Building Data Marts in a Manufacturing Environment

Jeffrey M. Stander
Principal Consultant
Director of Technology
jeff@anzustech.com

ANZUS Technology International (USA)

ANZUS Technology International Pty. Ltd (Australia)

Introduction

This paper presents solutions adopted in building a data warehouse system for a materials manufacturer. It will present an overview of the data components for a warehouse composed of Central Data Warehouse and data marts and the methodology used for distilling and loading operational data. Discoverer 3.0 is designated as the end-user query and analysis tool.

It is reported that over 90% of medium-sized to large-size companies have established or will establish a data warehouse. Why would they do this? The answer is that data warehouses provide a centralised distinct repository of vital business data that has been extracted from a variety of corporate databases. Fast and accurate analysis of this data provides a strategic, competitive advantage. Properly analysed and presented, this information can impact the long-term survival of the company.

The Company

Making Wafers

This particular enterprise, which manufactures silicon wafers for the semi-conductor industry, consists of a North and a South campus. Each campus has three fabrication plants – one grows crystals and cuts them into ingots, the second slices the ingots into wafers and polishes them, and the third adds an optional epitaxial layer to the wafer. Shipment occurs at the end of the polishing stage, either to the customer or to the epitaxial plant. The "epi" plant ships finished product back to the polished wafer plant for final shipment to the customer. The product is shipped with a "Certificate of Compliance" which attests that the product shipped meets the specifications agreed upon when the order was taken. Thus, the product of the company actually consists of both wafers plus test data. (See Figure 1, right)

Out with the Old, In with the Old

The North plant is a legacy system composed of a mainframe MES system, a number of third-party mainframe software packages (all heavily customised with Cobol patches and special software). This chaos is supported or supplemented by numerous applications developed in FoxPro, the single-user stand-alone desktop "database" package that goes back to MS-DOS days. Enthusiastic end users developed these programs in the IS vacuum which existed during the start-up. Many of the FoxPro applications remain at the core of the North plant's operations, undocumented, and supported with great difficulty. The relatively large IS department has spent most of its time supporting this patchwork of software. Although IS was not always responsible for the products it supported, it still took the heat for them.

The South campus was to be different. The goal was a modern, fully automated system using robotic technology. The data for the Certificate of Compliance was to be compiled from the output of the automatic testing equipment. What is amazing is not how difficult the old system was, but how few lessons had been learned from the experience.

The Current Situation

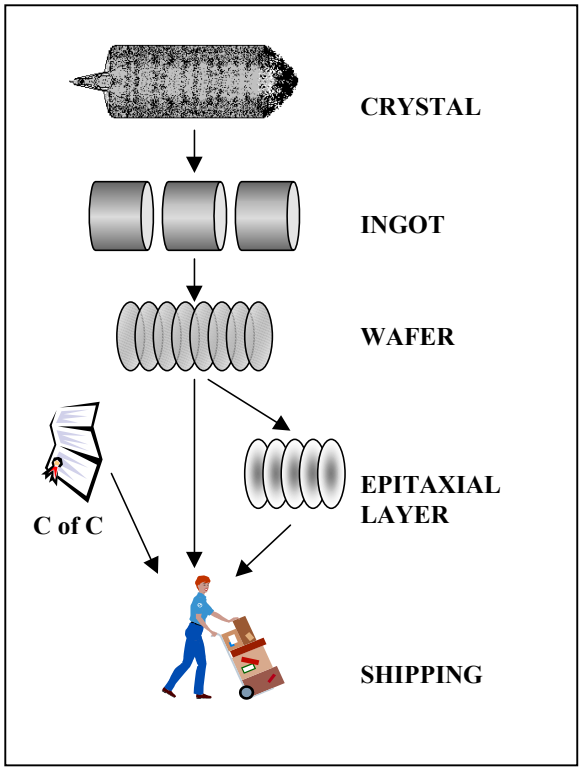

Figure 1. Silicon Wafer Manufacturing

The continuous need for rapid development has led to a "code-like-hell" philosophy of software development and machine integration. Many of the developers had grown up with the start-up company. Many were contractors who came and went. User demands for rapid deployment kept the pressure high. The "cowboy coder" mentality perceived project management, source code documentation, change control, and security as either an unnecessary or, at best, an unaffordable. No formal management of any IS project had ever been performed in the past. What management there was had been done by the user-owners of the software under development.

The Fortress Mentality

Rapid and unstructured development led to the creation of virtual enclaves among the individual fabrication plants. Each has become a fortress unto itself with its own computers, its own dedicated IS personnel, and its own unique software. As a result, the systems are not integrated, common problems have different solutions, and the common perception is to "avoid the mess in the other plants."

Revolution and Evolution

The hiring of a new CIO and appointment of an experienced and capable IS director with a strong belief in enterprise modelling and planning has led to a core group who are in the process of trying to effect a paradigm shift.

The Cultural Paradigm

In enabling this movement it is critical to understand that development issues are essentially cultural and not technical. The enclave mentality precludes an enterprise viewpoint. It is necessary to introduce modern methods very carefully because change is often perceived to be threatening both from a personal and an operational viewpoint.

Individuals are defensive about their work. To subtly change the culture a new infrastructure must be provided in almost a subversive manner. With the infrastructure fixed strategic approaches can be taken. Note that this change is not just the use of modern database design methodologies such as data modelling, data integrity, and data warehousing, but includes more commonplace Information System methodologies such as the SDLC, data integrity, documentation, and source code management.

The New Paradigm

It would be ideal to drive the data structures of the operational software from the enterprise viewpoint, but this simply is not possible. What we have done is to create the Enterprise Information Architecture (EIA) initiative. This includes an Enterprise Data Model that is the design paradigm for the Central Data Warehouse. Both data and process models are included. As legacy systems are replaced, the EIA becomes the driver for using architected data structures that integrate over the entire company. With executive buy-in and sponsorship there is a chance to overcome the cultural issues.

At the core of the EIA is the development of enterprise policies, procedures and standards (there were none, not even a preferred SDLC model – and this is not a tiny company). We are not only promulgating industry-accepted development standards but also requesting a shift to using standard software tools and methods. Although Oracle is perceived as expensive by management there are long-term unperceived savings in maintenance that are to be gained by standardising on the Oracle design, development, and analysis toolset as well as the database server.

What is a Data Mart and Why Do We Need One?

Data warehouses are a device derived from decision-support systems. They aren't really new. While the data warehouse can be looked at as just another database, it's purpose and structure is very different from the Operational Data Store (ODS). While the ODS is designed for OLTP, the purpose of the Enterprise Data Warehouse is to hold data in a structure designed for analytical chores. (Figure 2)

The data mart is a subset of the data warehouse that has crept from behind the skirts of the Enterprise Data Warehouse. A number of definitions have been proposed for the concept. Douglas Hackney states that

> The term "data mart" was primarily driven into legitimacy through the efforts of application vendors who needed a legitimate platform to host their three tier applications... Because of this heritage, some analysts are still application focused in their criteria for data mart legitimacy. In their view, if a data mart does not have a dedicated application for its data set, it does not qualify… In the quickly evolving world of data marts, this criteria is losing its relevancy, as data marts become the general purpose DSS (Decision Support System) information repository it becomes increasingly difficult to assign specific dedicated applications to specific dedicated data sets.[6]

Specific criteria of size, aggregation level, number of users, location, topology, and audience are no longer relevant. Today's data mart can be an information source for a small workgroup or an entire corporation. Detail can no longer be excluded. Many users require detail. At our organisation, engineering and customer reporting applications require highly detailed data, while plant management require aggregated yield points by manufacturing stage. The key element is that the data mart is an architected subset of the enterprise data and that the data mart may well be an evolutionary object. It is no longer "application-centric".

Hackney[6] defines data mart users as sharing one or more of the following criteria:

- common business problems
- common semantics
- common metrics

- common roles/levels
- common mission
- common domain
- common geography

We all have data sets that belong to one of these categories, but if they have not been designed into an industry-standard architecture, all you have is a "data dumpster". If the data set is integrated and built to a valid design and/or represents a valid enterprise information architecture, then you can call that data set a "date mart".

Role of the Enterprise Data Model

It is important to have a conceptual model in place that accurately represents the company's business. An ERD (Entity Relationship Diagram) is the usual tool for accomplishing this. The important part of the conceptual model is that it is normalized. Also, an ERD implemented as a corporate database will be reasonably stable, since its structure reflects the inherent structure of the business. Note that almost never should the modeller show an ERD to a user. They won't understand it. Function hierarchy diagrams and, most important, process diagrams more easily understood by the user community.

The data warehouse "phenomenon" is recognition of the fact that we really do have to move data from the conceptual model to external views. While the conceptual model might be the best way of storing data it may not be the best structure for retrieving them.

Data Stores, Warehouses, and Data Marts

Data is collected from testing and measurement machines, the MES system, and user input and is stored in the *operational data store.* Ideally the ODS is based on a conceptual schema, but if not, it is a good idea to implement a variation on the ODS which could be called a *central data warehouse.* Then the task of creating each data mart can be defined in terms of clearly understood concepts. (See Figure 2 below).

The *central data warehouse* is the physical embodiment of the conceptual model. A *data mart* is a representation of the user's view of the business and as such is not necessarily normalised. The data stored in the data warehouse, unlike the Operational Data Store, is optimised to allow fast analysis and retrieval. Therefore the overall data ware-

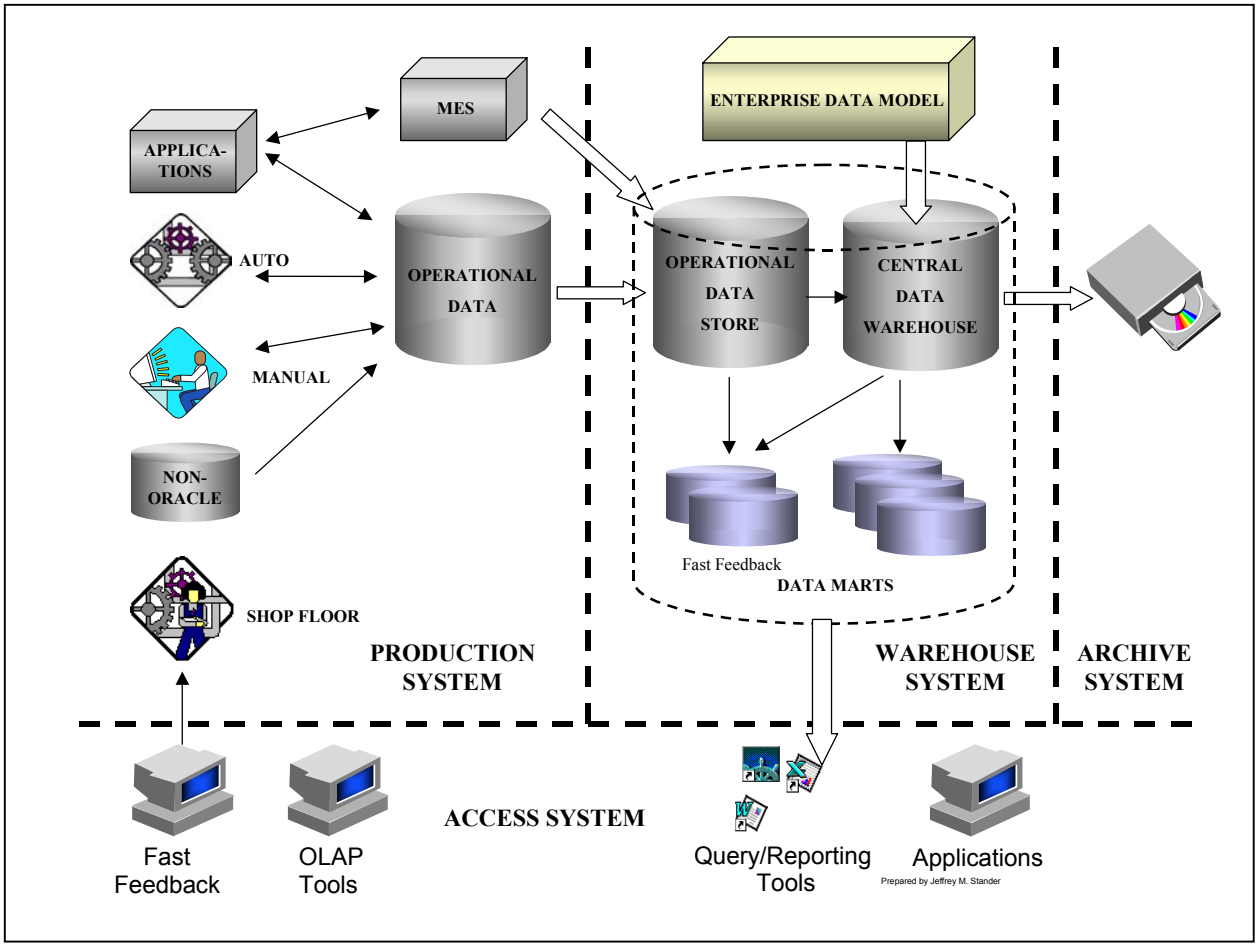

Figure 2. Production and Warehouse Data Architecture Showing role of the Enterprise Data Model.

house effort consists of two parts, the *central data warehouse* and *data marts.*

A data mart is a representation of the user's view of the business and as such is necessarily denormalised. Each data mart aggregates data in terms of the particular expectations of what someone will want to see.

The data represents a time series and is updated regularly. Each data mart aggregates data in terms of the particular expectations of what someone will want to see. Operational data that will not be used for decision support will be excluded. The use of multi-dimensional data browsing tools like Oracle Express™ and the redesigned and newly released Discoverer 3.0™ have changed the nature of decision support systems, allowing business users at all levels of an enterprise to gain immediate access to information. As the user's capabilities and expectations evolve, so will the data marts. The underlying data store should not.

The Role of the Warehouse and Data Marts

Because we want to deal with the legacy issues without legitimising them we have chosen an evolutionary, event-driven approach to data warehousing. This caters to the diversity of operational systems and application-specific information bases currently installed.

At the other end of the data stream, we use the data mart approach to cater to the diversity of the end-user community. The data marts are part of the overall data warehouse strategy based on the EIA concept, except that each data mart targets a specific user group. These can support fast feedback to the shop floor, engineering analysis, product yield analysis for production management, activity and inventory reporting, sales, corporate management and human resources.

To summarise the above:

- the cultural problems internal to the enterprise require an evolutionary or stepwise approach to change
- the paradigm for structural change is the Enterprise Information Architecture
- the data warehouse is an instantiation of the Enterprise Data Model
- the data marts cater to end-user diversity

Data Stores in the Manufacturing Environment

Data warehousing reflects the need for corporate users to view enterprise data from their own perceptual viewpoint. The user's perception of his business objects is not as entities and relationships. In other words the analysis environment is different from the transaction-processing environment. Although the relational model facilitates transactional processing and storage of data, it is usually not the best structure for retrieving data.

In the materials manufacturing environment we are faced with a great deal of operational data. At each process stage we have inventories of materials which are arrive as raw materials from upstream stages, are in process, are rejected, or are exported to the next stage (or shipped as finished goods). Usually an MES, or Manufacturing Execution System (not necessarily supported by an Oracle-compliant database or even a relational database) handles the process data, including test metrics, manufacturing recipes, automatic controls, activities and inventories. This mass of data is not structured to facilitate either *ad hoc* or canned queries against the operational database, often simply because the plant is running 24 hours per day, 7 days per week, and any slowdown in the computer or databases can impact production.

The classic method of the engineering-centric fabrication plant is to construct painstaking reporting and analysis based on one-off flat file extractions from operational data. The engineering mentality is such that the engineers request (No, demand!) access to the raw data and will attempt to process it using any tool at their disposal including spreadsheets, PC "databases", word processors, statistical packages, and custom applications written in a plethora of obsolete languages and packages. This often leads to down-stream dependencies of which the IS department has no knowledge. Until they change something, THEN the IS Department will hear about it. Almost worse, the user will deliver their pet package to IS and demand deployment and support. This brute-force "can-do" approach is part of the culture of the manufacturing company. Remember their motto: "Design is a waste of time so just code-like-hell".

Generating and Populating the Warehouse

We are all stuck with legacy systems. It seems that the concept of the data warehouse is a means where to overcome this problem. In our manufacturing plant, and in others, ERP (Enterprise Resource Planning) systems promise common data stores for all elements of the business process. This has not happened yet; so in order to integrate process and data in the enterprise we build warehouses. There are normally three approaches:

1. Build a warehouse that is driven by events.

2. Built a virtual data warehouse which is a "warehouse of warehouses"
3. Use the data warehouse model as the basis for creating a universal model with a standard API. This API is then used to integrate applications as well as to standardise the data in individual warehouses. Each warehouse models part of the business but together they model all of it.

Although the third method has appeal (it springs directly from the data model) it requires a great deal of co-operation between the owners of the data. In the environment of the medium-sized production-centric manufacturing plant the event-driven approach works the bests. If applications can be integrated then the data warehouse is built as a consequence of that process. The events that drive the business processes are actually the shared data that gets passed between the business processes. In the beginning the data warehouse appeals to the users as a literal instantiation – it is a place to store the historical mass of data generated by these processes.

While this approach satisfies an immediate need, future requirements must be considered. The goal is an enterprise architecture that integrates the business processes and the data assets.

Problems

Once more I stress that the data warehouse is built expressly for end users. The data source and the end-use often span multiple departments. Data ownership is usually a sensitive issue in an enterprise. Often the end user requirements are different, e.g. one group wants summary data and another wants detailed data. At the other end of the pipeline the owners of the OLTP system are concerned about being locked into providing the source data in a specific manner or that the extract will impact their systems. It is necessary to put a strong focus on data management and ownership and to resolve conflicts up front.

In practice, the OLTP systems feed the Operational Data Store on a transactional basis. Any immediate feedback to the production line is drawn directly from the ODS. The Central Data Warehouse is populated on a regular schedule.

We normally take a three-step approach to actually defining, creating and deploying a warehouse.

1. **Obtain and model user requirements.**

   The business needs of the users are extracted as a set of requirements to be transformed into a warehouse/data mart. More than one business area can be defined.

2. **Define data sources.**

   This is the meta-data model. This defines the data and the rules associated with this data

   a. **Describe data sources.** These are the rules for extracting, cleaning and moving data to the warehouse.

   b. **Describe business uses of the data.** This is the model of the business use of the information in the warehouse. Users use this plan to access and search their data.

3. **Describe Stocking Methods.** Here we make the final decision on the tools and methods to stock the warehouse and data mart(s). This step depends upon choosing the access methods that the users will use to access the data.

Accessing the Warehouse – End User Tools

Oracle tools span the full range from ad hoc query and reporting to multi-dimensional analysis. At the present time there is no requirement for developing full OLAP capability. The user community is simply not ready for a full OLAP against a MDDB. The current situation is users taking printed reports (often a single report) and hand keying the information into spreadsheets, often the same information. What we needed was a simple, plug-in, front-end tool that supported standard reporting and ad-hoc queries. Our focus has been on Discoverer 3.0™. Because the existing culture assumes a warehouse is only a place to keep data (the "data dumpster") we have adopted the Field of Dreams approach – "build it and they will come". This is not necessarily the best methodology but it does satisfy the "show me" mentality, and in this case is easy to deploy.

Discoverer 3.0™

Data marts not only do not have to be normalized, they *should not* be if they have to satisfy requirements for fast retrieval and analysis. The model of the data mart is not two-dimensional (like the relational model) but often cubic, or multi-dimensional. The data is collapsed into one or more *fact tables*, related to each of the dimensions. With the fact entity in the centre and the dimension entities ("product", "time", "location") distributed around the fact entity, we have the design which is called a "star schema".

Special tools such as Discoverer 3.0 and Oracle Express are required to view this data. Discoverer 3.0 is not a new version. It has been entirely rewritten and has a lot of functionality that was not present in previous releases. Its focus is on the

user's who have business needs, not on the data structures.

Because Discoverer 3.0 has a server based administrative mode it is easy to manage. The users find it simple to reference the data in terms they understand. As well, the tool allows reports to be exported as HTML for web deployment, or to "drill out" to tools like MS Word or the Excel spreadsheet.

Conclusion

The Main Issue

- To provide a shared corporate intellect.

Action Items to Achieve This

- That we consider the manufacturing enterprise as the system that it really is — an entity that uses human, financial and physical resources to convert raw materials into products.
- That as part of this approach a standard development methodology be defined and applied to all IS projects.
- That we build a conceptual data model with an ERD as the prime deliverable.
- That we use this conceptual model (ERD) to design a **central data warehouse** (CDW) that captures data from the **operational data store** (ODS) or other sources.
- That we use the CDW as a source of information to construct **data marts** tailored to specific production and business management needs. (Not an easy task)
- That we design and build **forms and reports** and configure **end-user tools** to control and use the information in the data marts.
- That we choose a tool or set of tools adequate for the users to browse, extract, and analyse this data (Discoverer 3.0)

References

1. The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, Ralph Kimball, John Wiley, 1996.
2. OLAP Solutions: Building Multidimensional Information Systems, Eric Thomsen, John Wiley, 1997.
3. Planning and Designing the Data Warehouse, Ramon Barquin and Herb Edelstein, chapter 10, Prentice Hall PTR, 1996.
4. Building a Data Warehouse for Decision Support, Vidette Poe, Prentice Hall, 1995.

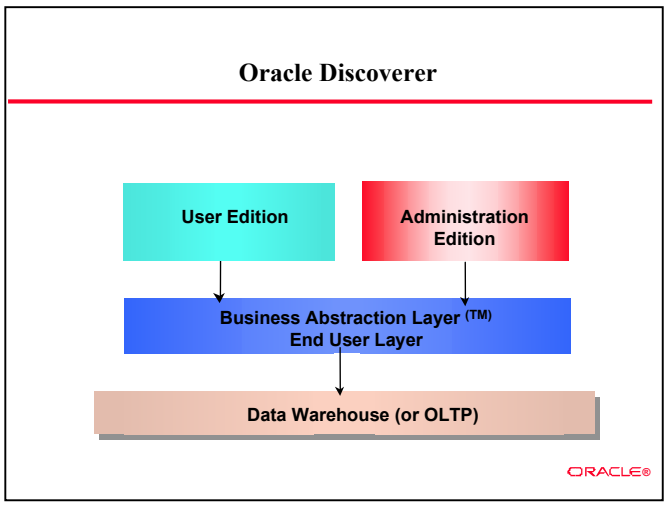

Figure 3. Oracle Discoverer 3.0 Architecture

5. Larry Greenfield's data warehousing home page http://pwp.starnetinc.com/larryg/index.html
6. Douglas Hackney, What is a Data Mart?, Data Management Review Magazine, February 1997. Excerpted from "Understanding and Implementing Successful Data Marts", a forthcoming book from Addison Wesley Longman Publishing. See http://www.entergroupltd.com.