Associative Analysis in Statistics

Mihaela MUNTEAN (email: mihaela.muntean@ie.ase.ro)
University of Economic Studies

ABSTRACT

In the last years, the interest in technologies such as in-memory analytics and associative search has increased. This paper explores how you can use in-memory analytics and an associative model in statistics. The word "associative" puts the emphasis on understanding how datasets relate to one another. The paper presents the main characteristics of "associative" data model. Also, the paper presents how to design an associative model for labor market indicators analysis. The source is the EU Labor Force Survey. Also, this paper presents how to make associative analysis.

Key words: *associative data model, in-memory analytics, associative analysis, statistical indicators*

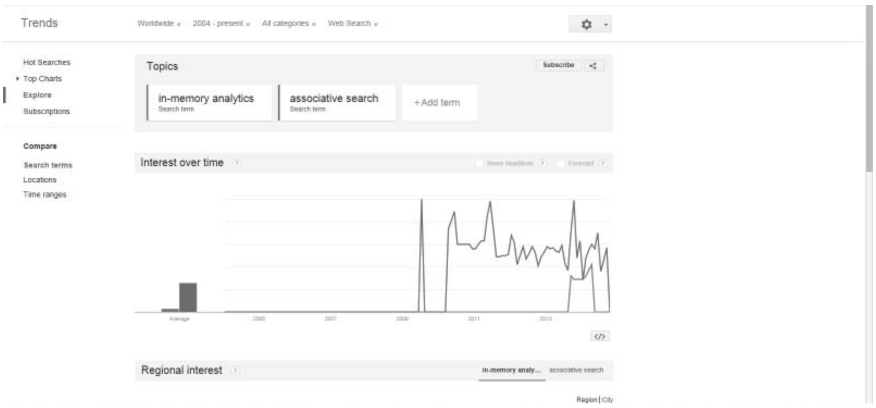
JEL Classification: *J6, C88*

Introduction

In the last years, the interest in technologies such as in-memory analytics and associative search has increased. The figure 1 shows the trend in the utilization of these technologies, using Google Trends.

In-memory analytics versus associative search. Interest over time

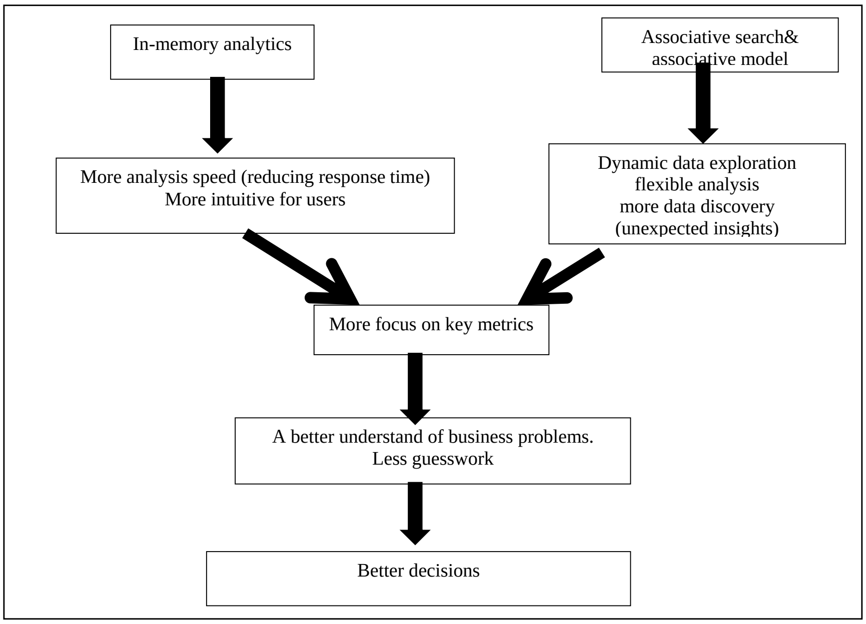
Figure 1



Also, figure 2 shows how these technologies affect businesses. These technologies allow business people to make basic exploration of large datasets and to find better answers to business problems. These technologies allow users to become less dependent on IT. The business user can act as an analyst. The speed of in-memory technology makes possible more analytics iterations within a given time. The speed is very important in analysis. In-memory technologies load the entire dataset into RAM before a query can be executed by users. Also, these technologies can save significant development time by eliminating the need for aggregates and designing of cubes and star schemas. Also, they can simplify a larger number of tasks in an analytics workflow. Associative search allows business people to make dynamic data exploration and to identify unexpected insights in large datasets.
There are many and different in-memory BI solutions such as: Qlikview, Tibco Spotfire, Tableau, Microstrategy, Microsoft Power Pivot, SAP Hana, Oracle Exalytics in-memory, etc. Most of in-memory BI solutions use the following data models: dimensional model (star schema, snowflake or combinations), hypercube and "associative" data model. The next section takes a look at the pros of the associative model.

How in-memory analytics and associative search affect businesses

Figure 2



THE MAIN CHARACTERISTICS OF ASSOCIATIVE DATA MODEL

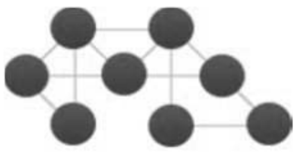
The model is implemented by Qlikview tool. Qlik Tech company is a leader in "Magic quadrant for Business Intelligence and analytics platforms" (Sallam et al., 2014). Qlikview implements three emerging technologies: interactive visualization, in-memory analytics and associative search. *"Associative"* data model free users from the paradigm of dimensions versus measures (figure 3). *"Associative"* data model makes no distinction between attributes that are facts and attributes that are dimensions. The word "*associative*" puts the emphasis on understanding how datasets relate to one another. This model is built around the concept of datasets with related logic tables. The datasets are loaded in memory, in a compressed and fully normalized format, via the Load script. The main characteristics of "*associative*" data model are (Morrison, 2012), (Muntean and Surcel, 2013):

- is based on the heterogeneous sources (databases, spreadsheets, Web pages and big data). This model is persistent and reacts as a whole to user "queries". A selection affects the entire schema. You can select any value for any attribute and all the related data from the entire data model will be displaying (associative search);
- eliminates the need to develop hierarchies, hyper-cubes and pre-aggregation of data;
- you don't have to use a data query language;
- you don't have to use data definition language;
- each load or select statements generate a logical table during the data load process. The associations between logical tables are generated automatically during the data load process based on matching column names across logical tables. Any fields with the same name in two or more tables will associate. The relationships among logical tables usually don't reflect foreign key relations. The associations between logical tables are similar to full outer joins. If there is more than one field with the same name a synthetic key is created. A synthetic key contains all possible combinations of common attributes among tables. It is resource intensive and makes data model difficult to understand.
- the aggregations can be done both in the load script (pre-defined) and at the user interface development stage. This enables a user to interact with a broader range of data than will ever be possible in SQL;
- adaptable to rapid business changes and flexibility in analysis. Any value of any attribute can be the starting point of analysis;

- faster model design. The common problems of the model are: synthetic keys and circular references. There are many ways to resolve a synthetic key such as:
    - We can rename the attributes of synthetic key that are not part of the association between two datasets;
    - We can delete the common attributes from one of the datasets;
    - We can create a composed key by concatenating the common attributes that represent the association between the two datasets. Then we delete the common attributes from loading script.
    - We can concatenate the logic tables, or we can use link logic tables (Qlikview Reference Manual, 2010), (Qlikview white paper, 2010). I used a link logic table.

Associative model

Figure 3



User Driven Adaptable to rapid business changes Associative search

"Associative" model is a bottom-up model and it is developed by each department and then adopted by the company. The model and the user interface are developed together using an agile development approach- Qlikview Project Methodology (QPM). This methodology changes the focus from data driven to the decision driven. QPM includes a description of all project management activities, documents and deliverables in all project phases. This methodology includes the following phases:

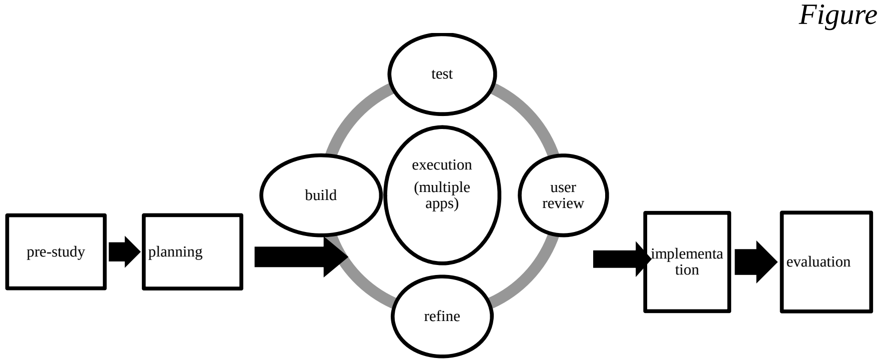
- Pre-study that includes the following main steps: define initial business objectives &scope, identify initial business requirements and data requirements, prepare enterprise platform;
- Planning phase with the following main steps: project management planning, prioritized business requirements and defining the data

staging requirements (QVD files for larger deployment), define enterprise platform and plan application cycles;

- Iterative execution phase that includes the following steps: build, test, user review and refine. Build step includes: build data reload process (configuring of the connections, create incremental load script), data model design, data provisioning, user interface development;
- Implementation;
- Evaluation (Sterry, B. and Nieuwenhuijs J., 2011), (figure 4).

Qlikview Project Methodology

Figure 4



The next section briefly presents how to design an associative model for labor market indicators analysis. Also, this section presents how to make associative analysis.

Associative data model design

The EU Labor Force Survey (EU-LFS) is the most important source of information on the situation and trends in the EU labor market. The EU-LFS provides quarterly and annual data on labor force. The EU-LSF data sources cover 33 countries: European Union members, Iceland, Norway, Switzerland and two EU candidate countries (Macedonia and Turkey). Data are acquired by personal visits, telephone interviews, web interviews and questionnaires.

The LSF contains information on the following subjects: employment and unemployment, job vacancy statistics, labor market policy, earnings and labor costs. The Employment and unemployment subject includes many variables such as: total population, employment, employment rates, employees, full-time and part-time employment, temporary employment, etc.). These variables depend on multiple dimensions: age group, gender, educational attainment, countries, year, etc. The address of data sources is: http://epp.eurostat.ec.europa.eu/portal/page/portal/employment_unemployment_lfs/data/database. Figure 5 shows only the employment and unemployment LFS main indicators that present the main aspects of the labor market such as: population, activity and inactivity, employment and unemployment.

We can access and extract data from database. Data can be downloaded in various formats such as XLS (Excel), csv, html, pdf, etc. I downloaded only the following source table (as excel file): *Employment (main characteristics and rates) - annual averages [lfsi_emp_a]*. This file has been customized. The figure 6 presents *lfsi_emp_a-final* (excel file with multiple worksheets).

I used the following variables:

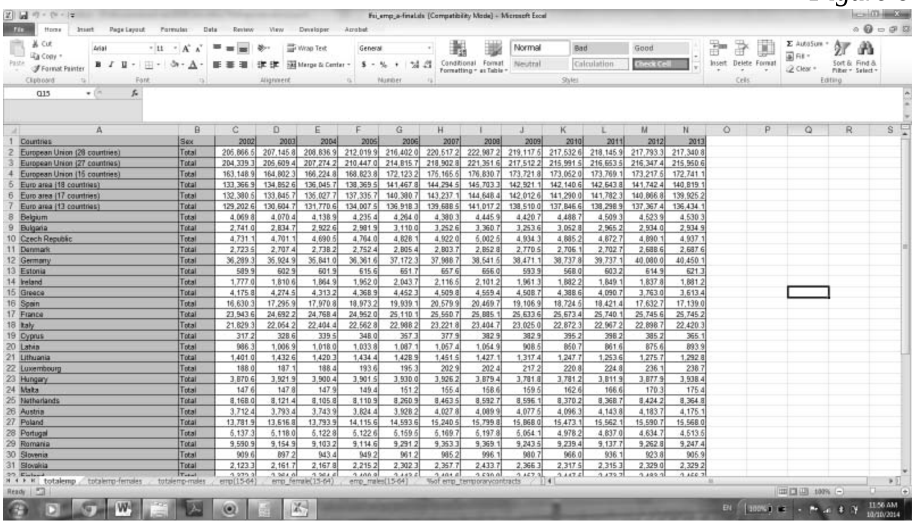
- *Totemp* (total employment -resident population concept - LFS) (Eurostat: http://ec.europa.eu/eurostat/data/database);
- *%emp_temporarycontract* (percentage of employees with temporary contracts). In the context of LFS, employee with temporary contact is "*an employee whose main job will terminate either after a period fixed in advance, or after a period not known in advance, but nevertheless defined by objective criteria, such as the completion of an assignment or the period of absence of an employee temporarily replaced*. „*An employee as an individual who works for a public or private employer and who in return receives compensation in the form of wages, salaries, fees, gratuities, payment by results or payment in kind. Professional military staff are also included*";
- *Emp* (Employment (15 to 64 years). In the context of the LFS, „*an employed person is a person aged 15 and over (or 16 and over in Iceland and Norway) who during the reference week performed work - even if just for one hour a week - for pay, profit or family gain. Alternatively, the person was not at work, but had a job or business from which he or she was temporarily absent due to illness, holiday, industrial dispute or education and training*" ;
- *Emp-rate* (employment rate). In the context of LFS "*employment rate is the percentage of employed persons in relation to the comparable total population. For the overall employment rate, the comparison is made with the population of working-age*" ;
- *%part-timeworkers* (part-time workers in % of total employment).

LSF database

Figure 5

Source: *lfsi_emp_a-final.xls* - worksheet *totalemp*

Figure 6



These variables depend on: gender, country, age group and year. Data source (*lfsi_emp_a-final.xls*) is loaded in memory, via the Load script. Database is not required. The data model and the associations between data sources are generated automatically during the data load process. The end users don't need to use a definition language. Also, the end users don't need to use a query language.

I will present only a part of the load script:

```
...
Employment:
CrossTable(Year, %emp_temporarycontracts, 2)
LOAD Countries, Sex,
[2002],[2003],[2004],[2005],[2006],[2007],[2008],[2009],[2010],[2011],[
2012],[2013]
FROM [lfsi_emp_a-final.xls]
(biff, embedded labels, table is [%emp_temporarycontracts_females$]);
// loading worksheet %emp_temporarycontracts_males
// loading worksheet %of emp_temporarycontracts
Part_timeworkers:
CrossTable(Year, %part_timeworkers, 2)
LOAD Countries, Sex,
[2002],[2003],[2004],[2005],[2006],[2007],[2008],[2009],[2010],[2011],[
2012],[2013]
FROM [lfsi_emp_a-final.xls]
(biff, embedded labels, table is [%part-timeworkers$]);
--loading worksheet %part-timeworkers-females
--loading worksheet %part-timeworkers-males
Join(Employment)
load * Resident Part_timeworkers;
drop table Part_timeworkers;
Employee:
CrossTable(Year, Emp, 2)
LOAD Countries,Sex,
[2002],[2003],[2004],[2005],[2006],[2007],[2008],[2009],[2010],[2011],[
2012],[2013]
FROM [lfsi_emp_a-final.xls]
(biff, embedded labels, table is [emp(15-64)$]);
// loading worksheet emp_female(15-64
// loading worksheet emp_males(15-64)
Join(Employment)
load * Resident Employee;
drop table Employee;
Totalemployee:
CrossTable(Year, Totalemp, 2)
LOAD Countries, Sex,
```

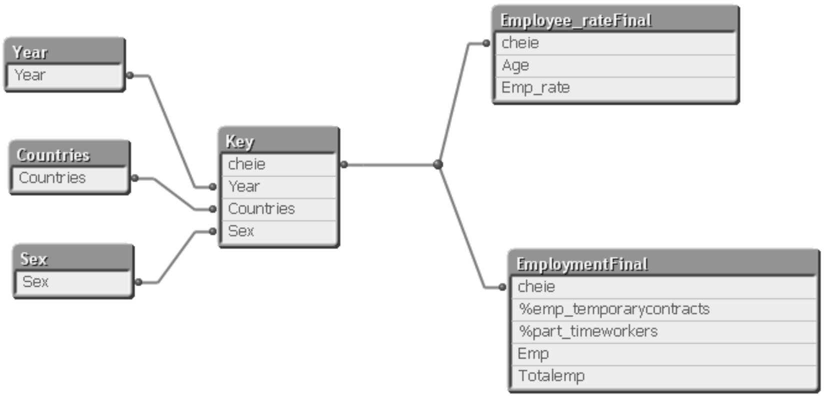
```
[2002],[2003],[2004],[2005],[2006], [2007], [2008],[2009],[2010],[2011],
[2012],[2013]
FROM [lfsi_emp_a-final.xls]
(biff, embedded labels, table is [totalemp$]);
// loading worksheet totalemp-females
// loading worksheet totalemp-males
Join(Employment)
load *
Resident Totalemployee;
drop table Totalemployee;
EmploymentFinal:
Load num#(Year) as Year, Countries,Emp,Totalemp,%part_
timeworkers,%emp_temporarycontracts, Sex, autonumberhash256(Year,Cou
ntries,Sex) as cheie
Resident Employment;
Drop Table Employment;
Employee_rate:
CrossTable(Year, Emp_rate, 3)
LOAD Countries, Age,Sex,
[2002],[2003],[2004],[2005],[2006],[2007],[2008],[2009],[2010],[2011],[
2012],[2013]
FROM [lfsi_emp_a-final.xls] (biff, embedded labels, table is [emp-rate(15-
24)$]);
// loading worksheet emp-rate(25-54)
// loading worksheet emp-rate(55-64)
// loading worksheet emp-rate-males(15-24)
// loading worksheet emp-rate-males(25-54)
// loading worksheet emp-rate-males(55-64)
CrossTable(Year, Emp_rate, 3)
LOAD Countries, Sex,Age,
[2002],[2003],[2004],[2005],[2006],[2007],[2008],[2009],[2010],[2011],[
2012],[2013]
FROM [lfsi_emp_a-final.xls] (biff, embedded labels, table is [emp-rate-
females(15-24)$]);
CrossTable(Year, Emp_rate, 3)
LOAD Countries, Sex, Age,
[2002],[2003],[2004],[2005],[2006],[2007],[2008],[2009],[2010],[2011],[
2012],[2013]
FROM [lfsi_emp_a-final.xls](biff, embedded labels, table is [emp-rate-
females(25-54)$]);
```

```
CrossTable(Year, Emp_rate, 3)
LOAD Countries,Sex,Age,
[2002],[2003],[2004],[2005],[2006],[2007],[2008],[2009],[2010],[2011],[
2012],[2013]
FROM [lfsi_emp_a-final.xls](biff, embedded labels, table is [emp-rate-
females(55-64)$]);
Year:
Load num(Year) as Year resident Employee_rate;
Sex:
Load Sex resident Employee_rate;
Countries:
Load Countries resident Employee_rate;
Employee_rateFinal:
Load num#(Year) as Year, Sex , Countries, Age, Emp_rate, autonumberhash2
56(Year,Countries,Sex) as cheie
Resident Employee_rate;
Drop Table Employee_rate;
Key:
Load distinct cheie, Year, Countries, Sex
Resident Employee_rateFinal;
Join(Key)
Load distinct cheie, Year, Countries, Sex
Resident Employmentfinal;
Drop fields Year,  Countries, Sex from Employee_rateFinal;
Drop fields Year,  Countries, Sex from EmploymentFinal;
```

The figure 7 shows the associative model. This model looks like a dimensional model (star schema).

Associative model

Figure 7



ASSOCIATIVE ANALYSIS

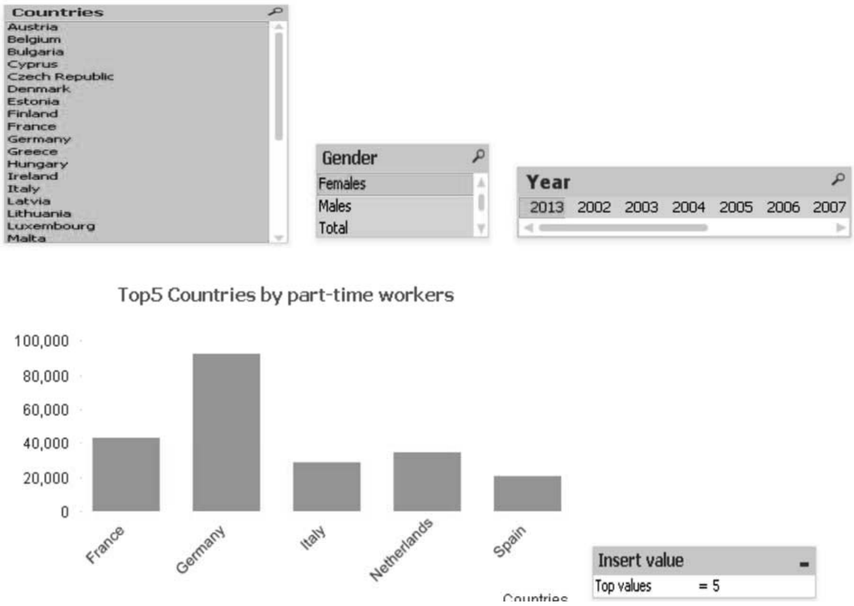
A large variety of powerful analytics are available with *"associative"* model such as: aggregations on-the-fly, set analysis, comparative analysis, conditional analysis, calculated dimensions, and so on (Redmond, 2013). Using this model, we can perform a diversity of analyses such as:

- employee rate analysis based on age group (15 to 24 years, 25 to 54 years and 55 to 64 years), gender (females, males and total), countries and year (2002-2013);
- total employment analysis based on gender, country and year;
- employees with temporary contracts analysis based on gender, country and year;
- employment (15 to 64 years) analysis based on gender, country and year;
- part-time workers analysis based on gender, country and year, etc.

I will now present some examples. Figure 8 shows the top 5 countries in 2013, using as variable: *part-time workers*. We observe that Germany is number one. The female part-time workers rate is 46.1% in 2013, in Germany. In Romania is only 10.8%.

Top 5 countries in 2013 using as variable *part-time workers*

Figure 8



I created three list boxes (*Year*, *Gender* and *Countries),* a bar chart and an input box. We can make any data selection (any combination of years, countries and gender). Also, we can insert into the input box, the number of countries that will be displayed (for example Top 5). We can select any value for any attribute and all the related data from the entire data model will be displayed (associative search). Selection is green, unrelated data are gray and associated data is white. The users can explore data knowing only associated facts. They can find data even when they don't know the data structure.

Also, employment rates can be calculated for a particular age group and/or gender in a specific country and a particular period. The figure 9 shows a comparative analysis using alternate states between *GroupA* (European Union, Year=2013, gender="Females") and *GroupB* (country="Romania", Year=2013, gender=Females) for each age group. We observe that the employment rate for Romania is lower than EU rate. We can select any country, any year and any gender. The following expressions are used:

Expression 1: avg({[Group A]<Age=p(Age)>} [Emp_rate]) with label Group A

Expression 2: avg({[Group B]<Age=p(Age)>} [Emp_rate]) with label Group B

Expressions set data sets used for the calculations, for examples selections of alternative state Group A / Group B and selections in *Age group* list box (Redmond, 2013).

A comparative analysis using alternate states

Figure 9

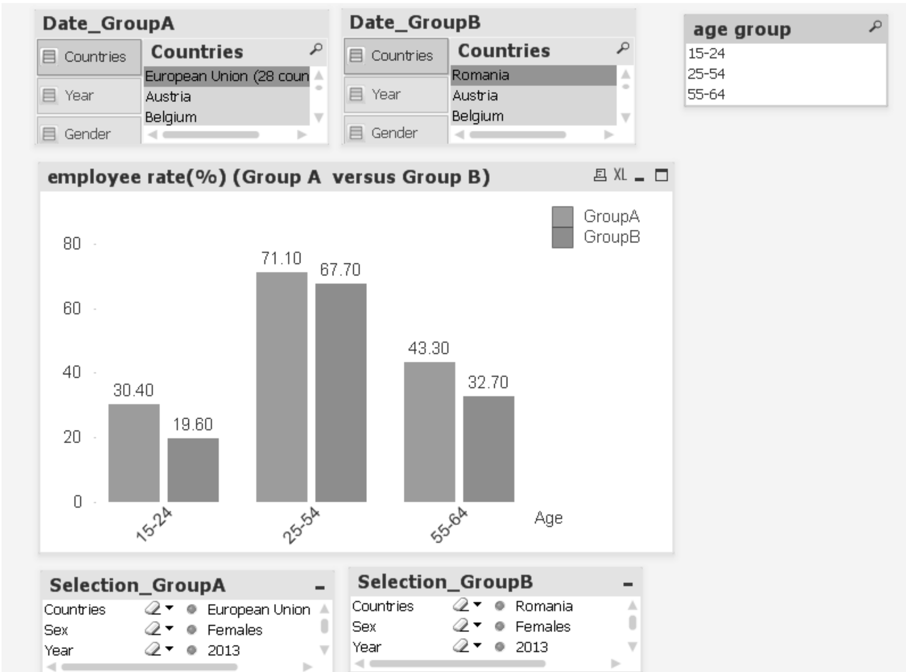


Figure 10 shows the comparative analysis between Germany and Romania in 2013, using the same bar chart and list boxes. We can select any variable from cyclic group and the bar chart is updated automatically. Also, we can select any value from any list boxes and the bar chart is updated.

A comparative analysis between Germany and Romania in 2013

Figure 10

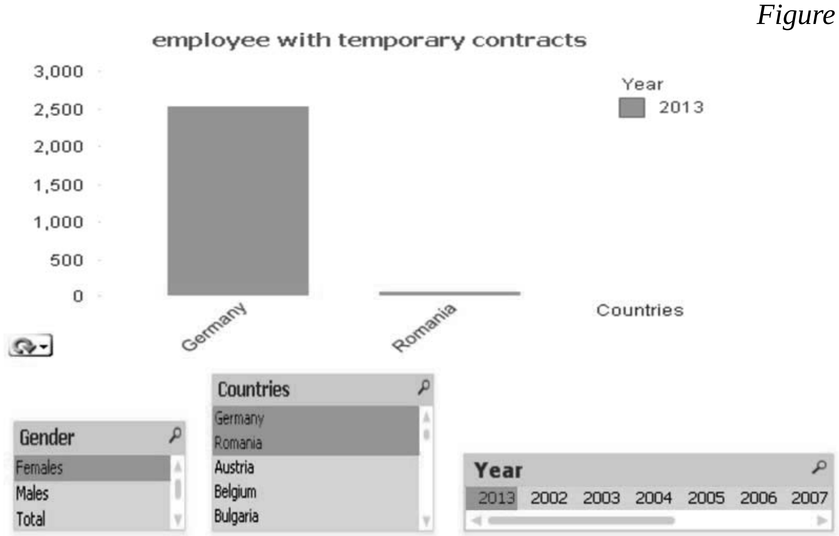
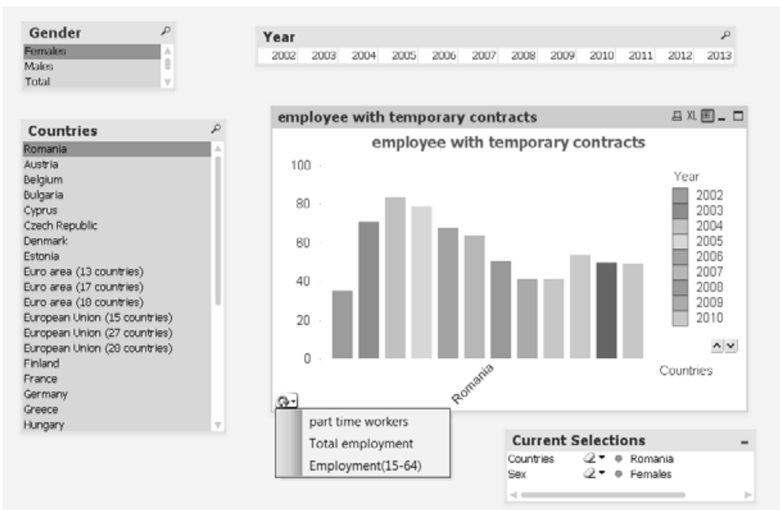


Figure 11 shows the evolution of female employees with temporary contracts in Romania, for period 2002-2013. We note that the number of female employees with temporary contracts is highest in 2004.

Evolution of female employees with temporary contracts in Romania, for period 2002-2013

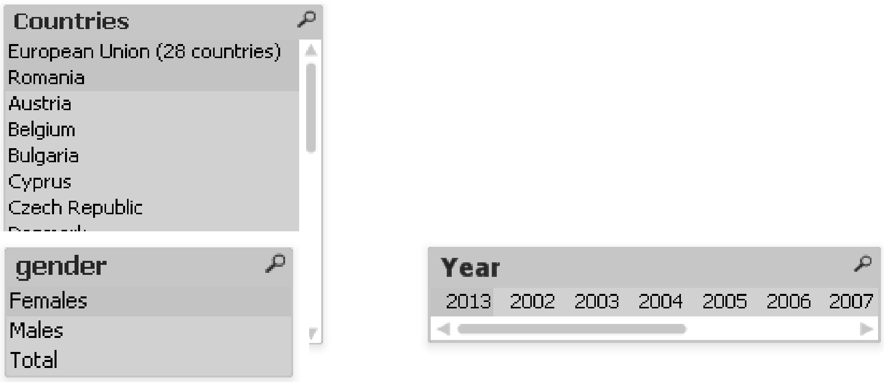
Figure 11



Time is really important in analysis. Comparative performance metrics over a period is a fundamental task from any analytical solution. Figure 12 shows a comparative analysis using different time periods: current year (for example, 2013) versus previous year (2012) versus two years earlier (2011).

A comparative analysis using different periods

Figure 12



Countries	employment selected year	employment previous year	employment two years ago	Employment growth over the previous year
European Union...	97632,7	97562,3	97476,9	0,0722%
Romania	3921,9	3933,9	3900,8	-0,3050%

We select from *Countries* list box only European Union and Romania. The indicator is *employment* (15 to 64 years). We compare results for three different time periods in one single view based on the same selection state. The comparisons are dynamic and based on the user's selections. Also, figure 12 shows employment growth over the previous year (for example, 2013 versus 2012). The female employment growth is negative for Romania, but positive for EU. The units of measure depend on the indicator, for example, people (thousands) or rates.

CONCLUSION

In conclusion, this paper explored how you can use in-memory analytics and an associative model to analyze labor market indicators. The flexibility and the advantages of associative model and in-memory technology for statistics are obvious: easy to use, speed of deployment and unexpected business insights via an associative experience. We can build an analytical solution faster than any other tool. Users and developers can remain focused on insights and outcomes. Also, more flexible data model allows data to be imported with fewer transformations.

Acknowledgement

This paper are based on the results of the EU Labour Force Survey (EU-LFS).

REFERENCES

1. Morrison, A. (2012), "The art and science of new analytics technology", Pwc Technology Forecast, issue 1, pages 30-44, available: http://download.pwc.com/ ie/ pubs/ 2012_technology_forecast_issue_1.pdf, accessed July 2014
2. Muntean, M, Surcel, T.(2013) "Agile BI – The Future of BI", Informatica Economică, volume 17, no. 3/2013, pages 114-123, available http://revistaie.ase.ro/content/67/10%20-%20Muntean,%20Surcel.pdf
3. Redmond, S. (2013), Qlikview for developer cookbook, Birmingham: [Packt] Publishing Enterprise, 2013, pp.106-108, 123-126, 153-157 and 369-400
4. Qlikview Reference Manual, vers. 10, (2010), pages 1434, available: http://www.qlik.com/, accessed January 2014
5. Sterry, B. and Nieuwenhuijs J. (2011), Qlikview project methodology handbook v.1.0, available: http://www.slideshare.net/AntoninoBarbaro/qlik-view-project-methodology-handbook-v-10-docx, accessed July 2014
6. Sallam, R. L., Tapadinhas, J., Parenteau, J., Yuen, D., Hostmann, B. (2014), "Magic Quadrant for Business Intelligence and Analytics Platforms", working paper, ID: G00257740, Gartner Group Inc., USA, 2014 Available http://www.ioz.pwr.wroc.pl/ Pracownicy/lubicz/DSS/ HOMEWORK/Magic-Quadrant-for-Business-Intelligence-and-Analytics-Platforms.pdf, accessed June 2014
7. Qlikview white paper (2010), "The associative experience: Qlikview's overwhelming advantage", 2010, pages 1-11, available: http://www.axisgroup.com/library/brochures-whitepapers-case-studies-new/whitepapers/29-qlikviews-overwhelming-advantage/file.html, accessed octomber 2014