LinkedIn Data Warehouse

Data Warehousing
ITCS/ITIS-6163
Project Report

By
Group 5

Pooja Passi
Madhukar Ganesh Chatra
Kiran Gaitonde

Table of Contents

1. INTRODUCTION

1.1. Motivation

The objective of the project is to build a LinkedIn profiles warehouse and to bring together available data related to profile like position title, company, location etc. and organize it. The organized data helps to retrieve, summarize and analyze the stored data easily and accurately.

1.2. Scope

The scope of the project includes:

- Extracting data from LinkedIn.
- Creating a star schema using data extracted.
- Analysis of the data using OLAP cubes and data mining algorithms.

1.3. Tools Used

- Informatica Power center express
- MS SQL Server 2014
- MS SQL Server Analysis Service 2013
- SAS E miner

1.4. References

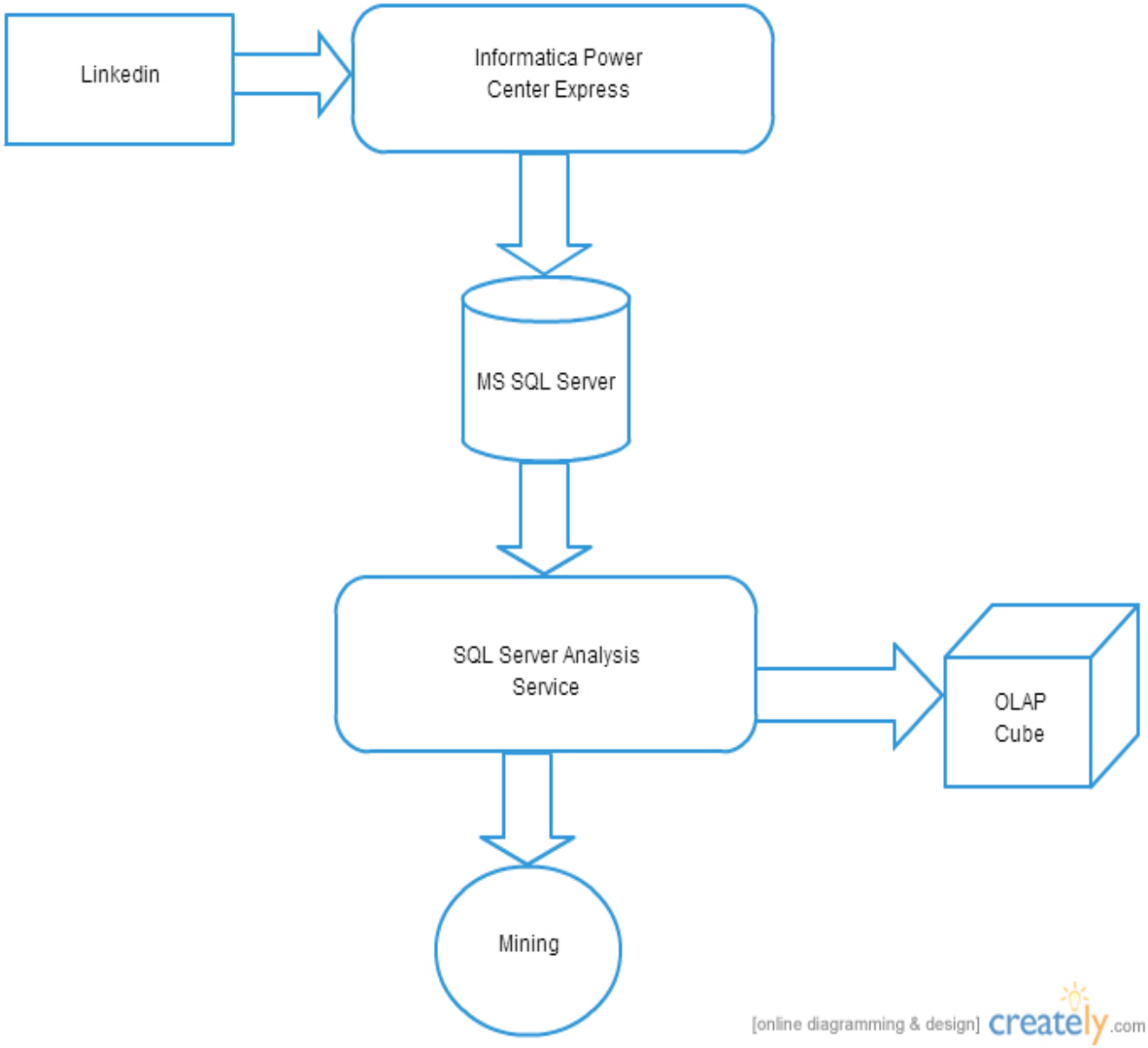
- https://community.informatica.com
- https://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf
- https://msdn.microsoft.com/en-us/library/ms175595.aspx
- https://www.youtube.com/watch?v=ctUiHZHr-5M
- https://stackovweflow.com

2. IMPLEMENTATION

2.1. SYSTEM ARCHITECTURE

Fig1 shows the overall architecture of the project. We extract, transform and load the LinkedIn data using the Informatica power center express tool. The data is loaded into the MS SQL Server database. We create the star schema using this loaded data. The tables of the star schema are used to create OLAP cubes and do data mining.

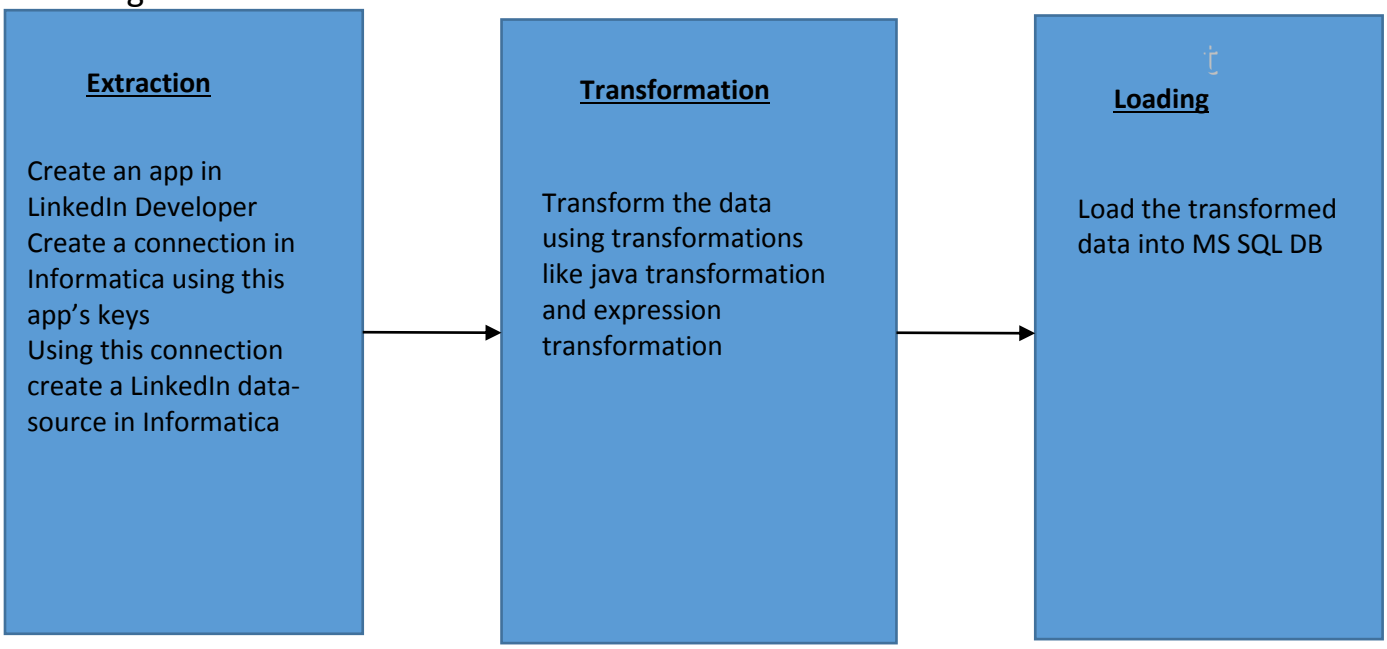
Fig1.



2.2. EXTRACT TRANSFORM LOAD (ETL)

Fig2 shows the ETL process followed in this project

Fig2.



2.3. Schema Creation

The fact and dimension tables are created using data imported from the LinkedIn source from Informatica power express tool, the data is mapped into a table by connecting into Microsoft SQL server through linked relational database connection object. The data is transformed using Informatica transformations before loading into Microsoft SQL server.

The data is loaded into AllProfiles table whose structure is given below. The data loaded is basically the profile information of employees.

```
CREATE TABLE AllProfiles( id int IDENTITY(1,1) PRIMARY KEY,
firstName VARCHAR(100),
lastName VARCHAR(100),
profileID VARCHAR(10),
headline VARCHAR(1000),
industry VARCHAR(100),
country VARCHAR(10),
location VARCHAR(100),
currentTitle VARCHAR(100),
currentCompany VARCHAR(100),
currentTitleStartYear VARCHAR(10),
currentTitleNoOfMonths VARCHAR(10),
 );
```

The first name and last name corresponds to name of the employee. Industry, current title, company and location corresponds to current job description and its field. Start year and no of months corresponds to number months of experience and the year in which he started to work for the current company under current position.

The fact and dimension tables are created using the main data source table Allprofiles. The below is the sample screen shot of the dataset obtained.

firstName	lastName	headline	industry	country	location	currentTitle	currentCompany	currentTitleSta	currentTitle
Ashutosh	Sinha	Group Head - Talent Acquisition & Strategy	Human Resources	in	Mumbai Area, India	Senior Vice President	Reliance Industries Lir	2015	3
Tina	Campbell	Sr. Technical Recruiter at Apple	Semiconductors	us	San Francisco Bay Area	Sr. Technical Recruite	Apple	2012	8
Eleni	Anderson	Technical Recruiter at Splunk	Information Technology	us	San Francisco Bay Area	Technical Recruiter	Splunk	2012	5
Andie	Borcz	Recruiting Consultant at Booz Allen Hamilton	Defense & Space	us	Baltimore, Maryland Ar	Principal, andie.borcz	Beratung Partners	2015	1
Mike	Ross	Helping Launch Careers at Weapply4u.com	Marketing and Advertisi	oo	Other	Helping Launch Caree	Weappl4u.com - Helpi	2013	1
Heather	Empson	Technical Recruiter at Intel	Information Technology	us	Phoenix, Arizona Area	Technical Recruiter	Intel Corporation	2014	11
Brent	Baker	Recruiting Manager at Websense	Computer & Network Se	us	Austin, Texas Area	Recruiting Manager	Websense	2013	12
Courtney	Manning	Campus Program Manager at AirWatch by VMware	Computer Software	us	Greater Atlanta Area	Campus Program Man	AirWatch by VMware	2014	4
Ellen	Parsons	Technical Recruiter at Two Sigma	Information Technology	us	Greater New York City	Engineering Recruiter	Two Sigma	2015	2
Brooke	Erickson	Technical Recruiter at Randstad Technologies	Staffing and Recruiting	us	Sacramento, California	Senior Technical Recr	Randstad Technologie	2011	2

The dimension tables are company, industry, location, position and employee. The company table contains company details, industry table contains industry details, location table contains location and country code, position table contains current position and the company name and employee table contains the employee name details along with their current title as headline and number of months of experience and the year in which they joined the current company. The fact table is associated with all the dimension tables which contains the industry, company, position, location and profile details of all employees. The below figure shows the overall structure of the fact and dimension table.

```
select * from AllProfiles ;
delete from AllProfiles where profileID='private' and profileID='NULL';
create table company(ID int IDENTITY(1,1) PRIMARY KEY, company_name varchar(100));
select * from company;
create table industry(ID int IDENTITY(1,1) PRIMARY KEY, industry_name varchar(100));
select * from industry;
create table employee(ID int IDENTITY(1,1) PRIMARY KEY,first_name varchar(200),last_name varchar(200),
headline varchar(200), date varchar(100),year varchar(100));
select * from employee;
create table location(ID int IDENTITY(1,1) PRIMARY KEY,location_name varchar(100),country varchar(20));
select * from location;
create table position(ID int IDENTITY(1,1) PRIMARY KEY,position_name varchar(100),company_name varchar(100));
select * from position;
create table profile(ID int IDENTITY(1,1) PRIMARY KEY,emp_id int,industry_id int,company_id int,position_id int,
location_id int,FOREIGN KEY (emp_id) REFERENCES employee(ID),
FOREIGN KEY (industry_id) REFERENCES industry(ID),FOREIGN KEY (company_id) REFERENCES company(ID),
FOREIGN KEY (position_id) REFERENCES position(ID),
FOREIGN KEY (location_id) REFERENCES location(ID));
select * from profile;
```

The dimension table company is inserted with unique company names from the Allprofile table, Industry table is inserted with unique industry names, position table is inserted with the unique positions along with their company names, location table with the location names and the corresponding country names and employee table with their names details and the their headline and the number of months of experience in the current position and the start year. Each table is identified by auto generated id while inserting the other details.

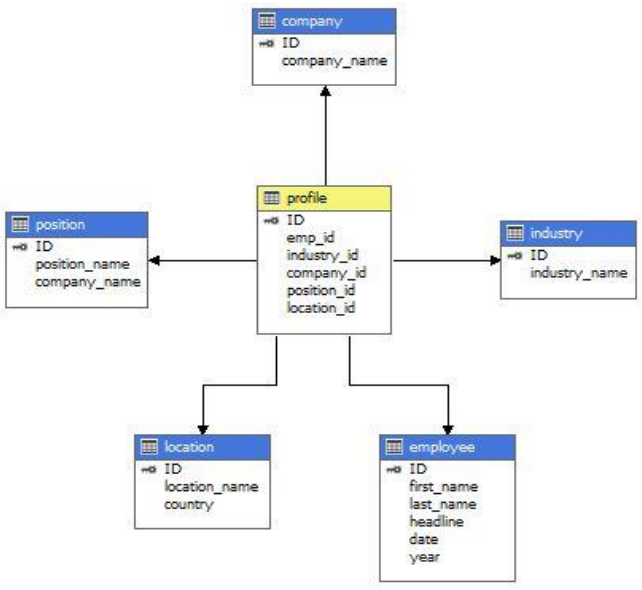
The fact is the combination of all dimension tables, it is created by using the left join on all dimension tables and then selecting the corresponding unique employee id, company id, industry id, location id and position id into the fact table called profile. The below figure shows the SQL queries through which fact and dimension tables are created.

```
insert into company (company_name)  select distinct currentCompany from AllProfiles;

insert into industry(industry_name) select distinct industry from AllProfiles where industry IS NOT NULL;

insert into location(location_name,country) select distinct a.location,a.country from AllProfiles as a where
exists(select b.location,b.country from AllProfiles as b where a.location=b.location);

insert into position(position_name,company_name) select distinct a.currentTitle,a.currentCompany from AllProfiles as a where
exists(select b.currentTitle,b.currentCompany from AllProfiles as b where a.currentTitle=b.currentTitle);

insert into employee(first_name,last_name,headline,date,year)
select distinct a.firstName,a.lastName,a.headline,a.currentTitleNoOfMonths,a.currentTitleStartYear
from AllProfiles as a where exists(select b.firstName,b.lastName,b.headline,b.currentTitleNoOfMonths,
b.currentTitleStartYear from AllProfiles as b where a.firstName=b.firstName);

select * from employee;
delete from employee where year IS NULL;
delete from company where company_name='-';
delete from industry where industry_name='';
```

The data cleaning and transformation is then carried out on the data to remove inconsistent data values like NULL and private.

Fig3 shows the star schema.

Fig3.



2.4. OLAP Scenarios

Scenario 1:

To find out the number of employees in a particular position of a company by taking location and country as filters.

SQL Query:

```
select count(*) as count,company.company_name,location.location_name
from company,location,profile
where  profile.company_id=company.ID  and
profile.location_id=location.ID and location.country='us'
group by company.company_name,location.location_name;
```

Output:

Company Name	Country	Profile Count
3DPLM Software Solutions Limited	in	1
3S Business Corporation Inc	in	1
Accenture	in	5
Accenture	us	2
ACI Worldwide	us	1
ADB Pvt Ltd	in	1
ADP	in	1
Aerotek	us	1
AirWatch by VMware	us	1
AksaTech Solutions	in	1
AL KHAN FOODSTUFF LLC	oo	1
Alcatel-Lucent	in	2
Alexander Mann Solutions	us	1
Allstate	in	1
Allstate Solutions Private Limited	in	1
Amazon	in	1
Amazon	us	2

Scenario 2:

To find out the number of jobs in particular industry at particular point of time.

SQL Query:

```
select count(*) as count,industry.industry_name
from industry,employee,profile
where  profile.industry_id=industry.ID  and
profile.emp_id=employee.ID and employee.year='2015'
group by industry.industry_name;
```

Output:

Industry Name	Year	Profile Count
Accounting	2012	1
Accounting	2014	2
Architecture & Planning	2013	1
Automotive	2012	2
Automotive	2013	2
Automotive	2014	2
Automotive	2015	3
Aviation & Aerospace	2013	3
Banking	2004	1
Banking	2010	1
Banking	2014	1
Biotechnology	2013	1
Capital Markets	2015	1
Chemicals	2013	1
Civil Engineering	2013	1
Civil Engineering	2015	1

Scenario 3:

To find out the distribution of work force industry and company wise.

SQL Query:

```
select count(*) as count,company.company_name,industry.industry_name
from industry,company,profile
where profile.company_id=company.ID and profile.industry_id=industry.ID
group by company.company_name,industry.industry_name;
```

Output:

Company Name	Industry Name	Profile Count
3DPLM Softwa...	Information Technology and Services	1
3S Business Co...	Information Technology and Services	1
Accenture	Computer Software	2
Accenture	Information Technology and Services	5
ACI Worldwide	Staffing and Recruiting	1
ADB Pvt Ltd	Architecture & Planning	1
ADP	Computer Software	1
Aerotek	Staffing and Recruiting	1
AirWatch by V...	Computer Software	1
AksaTech Solu...	Human Resources	1

Overall cube structure after Drill Down on all possible attributes:

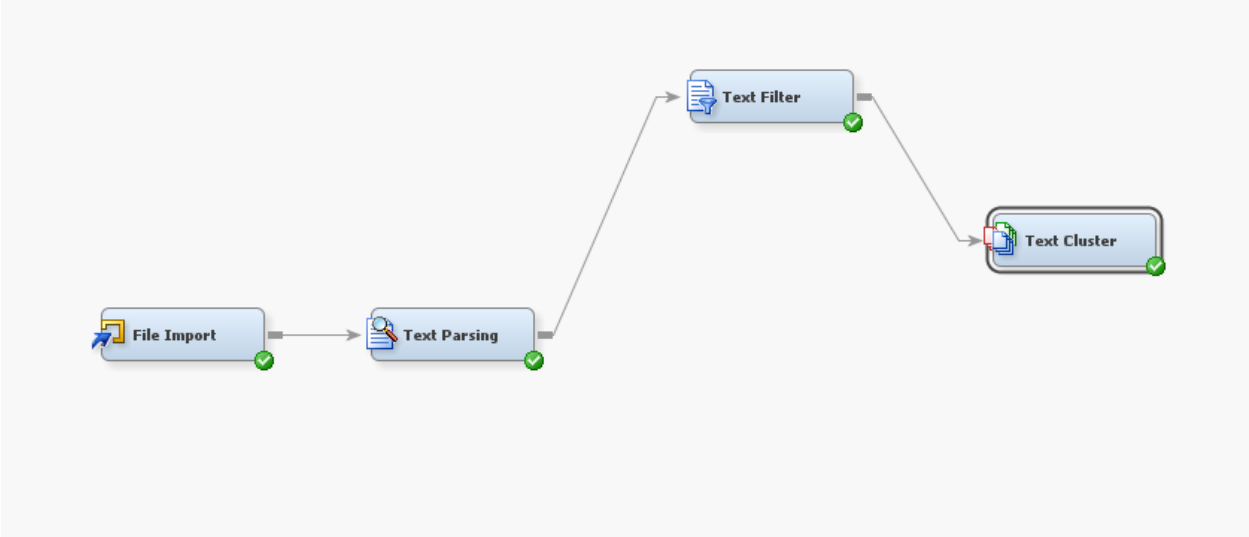
Company Name	Industry Name	Year	Country	Position Name	Profile Count
3DPLM Software Solutions Limited	Information Technology and Services	2013	in	Software Developer	1
3S Business Corporation Inc	Information Technology and Services	2014	in	Talent Acquisition Co-ordinator	1
Accenture	Computer Software	2010	in	Software Engineer	1
Accenture	Computer Software	2013	in	SAP ABAP Consultant	1
Accenture	Information Technology and Services	2010	us	Senior Software Engineer	1
Accenture	Information Technology and Services	2011	in	Oracle BPM Developer	1
Accenture	Information Technology and Services	2011	in	Software Engineering Analyst	1
Accenture	Information Technology and Services	2015	in	Software Engineering Analyst	1
Accenture	Information Technology and Services	2015	us	Sourcing Recruiter	1
ACI Worldwide	Staffing and Recruiting	2014	us	Director Global Talent Acquisition	1
ADB Pvt Ltd	Architecture & Planning	2013	in	Architect	1
ADP	Computer Software	2014	in	Software Engineer	1
Aerotek	Staffing and Recruiting	2014	us	Financial Recruiter	1
AirWatch by VMware	Computer Software	2014	us	Campus Program Manager	1
AksaTech Solutions	Human Resources	2015	in	Director of HR	1
AL KHAN FOODSTUFF LLC	Logistics and Supply Chain	2014	oo	Supply Chain Analyst	1
Alcatel-Lucent	Computer Software	2012	in	Software Engineer	1
Alcatel-Lucent	Telecommunications	2013	in	Senior Software Engineer	1
Alexander Mann Solutions	Staffing and Recruiting	2014	us	Sourcing Specialist	1

2.5. DATA MINING

Clustering by Text mining:

The text mining is carried out using SAS E miner. The given dataset is imported using SAS file import module and then the data set is parsed using default text parser and then filtered and divided into clusters. The steps involved in clustering are given below.

- Text Import – reads text files stored in a single directory
- Text Parsing – creates the Term/Document matrix
- Text Filter – allows analysts to eliminate non--useful words
- Text Topic – identifies document groups associated with topic words
- Text Cluster – identifies document clusters from their word structures.



The clustering process forms the ten clusters where each cluster is the group of similar profiles. The clustering is carried out based on considering the factors like type of industry, company and location in which employee is working and the position.

The below figure gives the detailed description of the each cluster with the cluster numbering.

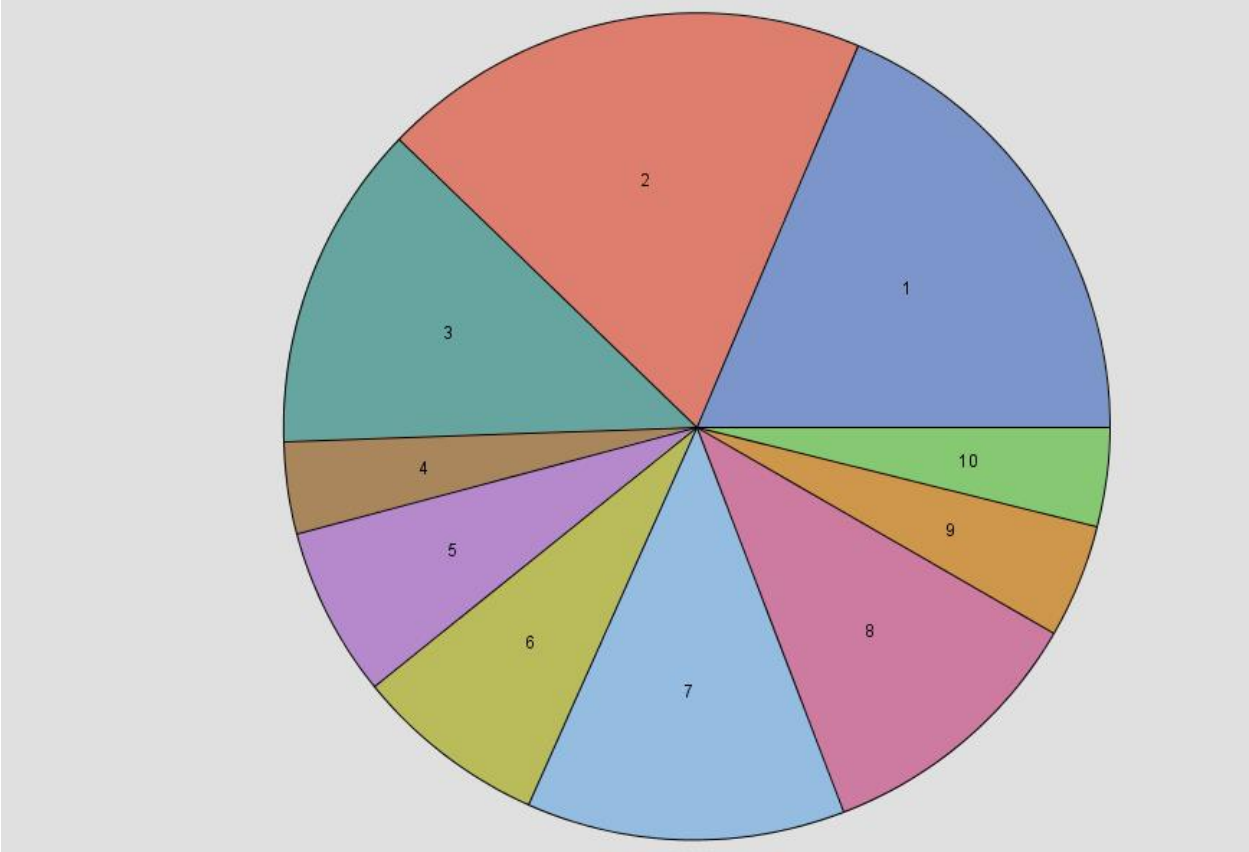
Clusters

Cluster ID	Descriptive Terms
1	tech +limit development group +solution senior +analyst executive director corporation hr bosch ibm +technology bank ...
2	+look arlington professor +master data research science institute computer +position 'computer engineering' assistant student college university ...
3	acquisition campus inc manager recruiter talent technical sr avidxchange collabera +staff corporate resource +program member ...
4	+consultant sap +solution +recruit career inc private member accenture ibm electronics +staff executive +limit development ...
5	ltd pvt sankalp toshiba +engineer software senior india design hcl intel +technology google associate information ...
6	+engineer software +system electronics communication infosys ge tcs corporation hcl +operation aerospace networks quality test ...
7	+internship +seek actively summer 'graduate student' graduate +'full time opportunity' +opportunity uncc +start information full time +'summer internship' unc ...
8	carolina charlotte north student university 'graduate student' graduate unc teaching assistant +embed 'computer engineering' electrical co-op college ...
9	+service consultancy tata +system +engineer java lead +analyst bank campus business assistant +consultant developer software ...
10	developer application intern web oracle java accenture software global tcs collabera experienced private +embed product ...

Cluster 1: The profiles related to Human resource department, director and senior analyst and other senior designations are grouped into one.
Cluster 2: The academic related profiles lie professor, teaching assistant, research assistant etc. are grouped together.
Cluster 3: The staffing and recruiter related profiles are grouped together.
Cluster 4: Consultant and staff executive which are related to talent management are grouped together.
Cluster 5: Hardware related profiles in specific that related to companies like HCL, Intel, Sankalp are grouped together
Cluster 6: Electronics, networks and operations related profiles are grouped together.
Cluster 7: The profiles of people seeking internships, graduate students, and summer interns are grouped together.
Cluster 8: The profiles that are related to North Carolina, charlotte and UNC Charlotte are grouped together.
Cluster 9: The profiles related to Analysts, Business analyst are grouped together.
Cluster 10: The profiles related to application development. Web development, java are grouped together.

The percentage distribution of the profiles among the different clusters are given below through table and the pie chart.

cluster	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	135	18.65	135	18.65
2	138	19.06	273	37.71
3	93	12.85	366	50.55
4	26	3.59	392	54.14
5	49	6.77	441	60.91
6	53	7.32	494	68.23
7	92	12.71	586	80.94
8	78	10.77	664	91.71
9	31	4.28	695	95.99
10	29	4.01	724	100.00

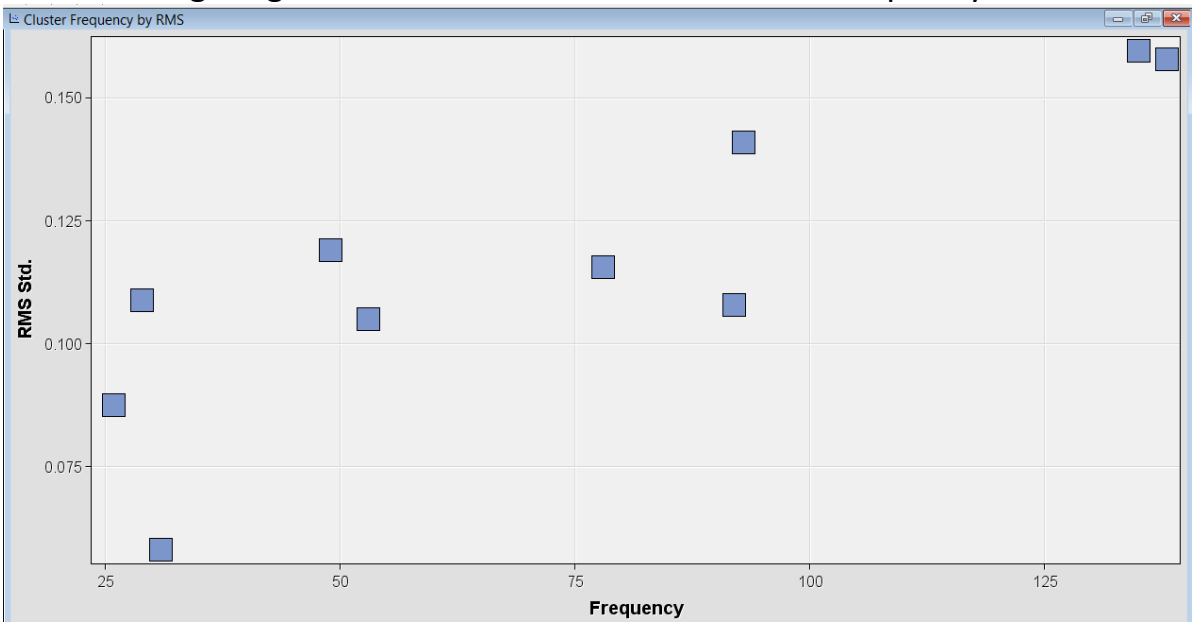


3.

The analysis of the clustering shows that most of the profiles falls under category related to either human resource, director and senior analyst or academic related professions like professor, teaching assistant, research

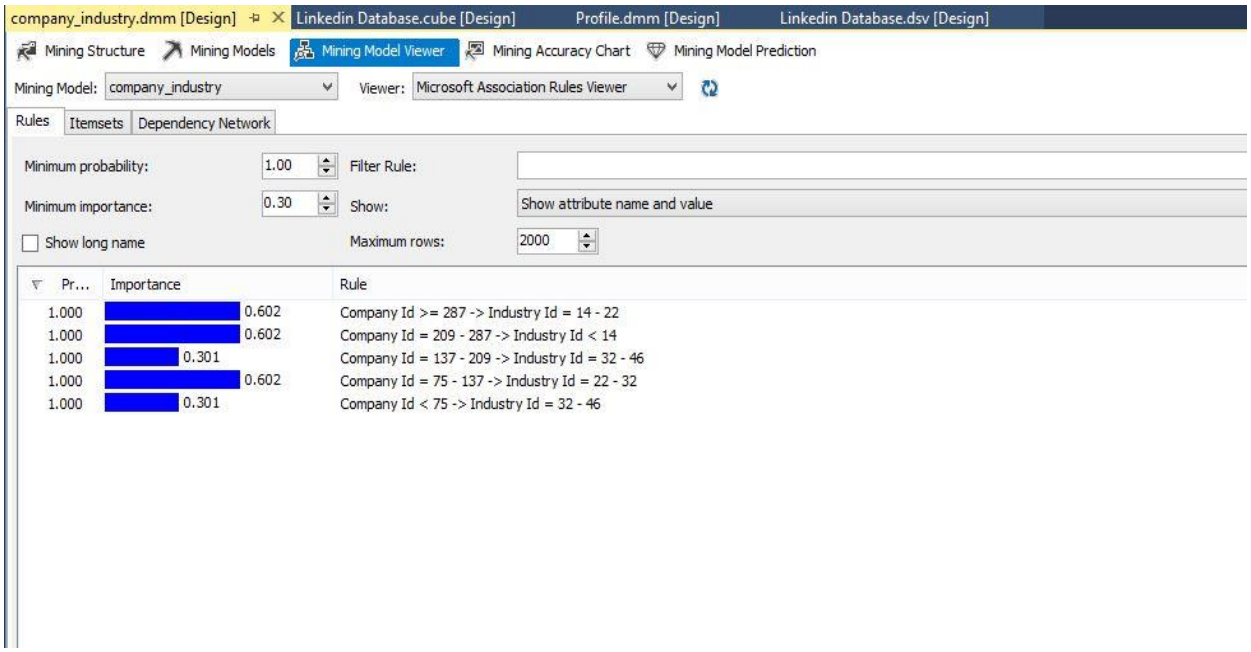
assistant etc., also very less number of profiles related to recruiter, staffing areas.

The below figure gives the variation of RMS VS Cluster frequency.



Association Rule mining:

The association rule mining is carried out to find out the association between the industries and the companies. The resulting graph is given below.



4. CONCLUSION and FUTURE WORK

The LinkedIn warehouse is built and analyzed on various factors like job industry, company, work location etc. The LinkedIn warehouse is constructed using the different dimensions like location, company, position, employee personal information and industry. The Rolling up and drill down operations are carried out on the cube created using these dimensions and the hierarchies like position inside a company, location inside a country and year and month for signifying the experience of the employee to find out and analyze the work force distribution based on different factors. The text mining and clustering revealed that facts like common job field, companies in each industry etc.

The Data Warehouse built can be used by the recruiters to find out the job trend and predictive analysis of the job market in future years. Also the warehouse can be used by the job seekers to find out the most demanding industry and the promising company and the skills required and the market expectation of the job.

The future work may consists of analysis other attributes like skills, certifications for more accurate prediction. These analysis can be used by the job seekers and IT staffing companies to find out the skills on which they need to train the workforce in order to easily market them and or job seekers to find out the field in which they need more training and certifications.

5. CODE LISTING

Informatica Java Transformation code:

Import:

```
import javax.xml.parsers.*;
import org.xml.sax.InputSource;
import org.w3c.dom.*;
import java.io.*;
```

Main Code:

```
try {
      DocumentBuilderFactory dbf =
         DocumentBuilderFactory.newInstance();
      DocumentBuilder db = dbf.newDocumentBuilder();
      InputSource is = new InputSource();
      is.setCharacterStream(new StringReader(connections_person_positions_position));

      Document doc = db.parse(is);
      NodeList positions = doc.getElementsByTagName("position");

      // iterate the employees
        Element element = (Element) positions.item(0);
        NodeList title = element.getElementsByTagName("title");
        Element op1 = (Element) title.item(0);

        Node child1 = op1.getFirstChild();
            if (child1 instanceof CharacterData) {
            CharacterData cd = (CharacterData) child1;
            curretTitle = cd.getData();
         }else{
              curretTitle = "noValue";
            }
       // System.out.println("Title: " + curretTitle);
       //get employees
        NodeList company = element.getElementsByTagName("company");
        Element companyElement = (Element) company.item(0);
        NodeList companyName = companyElement.getElementsByTagName("name");
        Element op2 = (Element) companyName.item(0);
```

```
        Node child2 = op2.getFirstChild();
            if (child2 instanceof CharacterData) {
            CharacterData cd = (CharacterData) child2;
            currentCompany = cd.getData();
         }else{
               currentCompany = "noValue";
            }
        //  System.out.println("Company: " + currentCompany);
        // Start date
         NodeList startDate = element.getElementsByTagName("start-date");
         Element startDateElement = (Element) startDate.item(0);
         NodeList year = startDateElement.getElementsByTagName("year");
         Element op3 = (Element) year.item(0);

         Node child3 = op3.getFirstChild();
            if (child3 instanceof CharacterData) {
            CharacterData cd = (CharacterData) child3;
            currentTitleStartYear = cd.getData();
         }else{
               currentTitleStartYear = "noValue";
            }

         //System.out.println("Year: " + currentTitleStartYear);
         NodeList month = startDateElement.getElementsByTagName("month");
         Element op4 = (Element) month.item(0);

         Node child4 = op4.getFirstChild();
            if (child4 instanceof CharacterData) {
            CharacterData cd = (CharacterData) child4;
            currentTitleNoOfMonths = cd.getData();
         }else{
               currentTitleNoOfMonths = "noValue";
            }
       //  System.out.println("Month: " + currentTitleNoOfMonths);

      }
    catch (Exception e) {
       e.printStackTrace();
    }
```