



Egypt's Census Bureau Goes Digital

A white paper by:
Dr. Ahmed A. Elragal
German University in Cairo
March 2011

Egypt's Census Bureau Goes Digital

Table of Contents

Executive Summary	2	BI Architecture: Layered-framework	9
What is the Egypt Census Bureau?	3	Project Implementation	10
Data that CAPMAS Collects, Owns, and Uses	3	Teradata Tools and Utilities at CAPMAS	12
Business Intelligence and Government Agencies	4	Data Mining Training I: What Went Wrong?	12
BI: Governments to Reap the Benefits	5	Data Mining Training II: the Triangle of Success	12
Problem Statement	5	Get Your Hands Dirty	13
Strategic Project	5	Problems Solved?	13
Project Scope	5	Data Mining Scenarios	14
Proof of Concept	6	Appendix I: CAPMAS Organization Structure	15
The Data Warehouse at CAPMAS	6	Appendix II: Data in the Subject Areas Census	16
Objectives of the Data Warehouse	6	Appendix III: Information On Demand Details	17
EDW: Technology and Business – the Inept Couple	7	Appendix IV: Understanding BI	18
Project Challenges	8		

Executive Summary

In this case study, Egypt's Census Bureau (CAPMAS) is described in terms of how it uses business intelligence to support its mission and objectives. The case is organized as follows: CAPMAS is described in terms of its mission, data used, agencies served, and the on-demand information services it offers. Then, the fundamentals of business intelligence (BI) will be explained, where potential benefits to government agencies will be introduced. Following that, the problem statement is introduced together with project scope and the proof of concept (POC). Project challenges and implementation details are then illustrated. In the end, there are future outlook and questions for discussion.

Egypt's Census Bureau Goes Digital

What is the Egypt Census Bureau?

CAPMAS¹, the Central Agency for Public Mobilization and Statistics, was established in 1964 by Egyptian Presidential Declaration number 2915. Its mission is to collect, own, analyze, and distribute all statistical data for Egypt. In addition, CAPMAS is responsible for Egypt's census data. The objective of CAPMAS is to standardize statistical measures and create information that supports strategic and developmental decisions.²

CAPMAS was experiencing difficulties fulfilling its mission because it relied on multiple data sources that were not easily linked together. Generating reports from these multiple sources was difficult. Also, there was a need to mine the data to gain information to support decision making. As a response to these challenges and needs, CAPMAS management decided to invest in an enterprise data warehouse (EDW) and data mining tools to analyze the data better.

With the help of Teradata Corporation as a partner, CAPMAS has been successful in developing an EDW and using data mining to support decision making. This journey has involved challenges, however, and there is much to learn from how the challenges were addressed and overcome.

Before discussing the development of the EDW and the use of data mining at CAPMAS, let's understand the importance of CAPMAS better, the kinds of data it collects and analyzes, the agencies it serves, and the potential of business intelligence in government organizations.

Importance of CAPMAS

CAPMAS provides important services to Egypt:

- It is the source of census data.
- It publishes important indicators (e.g., census, industry, services, tourism, water resources, justice and security, communications, housing, and social services).
- Based on its indicators, many government decisions are made.
- It provides information to government and private organizations.
- Investors use its publications to make business decisions.
- Its on-demand information service provides a resource for information seekers (individuals and organizations) to get any information they need.

Data that CAPMAS Collects, Owns, and Uses

CAPMAS maintains a large repository of data that is collected from various sources. The data are used to provide information to government agencies through either periodic or on-demand reports.

CAPMAS collects and operates many databases, including those for census, buildings, businesses, and many others. CAPMAS collects data through the use of nearly 100 different forms. Generally speaking, these forms fall into two broad categories: static and dynamic. The static forms collect snapshot data (e.g., census forms), while dynamic forms collect data continuously throughout the year (e.g., birth and mortality).

Figure 1 shows a model of the various categories of data that CAPMAS collects and maintains. This model also reflects the subject areas in the EDW.³

1 Web: www.capmas.gov.eg

2 Appendix I describes the organization structure for CAPMAS.

3 Appendix II provides detailed descriptions of the data in each subject area.

Egypt's Census Bureau Goes Digital



Figure 1. CAPMAS Model Subject Areas.

Agencies Served

CAPMAS provides information services to a variety of organizations. Government agencies and public administration units are the principal service recipients. Private sector organizations, researchers, and individuals can also apply for CAPMAS' information services (e.g., data collection, periodic reporting, and national statistics).

Information On Demand

CAPMAS publishes periodic reports and statistics for different areas. Some of them go directly to specific beneficiaries, while others are published on the CAPMAS website (www.capmas.gov.eg). Most of CAPMAS' publications rely on data collected through forms. Sometimes CAPMAS does not have required data, which leads to the design of a custom form that enables collection of the needed data. This is referred to as the on-demand information service.

Researchers and investors can apply for the on-demand information service to obtain information about any of the following: distribution of population, classification of

buildings, and classification of enterprises.⁴ CAPMAS also provides on-demand information service upon the request of any government agency or administration unit.

Business Intelligence and Government Agencies

Information and communication technologies bring considerable benefits to government agencies, including better service delivery to citizens, improved interactions with business and industry, citizen empowerment through access to information, and more efficient government management. BI is able to support the decision-making processes and enhance operations.

Experience shows that integrating BI⁵ into government agencies and public administration units can result in these benefits:

Integrated data sources

A data warehouse facilitates access to integrated data. For example, the Family Database is one of the projects in Egypt that aims to link all relevant family information from various government agencies and make it available for decision support. The Family Database is based on data collected from the ministries of education, higher education, interior, social solidarity, and other administrative agencies. Making a decision based on data gathered from these agencies without a data warehouse would take longer and cost much more, if feasible at all.

Classification

Classification means structuring unclassified data and can be achieved using cluster analysis (a data mining technique). For instance, if we want to predict the labor market needs, it is not sufficient to say we need 3,000 graduates. Instead, we need to cluster those by discipline, living areas, experience, and verticals (e.g., banking, retail, oil and gas).

⁴ Appendix III identifies the kinds of data available in each category.

⁵ Appendix IV provides detailed description of BI.

Egypt's Census Bureau Goes Digital

Predictions

Predictions and forecasts can be made using various mining techniques (e.g., regression analysis) that utilize data stored in the data warehouse. For example, data mining techniques can be used to predict the number of patients in a specific rural area, the number of students graduating from a university, or labor market requirements.

Pervasive analytics

It is important to provide BI to administrative decision makers throughout the organization in order to support decision making.

BI: Governments to Reap the Benefits

Governments are at the top of the list when it comes to making strategic decisions, and government decisions need to be supported by BI. For example, let us assume that a government of a developing country wants to make a strategic decision regarding the number of schools that should be established over the next two years. Without BI, the government may allocate the funds available to governorates by population, or illiteracy, or some other variable(s). However, if there is a central data warehouse that collects data from the various governorates and ministries involved, and users are given access to that central data store (i.e., the enterprise data warehouse), government officials can look at the problem under investigation from various perspectives and provide visual aids and prediction techniques. Of course, to reap the benefits, the BI challenges must be properly addressed. Among the challenges are data integration, platform heterogeneity, scalability, availability, and security.

Problem Statement

Generating strategic reports out of multiple data sources has always been a challenge for CAPMAS. Linking the different

database systems to each other took considerable time and effort. Dealing with too many scattered data sources had impaired CAPMAS' ability to fulfill its mission. Consequently, there was a consensus within CAPMAS to integrate these various data sources in a state-of-the-art platform designed to support decision-making processes. This was the beginning of CAPMAS' journey towards data warehousing and data mining.

Strategic Project

CAPMAS decided to extend its mission beyond the distribution of data to include the use of advanced analytics. CAPMAS management identified the creation and sponsorship of a data mining unit as a strategic objective.

Project Scope

The project included two major milestones:

1. The design and realization of a data warehouse.
2. The use of data mining techniques in problem solving and decision making.

Feet on Data Warehousing, Head in Data Mining

While the project started with the data warehouse, data mining was the ultimate goal. CAPMAS management made it clear that the project's success would be measured by the ability to successfully complete data mining projects.

The CAPMAS team went through multiple training sessions on the use of data mining techniques. However, the results did not meet the expectations of either Teradata or CAPMAS management. After several meetings between the two parties, it was concluded that case studies were needed that included (1) a specific problem, (2) data related to the problem, and (3) an opportunity to apply data mining techniques. An external consultant was hired to develop these specialized training materials.

Egypt's Census Bureau Goes Digital

Proof of Concept

CAPMAS was aggressive in its efforts to find appropriate, qualified partners. It invited major companies specializing in data warehousing and/or data mining to perform a proof of concept (POC). The POC was designed by CAPMAS consultants to assess (1) the competency of potential partners and (2) the power and ease-of-use of possible data mining tools.

Microsoft was invited, but did not participate in the POC. Sybase completed the POC, using Sybase as the DBMS engine and Clementine as a data mining tool.

Teradata was invited and participated in the POC. Throughout the POC, Teradata was able to demonstrate its powerful technology and the ability of its Professional Services (PS) consulting team to complete complex projects. Based on the outcome of the POC, CAPMAS selected Teradata as its partner.

The Data Warehouse at CAPMAS

Because CAPMAS has many data sources, the integration of all of them was not possible in a single phase. Instead, the data sources were grouped by subject areas and were added to the EDW in two phases (See Figure 2.).

The first phase, the foundation phase, included the Census data, which CAPMAS considers the *core and crux* of their data warehouse. The second phase is still under development as shown in Figure 2.

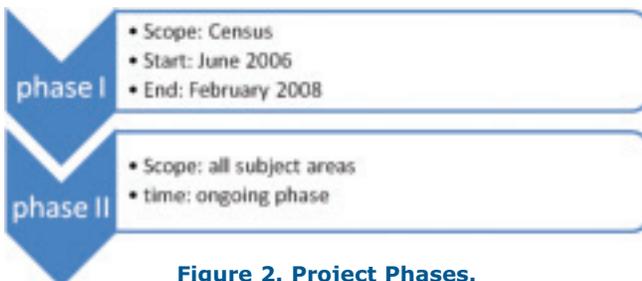


Figure 2. Project Phases.

In 2007, after the success of establishing a data warehouse for the census, Teradata completed the data warehouse by adding more sources to the data repository (e.g., industry and services) and has linked it to the previous census data in a way to maximize the use of the data warehouse.

Objectives of the Data Warehouse

The objectives of the data warehouse are to:

- Provide accurate, up-to-date information to support decision makers at the various government agencies cross levels. This will enhance the decision quality and planning efficacy.
- Provide information on demand.
- Provide investors with investment-related intelligence (e.g., demographics, age distribution, investment map (current production of each product, pattern of production, adequacy of production), and export opportunities).
- Rationalize government spending in the areas of public services by identifying potential service beneficiaries based on eligibility reports.
- Based on the records of the Ministry of Finance, items held in inventory will be made available to all other government agencies to consider before making purchase decisions.
- It is also among the objectives of the CAPMAS' data warehouse to serve these government activities by providing the required intelligence:
 - > Defense and public safety
 - > Public assets
 - > Register citizens
 - > Establish and maintain public infrastructure
 - > Manage social development programs
 - > Collect public revenues and associated expenditures
 - > Enact and enforce laws

Egypt's Census Bureau Goes Digital

- It is also among the objectives of CAPMAS data warehouse to support the decision makers in all government agencies across levels in performing these functions:
 - > Planning
 - > Monitoring
 - > Providing the required intelligence to support decisions in solving economic, social, health, education, and population problems

EDW: Technology and Business – the Inept Couple

Building an EDW is not a typical system development project. It requires a blend of both technical and business experience. Business people's involvement in the EDW is a major success factor. Teradata PS consultants realized this fact and understood the importance of involving business people and succeeded in getting them and top management involved in the project and setting expectations properly.

The Teradata PS team studied CAPMAS' business even before the POC, which allowed them to efficiently plan and conduct the implementation.

The CAPMAS EDW is a multi-subject, centralized data warehouse. It is considered a back-office system, not an "active"⁶ data warehouse. Data feeds to the EDW come from surveys, not transactions. Surveys are done on a regular basis or upon request. The frequency of the surveys differs based on the nature of the survey; surveys can be done quarterly, twice yearly, annually, every two years, or every ten years. The CAPMAS EDW is connected to CAPMAS network; however, data are pushed to the EDW on demand.

So far, the CAPMAS data warehouse is the sole organization source for analysis and advanced analytics with respect to

census data. Eventually, as more subjects are added, it will be the source for all analysis and advanced analytics at CAPMAS.

Currently, only the data warehouse (DW) team has access to the EDW, and all analysis required must go through this centralized team. The team has all the resources required to manage the data warehouse, including two DBAs and an operational team responsible for normal operation and regular backup. The CAPMAS DW team has personnel with different backgrounds. It includes developers and analysts with solid business knowledge. Teradata training was bundled with the solution, which positively impacted the development and use of the data warehouse.

The CAPMAS team is capable of enhancing the system following the same methodology that was used in expanding the logical and physical models, and developing ETL scripts and reports. Teradata PS consulting skills will be required for new business requirements that require major changes to the EDW architecture and also for advanced analytics.

Figure 3 explains the process end-to-end.



Figure 3. CAPMAS BI Project – the Big Picture.

⁶ An active data warehouse (ADW) is a special type of data warehouse that supports real-time or near-real-time decision making. It is featured by event-driven actions triggered by a continuous stream of queries (generated by people or applications) against a broad, deep granular set of enterprise data. Teradata defines ADW as a traditional data warehouse extended to provide operational intelligence based on historical data combined with today's up-to-date data.

Egypt's Census Bureau Goes Digital

Project Challenges

The project faced many challenges including time, budget, technical, and people issues:

- **Time:** the challenge was to get the data warehouse ready for the 2006 census. That was a time constraint since the project started in 2006, and the census data come in handwritten documents. Transferring the manual into an electronic format, designing the DW, and populating the data into the DW with all the necessary ETL and cleansing efforts was a challenge in terms of the timeframe available. The challenge was met by a decision from Teradata to classify the project as a strategic one and to put all needed resources on the project to finish on time.
- **Budget:** CAPMAS is a huge organization with many key arms that compete for a limited budget. For instance, the project required advanced ETL tools (e.g., IBM Data Stage or Oracle Data Integrator), but due to insufficient budget, the team had to use the basic Teradata ETL tools (e.g., Teradata FastLoad, Teradata FastExport, and Teradata MultiLoad).
- **Technical:** CAPMAS has no formal data quality (DQ) process. However, the Teradata team together with the CAPMAS team put a process in place for DQ issues that may arise when loading census data into the data warehouse, and along with the data collection departments, they agreed on a process for data quality and cleansing. The DW team has a good understanding of the nature of the data so they developed scripts to examine the quality of the data. See Figure 4.

The DW team has a good understanding of the nature of the data and how they are collected and processed. CAPMAS data collection departments collect data about each census subject area (i.e., buildings, people, public houses, and businesses) separately. Subsequently, the integration of the sources takes place in the DW.

The DW team developed scripts to examine quality of data. For example, during the loading of business buildings data records, we make sure that their associated business does exist, otherwise data are rejected and sent back for verification.

People: CAPMAS management made it clear that they would like to utilize CAPMAS resources to implement this project led by Teradata consultants and use phase one as hands-on training for the CAPMAS team. The Teradata team was to take all the necessary actions to ensure a successful implementation of phase one, including training the CAPMAS team to implement further phases with Teradata support. The introduction of analytics to CAPMAS was a challenge for the Teradata team. The CAPMAS team used to produce a large set of well-defined static reports and was not used to business questions that required evaluation and planning, plus knowledge of analytical skills, including drilldown and "slicing and dicing" data. The CAPMAS team also did not have the basic knowledge of data mining that is required for answering the business questions that require the use of data mining techniques.

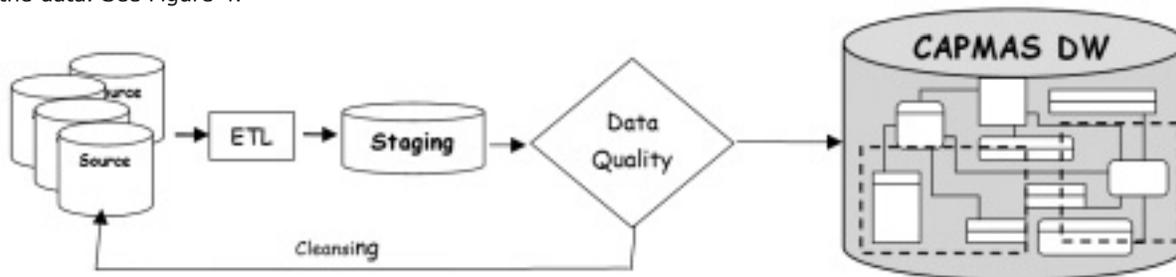


Figure 4. The Approach to Data Cleansing.

⁷ Slicing and dicing is the ability to move between different combinations of dimensions when viewing data with OLAP tools. It is also a term which refers to the ability to combine dimensions to see different slices of information.

Egypt's Census Bureau Goes Digital

BI Architecture: Layered-framework

Figure 5 shows the end-to-end CAPMAS BI architecture:

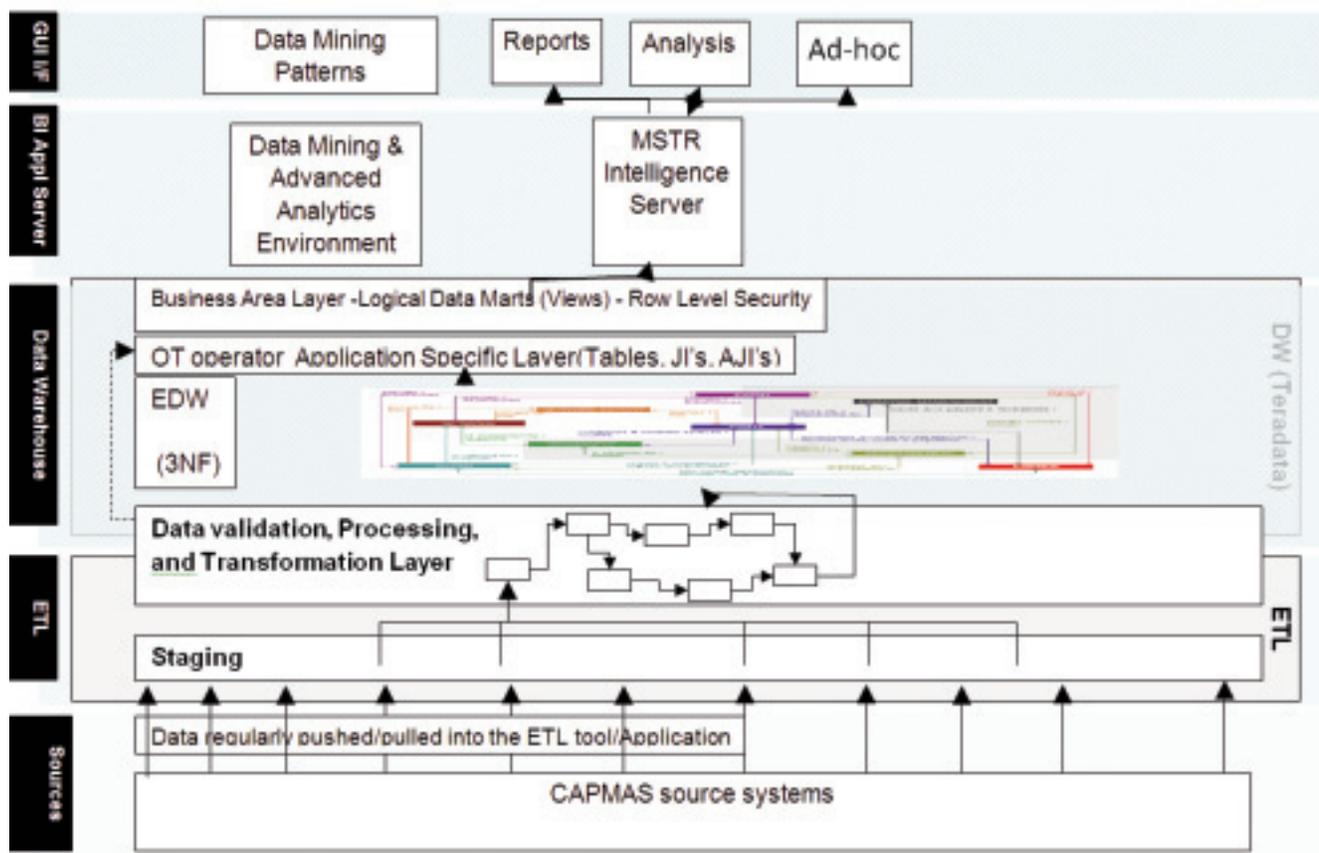


Figure 5. BI Architecture.

The following components are used in the BI architecture:

Source systems

This layer represents all interfaces with the source systems; it manages and controls the receipt of and the validation of the files. Each of the source systems sends files directly to the file storage server and the ETL framework. Also, it accepts direct access to the sources, depending on the detailed data acquisition requirements. The management of all ETL processes is done manually.

ETL tools

Teradata tools have been used during the data staging and as a transformation hub. Data are accepted from all the source systems as required, then bulk Teradata load utilities are used to load the data into staging tables within Teradata Database. The ETL application developed using Teradata tools performs the transformations and applies them to the data.

Data warehouse platform

The Teradata Database used a 3rd normal form (3NF) data model. This layer, staging, serves as the Integration layer,

Egypt's Census Bureau Goes Digital

which is implemented using the Teradata RDBMS platform. The CAPMAS EDW is populated by executing as much as possible on the Teradata system, to utilize the "heavy lifting" capabilities of Teradata Database. The other layer, integration, represents implementation of the data model in 3NF, and data are stored at the lowest granular level. On top of this 3NF layer, users have sets of views that support business usage of the data (to simplify access).

BI application server

MicroStrategy: This layer represents all application servers, MicroStrategy, and the sets of virtual data marts used for data mining.

GUI reporting and analytics

This layer represents tools that are utilizing application servers and a combination of fat and thin-clients⁸. In addition, this layer represents tools that are directly accessing Teradata Database, such as Teradata Warehouse Miner, SQL Assistant, and other direct query tools.

Project Implementation

CAPMAS top management considered the data warehousing and mining project to be strategic. From the beginning, management made it clear that they wanted to use CAPMAS personnel to save expenses. Teradata consultants managed the project and helped train the CAPMAS IT team. The Teradata team took the necessary actions to ensure a successful implementation of phase I, including training CAPMAS personnel so that they could implement later phases with minimal support from Teradata.

The scope of phase I included building the CAPMAS data warehouse, populating it with detailed data from the 2006 census, and using these data to find hidden patterns of new trends or unknown relationships. At the end of phase I, an

appropriate data mining environment was in place. The technology components for phase I included:

- An SMP server for development
- Teradata MPP Server Model 5500 EC for production
- Teradata Database V2R5, Windows OS
- Teradata tools and utilities
- MicroStrategy as a BI tool
- Teradata Warehouse Miner as a data mining tool

Building the logical data model (LDM) was considered a core project milestone. The LDM was designed not only to integrate Census data, but also to take into consideration future phases of the project. The development of the LDM and design of the CAPMAS data warehouse was performed by Teradata consultants.

One of the challenges the consultants faced was that CAPMAS source systems use different database management systems, including Oracle, Sybase, and Microsoft SQL. Teradata ETL tools were used to extract data from these sources and then transform and load them into the data warehouse.

Another challenge was the lack of formal data governance process at CAPMAS, which made it difficult to address data quality issues. To overcome this problem, Teradata, together with the CAPMAS team, created a process for ensuring data quality. This process included implementing data quality and cleansing procedures agreed to among the departments that are responsible for the source systems.

Introducing data analysis tools and analytics to CAPMAS proved to be a challenge for the Teradata consultants. Previously, the CAPMAS team produced a large set of well-defined reports. Also, CAPMAS personnel did not have a basic knowledge of data mining. Consequently, the team had to go through extensive data mining training.

⁸ A thin client is a machine or a program which depends on server computers to fulfill its computational duties. In contrast, the fat client is designed to carry on these duties on its own.

Egypt's Census Bureau Goes Digital

When building the data warehouse, the Teradata team emphasized the importance of having a structured presentation layer over the normalized logical data model in order to make the organization of data in the warehouse easier for users to understand. This layer was used for building reports.

MicroStrategy's software was used for OLAP-type reporting applications. After training, the CAPMAS team was able to master the use of the software. They were also able to use Teradata SQL tools and MicroStrategy to respond to ad-hoc queries for information.

The CAPMAS data warehouse environment is shown in Figure 6. The major components include the ETL server, Teradata data warehouse platform, and MicroStrategy Intelligence Server.

Each source system sends data files to the ETL server. The ETL server is a file storage server with enough space for staging the data. The ETL application runs on the ETL server

to validate, transform, and load the data to the data warehouse. The data warehouse is a Teradata MPP server model 5500EC running the Teradata Database and is connected to Teradata 6843 external storage that has 3.5TB of raw data. Teradata Administration Workstation (AWS) is used for administrating the MPP system.

Information delivery in CAPMAS takes different forms, including reporting, analysis, and data mining. For reporting and analysis, CAPMAS uses the MicroStrategy OLAP tool and MicroStrategy Intelligence Server for core analytical processing and job management. MicroStrategy Intelligence Server makes it easy to standardize on a single, open platform for all enterprise reporting and analysis needs through a number of data access channels such as Web browsers, Microsoft Office, desktop clients, and emails. The OLAP application development team uses desktop clients for delivering required information, and the data mining team uses Teradata Warehouse Miner to develop models to detect patterns.

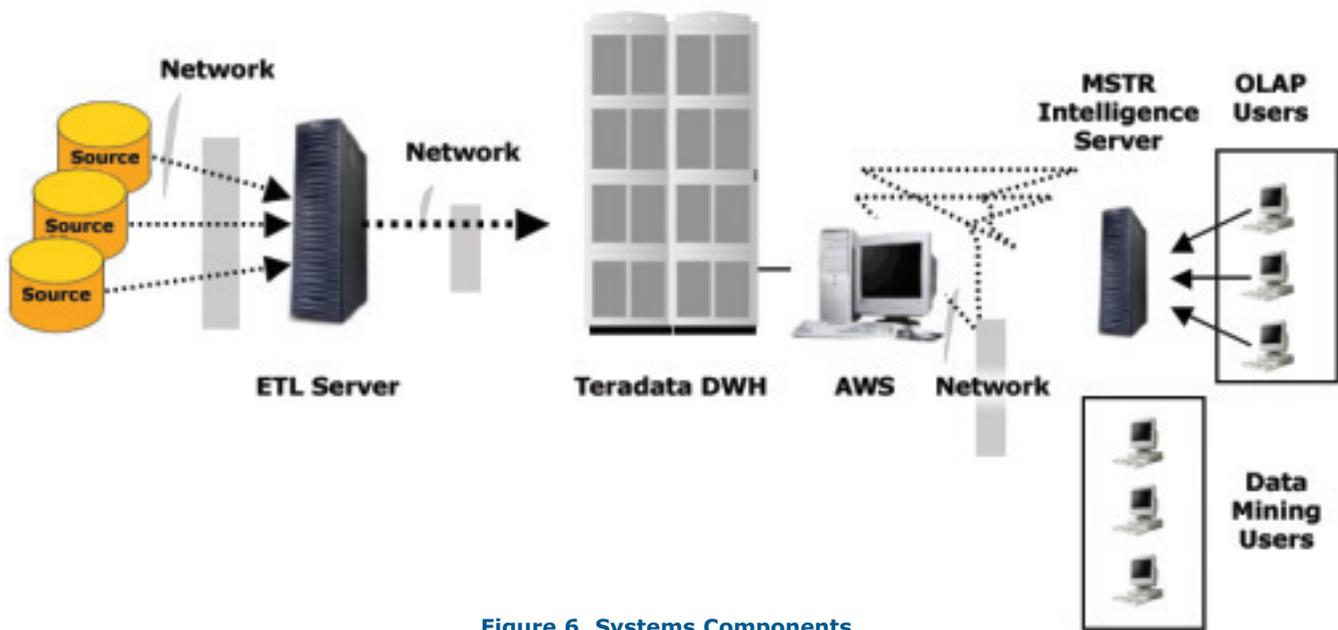


Figure 6. Systems Components.

Egypt's Census Bureau Goes Digital

Teradata Tools and Utilities at CAPMAS

These tools and utilities have been used by CAPMAS:

- Teradata TTU 8.2 - Teradata Utility Pack and Teradata Manager
 - > Access Teradata DBMS
 - > Query database
 - > Administration
 - > Backup
 - > Monitoring
- Teradata TTU 8.2 – Teradata FastLoad
 - > Load data for the first time to empty tables; FastLoad is a bulk load utility used to load the data into staging tables within Teradata Database.
- Teradata TTU 8.2 – Teradata MultiLoad
 - > Add data to an existing table; MultiLoad is used to integrate the new data loaded in the landed area to CAPMAS' logical data model.
- Teradata TTU 8.2 – Teradata FastExport
- Export data from a table; FastLoad is used to export specific sets of data that are required by other entities within CAPMAS.

Data Mining Training I: What Went Wrong?

As part of phase I activities, the CAPMAS data mining team received Teradata Warehouse Miner training. The training mainly focused on the tools and techniques. Unfortunately, once the training was finished, the team found that they were incapable of doing any mining analysis. This came as a

surprise to the CAPMAS top management officials, and they started an investigation. An external data mining expert was called in for a meeting with both CAPMAS and Teradata. The consultant had a chance to interview key members of the data mining team and wrote a report in which he highlighted the need for conceptual training and a proper analysis methodology.

Data Mining Training II: the Triangle of Success

The consultant's report suggested six training sessions based on a mixture of data, problems, and techniques. The last session was a workshop where the data mining team members were divided into two teams to study two different problems and provide mining solutions. The six sessions were successfully conducted. The following topics were among those explained in the sessions:

- Data Mining: an overview
- Supervised versus unsupervised mining
- Evaluation of mining outcome
- Association rules
- K-Means
- Clustering
- Decision trees
- Regression analysis
- Artificial neural networks (ANN)
- Advanced techniques
 - > Bayes theorem
 - > Genetic learning (GL)
- Case study workshop: "Presentations by Analysts"

Egypt's Census Bureau Goes Digital

Get Your Hands Dirty

Getting their hands dirty on linking problems, techniques, and data together was the objective of the workshop. That was achieved by dividing the team members into two groups:

- **Group A** consisted of five team members who studied the "illiteracy problem" using various data mining techniques including decision trees, association rules, clustering, and regression analysis.
- **Group B** consisted of four team members who studied the "ideal family" using various data mining techniques, including decision trees, association rules, clustering, and regression analysis.

Problems Solved?

During the meetings with Teradata, CAPMAS, and the external expert consultant, two problems were detected – the conceptual problem and the methodology problem. Here is what has been done to solve each:

The conceptual problem

The team members were trained on the algorithms behind the techniques, and that was followed by a training session on Teradata Warehouse Miner with examples. For each technique, the core concepts were numerically and algorithmically illustrated:

- **Decision trees:** information gains, expected value of information, confusion matrix
- **Regression analysis:** independent variables, dependent variables, least squares, R-square, assumptions, t-test, p-value, stepwise-regression, degree of freedom, f-ratio, multi-collinearity, logistic regression
- **Clustering:** k-means, gauss, proximity measures, Euclidean distance
- **Association rules:** Apriori algorithm, support, confidence, lift, z-score
- **ANN:** feedforward connected network, MLP, RBF, training and evaluation sampling

- **GL:** fitness function, operators: cross-over, mutation, selection, combining GL with unsupervised clustering [not in Teradata Warehouse Miner so team was directed to WEKA (Waikato environment for knowledge analysis)]
- **Bayes theorem:** prior and posterior probabilities [same as previous]

It is worth mentioning that before the sessions started, the team was mainly using decision trees. However, during the training, the other three basic techniques (association rules, regression analysis, and clustering) were used by the team extensively. This change is a direct result of the training sessions.

The Methodology Problem

The team was trained on the basic concepts of data mining. Mainly, the link between data, problem, and techniques was highlighted, and team members were frequently advised to follow an appropriate methodology. Figure 7 shows what the team was trained to follow:

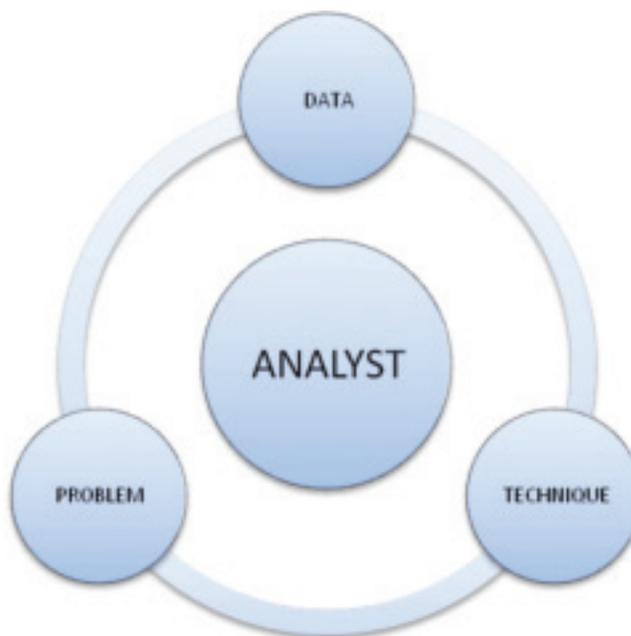


Figure 7. The Data Mining Triangle.

Egypt's Census Bureau Goes Digital

Data Mining Scenarios

In an attempt to use the pattern recognition capability of data mining, CAPMAS studied different scenarios to explore what data mining could add in terms of analyzing important national problems. Below are some of the data mining scenarios used at the exploratory phase by CAPMAS analysts to dig deep for hidden patterns and unknown information.

Child Labor⁹

There is an intense interest by many government agencies in understanding this phenomenon so that they are able to treat it. The analysts knew that child labor differs by geography, family income, parents' level of education, and other factors as well. With standard SQL, analysts are unable to identify important relationships; therefore data mining techniques were adopted. Table 1 shows how the analysts used data mining techniques in order to deeply understand the child labor problem, and hence take corrective actions.

Mining task	Independent variables	Dependent variable	Technique(s) used
Supervised and unsupervised	Geography, family income, education level, number of children, parents age	Child labor rates	Regression analysis, decision trees, association rules

Table 1. Data Mining and Child Labor.

Death-disease

The relationship between death and disease is of importance to many government agencies including governorates and the Ministry of Health (MOH). At CAPMAS, analysts started studying what are the most frequently prevailing diseases causing death in Egypt? Do they differ by geography? Are there job-disease-death significant associations? Data

mining techniques were adopted to help explore the relationships and answer those questions. Table 2 shows how the analysts used data mining techniques in order to explore the relationship between death and diseases.

Mining task	Variables	Technique(s) used
Unsupervised	Disease causing death, geography, job, age	Association rules, cluster analysis

Table 2. Data Mining and Death-Disease.

Unemployment

Knowing the areas that suffer the most unemployment has been the objective of some government agencies. Data mining was used to associate certain geographies, age categories, education level, and skills with unemployment. Table 3 shows how the analysts used data mining techniques to gain insights into the unemployment problem.

Mining task	Independent variables	Dependent variable	Technique(s) used
Supervised and unsupervised	Geography, age category, education level, skills	Unemployment rates	Regression analysis, association rules, cluster analysis

Table 3. Data Mining and Unemployment.

Divorce Rates

With the increase in the number of divorce cases, some agencies want to understand why divorces are occurring before taking corrective decisions. Mining techniques were used to find out: Which areas have highest divorce rates? What age categories? Which education levels? Which jobs contribute most? Table 4 shows how the analysts used data mining techniques to establish relationships and understand the divorce rate pattern.

⁹ Egyptian Law prohibits the employment of minors under the age of 14. Minors below 17 years are protected from employment in hazardous occupations. However, a 2001 national survey on child labor commissioned by the National Council for Childhood and Motherhood, and CAPMAS revealed that 2.7 million children aged 6-14 (21% of all children in that age group) work.

Egypt's Census Bureau Goes Digital

Mining task	Independent variables	Dependent variable	Technique(s) used
Supervised and Unsupervised	Divorce, geography, family income, education level, number of children, parents age	Divorce rate	Step-wise regression, Cluster analysis

Table 4. Data Mining and Divorce Rates.

Family Classification

It is in the favor of the whole society to increase the number of what CAPMAS calls the "Ideal Family." An ideal family is one where both parents are university graduates, they work, have a maximum of three kids, and own a flat with sufficient income. Various data mining techniques were used to answer: How many ideal families exist? Where? How to increase that number? How to predict the number of those families in the future? Table 5 shows how the analysts used data mining techniques to group together similar families, and hence make predictions.

Mining task	Variables	Technique(s) used
Unsupervised	Geography, family income, education level, number of children, parents age, health status	Cluster analysis

Table 5. Data Mining and Family Classification.

Appendix I: CAPMAS Organization Structure

The organization structure of CAPMAS consists of five sectors and four assistant-level units. Figure 8 shows the organization structure of CAPMAS followed by a brief explanation of what each unit does to fulfill its mission.

Assisting Units

- Consulting Committee:** The consulting committee plays a role in unifying standards of data, results, and statistics. It also has an advisory role to the president as well as the role of coordination between ministries and data owners.
- President's Office Administration:** This is a connection point between the president and all other sectors. It also works as a hub connecting CAPMAS president with domestic and international organizations.
- Legal Affairs:** The legal affairs group studies the legal aspect of all subjects submitted to the president's office. It also issues governing laws and procedures to ensure that work and projects in all sectors are performed on time and meet their objectives.
- Technical Office:** The technical office prepares the president's agenda of meetings, committees, and other ceremonial activities.

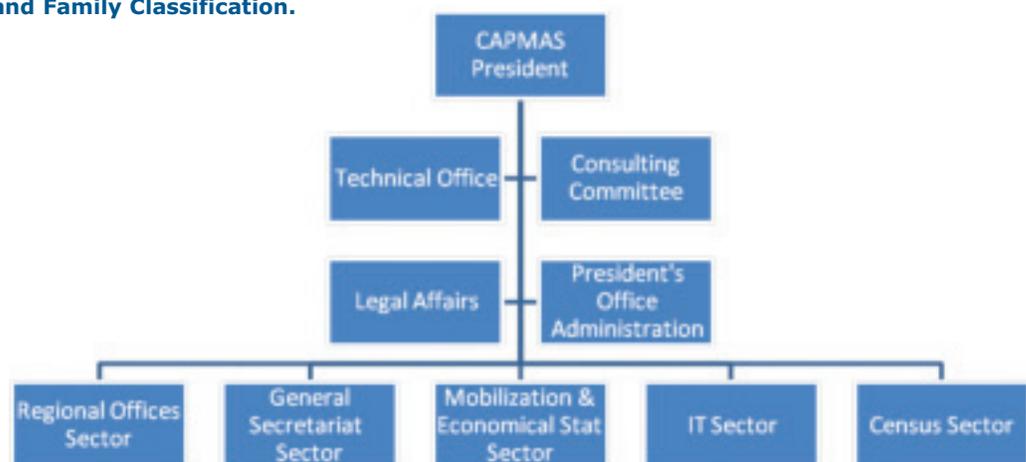


Figure 8. CAPMAS Organization Structure.

Egypt's Census Bureau Goes Digital

Sectors

- **Regional Offices Sector:** This sector is responsible for planning regional statistics needed by every region in the country according to preprogrammed plans. It also helps prepare all mobilization data about every region. Providing information specific to regions to those who need it is also among the responsibilities of this sector.
- **General Secretariat Sector:** This sector is responsible for the organization structure and personnel at CAPMAS. Training plans, employee fringe benefits, and social activities are part of its core duties. The sector also prepares the CAPMAS budget. The sector also prepares requests for proposals (RFPs) and manages the engineering and power projects.
- **Mobilization and Economical Stat Sector:** This sector is responsible for the annual plan of activities related to the mobilization and economic statistics. This includes issuing periodical reports related to all economic and mobilization activities. Further, the sector is also in charge of issuing related statistics and reports.
- **IT Sector:** This sector is responsible for designing integrated information systems databases, and geographic information systems (GIS). Technical support activities are also under the IT sector's umbrella. IT training of CAPMAS employees is another responsibility of the IT sector. The sector also takes full responsibility for the data warehouse.
- **Census Sector:** This sector is responsible for planning and executing census activities. This includes the census for people, businesses, and buildings.

Appendix II: Data in the Subject Areas

Census

"The census procedure takes place every ten years"

- Buildings: any house, villa, apartment, shop, tower, or business building
- People: families and their members

- Public housing: people living in public houses (e.g., hotels, dormitories, hospitals, etc.) during the census
- Businesses: detailed information about the business

Transport and communication

- Vehicles used in river transportation (public and private)
- River bus transportation statistics
- Public transportation statistics
- Railways transportation statistics
- Transportation activities in the unregulated private sector
- All other transportation statistics

Services

- Health services
- Youth care facilities
- Sports and social activities in education
- Nursery statistics
- Social solidarity indicators
- Hospitality indicators

Industry

- Annual industrial and commodity statistics in public and private sectors
- The quarterly statistics of industrial production
- Manufacturing in the unregulated private sector
- Production and distribution of electricity, gas, and steam
- Construction statistics in the public and private sectors
- Mining and quarrying in the unregulated private sector
- Financial indicators
- Monthly consumption of cotton
- Maintenance business statistics

Agriculture

- Stock of (raw) cotton
- Syndicates of agricultural credit
- Agricultural statistics
- Syndicate of water

Egypt's Census Bureau Goes Digital

- Average monthly levels of water in the Nile and inventory in the High Dam as well as Aswan High Dam
- Water needs of crops
- Lengths of canals
- Land reclamation
- Agricultural business of the private sector
- Production of honey and wax
- Hatcheries
- Leather production
- Animal diseases

Commerce

- Wholesale and retail statistics
- Wholesale trade activities in the unregulated private sector
- Storage facilities (for the barns and stores)
- Statistics on warehouse facilities (silos)
- Hospitality industry indicators and statistics
- Financial statements of private sector companies

Education

- Statistics of patents
- Higher education statistics
- Educational statistics (pre-university)

Human Resources

- Employment
- Salary surveys
- Injuries at work
- Public employments

Culture

- Movies
- Cinemas
- Theaters

- Foreign theater groups
- Books printed in Egypt
- Book shops and libraries
- Cultural centers
- Cultural associations
- Museums, zoo, aquarium, archaeological sites, and natural reserves
- Newspapers and periodicals

Birth and Mortality

- Mortality and birth
- Causes of death
- Marriage records
- Divorce records
- Court rulings on divorce

Appendix III: Information On Demand Details

Distribution of population according to

- Geography
- Gender (male – female)
- Income
- Age
- Type of housing (rent or own)
- Educational status
- Working status
- Nature of work
- Marital status
- Religion
- Nationality
- Main economic activity
- Health status
- Disability

Egypt's Census Bureau Goes Digital

Classification of buildings according to

- Building type
- Building ownership
- Building use
- Status of state buildings
- Unit type
- Inhabitants type
- Family classification

Classification of enterprises according to

- Work status and number of employees
- Labor sector
- Nationality of owner
- Legal entity
- Number of workers
- Business sector
- Type of facility being exploited
- First year in business
- Number of branches
- Connectivity with public infrastructure

Further, CAPMAS provides the following on-demand information service, as per the request of any government agency or administration unit:

- Governorates population according to gender and age, marriage and divorce statistics, employment versus unemployment rates, in addition to illiteracy rates
- Trade union statistics
- Medical statistics: number of doctors, nurses, number of beds (at both governmental and private hospitals), and the cost of state medical services
- Educational statistics: number of schools, classrooms, and students at all schools
- Public spending on education and health

- Number of telephone lines and vehicles in each governorate
- Statistics about traffic accidents and human injuries/death
- The volume and value of agricultural production by commodity groups. This is in addition to per capita food consumption
- Areas of reclaimed land
- Fisheries and fish production
- Manufacturing companies and associated manpower
- Number of vessels crossing the Suez Canal broken down by their capacity and destination
- Electricity generation and distribution
- Number of tourists classified by nationalities and number of nights spent
- Value-added activities
- GDP, foreign investment, and expatriates remittances
- Egypt's foreign trade
- Egypt's external debts
- Consumer prices' monthly indices
- Indicators in the areas of housing, health, environment, education, industry, transportation, utilities, energy, communications, media, banking, and capital markets

Appendix IV: Understanding BI

Recently, BI has attained a prominent role in operational and strategic decision making in many organizations, especially when there is a need to exploit the huge amounts of data that they own. The term BI was coined by the Gartner Group in the early 1990s. BI is an umbrella that includes tools, technologies, databases, applications, and methodologies. A major use of BI is to support the real-time access to and manipulation of data. By analyzing historical and current data, BI allows decision makers to gain valuable insights that support more informed decision making.

Egypt's Census Bureau Goes Digital

The BI process is transformative; wherein data are transformed into information and then into decisions. BI is not a product that is bought off-the-shelf. Its goals are achieved when different information consumers within an organization can leverage vast amounts of data that are collected and created to improve business performance. BI has these basic components:

- **Data warehouse:** a data warehouse is an archival data store used in support of the decision-making process. A DW has the power to integrate data from multiple sources to facilitate reporting and analyses.
- **Data mining:** data mining is a process of finding and interpreting hidden information in large data sets. A collection of tools and techniques is used in the mining process. Data mining uses a variety of techniques, such as association rules, cluster analysis, decision trees, and genetic algorithms.
- **Data presentation and visualization tools:** these are the tools that users run to display analyses' outcomes in an interactive manner (e.g., scorecards and dashboards, OLAP, and alerts and notifications).

Nowadays, government databases receive gigabytes of data that must be analyzed for decision making and other strategic purposes. Although IT has been used to supporting data analysis and decision making, applying BI with all its components in administration is only at the beginning. Neither research nor practice yet has a mature approach of how to realize the potential of BI in public administration. Linking BI to government decision making in an integrated framework is expected to deliver added value to public administration.

Questions for Discussion

1. What were the major project challenges?
2. How can CAPMAS, a government organization, measure the ROI of implementing data warehousing and mining technologies?
3. The case indicates that the data warehouse of CAPMAS is not spontaneously updated, and hence, it is not an "active" data warehouse. Evaluate this decision in terms of: why do you think they have made such decision? Do you think they have made the right decision? Why, or why not?
4. According to the data mining scenarios listed in the case, suggest other scenarios that you think could be useful for the census bureau.
5. Having studied the project phases and activities associated, what could have been done differently to achieve quick wins?
6. There are many data mining techniques, and some of them are not available with Teradata Warehouse Miner (e.g., time series). CAPMAS complained that the unavailability of those techniques has been a barrier to them. How do you evaluate this argument?
7. Hands-on: use any data simulation software to generate sufficient data and load them into Teradata Database as Census Bureau EDW. Then implement the before mentioned data mining scenarios and interpret results.
8. Web exercise: compare the mining scenarios taking place at CAPMAS with any other mining scenarios of other Census Bureaus worldwide.

About the Author

Dr. Ahmed A. Elragal is an Associate Professor of Information Systems at the German University in Cairo (GUC). You can reach Dr. Elragal at: ahmed.elragal@guc.edu.eg. Dr. Elragal was a winner of the 2010 Case Competition sponsored by Teradata University Network.

Teradata is a registered trademark of Teradata Corporation and/or its affiliates in the U.S. and worldwide. Teradata continually improves products as new technologies and components become available. Teradata, therefore, reserves the right to change specifications without prior notice. All features, functions, and operations described herein may not be marketed in all parts of the world. Consult your Teradata representative or Teradata.com for more information.

Copyright © 2011 by Teradata Corporation All Rights Reserved. Produced in U.S.A.