Project Report on the steps taken during the implementation of the Data Warehousing Project.

I. Operational data sources inspection and profiling.

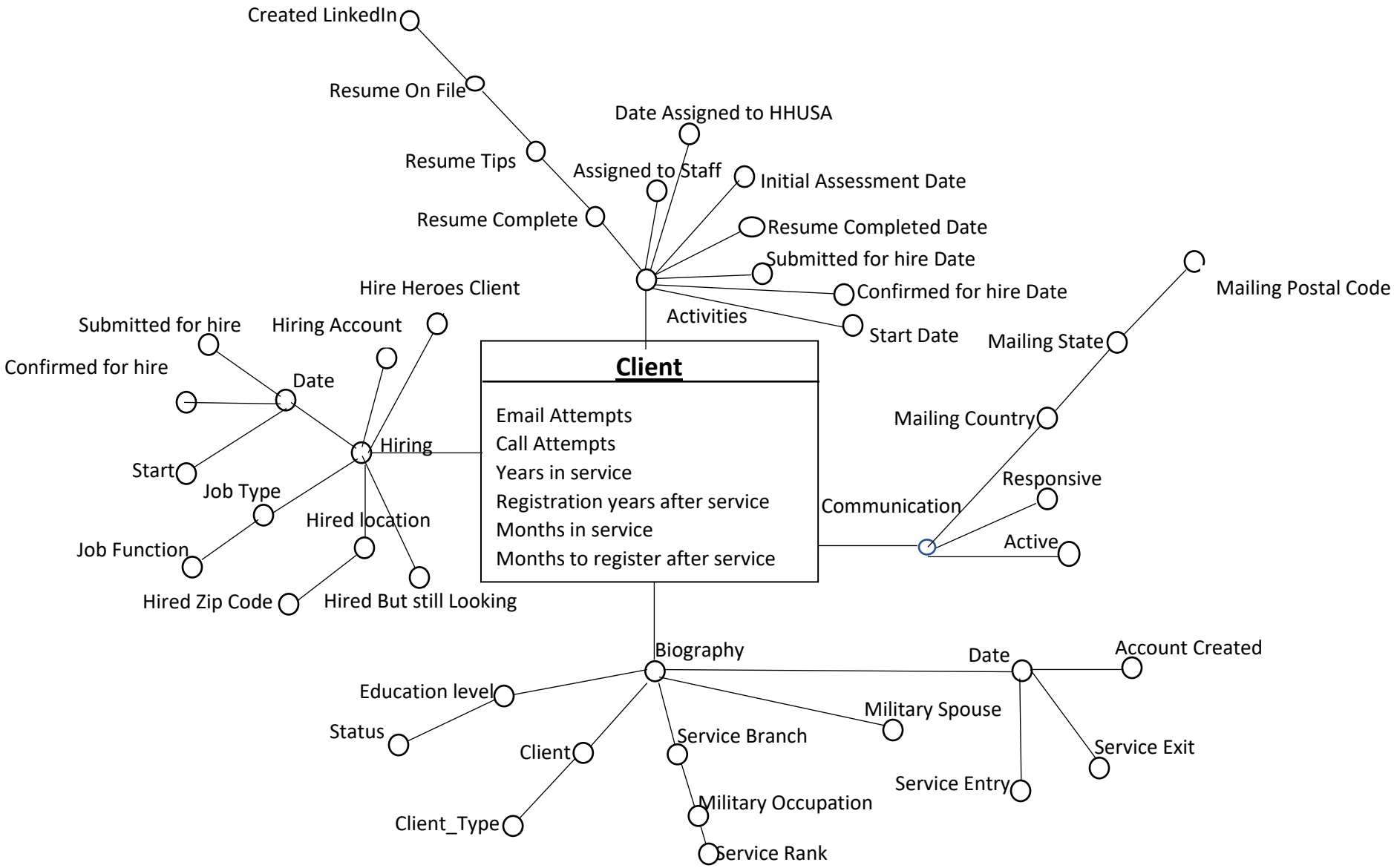
There were two sources of data provided, that is the hire heroes dataset and the adventure works dataset from Microsoft products. I chose to use the hire heroes dataset because it involved real people and real activities happening to real human beings in an organization trying to make a difference in the lives of service men after service in the army. The files came in as csvs so I had to put them all in a database in order to check their contents, I used postgresql as the relational database software. To speed up the importation process, I used dbeaver[1] a database management software interfacing with postgresql. Using this process I imported in the data in batches and the initial tables were created using the names of each of the original csv files provided.

Of all the provided csvs I ended up using the salesforce_contact relation because the questions I was going to answer were based on it. During the inspection of this relation I found out that there were some weird values represented in the table, so I did some cleaning. Some of the examples of inconsistent values included say service men that had entered the army is 1800s but had just been released recently after 125 years. Yes the date of service entry is correct but the date of exit is wrong. The years of service entry in contest included '1885-10-27 00:00:00';, '1892-10-19 00:00:00'; '1893-11-02 00:00:00'; '1884-09-30 00:00:00'; so I dropped them.

Because the data had come in as texts I had to do some transformations to change it to the types of my liking, all the dates came in text, so I had to covert these to date time. To decide which data to bring in and which data to throw a way I had come up with a major area of focus and this was a mashup between what was presented during the contest in terms of the business questions and what I would like to do. This mashup landed on the areas of employment with the army, registration with HHUSA, communication and job placement of servicemen that had finished their time with their individual branches now transitioning into the civilian. Some filters were created based on the later, a new column hired was created with the following conditions “start date of work”, “confirmed hire date” and “date submitted for hire” and had to be “clients with HHUSA” denoted with a “1”. Those that are submitted for hire have generally passed into the hiring process of HHUSA. During this process, some initial staging tables were created. After this process I had to define the initial logical schema of the DW based on the business questions.

[1] https://dbeaver.io/

II. Data warehouse conceptual design using the DFM notation.

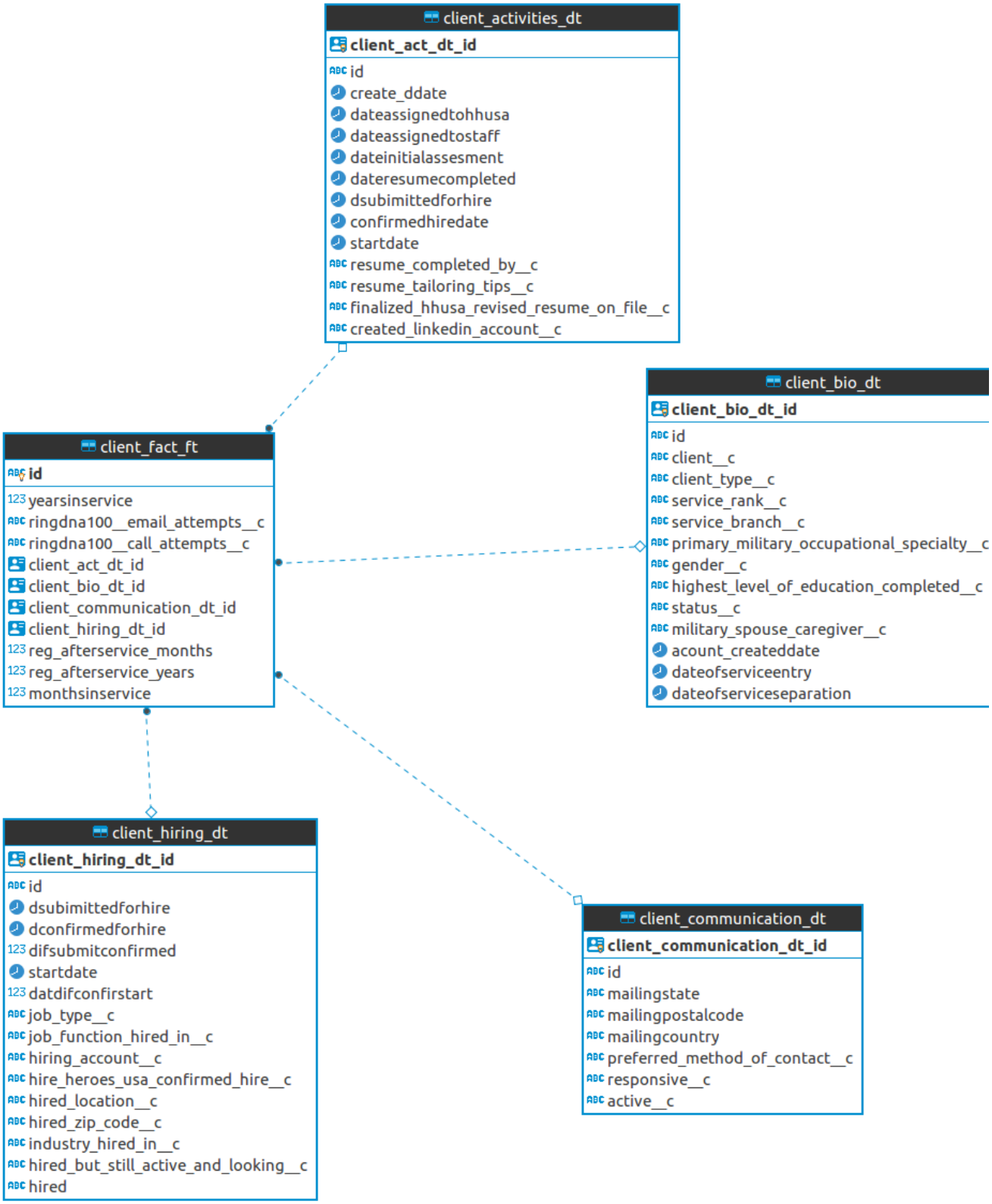


The motivation behind the choices I chose for the Datawarehouse were based on the business questions that I formulated. The questions that had to be answered were based on.

1. The overall registration trends of service men with HHUSA by service branch and rank.
2. The client registration trends over the years.
3. **How successful is the program in placing service men into jobs considering service branch and rank?**
4. The most effective communication avenue with the enrolled clients in the program.
5. Analyzing biographical information and how it plays into the employment outcomes of the clients.

From the DFM above the fact that I have chose to tackle in the project was client. The measures are time spent in the army, duration between which former service men register for services and communication data. The dimensions I have chosen to use in this data warehouse are based on biography, communication, hire, and activities attached to and individual client. The choice of contents for each of the dimension is based on the information I will extract to answer some of the questions above. Where applicable some dimensional attributes have hierarchies, this is because some dimensions got a one to many association with each other.

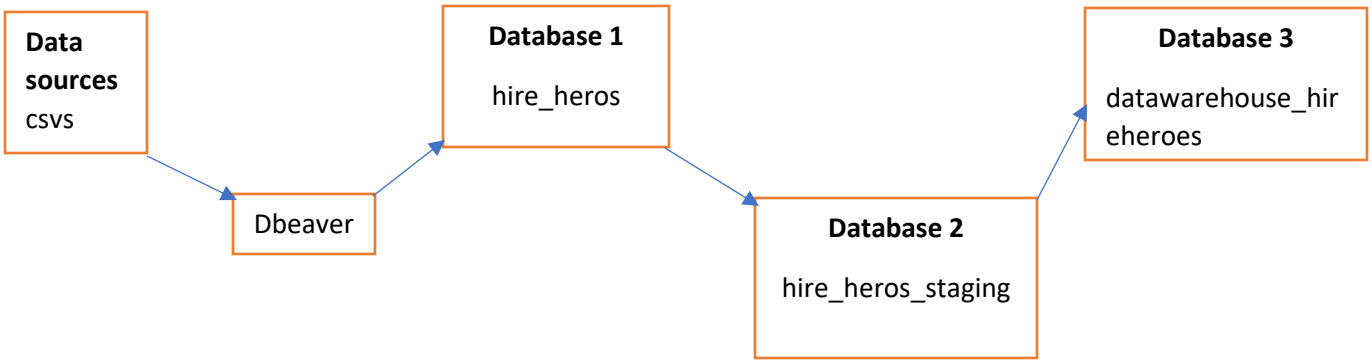
III. Data warehouse ROLAP logical design



During this stage of the process I had to put into practice what I had designed in the conceptual design phase based on the available data sources. The star schema was used in this process because its quicker to implement, since I had time constraints. The developed schema can be viewed in the image above as generated from dbeaver. This design is based on the DFM model above. From the schema above we can view that there is one fact table "client_fact_ft" that brings in 4 dimensions as a unit " client_activities_dt , client_bio_dt , client_communication_dt , client_hiring_dt " . The name of the data warehouse is called "datawarehouse_hireheroes". To link up the dimensions with fact table I used uuids in each of the dimension table and then placed them in the fact table.

IV. The system architecture.

The whole system onto which the data warehouse was built runs on the following architecture.



As mentioned earlier the data is brought into the system using Dbeaver and then inspected in database called hire_heros, some cleaning is done during this stage and then its sent to Database 2 which is the staging database further refinement is done and then its sent to the final Database which hosts the data ware house called “datawarehouse_hireheroes”. The extraction in sending of data between the three databases is done using dblink in posgtresql.

V. The OLAP queries.

The OLAP queries were performed as required. Please check the attached file called Olap_Queries.sql.

VI. Hive/SparkSQL.

In this part of the assignment I used SparkSQL on the DLTM cluster as user17. Data was uploaded on the cluster as hdfs and queried successfully. Each of the tables in the database was put on the cluster. The python script that runs on the cluster is attached and everything can be reproduced. The data was stored using the following paths on the DLTM cluster and appropriate dataframes were constructed from the inputs.

inputFileHire ='/user/user17/mosedata_proj/input/client_hiring_dt.csv'
inputFileBio ='/user/user17/mosedata_proj/input/client_bio_dt.csv'
inputFileCom = '/user/user17/mosedata_proj/input/client_communication_dt.csv'
inputFileAct = '/user/user17/mosedata_proj/input/client_activities_dt.csv'
inputFileFact = '/user/user17/mosedata_proj/input/client_fact_ft.csv'

I used windows with putty as the link to the DLTM linux system. The name of the file that can be used to reproduce the whole process is called dataframes.py.

VII. Tableau

During this stage in time I used the tableau trial version to analyze the areas of interest. Attached is the tableau workbook and power point presentation of the same which is much more refined. The name of the workbook is DWProjectFinal.twb. The first five dashboards have the summarized information for the analysis while other sheets are standalone's but the dashboards are based on the individual sheets.

VIII. Effort.

In terms of time spent on the project, since I have been working on it alone it has taken me over 60 hours, I was meant to hand in earlier but could not. There have been many tasks to implement but also this being an uncertain season I was slowed down by the whole happenings around me. I have had to dig deep so that I get this done.