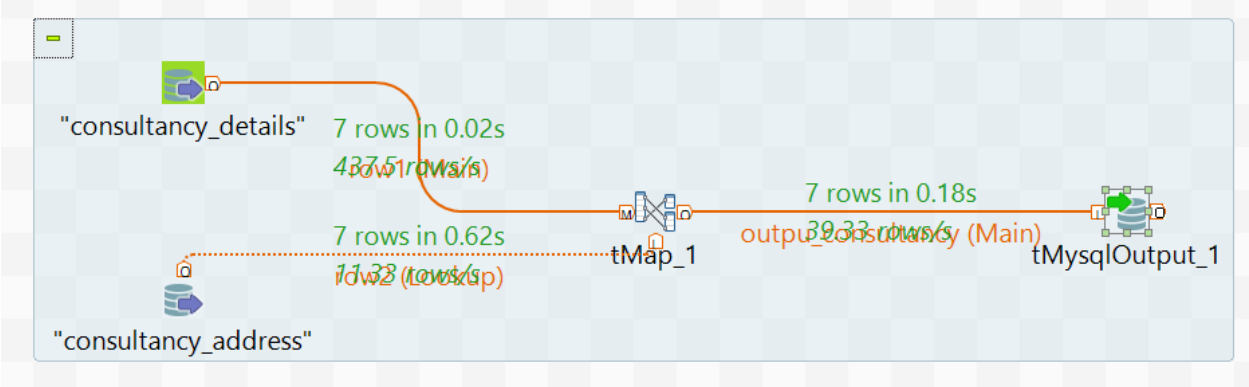
UML Diagram for Data Warehouse

After creating the data warehouse ETL comes into the picture. The ETL process is helpful in extracting the data from source databases and perform the required transformations on the data according to the requirements and loading the data into the target database. For the ETL process we have TALEND a data integration tool. In Talend we will design the jobs for ETL process. In jobs we used MySQL input and ouput connector's for extracting and loading of the data. In the staging process we 'tMap' to perform required transformation on the data. This is

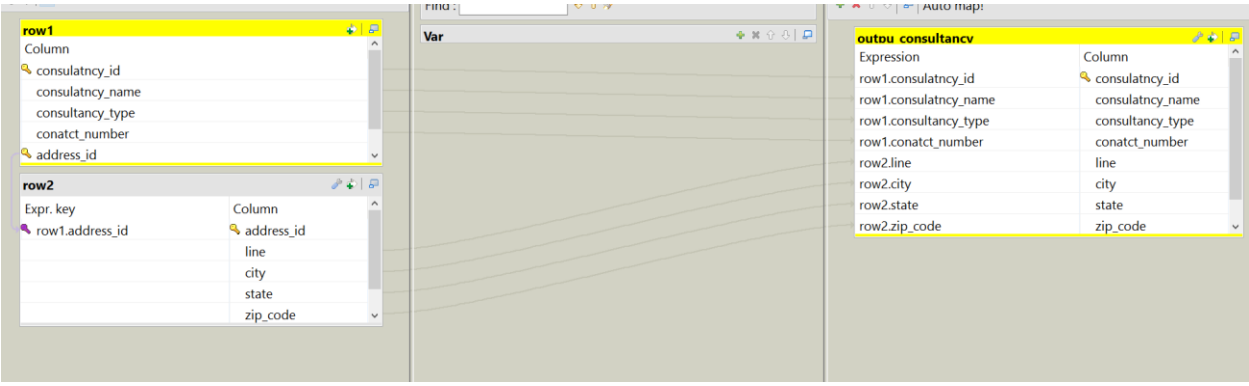
how we performed ETL process. After the completion of job design we execute the job to complete ETL process.

Let's have a one example on ETL process.

Job for loading data into 'dimconsultancy' database:

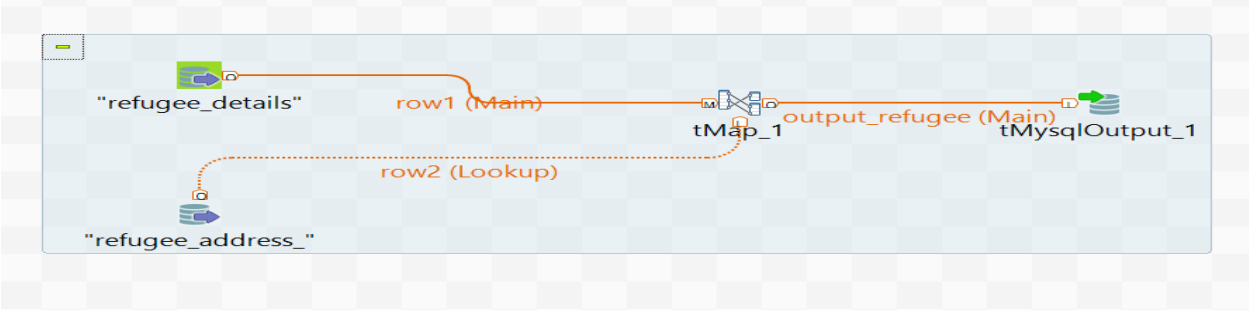


Mapping inside the 'tMap'

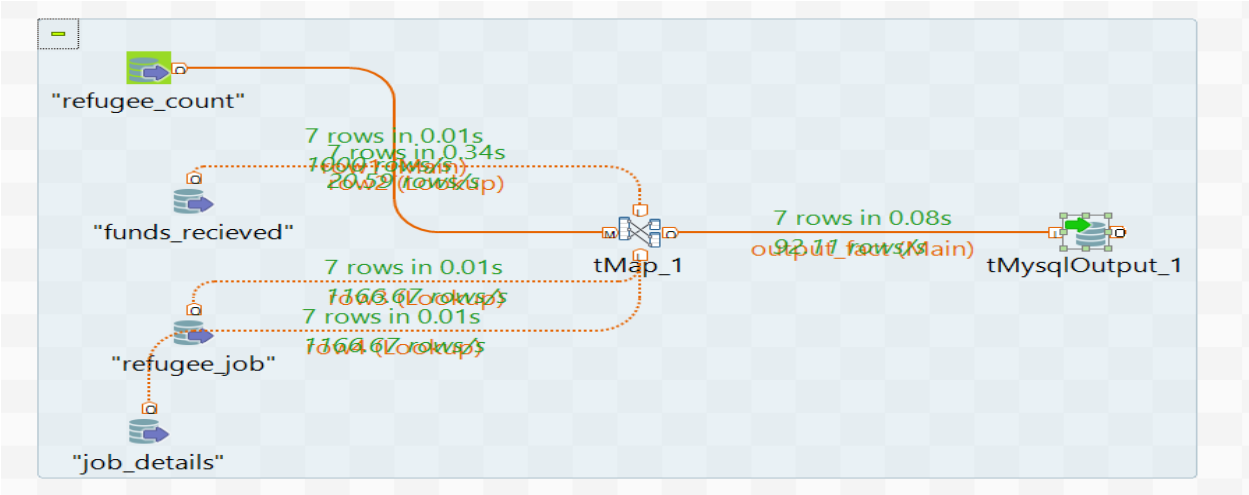


In the above figure the data from the consultancy_details and consultancy_address extracted and using tmap the two table are joined and the required columns are selected. The required columns from the two tables are loaded into the target database that is dimconsulatncy database as shown in the above mapping diagram. similarly we have designed five jobs for loading the data into our four dimesion and one fact table. Below are images shows the all jobs

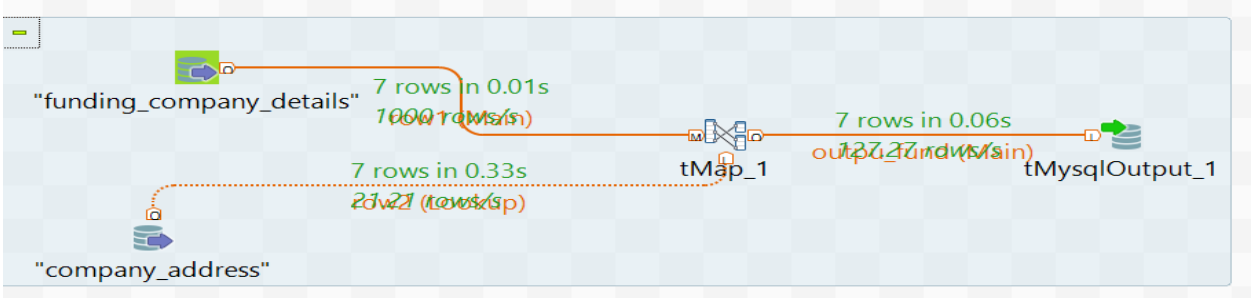
Job for loading data into 'dimrefugee' table in data warehouse:



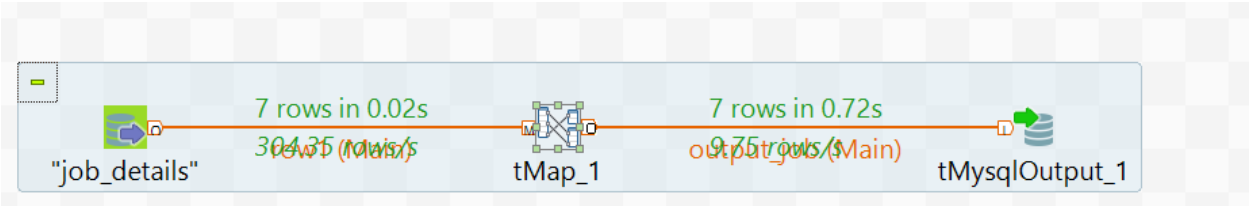
Job for loading data into 'fact' table in data warehouse:



Job for loading data into 'dimfund' table in data warehouse:



Job for loading data into 'dimjob' table in data warehouse:



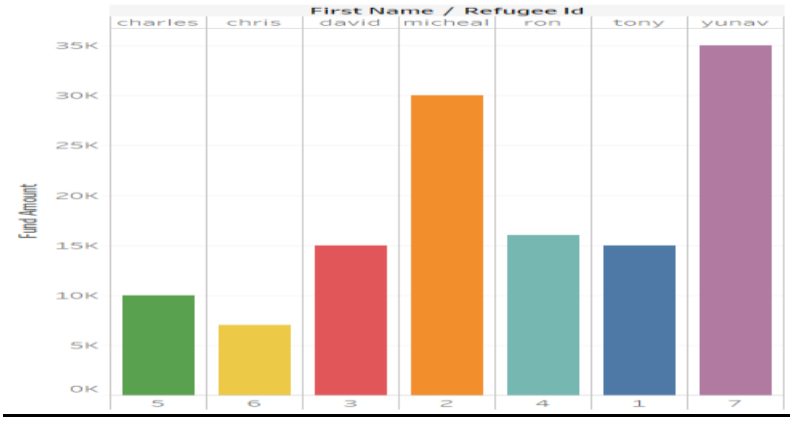
Now we are ready with the data warehouse. we got the required data into all the tables. So based on the problem statements we can perform the analysis. Below are the few problem statements.

1.Who received the maximum funding from the sponsors?
2. On Which skill sets job agents are providing jobs more refugees?
3. Which refugees got the highest pay?

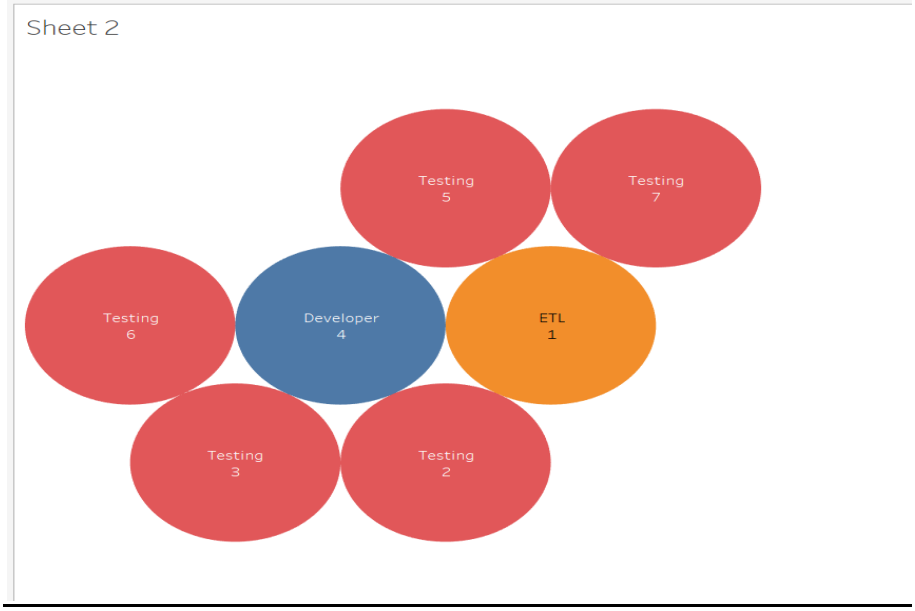
Based on the above problem statements we have performed analysis on data warehouse and given output to **Tableau** to visualize the data. Apart from this we can also gain many other insights from the data warehouse.

Results

The below figure gives us answer for the first question. It tells us that Refugee named yuvan got highest funding.



The below figure gives us answer for the second question, It tells us consultancies are providing more jobs on Testing.



The below figure gives us answer for the third question, It tells us which refugee got highest after getting a job through consultancy.

Conclusion:

- In this project, we presented the design and implementation of a data warehouse for Refugee Admissions System.
- It is the first step of a long journey towards a comprehensive data warehouse solution.
- A data warehouse is not a system that is designed once and installed; it needs to be maintained and developed according to the needs..
- And the results says that there are more jobs on testing so it's better to train the refugee's on testing so that we can provide a job to most of the refugee's. It also tells that which refugee got highest funding so that we can have a look over his background and try to develop his living standards. With the above example as a reference we can treat the refugee's coming the future.
- This helps in allocating jobs and providing funds to the refugee's based on their skill set and family background(physical status and number of people in the family).

Challenges faced during the project:

- In the design process of data warehouse we found difficulty in establishing the relationship between tables in data warehouse.
- Faced many issues during the ETL process but with the help of Talend documents we overcome the ETL issues.