Distr.
GENERAL

Working Paper No.8
11 April 2007

ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Meeting on the Management of Statistical Information Systems (MSIS 2007)
(Geneva, 8-10 May 2007)

Topic (ii): Statistical information systems architecture

DATA WAREHOUSE ARCHITECTURE TO SUPPORT ANALYTICS

Invited Paper

Prepared by Karen Doherty, Ann Stoyka and Michel Girard, Statistics Canada, Canada

I. INTRODUCTION

1. Statistics Canada (STC) collects information on a wide range of subjects and has amassed a large amount of statistical data mostly stored in stand-alone operational databases. This tendency to build silos has impeded the Agency's efforts to reduce response burden and eliminate duplication of effort during the data processing and analysis phases. As our reliance on administrative data grows and we look for ways to better use the data we collect, our need for tools to help analysts merge and compare data from multiple sources increases.

2. To address these issues, Statistics Canada embarked on an initiative to build a flexible architecture for analytical data warehouses. This initiative proved to be extremely successful and warehouses have proliferated across the Agency, not only for analytical purposes, but also as tools to provide up-to-date information to managers in traditional management roles and operations. This paper describes the evolution of data warehousing at Statistics Canada and the development of our data warehouse framework.

II. THE EARLY YEARS

A. The System of National Accounts Embarks on Renewal

3. The Canadian System of National Accounts (SNA) provides a conceptually integrated framework of statistics and analysis for studying the state and behaviour of the Canadian economy. The accounts are centred on the measurement of activities associated with production of goods and services, the sales of goods and services in final markets, the supporting financial transactions, and the resulting wealth positions.

4. In 2001 the System of National Accounts articulated a strategy for analytical renewal aimed at enhancing the analytical capacity of the SNA framework. The strategy included the resolution of data inputs and supplier issues as well as the review and improvement of data outputs and products.

5. At the time the SNA Branch was comprised of six Divisions: Income and Expenditure Accounts, Balance of Payments, Public Institutions, Input-Output, Industry Measures and Analysis, and Environment Accounts and Statistics. In 2006 the Input-Output Division and Industry Measures and Analysis Division merged to form the Income Accounts Division.

B. The Input-Output Division Faces a Business Challenge

6. At the time, the Input-Output (I-O) Division was responsible for producing the input-output accounts of Canada and its provinces and territories, in particular the output table, input table, and final demand table for each provincial and territorial jurisdiction. The tables from the various jurisdictions are linked together through an interprovincial flows table that shows each jurisdiction's exports to, and imports from, other provinces and territories as well as abroad.

7. Input-Output tables cover all economic activities conducted in the market economies of each province and territory, encompassing persons, businesses, government and non-government (non-profit) organizations, and entities outside its jurisdiction that give rise to imports or exports (interprovincially or internationally.) The Input-Output tables represent the most detailed accounting of the Canadian economy available and thus serve as benchmarks to the Canadian System of National Accounts.

8. The Division's capacity to produce accurate Input-Output Tables was in serious jeopardy because of the vulnerability of the I-O Division's production systems and work processes. In particular, the lack of appropriate automated tools for data verification and table balancing imposed a heavy burden on staff and the lack of integration (among others, the problem of not using a uniform set of classifications within the division, the SNA Branch and the rest of STC) and standardization of processes and procedures impeded the division's ability to handle the growing amount of input data.

9. A project was launched to modernize the I-O Division's business and information management processes. The specific objectives of the project were:

- maximize knowledge retention and the potential for reuse through the introduction of software to specify derivation and balancing methodologies;
- maximize operational integration potential of the various divisions of the SNA through the introduction of a data management system to integrate data and meta-data, and facilitate data reconciliation between I-O and other SNA divisions;
- maximize analytical potential through the introduction of main stream analytical tools to detect problems in the I-O Tables and source data.

C. A Vision Takes Shape

10. Following an architectural analysis of the business requirements, the team proposed a system comprised of:

- a centralized data base consisting of two logical components:
    - a data reception area for user-supplied micro and aggregate data that, together with improved analytical tools, to facilitate data confrontation;
    - an analytical database for aggregate (macro) data to support data reconciliation with the GDP outputs from other SNA divisions;
- a suite of tools for analysts containing both standard statistical functions and a number of I-O specific tools to calculate industry and commodity specific ratios.

11. After several attempts at defining an architecture for the new system, the design team settled on a solution based on a formal application of data warehouse concepts and technology. The first functional warehouse was delivered in 2004. The warehouse contained final results and source data maintained according to SNA concepts and classifications. This permitted analysts to compare data from different sources, ensuring consistency of estimates by making dissimilar classifications comparable.

12. The system provides analysts with the tools to reconcile information across divisions in the SNA and make effective decisions during the annual production cycle. The need to define what was to go into the warehouse and how the data was to be harmonized led to a standardization of the analytical process.

13. The warehouse allowed users to compare statistics in terms of ratios, proportions, growth rates, by region and in chronological series. It was particularly helpful in enhancing data coherency and in permitting comparisons of data taken from different sources, while applying SNA or other Agency classifications, definitions and concepts for any aggregate level. Access to metadata was provided, as was information on how the data was established, concepts and definitions, classifications and concordances and best practices with respect to processing or analytical procedures.

14. Analysts used the warehouse to generate Excel reports, which were automatically updated whenever they were opened by the analyst. The warehouse also contained analytical functions and permitted graphic analysis. Above all, the warehouse allowed the division to operate with a far less experienced staff, because analytical procedures had been standardized and normalized. Analyses were more transparent, more repeatable and more efficient.

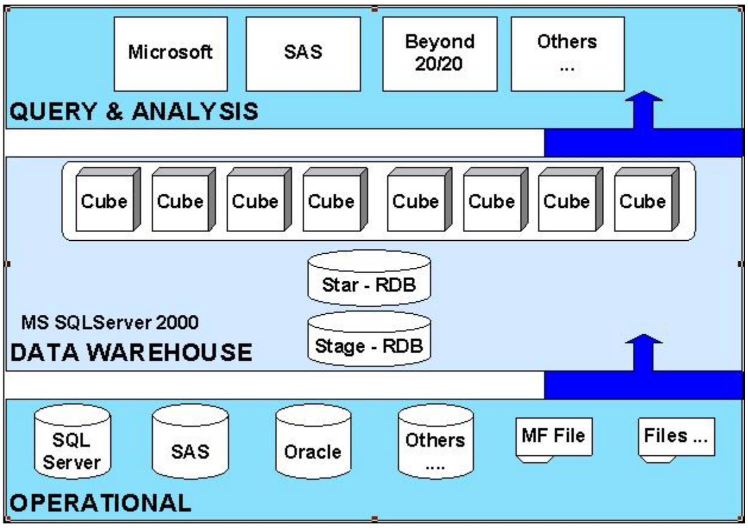
D. A Data Warehouse Architecture Is Born

15. Ralph Kimball, a leading expert in warehouse concepts, defines a data warehouse as a "*copy of transaction data specifically structured for query and analysis*". There are two important concepts covered by this definition:

- *Copy of transaction data*: a warehouse should not be used to perform operational functions. Operational data, data subject to immediate change, should be housed in operational databases and maintained appropriately. The warehouse should be created through imports or data loads from the operational sources. All changes to data should be made within the operational environment to ensure that the operational environment always contains the most recent information. This strategy also ensures that the business rules governing the operational process remain consolidated in one authoritative source.
- *Specifically structured for query and analysis*: The data in the warehouse should be structured to support the types of queries and analysis that the warehouse targets. The primary method of output for the data warehouse is On Line Analytical Processing (OLAP), i.e. output designed for human rather than machine consumption. This means that response time must be taken into consideration and therefore the system strives to produce the requested information in less than 15 seconds.

16. The project in the I-O Division, along with a similar project undertaken for the Centre for Education Statistics, led to the development of a formal approach to building a data warehouse, based on industry-supported concepts, standards and technologies. A centre of expertise was formed within the System Development Division, originally to develop a depth of knowledge and experience in the data warehousing field and build up a client base, and eventually to ensure that all warehouse projects at Statistics Canada use a common approach and development framework.

17. The Data Warehouse Centre originally described the warehouse architecture as follows:



18. At the bottom of the diagram are the operational databases. These databases can be built using any number of database products including, but not restricted to, Microsoft SQLServer, Oracle and SAS. These databases are specifically structured for input, update and delete activities and all updates to the data are performed at this level.

19. In the middle of the diagram is the warehouse proper. Data is loaded through an *Extract, Transform and Load* (ETL) operation into a relational database to stage the data. From the staging database, the data is moved to a second relational database where it is reformatted into a Star Schema. This database consists of a fact table surrounded by, and linked to, multiple dimensional tables in a star like pattern. This database contains the data at its lowest level of aggregation, which in many cases means micro data. The data residing in the star database may be transferred to data cubes or may be accessed through the cube in an operation called a “drill-through”.

20. From the star database, selected sets of data are transferred into a multidimensional database consisting of a variety of cubes. This physical restructuring of the data according to a predefined set of business rules allows for very fast data manipulation and retrieval. Cubes also provide for all necessary data security.

21. Finally, at the top of the diagram, is the query and analysis layer. Users can choose from a wide range of products to either view or extract data from the cubes for the purpose of analysis. At STC the most popular products for viewing data in the cubes are Microsoft Excel XP, SAS and a home grown web viewer called EzWeb, however there are any number of other query tools that can be used. Data can also be extracted from the cubes and manipulated through products such as SAS and Beyond 20/20.

E. The Warehouse Delivers

22. The original objective of the I-O Renewal Project was to bolster its analytical capacity through the redesign of the data processing and analytical support systems and streamline its production processes to improve quality and gain efficiencies. The project also included the development of a human resources renewal strategy. Although these ambitious objectives had to be scaled back somewhat, the project did achieve tangible results. These were especially meaningful in the improvements seen in the analytical capacity of operations.

III. MOVING ON TO THE PRESENT

A. The Business Challenge Expands

23. Having demonstrated that a data warehouse provided a measurable analytical benefit for the Input-Output Division, a proposal was submitted for funding to expand the warehouse to cover data from all the divisions in the System of National Accounts Branch. The objectives of the project were:

- improve access to Branch data by consolidating many sets of inter-related account information physically distributed throughout the Branch;
- support analysis and reconciliation of estimates prepared by the various divisions;
- harmonize the data to support data comparison and reconciliation.

24. For users unfamiliar with what a data warehouse offers, the advantages of building a warehouse are not apparent. The key was to develop the concept in an iterative way, building the components that were easily understood first and expanding the vision as each component was integrated into the whole. This implied that the design approach needed to incorporate a combination of strategic vision coming from management and more technical and practical ideas coming from the analysts and technical staff on the project.

25. To demonstrate the value of the project it was imperative that the team be able to quickly create a useful data warehouse using data in its native form despite a proliferation of coding rules and variable definitions.

B. The Data Warehouse Concept Evolves

26. As the teams developing warehouses gained experience, and clients articulated more ambitious business needs, the concept of a data warehouse evolved to cover a wider range of functionalities and business opportunities. A broader definition of a warehouse was adopted, a definition that is well articulated in Wikipedia as “a repository of an organization's data, where the informational assets of the organization are stored and managed, to support various activities such as reporting, analysis, decision-making, as well as the optimization of organizational operational processes."

27. The new vision of a data warehouse included the storage of virtually all transactional data, master data (customer, material), and metadata at a very detailed level. It was seen as the natural evolution from the traditional concept of information technology at STC, which is mainly grounded in the automation of operations. Data warehousing seeks to optimize data management and, through further manipulation by experts, support its deployment and exploitation as information.

C. The I-O Warehouse Grows Into the SNA Warehouse

28. The I-O warehouse provided a strong incentive to the I-O Division to review their work methods and standardize their classification systems, however, since the Division was responsible for only a part of the data that makes up the Canadian SNA, the value was limited to areas over which the division had direct control.

29. The SNA pulls together data from across the Branch to form a single consistent product. Individual divisions within the SNA pull in data from other divisions at STC, in particular from the business survey program areas. Before the warehouse was created, when two divisions needed data from the same file, they worked on the data in isolation, often duplicating work. Although the results published by the SNA were consistent, there were often inconsistencies in the way data was defined and processed, resulting in inconsistencies in the detailed data.

30. The SNA Warehouse is a vehicle for helping the SNA reconcile and resolve differences in classification systems, definitions and concepts across all the divisions. It increases the analysts’ ability to confront inconsistencies and agree on common definitions and processes, thus improving the quality of the data and increasing the productivity of staff.

31. Improvements to the data warehouse architecture have also resulted in increased technical functionality. A web-based viewing tool provides users with access to a range of data sources, organized as a data mart. Direct access to data stored in the Agency's metadata system reduces time lost looking up definitions and allows users to identify errors and deficiencies in the metadata itself.

IV. THE DATA WAREHOUSE FRAMEWORK

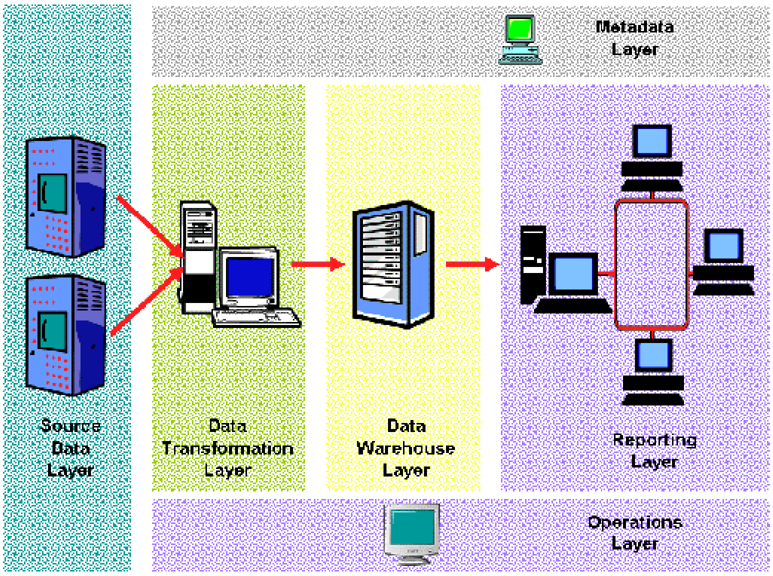
32. The STC Data Warehouse Framework grew out of the experiences of the project teams who developed the I-O Warehouse and the Education Statistics Warehouse. The framework came to maturity mainly through the experience gained building the SNA Warehouse and the Agency's Financial Reporting System. The framework provides and defines the tools, building blocks, procedures, common methods, etc. used when building a data warehouse. Experience shows that effective use of the Data Warehouse Framework ensures success regardless of the type of information content. The components of the framework are described below.

A. Technology

33. The original STC framework was built using Microsoft Data Warehouse Framework for SQLServer 2000. The STC framework is being ported to MS SQLServer 2005 which will result in a simpler environment since many of the components built by the Data Warehouse Centre for the SQL Server 2000 environment are built into the MS SQLServer 2005 product offering.

34. Before settling on a Microsoft-based warehouse solution the I-O Warehouse project team explored alternate solutions. The team first attempted to build a system with the Oracle toolset, then spent considerable time trying to develop a solution using SAS. At the time, neither Oracle nor SAS offered a rich enough warehouse development environment to meet the requirements of the project. Both vendors now offer warehouse toolkits that are similar to the MS SQLServer 2000 solution, however, the DW Centre believes that Microsoft's SQLServer 2005 offering still gives them the edge over the competition.

35. Microsoft describes its Data Warehousing Framework as an open, scalable architecture that speeds, simplifies, and reduces the cost of building, managing, and using today's business intelligence applications. The MS framework includes a number of Microsoft components in addition to SQLServer including Enterprise Manager, Data Transformation Services (DTS) and Analysis Services, and is fully integrated with Microsoft Excel XP.



36. The STC framework expands the MS framework by incorporating a set of tools and processes developed by the STC Data Warehouse Centre to support the types of applications that are built at STC. It also

addresses specific requirements of the Agency and the Canadian Public Service such as support for both official languages (English and French) and data security and confidentiality.

B. Standards – On Line Analytical Processing (OLAP)

37. The framework creates cubes that conform to the family of OLAP-related standards. It is an approach to quickly provide the answer to analytical queries that are dimensional in nature. It is part of the broader notion of Business Intelligence, which also includes Extract, Transform and Load (ETL), relational reporting and data mining. Databases configured for OLAP employ a multidimensional data model, allowing for complex analytical and ad-hoc queries with a rapid execution time. The output of an OLAP query is typically displayed in a matrix (or pivot) format. The dimensions form the rows and columns of the matrix, and the measures populate the cells of the matrix.

38. It has been claimed that for complex queries OLAP cubes can produce an answer in around 0.1% of the time for the same query against a traditional relational data. The single most important mechanism in OLAP which delivers this level of performance is the use of *aggregations*, which are the pre-defined roll-ups on pre-selected dimensions in the cube. An aggregation is often referred to as a view from the user's perspective.

39. The base data plus the combination of all possible aggregations contain the answers to every query which can be answered from the data. Due to the potentially large number of aggregations to be calculated, often only a predetermined number are fully calculated while the remainder are solved on demand. The problem of deciding which aggregations (views) to calculate is one of the challenges faced by cube designers since the choice of supported aggregations directly affects both the time a user has to wait for the results of a query and the time it takes to update the cube from the operational databases.

C. Data Transformations

40. The Data Transformation layer receives data from the data sources, cleans and standardises it, and loads it into the data repository. This is called "staging" data as data often passes through a temporary database during the transformation. This activity of transforming data can be performed by manually created code, an commercial ETL tool or a combination of both. Regardless of the nature of the software used, the following types of activities occur during data transformation:

- comparing data from different systems to improve data quality (e.g. Date of birth for a customer may be blank in one system but contain valid data in a second system. In this instance, the data warehouse would retain the date of birth field from the second system);
- standardising data and codes (e.g. if one system refers to "Male" and "Female", but a second refers to only "M" and "F", these codes sets would need to be standardised);
- integrating data from different systems (e.g. if one system keeps orders and another stores customers, these data elements need to be linked);
- performing other system housekeeping functions such as determining change (or "delta") files to reduce data load times, generating or finding surrogate keys for data, etc.

D. Application Programming Interfaces (API) and Query Languages

41. Unlike the world of relational databases, which has SQL as the standard query language and a range of APIs to choose from, there was no such unification in the OLAP world for a long time. Microsoft introduced the MDX query language in 1997. In 2001 Microsoft and Hyperion announced the XML for Analysis specification which was endorsed by most of the OLAP vendors. Since this standard used MDX as a query language, MDX became the de-facto standard in the OLAP world.

E. Metadata

42. Metadata, or "data about data", is used not only to inform operators and users of the data warehouse about its status and the information held within the data warehouse, but also as a means of integrating incoming data and updating and refining the underlying Data Warehouse model.

43. Examples of data warehouse metadata include table and column names along with their detailed descriptions and names in common use by subject-matter experts, the business meaning of a data item, the most recent data load date and the number of users that are logged in currently.

F. Reporting and End-user Tools

44. The data in the data warehouse must be available to the organisation's staff if the data warehouse is to be useful. There are a very large number of software applications that perform this function, or reporting can be custom-developed. Examples of types of reporting tools include:

- *Business Intelligence tools:* These are software applications that simplify the process of development and production of business reports based on data warehouse data.
- *Executive or Management Information Systems:* Known more widely as Dashboards in the business context, these are software applications that are used to display complex business metrics and information in a graphical way to allow rapid understanding.
- *OLAP Tools*: OLAP tools form data into logical multi-dimensional structures and allow users to select the dimensions by which data is viewed.
- *Data Mining*: Data mining tools allow users to perform detailed mathematical and statistical calculations on detailed data warehouse data to detect trends, identify patterns and analyse data.

45. The DW Centre has made use of Microsoft Excel, SAS Enterprise Guide and Beyond 20/20. The current standard is to display report using the EzWeb OLAP Report Browser and EzWeb OLAP Report Designer developed by the System Development Division’s Web Technology Center. If EzWeb cannot meet the client’s needs other tools can be considered.

G. EzWeb

46. Developed and maintained by the System Development Division’s Web Technology Centre, EzWeb is a product used for web site construction and management. Using a simple “what you see is what you get” (WYSIWYG) approach, EzWeb provides users who have no web skills with a user-friendly environment in which to create web pages that are compliant with the Canadian Government’s Government On Line (GOL) standard, including the need to make pages available in both official languages.

47. Microsoft Office Web Components (OWC) Pivot Table is incorporated into EzWeb and provides the OLAP functionality needed to view data in warehouse cubes. EzWeb also has the ability to navigate from one OLAP report to another, thus providing users with easy and intuitive information exploration and discovery through a single interface.

H. STCWiki

48. One of the challenges faced by the warehouse architects was how to provide warehouse users with access to metadata in a way that easily allowed them to update metadata that was incomplete or inaccurate and to navigate quickly through the rich metadata environment. The DW Centre resolved this issue by leveraging a wiki application to act as a link between the Agency’s Integrated Meta Data Base (IMDB) and the Data Warehouse Framework.

49. A wiki is a tool for collaborative authoring and knowledge sharing. The most well-known wiki application is Wikipedia, the online collaborative encyclopedia which contains over 1.7 million articles. There

are a number of wiki engines available in the Open Source domain – STC chose MediaWiki (used by Wikipedia) to build STCWiki.

50. The wiki is still in pilot mode. The environment is designed for collaboration and provides users with access to a large amount of information including all the information in the IMDB. In 2007, the team plans to build in the functionality required to allow users to update the IMDB directly from the warehouse.

I. Data Marts

51. The use of EzWeb as the standard viewing tool has allowed STC to implement the concept of a data mart which provides users with access to a broad range of data sources organized to meet their needs.

- *Dependent Data Marts*: A dependent data mart is a physical database (either on the same hardware as the data warehouse or on a separate hardware platform) that receives all its information from the data warehouse. The purpose of a Data Mart is to provide a sub-set of the data warehouse's data for a specific purpose or to a specific sub-group of the organisation.
- *Logical Data Marts*: A logical data mart is a filtered view of the main data warehouse but does not physically exist as a separate data copy. This approach to data marts delivers the same benefits but has the additional advantages of not requiring additional disk space and it is always as current with data as the main data warehouse.

V. LESSONS LEARNED

52. The existence of a data warehouse framework removes the technology related challenges from projects trying to provide an integrated repository of data to meet their business needs. The tools highlight the conclusion that the real challenge lies with how data should be processed, analysed and classified. A data warehouse is an excellent tool for data validation and certification in the early stages of the survey life cycle as low data quality can be immediately revealed and therefore corrected. Further along in the survey life cycle, a data warehouse underlines the value of harmonization as the degree of success in the comparison of data depends directly on the degree of harmonization that exists.

53. The recognition of the value gained by harmonization usually results in modified working procedures. The I-O project ultimately led to an initiative to completely alter the methods used to establish baseline consumer expense data and a review of the public sector information passed between the Public Institutions Division and the I-O Division.

54. The first major lesson learned is that business units must separate the tools and systems used for the processing versus the analysis of data and must start by having good analytical tools before they invest excessively in data processing and production systems. These analytical tools allow business units to detect problems and improve and adapt the methods used to ensure that the data is of high quality.

55. The second lesson learned is that providing users with a customized data mart that can be seamlessly connected to other relevant data marts, is more effective than trying to build a single warehouse containing all Agency data. The data mart approach effectively partitions the effort involved in harmonization and the management of security and access rights, while allowing users to customize their personal portal to list only those sources that are of business interest to them.

VI. CONCLUSION

56. Data warehouses can contain data from any source and their application covers a wide range of business requirements. Although Statistics Canada’s first warehouses were specifically aimed at providing structured sources of data for the purposes of statistical analysis, the warehouse approach has also been effective for a range of other applications including financial reporting, management information systems, census progress reporting and case tracking.

57. Warehouses should not be viewed as merely a means of simplifying the reporting function, although they do a very good job of providing users with a flexible reporting capability. Nor are they simply an analytical tool for analysts looking at reported data. Rather, their most striking benefit, lies in the ability to provide production analysts with a wealth of information for comparing source data and identifying shortcoming and errors in the creation of final results. They enable business to make major strides towards reducing the effort associated with production processing while delivering a higher qualify product at the end of the exercise.
