From Data Processing To Information Analysis - A Complete Oracle Solution

Howard Ong

Senior Consultant

Aurora Consulting Pty Ltd

Abstract

Since the dawn of the data processing age, organisations have been building application systems aimed at addressing their operational needs. The IT industry has reached a stage where many organisations now possess years, if not decades, of detailed operational data, buried in a variety of legacy systems. The constant quest for a competitive edge has driven many organisations from sheer data processing to the new exciting arena of information analysis. In meeting these new challenges, traditional legacy systems typically suffered from a number of serious shortcomings.

The advent of Data Warehouses provided a ray of hope in overcoming the shortcomings. Data Warehousing has come a long way since the days of what-if analysis running on Lotus 1-2-3 spreadsheets. The barrage of new terms and acronyms that have emerged in recent years - DSS, EIS, ETT, OLAP, Data Mining and the term “Data Warehouse” itself, signifies a fast expanding field and the subsequent rapid gain in interest by the Information Technology (IT) community. Recent advancements in both technological and academic fronts have brought the reality ever closer to the ideal information analysis environment.

This paper shares the experiences of the UNIPHI Project at Hamersley Iron. It covers topics such as the motivation that led to UNIPHI’s inception; its evolution path; technical architecture; its innovative implementation of Oracle RDBMS and Express technology; and future challenges that it faces.

About The Author

A principal of Aurora Consulting Pty Ltd and a frequent presenter at Oracle Conferences throughout the region, Howard has been working on Data Warehouses since 1991, before the term "Data Warehouse" was even coined. Howard possesses in-depth experience in the planning and development of Data Warehouses; and has been an expert user of Oracle database and tools for the past 6 years. Harbouring a keen interest in the application of Oracle Technology in Data Warehouse development, Howard and his team of consultants have helped many organisations deploy Data Warehouses using technology such as Oracle Express, Discoverer, advance features of the Oracle8 Server and Aurora's proprietary DataWare and DataView products. Aurora’s most recent success includes a major Data Warehouse initiatives at Education Department Western Australia. In other projects throughout his many years of consulting experience, Howard has worked with a wide variety of organisations such as government departments; and companies from mining, finance and transportation industries. In addition to Data Warehouse development, other services that Howard rendered to these organisations include business analysis, relational application development, multidimensional database design and database administration.

For more information on Howard and Aurora Consulting, visit http://www.aurora-consult.com.au or email to info@aurora-consult.com.au.

Introduction

Walk into any computer book store today and you will likely stumble upon a barrage of books written on Data Warehousing. What exactly is a Data Warehouse ? Why has it become such a popular subject not just in the Information Technology (IT) community, but in the general business community as well ? Why do we need it ? How can we build one ? What software and tools are available to help us ?

This paper presents the UNIPHI Data Warehouse Project at Hamersley Iron. The authors hope to shed some light on some of the questions raised by sharing the experiences gained from the UNIPHI Project. The paper covers topics such as the motivation that led to UNIPHI's inception; its evolution path; technical architecture; its innovative implementation of Oracle RDBMS and Express technology; and future challenges that it faces. In particular, the evolution of the UNIPHI Project from a single-subject data mart to a full-fledge relational and multidimensional Data Warehouse serves as a guide to organisations seeking to follow the same path. The UNIPHI infrastructure addresses key concerns pertaining to the Extraction, Transportation and Transformation (ETT) of data, as well as the issue of presenting information from the Data Warehouse to the management.

What Is A Data Warehouse

What exactly is a Data Warehouse ? As with any new discipline, several authors have suggested several different definitions. In this paper, we will examine 2 of these. Corey and Abbey defined a Data Warehouse as "a collection of corporate information, derived directly from operational systems and some external data sources".

This is an excellent definition in that it captures the essence of a Data Warehouse from a "content" perspective. The early 1960s saw the advent of massive batch-processing systems, where transaction files stored on punch cards are periodically processed against tape-based master files. With the introduction of new technology such as RDBMS and client-server computing, increasingly larger and more complex application systems were built, all aimed to address the operational needs of organisations better. The IT industry has reached a stage where many organisations now possess years, if not decades, of detailed operational data, buried in a convoluted heap of legacy systems. Most of these systems represent a showpiece of the prevailing technology at the time they were built, sort of like a "live" time capsule. Sentimental values asides, they present a ultimate nightmare to the today's system developers. System developers today are under increasing pressure to harness the wealth of data encapsulated within these legacy systems, steering them from merely supporting business operations to supporting *business decisions*; from sheer data processing to the new exciting arena of *information analysis*.

Until recently, the demand for information analysis has been met by standalone data extraction programs, which extract data into spreadsheets where what-if analysis are performed. Alas, like the effect of a pain-killer, this approach only offers temporary relief. Organisations are now feeling the pain, yet again. As standalone extraction programs proliferate throughout an organisation, more problems result. W.H. Inmon, in his book *Building The Data Warehouse*, highlighted 3 of these :

- Credibility of data.
- Productivity.
- Inability to transform data into information.

One of the authors could recall a painful experience where he spent 2 months investigating the reasons a report based on data in the Data Warehouse was different from that produced from a standalone extraction program. In the end, it boiled down to differences in time basis, algorithm and data classification.

Inmon defined a Data Warehouse as "a subject-oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions". This definition examines the Data Warehouse from a qualitative perspective. Readers are highlighted to the 4 qualities of *subject-oriented*, *integrated*, *non-volatile* and *time variant*. With these qualities, a Data Warehouse sought to provide consistent and reliable information to help decision-makers make better decisions, from human resource policies to advertising campaigns. A recent IDC study revealed that the average return on investment of Data Warehousing over a 3 year period was 401%. Judging from that, it is not hard to conclude that Data Warehouses have indeed helped improve decision-making in organisations.

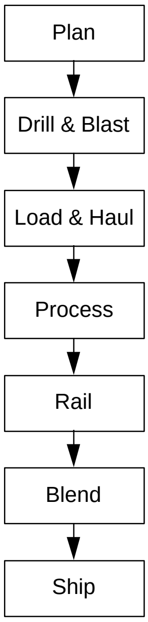

Figure 1 Overview of Hamersley Iron's Operational Process

About Hamersley Iron

Hamersley Iron Pty Ltd is a wholly owned company within the Rio Tinto Group. 1996 marked the thirtieth year in which Hamersley Iron has been mining, processing and shipping iron ore. It has been one of the most profitable companies for some time and has had record-breaking performance for a number of years. Hamersley Iron has been Australia's largest iron ore exporting company for some time. The ore is exported principally to Japan, China and other South East Asian Countries. Europe is the other market that is serviced by the company.

Hamersley Iron has a base of five mines in the Pilbara region in Northern West Australia. The operating mines are Mount Tom Price, Paraburdoo, Brockman No 2 Detritals, Marandoo and Channar (a joint venture 60% owned by Hamersley Iron). At each of the mine sites, ore is processed in order to improve the percentage of iron ore and to remove various impurities. Hamersley Iron currently markets two type of products. The first is lump, which is the premium product. The second is fines which are smaller lumps of iron ore. The lump iron ore is more suitable for direct feed into iron blast furnaces. The fines product require some preliminary processing and is therefore not always as attractive.

The company has one of the longest privately owned and operated railway in the world. The rail system is used to transport the ore from the mine sites to the port at Dampier. The ore is dumped from the trains and blended onto stockpiles to produce a known quality mix ready for shipment. Two berths, Parker Point and East Intercourse Island, are used to load the iron ore onto ships for transport to the respective markets. Figure 1 illustrates the broad extraction process used by Hamersley Iron going from mining the ore to shipping the ore.

A key element of Hamersley Iron's success has been its ability to maintain a consistent blend of iron ore from the various mines to achieve the required quality in relation to the percentage of iron ore and other impurities. The consistency of product is important for the operators of blast furnaces to ensure that the most appropriate product mix is used for the production of iron. As testimony to its success, in 1996 Hamersley produced and shipped about 56 million wet Mt which was the third production record year in succession.

Hamersley has continued to invest in technology to improve quality and reliability. Some examples of this include the use of robotics for product sampling and analysis at the mines and the port; and the introduction of a fleet of new GE Dash-9 locomotives. These locomotives only require single man operation and have resulting in a considerable cost saving.

Approval has been given recently for the development of a mine at Yandicoogina. Construction of the mine site and extension of the existing rail system has yet to begin. Yandicoogina is a different ore type to that mined from the existing mines. This will create some challenges in the future in terms of marketing, stockpiling and shipping.

Why UNIPHI ?

Prior to Uniphi, the business identified the importance of information as key to the effective running the business and improvement its market position. In 1995, the Executive Committee made the provision of timely, accurate and easily accessible information one of the business' key strategic goals.

A business measures project was initiated in 1995 to look at developing a consolidated and consistent set of measures across the business. The project used a methodology called Customer Focussed Process Management based on the SIPOC (Supplier-Input-Process-Output-Customer) TQM model. The project developed a framework which provided a starting point for determining the measures that would be included in the Data Warehouse.

In addition, a prototype business measures application was developed using Microsoft Access. This prototype demonstrated the viability of using a tool to access, relate and graph different measures in a similar fashion to an OLAP query tool. Some of the underlying thinking that went into the design and features of this prototype was reflected in the subsequent front-end application developed using Oracle Express Objects later.

In order to provide information to senior management, a paper based reporting system had evolved for some time. This reporting process was managed by the Management Accounting Section and drew data from across the whole business. Even though the reports had evolved over time, the increased focus on better information precipitated a review of these reports. The two main types of reports that were targeted was the Managing Director's monthly report, which focussed at the business unit level; and General Manager's monthly reports which had a more detailed operational focus. These reports were largely compiled manually through the use of a maze of inter-related spreadsheets and manual entry. Due to the intense manual effort and data collection required, the reports were often not completed until the 20th day of the month - just before the next reporting cycle would begin.

In order to give the reader some appreciation of the size of reports, the Managing Director's reports covered the three business units within the company - Development, Operations and Marketing. The reports were typically two centimetres thick and covered a wide range of information such as:

- overall business performance;
- financial summary (balance sheet, cash flow, profit and loss);
- mine performance (e.g. actual against plan for expenditure, actual against plan for tonnes mined, quality, waste, equipment utilisation, ore enrichment yields from the processing plants, and so on);
- port performance;
- rail performance;
- market overview and performance;
- health and safety;
- capital project summary;
- manning levels; and
- service and support areas performance (e.g. power).

The General Manager's reports varied in size depending upon the particular area of responsibility and preference of the general manager. These reports tended to be more detailed and hence required more data collection.

The presentation of the information in paper format also meant that the reports were voluminous. Different business measures had to be represented by a series of tables or graphs each reflecting a different aspect (or dimension) of that measure. It also did not allow information that could be related to be viewed together unless a specific table or graph was created.

In order to address the information needs at the senior management level, some environment needed to be developed to provide the key business performance information in a simple but powerful way, and to streamline the reporting process

From an IT perspective, with years of operational data buried in legacy systems, Hamersley Iron, like most organisations, possess all the data necessary to support strategic decision making. However, operational systems were never built to support high-level business decision-making. The attempts to make these systems fulfil the decision support role were like trying to fit a square block in a round hole.

In more concrete terms, the problems can be summarised as follows :

Data Summarisation

Strategic decision-making often requires summarised data of various degree. Often, the problem is complicated by the fact that a business measure cannot be directly obtained as a base measure from the operational systems, but has to be derived from several base measures. In addition, summarisation is often performed on the fly and is thus lacking in performance and flexibility.

Islands Of Information

Frequently, sound decision-making requires cross-analysis of business measures from different operational systems. For example, to ascertain the average cost of mining one tonne of iron ore, measures from the 2 operational systems - the One Mine System and the Costing System are needed. Such situations of cross-system analysis can be complicated, if possible at all. Among the issues are structural differences in the measures, classification differences in the dimensions, hardware and software differences among the operational systems.

Productivity

Traditional query tools are unsuitable for multidimensional analysis, the type of analysis often required of Decision Support Systems. Often, major workaround is required before any multidimensional analysis can even be contemplated. A spreadsheet interface enhances the analysis capabilities but suffers from other limitations :

- The awkwardness in analysing measures of more than 2 dimensions.
- The additional extraction step required.
- The tediousness in presenting the measure at the required level and format.

Thus, the direct analysis of data from operational systems typically suffers from productivity problems due to its labour intensity and long turn-around time.

Inflexibility

One of the key characteristics of Decision Support analysis is that the user frequently cannot predict what he or she would like to know about the data. Often, there are follow-up queries that are even more unpredictable. For example, a typical question might be “for 1996, break down the total tonnes mined by mine sites and months”. Upon analysing the answer to this question, a follow-up question might be “for Tom Price (a mine site), break down the tonnes mined figures by the individual plants, the grades of ore mined and months”. Alternatively, the follow-up analysis could follow a drastically different track, with a follow-up question such as “while keeping the same break down by the mine sites and months, display the figures for tonnes blasted in addition to those for tonnes mined”.

With traditional query tools, each of the aforementioned question often requires a separate customised report to be written. Recalling these reports usually suffer from long turn-around time, it is not hard to conclude that using traditional tools for Decision Support is an extremely frustrating experience indeed.

The UNIPHI Project sought to address these issues by building an integrated repository of consistent and reliable information, in contemporary term, a Data Warehouse. (And that is yet another definition of a Data Warehouse - “an integrated repository of consistent and reliable information”.) UNIPHI stands for “Unequivocal Information Pool for Hamersley Iron”. With the innovative use of Oracle RDBMS and advance PL/SQL features, much of the cleansing, derivation and integration of operational data from heterogeneous sources were automated. By deploying appropriate Online Analytical Processing (OLAP) tools at the front-end, multidimensional queries are supported in a versatile and productive manner.

The Evolution

Rome was not built in a day, neither was UNIPHI. Like most Data Warehouse Project, it started with a Proof-Of-Concept (POC) Phase, where a small subject-specific Data Mart was built. Between April to June 1996, a prototype system was developed, using data from the Costing System. 2 measures were defined for this POC Data Mart. Oracle Express Server formed the back-end while Oracle Sales Analyzer (OSA) was chosen as the front-end presentation tool. OSA was chosen for the following reasons :

- Its rapid development capability allows the power of Express multidimensional capabilities to be harnessed quickly.
- More powerful tools like the object-oriented Oracle Express Analyzer and Oracle Express Objects were not available at that time.

Figure 2 illustrates the architecture of this initial prototype.

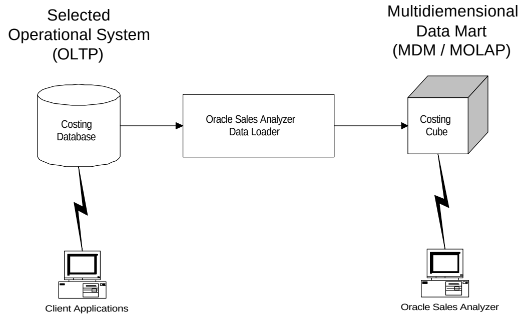

Figure 2 UNIPHI System Architecture - Stage 1

The prototype serves a very important function of getting the users acquainted with the new IT paradigm called OLAP analysis. As users' confidence grew, what Corey and Abbey called "power users" started to emerge. These are users who have "a desire to learn technology and a willingness to teach others". Power users could be anyone from an IT manager to a technically oriented accountant. Power users help the Data Warehouse developers promote the Data Warehouse to the wider user-community, a crucial role in the success of the Project.

After the stage 1 had been completed, the conceptual design of the Data Warehouse architecture was conceived and completed by one of the authors. Figure 3 illustrates the underlying thinking. This design then evolved as outlined in the remainder of this paper.

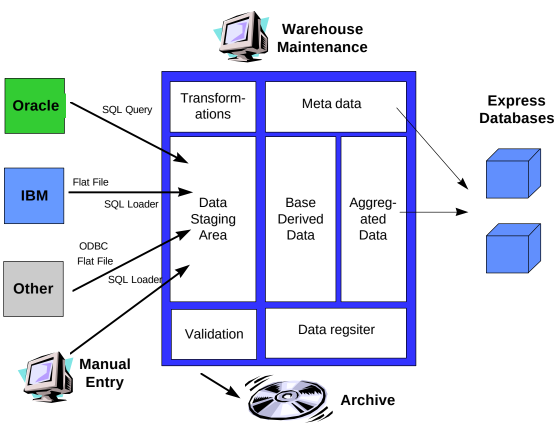

Figure 3 Conceptual design and architecture.

In August 1996, the UNIPHI Project was ready to take its second step - the building of a second prototype. The Financial Statement Data Mart consisted of 7 measures and was developed as an alternative to analysis that was performed on Lotus spreadsheets. It drew its data from the spreadsheets, which in turn loaded data from an old financial database developed in Btrieve. OSA was by then a familiar interface to many and was again chosen as the tool for implementing this new Data Mart. The new architecture is illustrated in Figure 4.

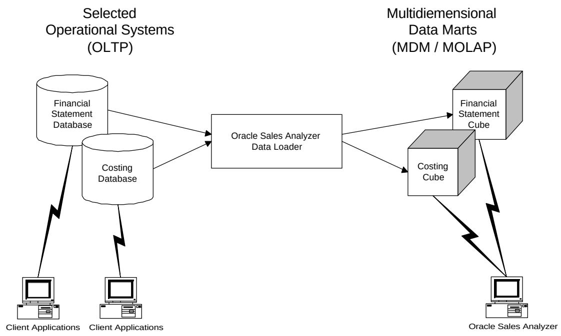

Figure 4 UNIPHI System Architecture - Stage 2

The next significant event took place in November 1996. A blue print was drawn for the development of a Multidimensional Data Warehouse (MDW). Incorporating the measures from the original Costing and Financial Statement Data Marts, the new Data Warehouse consisted of 89 measures, covering 3 major areas of Key Performance Indicators (KPI), Operations Analysis Database (OAD) and Marketing Analysis Database (MAD). This represented a major rationalisation effort on

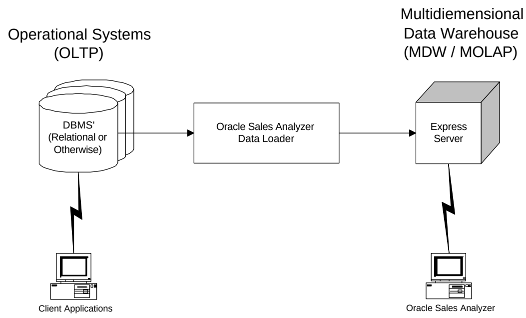

Figure 5 UNIPHI System Architecture - Stage 3

the functional structure of the Multidimensional Data Warehouse. Figure 5 illustrates the architecture of this first-cut Data Warehouse architecture.

It was soon apparent that while the development of the MDW represented a major milestone, the architecture lacked versatility and flexibility in a number of ways :

- All scripts used by the OSA Loader, what OSA terminology called *fixed format files*, were manually written and maintained. This made the maintenance of the MDW rather laborious. The person performing the maintenance would also need in-depth knowledge of the OSA Loader.
- The loading of the cubes were performed manually, which added further to the labour-intensity of the maintenance activities.
- Data extracted from the source operational systems were directly loaded into the MDW. This gave little opportunity for any data cleansing activity to be performed. In fact, the only data cleansing available was the default structural validation performed by the OSA Loader. Semantic data cleansing was not supported by the architecture.
- In any case, leaving structural validation to the OSA Loader was not recommended. As some of the cubes were quite large (up to 4 GB in size), their loading was rather time-consuming. Leaving the OSA Loader to do the validation implied that error detection is left till a rather late stage in the loading cycle. The multiple reloads that were required before the data could be correctly loaded raised some concerns on the efficiency of the architecture.
- The architecture supported 2 forms of derived measures - Source System Derivation (i.e. derivation performed within a source operational system); and MDW Derivation (i.e. derivation performed within the MDW using Express formulae). Source system derivation suffers from complexity and feasibility problems where the base measures are from different source systems. Formula derivation, on the other hand, does not support complex derivation. With these limitation in mind, it became apparent that the system needed a procedural form of derivation, to be performed after the data are extracted from the source operational systems but before they are loaded into the MDW.
- Data from external sources such as CPI figures and exchange rates needs to be captured by manual entry. The architecture did not support any manual mean of data entry.

These shortcomings paved the way for the incorporating of a Relational Data Warehouse into the architecture.

Figure 6 illustrates the present architecture of the UNIPHI Data Warehouse. The architecture consists of a separate Relational Data Warehouse (RDW) and a Multidimensional Data Warehouse (MDW).

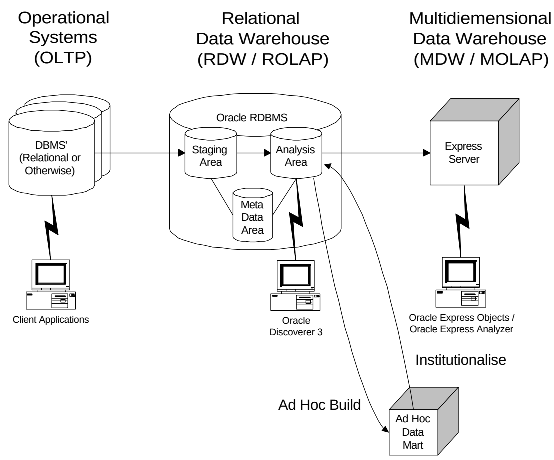

Figure 6 UNIPHI System Architecture - Stage 4

At the multidimensional end, OSA is superseded by Oracle Express Objects (OEO) and Oracle Express Analyzer (OEA), Oracle's object-oriented development environment for OLAP applications. Complete with DDE and OLE capability, OEO/OEA provided the much needed upgrade to the rather mediocre user interface that OSA provides.

The RDW consists of 3 distinct components - a Staging Area, an Analysis Area and a Meta Data Area.

Staging Area

The Staging Area is a work area where data entering the Data Warehouse are pre-processed before they are moved to the Warehouse proper. Presently, 2 types of pre-processing are supported :

- Data cleansing and validation, which includes both structural and semantic validation.
- Derivation of data. Staging Area Derivation are not subjected the aforementioned limitations of Source System Derivation and MDW Derivation. It is procedural in nature, and is thus able to support very complex derivation. It is performed outside the source systems, which makes cross-system validation possible. Last but not least, it is performed after data cleansing, which means that reliable source measures are used for derivation.

The other role of the Staging Area is that it provides a logical location, where manually captured data can be entered into. Once entered, manually captured data are treated just like any other data, subjected to data cleansing and can act as a derivation source.

Note that once the data loaded from external source system or manually entered into the staging had been processed, the source data is archived in external files to create an archive log. The location of the data archive depends upon whether there was any problems identified with the data or not.

Analysis Area

The Analysis Area can be considered as the Relational Data Warehouse proper. It houses cleansed data that are ready to be queried by a Relational OLAP (ROLAP) tool like Oracle Discoverer 3. It is also from the Analysis Area that data are extracted and loaded into the MDW.

Meta Data Area

The Meta Data Area stores meta-level information about the Data Warehouse. Examples of meta-level

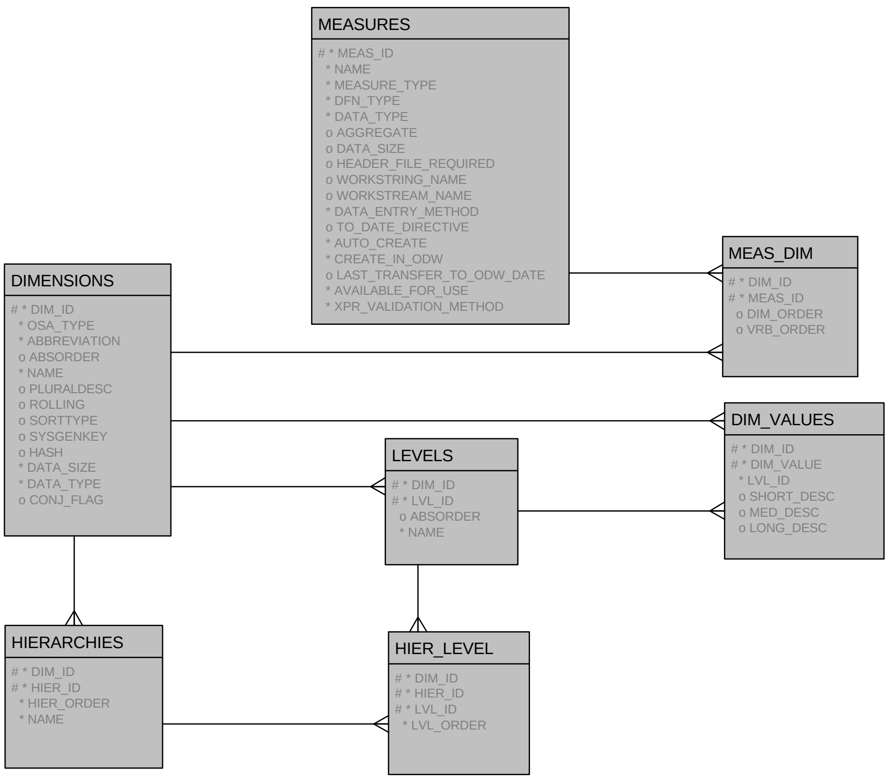

Figure 7 Selected Meta Data Area Tables Supporting Data-Driven Maintenance

information stored in the Meta Data Area are : Data Warehouse structure derivation logic, validation procedures, status of data loading processes, etc. These meta data support the administration and maintenance of the Data Warehouse in 2 significant ways :

- It provides a platform from which data-driven maintenance of the Data Warehouse is supported. Activities such as the creation of a new measures, changing the dimensional structure of a measure, modification of derivation or validation logic are all administered through the Meta Data Area. Figure 7 illustrates selected Meta Data tables used to support data-driven maintenance.
- The Meta Data Area also provides the data that support the Process Control Monitor. The Process Control Monitor is an utility that monitors and administers all ETT processes of the Data Warehouse. The Process Control Monitor is described further in the next section.

The UNIPHI ETT Processes

The UNIPHI ETT Processes are performed by an utility called the Process Control Monitor. The Process Control Monitor consists of the following components :

- Meta Data Area tables
- PL/SQL packages
- Unix shell scripts
- Front-end screen-based application

Workstream

The Process Control Monitor uses the concept of a *Workstream* to control and monitor the ETT processes. Each Workstream is defined as a series of Workstream Steps which specifies how a batch of data are processed in the Data Warehouse. Typical Workstream steps are initialisation, validation, data loading and data archival. A Workstream Step can either be automatic or manual. An automatic step is automatically invoked when the preceding step is completed successfully. A manual step, on the other hand, is manually invoked by the Data Warehouse Administrator through a Process Control screen.

Three generic Workstreams have been created to cater for the processing needs of most of the measures. These are :

- AUTO. This Workstream caters for measures whose loading process are fully automated. Automatic measures are the easiest to administer. The AUTO Workstream consists of the following automatic steps :

  ◊ Initialisation

  ◊ Initial Validation

  ◊ Main Validation

  ◊ Loading To Analysis Area

  ◊ Archival Of Staging Area Data

  An AUTO batch is extracted periodically from one of the operational systems in the form of a text file. This is placed in a designated directory on the host machine that houses the Data Warehouse. A Unix shell script detects the presence of the input text file and invokes SQL*Loader to loads the data into the staging area. A PL/SQL procedure then process the Workstring Steps according to their dependency and invocation methods.

- MANUAL. This Workstream caters for measures that are manually entered to the Data Warehouse. It is structured very similar to AUTO, with the only difference being that its first Workstream step is a manual step.

  ◊ Initialisation (manual)

  ◊ Initial Validation (automatic)

  ◊ Main Validation (automatic)

  ◊ Loading To Analysis Area (automatic)

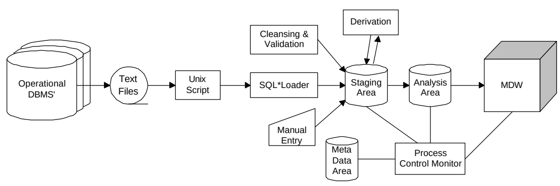

Figure 8 UNIPHI ETT Processes

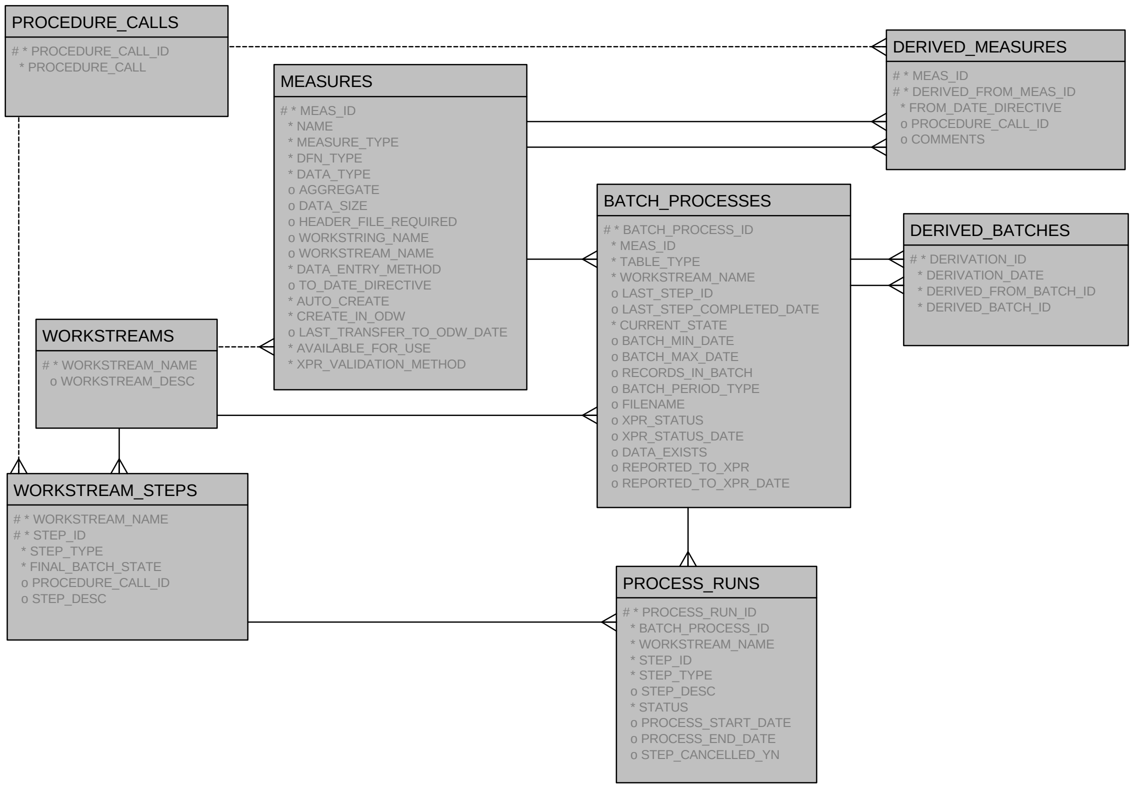

Figure 9 Selected Meta Data Area Tables Supporting ETT Processing

   ◊ Archival Of Staging Area Data (automatic)

   For the MANUAL Workstream, the Initialisation step is manually invoked once the user has finished entering a batch of manual data. From this point on, the batch of data is processed just like an AUTO batch.

- DERIVE. This Workstream caters for derived measures. It consists of an addition Derivation step which, upon the detection of all required source measures, invokes the derivation process. All its Workstream Steps are automatic.

   ◊ Initialisation

   ◊ Derivation

   ◊ Initial Validation

   ◊ Main Validation

   ◊ Loading To Analysis Area

   ◊ Archival Of Staging Area Data

It is also worth noting that while the generic Workstreams described above meet the needs of most measures, custom-built Workstreams have been created for special cases.

Figure 8 illustrates the ETT Processes of the Data Warehouse. Figure 9 is a database diagram of selected tables in the Meta Data Area, used for ETT processing.

Front-End Screens

Figure 10 is the Workstream maintenance screen. The upper part of the screen displays the Workstreams, while the lower part of the screen displays the Workstream Steps associated with the selected Workstream. The item "Procedure Call" identifies the procedure to be called when invoking the Workstream Step.

Figure 11 is a screen used to maintain derived measures. For each derived measures, 2 key pieces of information are entered : the measures used for derivation and the procedure used to perform the derivation.

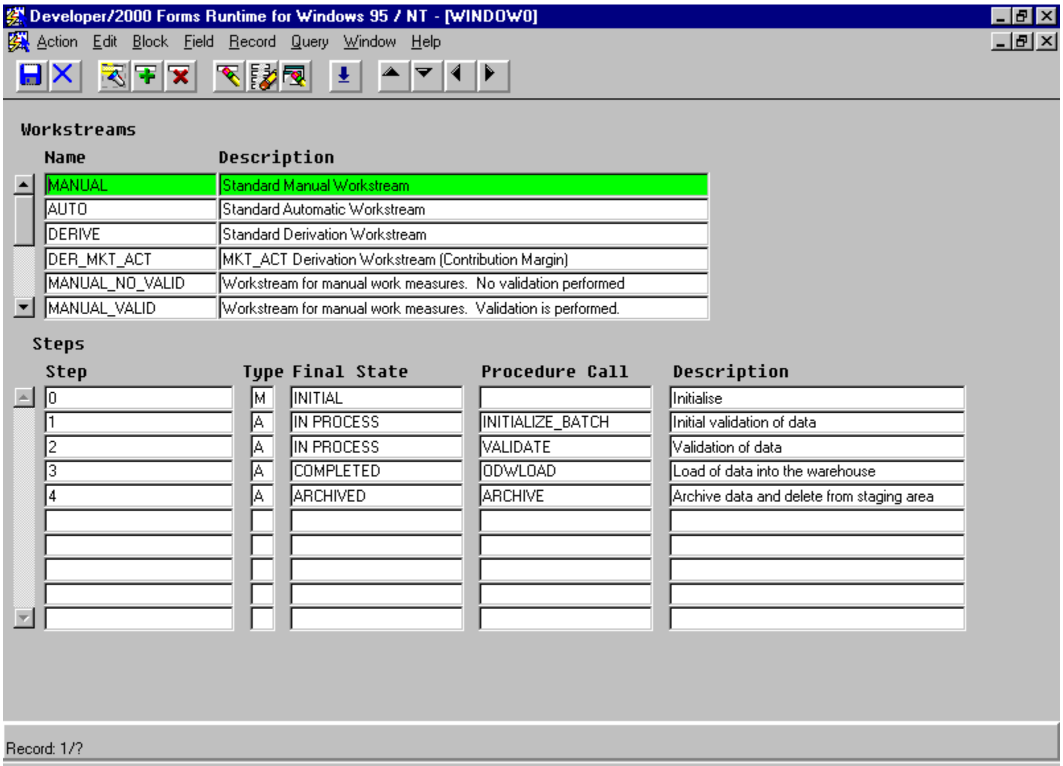

Figure 10 Workstream Maintenance Screen

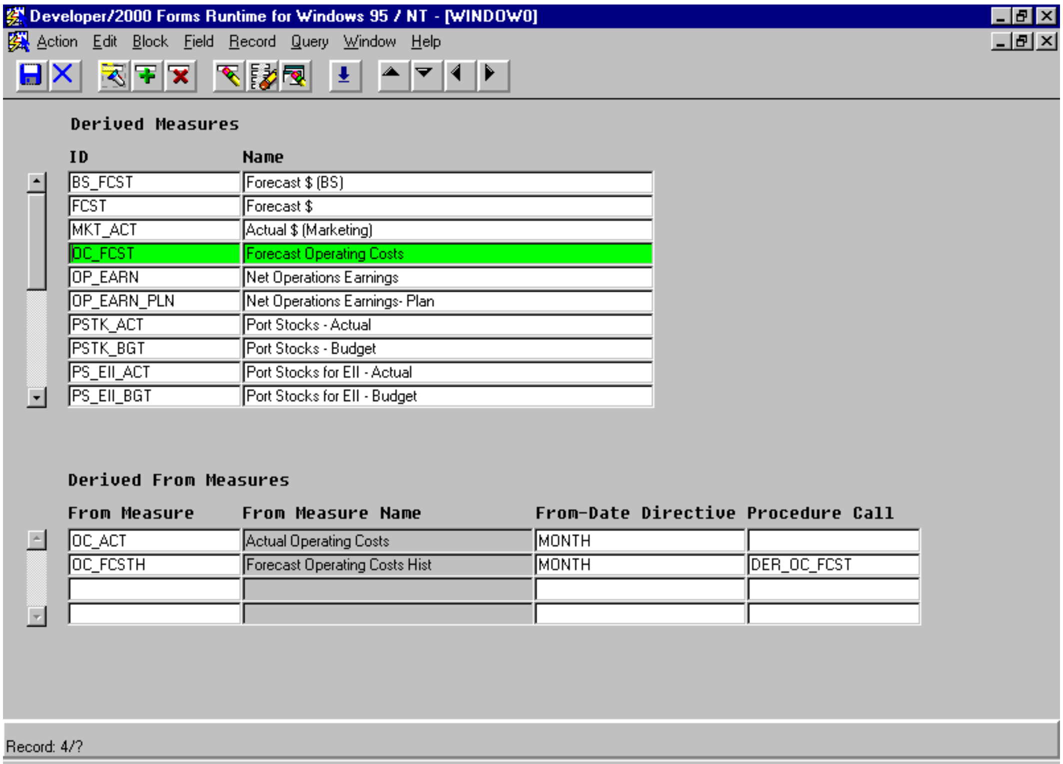

Figure 11 Derived Measures Maintenance Screen

The Procedure Call screen illustrated in Figure 12 maintains all procedures used in the ETT processes. Entered against each procedure is a piece of PL/SQL code. This piece of code is executed as dynamic SQL by the Process Control Monitor when the procedure is invoked. As demonstrated in the previous screens, these procedure can be used in any instance where a piece of programming logic is to be assigned to an ETT process, such as Workstream Step processing and measure derivation.

When Worksteams are invoked, they manifested in the Process Control Monitor as Batch Processes. Correspondingly, an instance of an executed Worksteam Step is manifested a Process Run Step. Figure 13 illustrates the Batch Processing screen. Notice that step 2 failed on its first invocation and was completed successfully on its second run.

Management of Meta Data

As indicated previously, the Meta Data Area dealt with data associated:

- with measures (auto, manual and derive);
- the dimensionality of measures;
- hierarchical relationships associated with the dimensions; and
- data used to support the Process Control Monitor and ETT processes.

All of which are required for the whole system to operate. However, this information does not cover the range of meta data that is required for managing the data warehouse environment effectively - particularly from a non-technical perspective. To complement the Meta Data Area, a Data Register was developed. The Data Register was designed to accommodate more general details about measures and dimensions. It contains details such as :

- description of the data;
- comments and notes relevant to the data;
- unit of measurement;
- the custodian of the data;
- historical information to aid in

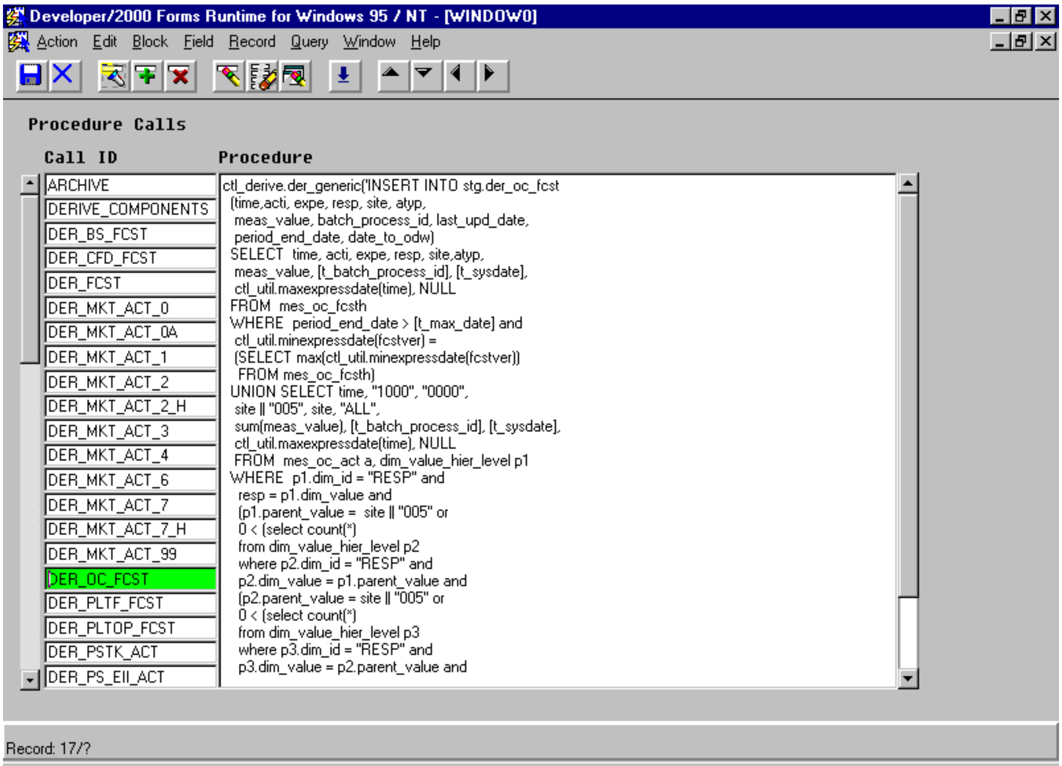

Figure 12 Procedure Calls Maintenance Screen

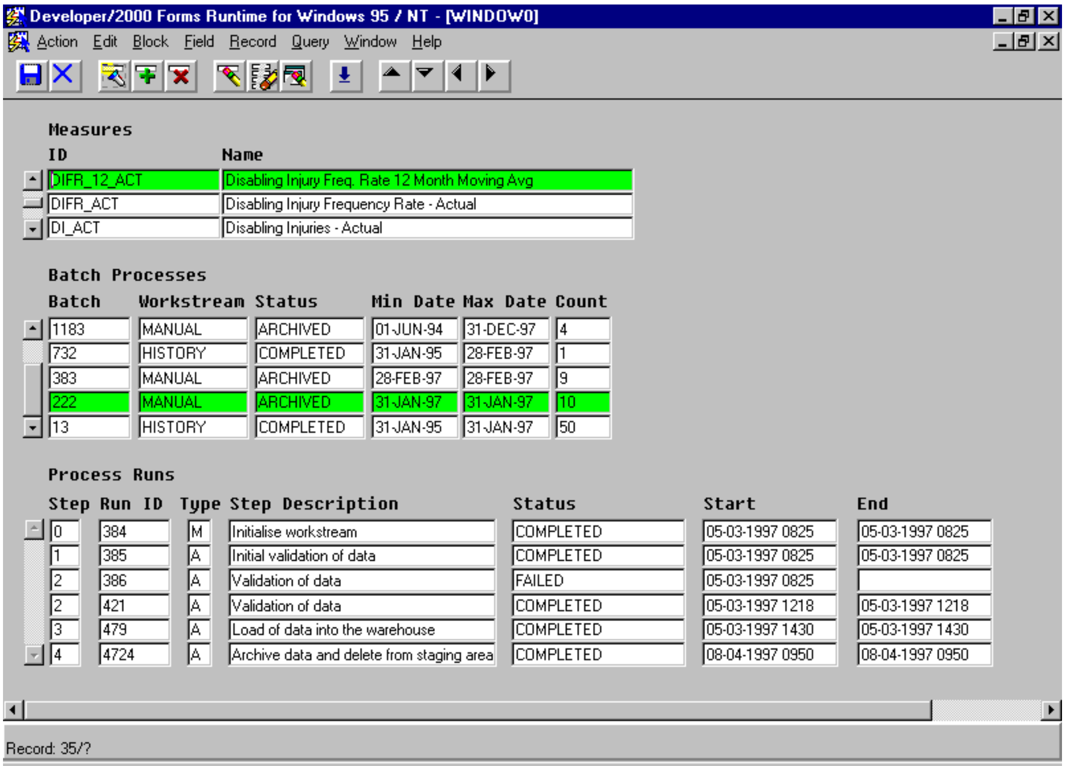

Figure 13 Batch Processes Screen

understanding changes to the data over time;

- the source of the data for loading into the data warehouse, and other sources of the data not used;
- details on the accuracy and reliability of the data;
- data validation details;
- frequency and the timing of data updates;
- granularity of the data; and
- why the data was included and the justification for the data.

In order to assist in the administration of the data warehouse, a Responsibility Allocation Matrix (RAM) was also developed in conjunction with the Data Register. The RAM identified which roles were responsible for the provision of data, the custodian of the data, who deals with data problems, and process outcomes from the ETT, and so on. Each role is associated with (usually) one person. In this way, the system is able to indicate who to contact about a particular issue or problem. In certain circumstances, an e-mail was generated to the relevant people fulfilling the roles based on a set of outcomes (e.g. an Express Cube failed to load for a particular reason).

The Meta Data and Data Register provide important data as to what the current status of the data is, and when data should be provided respectively. However, this information does not provide an adequate management tool for planning and co-ordinating the uploading of data into the data warehouse. Due to circumstances, changes in the business, or for a variety of other reasons, standard schedules are changed. The ability to allow flexibility in the system to allow this change is identified. This is particularly important during the month end process where there is a large volume of disparate data being loaded from multiple sources. As a result, a further module was developed which provided a data loading planning and scheduling tool. The information in the Data Register provides the seed data, but the Data Warehouse Administrator could modify the "standard" data loading timing and/or add additional data loads that occur on an ad-hoc basis (such as an update to a hierarchy) to suit the particular period of interest. This information provided a more detailed project schedule for that particular period facilitating the management of the process. The outcomes from the ETT provides the status of data loading. In doing so, the administration has a mechanism to determine whatdata has and has not be loaded against the schedule, has the required dependencies been completed, and the status of the data in terms of its progression from the staging area, to the data warehouse and to the appropriate express databases.

Future Challenges

While the UNIPHI Data Warehouse has come a long way since its inception, several areas of potential improvements has been identified, particularly in face of new emerging technologies. The key improvement areas are summarised below :

OFA

Oracle Financial Analyzer (OFA) was deployed in Hamersley Iron as a Financial Modelling tool at about the same time as OSA. While OSA has since been replaced by other tools, OFA has survived as an effective Financial Modelling tool. To date, OFA has existed as another multidimensional tool independent of UNIPHI. To promote centralised management of all Data Warehouse resources in the organisation and to leverage on the data resources and ETT capabilities of UNIPHI, OFA's incorporation into the UNIPHI Data Warehouse should be considered. There are 2 ways this can be achieved :

- In recent time, no bulk introduction of data has been performed on OFA. This could change in the near future as Hamersley Iron contemplates the construction of a new Financial Model. While a similar spreadsheet-based data loading mechanism could still be employed, the UNIPHI ETT engine offers a more powerful and attractive alternative.
- With the advent of object-oriented development tools such as OEO and OEA, it is possible to develop future Financial Model using these new technologies. Needless to say, cost-benefit analysis on the good-old custom-built versus off-the-shelf option would have to be performed. If and when a custom-built Financial Modelling application takes shape, it would be natural to incorporate that as part of the UNIPHI Data Warehouse.

Deployment Of ROLAP Tool

To date, the Analysis Area of the Relational Data Warehouse serves little more than a data feed to the Multidimensional Data Warehouse, even though the use of the ROLAP was considered as part of the of the original conceptual design. With some remodelling, the Relational Data Warehouse can serve as an effective alternative source of Warehouse information. Advance database features such as Star Query and Bitmap Indexing can be introduced to improve performance. ROLAP techniques such as data summarisation, employed by a number of ROLAP tools such as Oracle Discoverer 3, can further improve performance of ROLAP queries, rivalling those of MOLAP queries.

Web Enabling

For an organisation as geographically dispersed as Hamersley Iron, the Web enabling of an application is always an attractive option to promote wider usage. Through the deployment of appropriate Oracle Web Server cartridges and CGI programs, UNIPHI data can be made available in the web in several ways :

- PL/SQL Cartridge provides native access to the Relational Data Warehouse.
- Discoverer 3 web capabilities present the Relational Data Warehouse data in a multidimensional format.
- Express Web Agent provides native access to the Multidimensional Data Warehouse.
- Express Web Publisher makes available the user-friendliness and power of OEO and OEA to the web users.

Conclusion

Being a relatively new IT discipline, we can expect Data Warehouse to continue experiencing an explosion of ideas and advances from both a technological and an academic perspective. In particular, an emerging trend to marry Data Warehouse technology with Web technology, itself a fast exploding field, is an encouraging sign. Through sharing the experience of the UNIPHI Data Warehouse Project, the authors hope that other innovative ideas can be generated, thereby further contributing to the dynamics of this enormously challenging climate.

References

Corey, Michael J, and Michael Abbey. *Oracle Data Warehousing.* Osborne McGraw-Hill 1997.

Inmon, W H. *Building The Data Warehouse.* Second Edition. John Wiley & Sons 1996.

Kimball, Ralph, and W H Inmon. *The Data Warehouse Toolkit.* John Wiley & Sons 1996.