MZ:2007:03

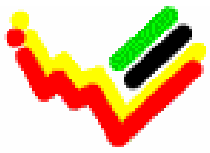
Mission Report

from a short-term mission on Data modelling and SQL

19 February- 1 March 2007

TA for the Scandinavian Support Program to Strengthen the Institutional Capacity of the National Statistics, Mozambique

Jesper Ellemose Jensen



Instituto Nacional de Estatística

Ref. no. 104.Moz.98 March, 2007

Jesper Ellemose Jensen
Statisitcs Denmark
Sejrøgade 11, 2100 Copenhagen Oe, Denmark
jej@dst.dk
+ 45 39 17 30 56

Table of contents

List of abbreviations

DBA	Database Administrator
CO	Scanstat Coordination Office in Statistics Denmark
Danida	Danish International Development Assistance
DKK	Danish Kroner
DSt	Statistics Denmark
EUR	European Euro
INE	Instituto Nacional de Estatística, Mozambique
INE-P	Instituto Nacional de Estatística, Portugal
IIS	Internet Information Server (A Microsoft product)
LTA	Long Term Advisors
STA	Short Term Advisors
MZM	Mozambique Meticais
NOK	Norwegian Kroner
PX	Family of Software produced by Statistics Sweden
Scanstat	Consortium between Statistics Denmark, Statistics Norway and Statistics Sweden
SCB	Statistics Sweden
SDMX	Statistical Data and Meta data exchange
SEK	Swedish Kronor
SSB	Statistics Norway
SQL	Structured Query Language
USD	US Dollars
XML	Extendable Mark up Language
ZAR	South African Rand

1 EXECUTIVE SUMMARY

Scope of the mission

The mission should be seen as a general introduction to and training in the use of the SQL language and data modelling. The Terms of Reference for the mission is included in the report as appendix 4.

For the training a combination of MS-Query and MS-SQL Server 2005 Express software was used. MS-SQL Server 2005 Express is a full database system and is provided free of charge from Microsoft.

Training sessions was conducted on 5 working days. The training focused on practical examples and the use of the software. A draft database, modelling the 2006 Census pilot questionnaire following the modified star-model by Kimball as suggested by Søren Netterstrøm was created by the consultant and the course participants. As the topic can be highly complex it is recommended that additional training is provided.

Background

The background of the mission was a need identified by INE for a series of practical exercises as a follow up to the more theoretical mission on data modelling conducted by Søren Netterstrøm in 2005. Also Lars Thygesen in September 2006 recommended that INE should focus on centralized storage of its data in databases instead of in fragmentized files.

Although the mission was build around Microsoft products, the use of the SQL language should be seen as platform independent. All database systems working with relational data models support the use of the SQL as its query language. Also dedicated software for statistical analysis like SAS and SPSS supports SQL as a tool to select and modify data. Also flat CSV files can be analysed in MS-Query using the SQL language.

If INE as mentioned in the 2008 – 2012 activity program (draft) is to be able to deliver data in the SDMX-ML format to internal organizations, then the data in question (Consumer Price Index) must be store in a relational database. It is therefore necessary for INE to improve its skills and knowledge in the area of relational database software.

Recommendations

INE should designate one or two IT persons to the role as DBA's. The persons should have knowledge and / or receive training also in Win 2003 server, as the database system integrates with the network operating system.

A version of MS-SQL 2005 Express should be installed on a dedicated server on the network for testing and development purposes.

INE should investigate the possibility of switching future developments from MS- Access to MS-SQL 2005 Express, as this will give better performance, help to improved data integrity, give stronger security and ease the backup procedures.

Further training in the use of MS-SQL Sever 2005 Express should be given. It is strongly advised to build the training around a specific application / questionnaire which INE either needs to develop from scratch or where modernization is needed. Case driven training is suggested as it is much easier to learn when the new knowledge is directly relevant and applicable to your daily work.

INE / DISI is recommended to improve and extend it's knowledge and understanding of relational database theory and systems if it is to be able to meet the SDMX challenge.

Ultimately a relational database with micro and macro data from the involved subject matter areas must be put in place. This will also significantly improve the possibilities for dissemination and analysis of INE data.

2 INTRODUCTION

Background

The background for the mission was a need identified by INE for a series of practical exercises as a follow up to the more theoretical mission conducted by Søren Netterstrøm in 2005. Also Lars Thygesen in September 2006 recommended that INE should focus on centralized storage of its data in databases instead of in fragmentized and often also decentralized files.

2.1 Present state of "data storage"

Data is decentralized and fragmented in different formats

Most data at INE is store in so called flat files, MS-Access database files or in files related to statistical software like SPSS or CSPro. The associated descriptive Meta-data is found and replicated in all the systems and not in a central meta data repository. From an IT- Management perspective this situation is hardly satisfactory.

Creates consistency problems

This challenges the quality of the stored data in terms of consistency. In a worst case scenario the spelling of province or district names can very from one survey to another. As what is clearly common meta-data is not reused across the different systems. If and when such inconsistency occurs either in printed material or on the internet it is hardly helping to promote INE's status and reputation as a professional statistical institution.

Logical –security of data

When data is stored in application files, access to the data is only limited and controlled by the security structure of the network. However this setup provides no real control over which users inside INE that either by design or by accident makes changes or corrections to data.

Role based access to data

The problem with data based in files (both in flat files and in application specific files) is that every person having access to the network also has access to the files. And therefore he or she can by accident change or delete data. Ultimately also the possibility of change by malign design has to be considered from an IT security perspective.

Data entered into a relational database system can however be effectively protected inside the database through a series of roles. The role based data access makes it possible to create a security situation were staff undertaking analytical work can only read but not change the data. Data-Entry people can only enter data accordingly to specification – a fact that also leads to improved data quality. And the dimensional tables with the meta-data describing the data can only be changed or entered by those people tasked with harmonizing the organizations meta-data.

Log data

Also database systems provides functionality to register / log all access to data. Furthermore database systems also provides "roll back" so that data can be rolled back to "as the where" at a specific point in time.

Backup - physical security of data

Also the data is fragmented in large number of files it is both difficult and time consuming for the IT staff to ensure that proper back-up procedures are in place. Again in a worst case scenario some data is stored on shared drives which the IT backups on a regular basis, but sometimes important data also by accident or perhaps for reasons of better performance ends up on local

hard drives. And these local drives are typically never included in backup plans. In a very formal way it can be said that although unintentional storage of data on local drives moves responsibility and ownership from the NSI to the individual employee.

Demand for more flexible analysis

By drawing data form a unified database system INE will also be in a position to offer its user better possibilities for data analysis.

The upcoming census can in some ways be seen as on gigantic source of information if it is entered into a unified database system together with INE's other surveys. It will then fx be possible to relate CPI data from the provinces to data from the census describing the family structure in the area.

The amount of data will only grow

As the amount of data inside INE continues to grow it will be increasingly important for INE to store and extract data from a unified and coherent system. A full database system will give INE opportunities to store and backup data in a more efficient manner and to assign different data access rights do different people inside INE.

2.2 SDMX

According to the draft of INE's 2008 – 2012 activity program there is an recognized need for INE to be able to deliver data in the SDMX-ML format to international organizations and perhaps also to other users of statistics internally in Mozambique, like the National Bank of Mozambique. The data most likely to be required by external partners is the Consumer Price Index and data from the quarterly national account system.

A number of different software solutions are available for transformation and transmission into the format know as SDMX-ML. Also the PX-family of tools already in use at INE, has a series of build in functions to support publishing according to the SDMX standard. However to efficiently use the available standard software solutions for transmission and presentation, the data to be transmitted must be available in a uniform and automated way.

So in order to accomplish this strategic goal the data in question must be store in a relational database.

Relational databases are and will be needed at INE

From the above mentioned factors comes a strong need / imperative for INE to move from the use of flat files to a more centralized and robust database system build around relational models.

2.3 Storing survey data in a relational database

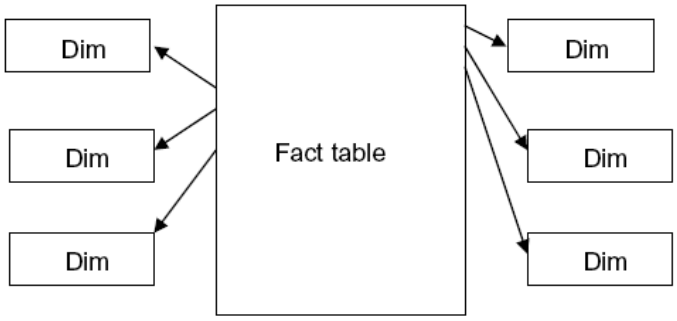
Centralized meta-data

In all NSI's we find that a large proportion of the relevant meta-data is shared across the statistical domains inside the organization. The most prominent examples are district, province, age, gender, and classifications of economic activity.

However as these shared meta-data populates different and non-relational system few if any statisticians inside INE and surely no external users have a coherent overview of the available data and the relations between them.

It has been shown by Søren Netterstrøm and the DISI staff that most statistical questionnaires with advantage can be stored using a modified star model (Kimbal).

Most survey questionnaires do in fact by default follow a structure / layout which approximates a relational database design.



Fact table

The illustration above demonstrates the thinking of Netterstrøm and Kimbal. The data collected through the survey is stored in a fact table. The fact table has one row for each observation unit / study object in the survey.

Dimensional tables

The dimensional table(s) in the model is used to described the variables and their values in the survey. Fx one dimensional table will hold the codes and values for the variable Gender, and another dimensional table will hold the codes and values describing the districts of Mozambique. It is important to note that the dimensional tables often if not always can be reproduced from a central meta-data repository as most of the meta-data will in fact be constant over time. And also will be constant and common over the subject matter areas inside the National Statistical Office.

2.4 Construction a preliminary model of the Census

From census pilot questionnaire to relational model

Following the concepts of Netterstrøm a model partially representing the data collected during the census pilot was constructed by the consultant and the training participants. Using the CSPro data dictionary it's straight forward to describe and create the relevant dimensional tables in the database.

On to different platforms

To demonstrate the universal application of the SQL language the database was first constructed in MS-Acccess which DISI staff is familiar with. The database was the queried and manipulated using MS-Query. After that the same database was created in the more robust MS-SQL 2005 Express package.

Using SQL

One of the advantages of SQL - or the Structured Query language – is that it should not be seen as normal programming language like Basic, C or Fortran instead it should be seen as the common language used to manipulate, store and extract data from all types of relational database systems. In our case all the command learned and use during the MS-Query part of the training could be directly applied also to the MS-SQL server version. So this reuse of commands spared the participants a lot of work. But more importantly it demonstrated the cross platform advantages of using SQL.

3 RECOMMENDATIONS

The consultant would like to make the following recommendations based on the work undertaken during the mission:

Additional training to be provided

It is recommended that additional training in the use of MS-SQL 2005 and data modeling is provided. The training should be constructed around an application that INE would like to construct in order to gain the maximum impact. Class room examples often tend to be abstract for real efficient learning. Alternatively an advanced introduction to SQL and MS-SQL Server 2005 Express can be given using an expanded dataset from the census. It is recommended to include the purchase of a dedicated server for experimental purposes in such a mission.

Training in VB

Training in Microsoft Visual Studio Express as tool to build client server applications interacting with databases should be provided.

Dedicated server

MS-SQL Express should be installed on a dedicated Internet server running on Windows 2003. The use of a dedicated server will ease access to the database and help INE staff improve their skills in database management.

DBA

On or to members of the IT staff should be officially designated to the posting as DBA (database administrator).

Experience with Win 2003

To take full advantage of the security model used on the INE network is recommended that the person(s) designated to the DBA role either has or is given training in Windows 2003 Sever.

Continuation of IT – Infrastructure consolidation

As INE moves to a more centralized and controlled storage of data and Meta-data, it is even more important that the IT-infrastructure is physically- and logically- safe and considered as such by the INE staff. It is recommended that priority is still given to the Windows 2003's security model in order to ensure logical safety and physical safety of INE data.

3.1 Concluding remarks

Thanks to all at INE

Finally I would like to express my thanks to all officials and individuals meet during the mission. They all provided me with the necessary information in a kind and open atmosphere which greatly facilitated my work in Mozambique. But specially, I would like to thank Mr. Lars Carlsson for being an excellent host and for a very constructive sharing of his thoughts on the project.

My best personal professional opinion

It should be noted that this report contains my best personal professional opinions as a consultant, and that they therefore do not necessarily correspond to the views of Statistics Denmark, Danida or INE.

4 APPENDIX 1. List of persons met

INE
Ms. Anastasia Honwana, Head of IT
Mr. Antonino Reginaldo A. Francisco
Mr. Anselmo Leonardo O. Nhane
Mrs. Beatris Maria Ismael Manjarte
Mr. Calado Pereira Fijanto
Mr. Celso Azarias Machava
Mr. Socrate Tiago
Mr. Tomas Bernardo

Scanstat Consortium, LTA:
Mr. Lars Carlsson, Team Leader
Mrs. Julia Cravo, LTA on Business Statistics
Mr. Jan Redeby, LTA on National Accounting

5 APPENDIX 2. List of Literature

All mission reports from the Scandinavian programme are available online on: *www.dst.dk/mozambique*

For this mission I would also like to refer to the report:

Mission Report from a short-term mission on Data Modelling 31 January – 4 February 2005 by Søren Netterstrøm MZ:2005:08

Alison Balter: Microsoft SQL Server 2005 Express in 24 Hours, Sams Publishing 2006, 436 pp

Jonathan Gennick: SQL Pocket Guide, O'Reily 2004 pp154

Ralph Kimball, Margy Ross: The Data Warehouse Toolkit – The complete guide to dimensional modelling, 2nd ed Wiley 2002 416 pp

6 APPENDIX 3. Terms of Reference

TERMS OF REFERENCE

for a short-term mission
on

Data modeling and SQL

19 February – 01 March 2007

within the Scandinavian Assistance to Strengthen the Institutional Capacity of INE/Mozambique

Consultants: Jesper Ellemose Skou Jensen
Counterparts: Tomas Bernardo and Anastácia Judas Honwana

D R A F T

Background ´

It was originally the plan for INE to develop a Data Warehouse system based on 3 database components: A Micro Data Warehouse, A Macro Data Warehouse and a Dissemination Database.

The Dissemination Database is already up and running based on the Scandinavian PC-Axis / PX-Web platform, also used by FAO.

The Data Warehouse strategy and a road map for implementation is described in greater detail in a *Short Term Mission on Data Modelling 31 January - 4 February 2005* by Søren Netterstrøm, MZ:2005:08. The mission then was directed at Anastacia Honwana, Clara Panguana, the developer group at DISI and the LTA on IT Karsten Bormann. The report by Lars Thygesen a *Short Term Mission It Management and Strategic IT use* from September 2006, MZ:2006:10, recommends increased focus on the Data Warehouse. However a need for a more practical / hands approach and training in the related subjects of Data modeling and working with Databases is recognized by INE (DISI) and therefor a mission on Data modeling and SQL was discussed and requested during Lars Erik Gewallis visit in November 2006.

All data at INE are at the moment stored in either flat data files, MS-Access databases (The National Account system) or in files relating to analytical software like SPSS. However it is and will be increasingly important for INE to store and extract data from a more robust Database system. A full database system will give INE opportunities to store and backup data in a more

efficient manner and to assign different data access rights do different people inside INE.
Also the “though” data management discipline which is a part of full scale database systems will help improve the data quality of INE’s surveys. Further more an increased used of databases will allow INE to develop more ad hoc Client – Server applications.

As a low cost but high tech introduction to database technology is suggested to use the Microsoft MS-SQL 2005 in the Express version. This is provided free of charge by Microsoft and has all the functionality of Microsoft’s commercial versions. There are limitations in the amounts of data that it is possible to store in Express version. However for training and familiarization purposes the Express version is more than sufficient.

INE has an ambition to follow the international SDMX standard for exchange of statistical data. Although the PX-family has a series of build in functions to support publishing according to the SDMX standard. SDMX files are usually best created from databases and therefore a supplement to the dissemination database must be created.

In order to provide INE with a practical skill building it is planed to build a series of missions around one or two development cases. Each case should lead to a working database.

SDMX is based on XML and XSLT is recommended that training is also provided in this field to INE staff before the end of the year. CPI and National Account data are the data most likely to be requested by international organizations in SDMX-ML. INE may therefore like to enter data from these two subject areas in to a Micro / Macro Data Warehouse model before the end of 2007.

Drawing on Visual Basic and XML it should then be possible for INE to construct a Web service with data in the SDMX-ML format.

Also the LTA on IT left INE by the end of August 2007. Instead it is planned to have a series of short term mission providing INE with gap filling, reflections, discussions and second opinions in the area of IT. This mission should also be seen in this context.

Objective

The objective of this mission is to strengthen the practical and theoretical knowledge of data modeling at INE, through the use of the Census Pilot as a case for building a database in MS-SQL 2005 Express.

To demonstrate basic functions in the SQL language to extract, manipulate and present data from the Census pilot database.

The mission must also lay the ground work for a future mission on Software Development using Visual Basic. This follow up mission should concentrate on the development of Client- Server applications with MS-SQL Express as the database engine. Also the mission should through hands on training show how data is entered and extracted to a database from other file formats.

Expected results

- Installation of MS-SQL 2005 Express at a number of workstations for training purposes
- An introduction to basic database concepts
- An implementation of a database structure reflecting data from the Census pilot, conducted in 2006
- Introduction to the SQL language
- Introducing MS-Query as a tool to extract data from MS-SQL, Access and CSV files
- Prepare for a follow up mission on application development using Visual Basic Express – this mission should concentrate on building client server applications with MS-SQL Express as the database behind.
- Mission report that comments on the objectives and achievements and include recommendation about of the next steps on improving the general modeling and programmatic skills inside INE

Activities

- A meeting with the counterparts on the objectives and expectations of the mission.
- Classes / Workshops on Data modeling, MS-SQL 2005 Express, basic SQL statements, and MS-Query will be conduct on 5 working days (Wednesday to Tuesday). The workshops will be from 8.30 to 12.00. They will consist of a combination of short technical / theoretical briefings followed by hands on experience
- A meeting towards the end of the mission with Counterparts to present and discuss the results and recommendations

Tasks to be done by INE to facilitate the mission

- Elaborate ToR for the mission
- Prepare and supply the consultant with necessary documents and information, such as mission reports, strategies, plans etc.
- Supply good working conditions for the consultant

Consultant and Counterpart

Consultant: Jesper Ellemose Skou Jensen from Statistics Denmark

Main counterparts:
Anastácia Judas Honwana, Tomas Bernardo

Timing of the mission

Two weeks (19 February – 1 March, 2007).

Report

The consultants will prepare a draft report to be discussed with INE before leaving Maputo. They will submit a final draft to INE for final comments within one week of the experts have returned to work. Statistics Denmark as Lead Party will print the final version within 3+ weeks of the end of the mission. The structure of the report should be according to Danida format.

The Counterpart has to ensure that the final printed report has at least a summary in Portuguese if the main report is in English – or vice versa.

These Terms of Reference were prepared by

Day / /
.....

Approved by/in the name of the President of INE

Day / / .....

Prepared by:

7 Appendix 4 ACTIVITIES DURING THE MISSION

The following activities where conduct during the mission:

Monday 19 February	Meetings whit Team Leader, Main counter parts, visit to the Census office
Tuesday 20 February	Converting census pilot data from CSPpro into MS-Access database and MS-SQL Server database. Discussions on the use of PX-web to create a stand alone CD-rom publications from the INE database.
Wednesday 21 February	Converting census pilot data from CSPpro into MS-Access database and MS-SQL Server database
Thursday 22 February	Converting census pilot data from CSPpro into MS-Access database and MS-SQL Server database
Friday 23 February	Training and discussions
Monday 26 February	Training and discussions
Tuesday 27 February	Installation of MS-SQL Server express on a number of workstations, Training
Wednesday 28 February	Training and discussions
Thursday 1 Marts	Training and meeting with main counterpart, discussion about future use of database in INE.