INTEGRATED UTILITY DATA WAREHOUSING - A PREREQUISITE TO KEEP UP WITH COMPETITION ON ELECTRICITY MARKETS

M Werner, U Hermansson

ABB Utilities, Sweden

ABSTRACT

The Power Industry is changing rapidly. Utilities on de-regulated markets are forced to focus on increased customer service on a very competitive market. In order to become successful the Utilities have to utilize all their resources. One such resource is the huge amount of information in different computer systems. This information can be refined and combined to show patterns and support conclusions that could be used to provide better service to the customers, gain market shares and increase profit.

The key is to store this information in a data warehouse that is flexible enough to provide powerful data-mining capabilities, and powerful enough to handle the big information pressure caused by the constant inflow of measured data from the power network.

To use data warehousing in the power industry needs special considerations. To be used as a reliable historian for process data, there are high requirements on performance, availability and redundancy. At the same time it must have all the flexibility and openness required to let the user perform e.g. ad hoc queries and data mining. This paper will discuss some of these considerations, and possible realizations.

HISTORY OF HISTORICAL DATA STORAGE IN THE UTILITY INDUSTRY

All industrial companies have a need to record information from their business processes. For power utilities the situation is the same, there has always been a need to report the performance of the power network operation, both for internal improvement and for reporting to authorities.

Therefore there is no surprise that most SCADA systems have functions for storing historical data, and for creating reports on e.g. production and power exchange. These systems do however have some limitations.

1. These systems are not designed for combination with analysis tools especially not general "off-the-shelf" tools.
2. It is more or less pre-defined at installation time which reports that should exist, and creating new reports is often not done in a flexible way.
3. Use of these reports loads the actual on-line SCADA system.
4. The data storage is strictly limited to the data that is really needed; no "nice-to-have" data exists.
5. Storing data from external systems requires costly integration work

The market developments along with the IT development have however both put new demands and possibilities to what can be done.

MARKET DEMANDS AND IT POSSIBILITIES

The deregulation process that is going on in a lot of countries is changing the business environment for the power utilities, and is also generating structural changes in the business. Mergers and acquisitions, as well as splits and diversification are changing established electric power companies, and a growing number of new actors are entering the business.

The deregulation and the increasing competition between power utilities are causing a higher focus on profit and competitive edge, for all types of power utilities. The split of traditional utility companies into separate power producers and distributors is also causing changes in processes and operational focus.

In parallel, the IT-development is invading the control rooms and offices of the power utilities. Common desktop applications, such as Microsoft Excel, are a commodity that makes almost every employee a potential trend curve analyst or application programmer. The constantly decreasing price/performance on computational speed, memory and bandwidth, together with new software products and de facto software standards, are not only changing the IT-business in general, but is also changing the different utility computer systems, such as SCADA-systems, Geographical Information Systems, Customer Information Systems and so on.

All these trends are all pointing in one direction. The utilities need competitive edge in order to achieve

profitability. It requires support for making wise decisions. The necessary information exists within the company, but is too fragmented and complex for a human mind to make efficient conclusions upon, and is too inaccessible and time consuming to gather.

The solution is decision support applications, based on a powerful Utility Data Warehouse, by which electric power companies can utilize the data they already own in ways that support decision-making. Decisions, that can achieve effective cost reductions, eliminate redundant procedures and improve marketing and profit margin.

The benefits of data warehousing are, according to market research by the IT market analysis firm Aberdeen Group:

1. increased profits
2. improved knowledge worker productivity
3. sounder decision making
4. harnessing of the unpredictable
5. subject-oriented information
6. distributed decision making
7. sparing operational systems the performance degradation of ad-hoc queries
8. cleaning up legacy systems, while moving the corporate systems architecture forward

Besides the powerful decision support for management on increasing the overall company economics, a data warehouse solution allows for new creative usage within a company. It is, for instance, possible for a user sitting at his desktop PC somewhere in the office computer network, to make his or her own reports and analyses using e.g. Microsoft Excel, without loading the real-time operational systems.

DATA WAREHOUSING FOR THE UTILITY INDUSTRY

Although most SCADA systems with self-esteem have had some sort of handling of historical (or time-tagged) data for a long time, these solutions were often based on closed-world proprietary databases or file structures. Today there is a need for open systems with standard access methods. The most well-known and accepted database query language is SQL, which consequently is the preferred interface to access data in an open database.

This is driving a lot of vendors to create open SQL-interfaces towards their historical data, e.g. by using the Microsoft ODBC interface specification. By doing that, it is possible to access historical data using e.g. the Microsoft Office suite of applications. One must of course bear in mind that ODBC in practice is not equivalent to having full SQL support, since the SQL usage through an ODBC interface in reality might be rather limited. If the database behind the ODBC interface is not a relational database, more complex SQL statements (for example the ones requiring table joins) might cause the SQL-query to completely stall the database for a long time.

However, providing openness to the historical data does not in itself make a Data Warehouse. The system must also be able to store information from other systems without excessive integration efforts, and most important of all: store the information in a way that is suitable for data mining. This means in practice an open relational database. Using a commercially available database also opens up for a wealth of third party database manipulation and data mining tools, which already exist for that particular database.

For most analyses, data is not interesting on a detailed level. A popular design of a Data Warehouse system is therefore to support On-Line Analytic Processing (OLAP). The basis for an OLAP system is data that is aggregated in all dimensions, so that summary data is always available on all levels (i.e. an OLAP "cube"). By doing this it is possible to easily access data on a high abstraction level, and to "drill down" in detail wherever data is interesting. Many vendors are not only providing summation, but also maximum, minimum, average, standard deviation etc.

Such calculations on time series of data is of course a basic need when analyzing historical data. In the past such calculations have often been made in the HMI client program (on demand), to limit storage cost and processing load in the historian. But today – due to decreased disk prices and increased processing power - such calculations can be made continuously in the historian and the results be stored back into the database. This is a huge benefit, even for such basic control room functions as being able to easily show a trend curve of e.g. daily maximum values or hourly average values.

The Utility Data Warehouse

The Utility Data Warehouse represents the merger between the traditional administrative data warehouse, and the classical storage of historical data from an on-line SCADA system. The real power of a Utility Data Warehouse comes when it is used for storing actual data from the power process, along with administrative data such as customer information and asset management information.

The Utility Data Warehouse can thus act as both an on-line historian in the control room, and as a general data

warehouse for decision support. Using a general Utility Data Warehouse as an on-line historian puts however special demands on it, compared to more administratively oriented data warehouses:

- It must have very good performance so that it is able to store a lot of small pieces of information very often, e.g. thousands of measured values every ten seconds around the clock.
- It must be as operationally reliable as an on-line SCADA-system, which means high performance and reliability both for the user (e.g. client response times for normal trend curve analyses) and for data sampling (e.g. to allow a certain down-time of the data warehouse without loosing any data, and if required also provide computer server redundancy solutions).

Just installing an off-the-shelf relational database does of course not make a Utility Data Warehouse. The implementation must solve the following requirements:

- A model of the data to be stored, including a data engineering environment. The model should be temporal, which means that all historical relations and calculations are retained, so that the historically correct relations are used when studying old data. Besides modeling the objects which values are to be stored, some of their physical (topological) relations should modeled as well, in order to provide the ability to query on such relations (e.g. all breakers in a certain power station).
- Actual data collection implementations, reading data from the appropriate source systems and storing them in the data warehouse, with a required performance. These are typically adapted to each source system, or adhering to an industry standard (such as OPC or the emerging HDAIS standard from Object Management Group, OMG).
- A data archiving model, to allow current data to be retained on-line and older data to be automatically archived, and seamlessly reloaded into the on-line data warehouse upon demand.
- Automatic calculations to be able to continuously store summations, maximums, minimums, average values etc. as new data items are inserted into the database.
- Openness with standard interfaces to other systems. Besides direct SQL access, ODBC/JDBC interfaces and Microsoft COM interfaces are popular. An application-programming interface is also beneficial.
- Short enough call up response times to be used in control room operation.

Other advantageous features that are often required by utilities are:

- Management of the measurement quality of each value, with quality code being propagated through all types of calculations and summations.
- Possibility for manual data entry (marked with a special quality code), that also causes subsequent re-calculations of the dependent calculations.
- Different data compression techniques (as discussed in more detail below)
- Possibility to create more or less complex calculations containing different instances, types of objects and points of time. Such calculations should also handle quality code propagation in an appropriate way.
- Integration with user interfaces of other utility information systems
- Automated database maintenance, eliminate the need for database administration specialists in the utility organization. This should also include automated on-line backup of the database, during which the system shall continue its normal operation.
- Storage support of planned values, i.e. to support entries of future data, such as forecasts and schedules, for planning purposes.
- Support for Daylight Savings Time

Design Considerations

One fundamental aspect of a Utility Data Warehouse is performance, and related to this is the aspect of how much data that is stored in the data warehouse.

Even when considering only data from the power process, it is easy to see that the database can become very large. Supposing a utility wants to store 10,000 measurements for three days, that alone would generate 260 million values in the database.

This points to two important design considerations: data compression, and performance in large databases.

Data Compression. The size of the database can be reduced significantly if values are only stored when they change, instead of cyclically at each sampling interval.

For digital values, such as indications, storage upon change is obviously recommended. For analog measurements storing upon change is in reality only interesting if there is some sort of dead band, either in the RTU:s or in the data warehouse, that eliminates very small changes from triggering a storage (since there are always small fluctuations due to measurement noise). How much data that can be compressed by storing on change depends on the behavior of the particular power network and how big dead bands that can be accepted in the data warehouse.

Compressing data using storage on change may be called "time-domain compression". Such compression makes general data mining a bit more complicated, since the value for a given point of time must be "decompressed" upon retrieval (in this case, retrieved from a previous point of time). This is however quite straightforward, and can typically be hidden behind interfaces.

Data can be compressed even more, by also compressing data in the value-domain, e.g. by approximating the incoming measured values by a mathematical function. Supposing the change in measurements is approximately partially linear, then the measurements can be represented in the database by piecewise polynomial approximation, i.e. for a given time interval the measured values may be stored as one data point and the slope of the line. Even more advanced mathematical approximations can be made, to reduce the amount of information that needs to be stored in the database.

However, all these compression techniques have one disadvantage. As mentioned before, a data warehouse should be used to analyze data in a free and "ad-hoc" way. This means supporting ad-hoc queries, independent analysis tools and easy creation of various reports. If data is stored cyclically, such queries and reports are easily made, since there is always data stored for the expected point of time. However, if data is compressed the queries become rather complicated – and in some cases virtually impossible to use in practice. This can of course to some extent be simplified by creating database views that enables the user to define queries with a more simple syntax, or by creating interfaces that cover the complexity. But covering such complexity behind interfaces can sometimes give so bad database performance that they are impossible to use in reality.

For time-domain compression (storage on change) the extra complexity is quite easily covered by an interface. Therefore might such compression be a practical solution since it will reduce the storage requirements. However, if data is compressed in the value-domain, time oriented queries get even more complex, and value-oriented queries become almost impossible to realize. For instance, a query (e.g. through an ODBC interface) asking for all points of time where a value was above a certain threshold, will be practically impossible to resolve if the database is compressed in the value-domain, since the whole database then have to be decompressed in order to evaluate the query.

Therefore, since storage prices per Gbyte are decreasing continually, cyclical storage and large databases might be preferred in many cases where ad-hoc querying is an important functionality. Time-domain compression might be preferred in cases where data access through interfaces is dominating. Value-domain compression should be used only when the end-user requirements on query capability are very limited, or in parts of data where free ad-hoc querying is not supported.

Performance In Large Databases. As databases are getting large, there is a problem with performance. The performance problem is however different in a Utility Data Warehouse than in a more purely administrative data warehouse. The reason is that in a Utility Data Warehouse there is a very large amount of small objects that shall be individually accessed.

The first problem is that the database tables can get very large, maybe several hundreds of millions of rows. To get good response times regardless on how large the tables are, the tables should be partitioned into equally-sized partitions of, say, half a million rows each. This way the search times are kept low and virtually independent of the database size.

The second problem is I/O performance. Disk accesses are much slower than accessing data in primary memory, and should be minimized. In a Utility Data Warehouse data is typically inserted in a time-oriented way. As time goes by, all relevant data for the current point of time is inserted into the database. However, data is typically read in an orthogonal manner: a few identities are chosen for which values of many points of time are queried.

To increase the I/O, buffering is used on different levels. This has however effect only if consecutive read operations are made in approximately the same order as data was written, which it is not in this case. Other techniques involve writing large chunks of data on several disks to gain parallelism ("striping"). This is however not effective in this case either, since each piece of information is small.

Instead, the data lookup behavior of the relational database system must be optimized. One way is to keep more of the interesting data in the search indexes, and thus preventing the relational database from having to access the data elements at all for certain frequent queries. This in turn requires a good knowledge on how the database system works, so that it is always choosing the correct index when evaluating such queries.
In summary, providing good performance in large databases with data from the power process requires a lot of optimization and configuration by the data warehouse vendor.

NEW APPLICATIONS BASED ON DATA WAREHOUSING SOLUTIONS

The applications of a Utility Data Warehouse are in many ways limited only by the imagination of its users. Many tools also helps in "harnessing the unpredictable", to find patterns that were not obvious before they were found by a data-mining tool. Previously, there were special solutions in the different systems that could solve some of these types of tasks. These systems had however limited analysis tools and to combine information from external systems was expensive if at all possible. The following chapters will sketch some examples on applications that might be interesting for a power utility operating on a deregulated market.

Examples For A Transmission Company

The economy of a Transmission Company is directly dependent on how much energy it is able transfer in the network. An evaluation of maximum power transfer capability in the interconnected power system is made, and contracts are entered with production and distribution companies.

Suppose process values are stored in the Utility Data Warehouse, e.g. from a State Estimator running in a SCADA/EMS system, it is possible to have the network state history stored for a very long time. This makes it possible to evaluate the supposed transmission capacity against the actual power transmitted through the network, as a long-term trend, and to find patterns such as systematic deviations or periodic deviations over the years.

This makes it possible to optimize power transmission contracts and to adjust the transfer capability estimation algorithms.

The Utility Data Warehouse is also a very powerful tool to store and analyze disturbance records recorded by protection relays or disturbance recorders.

Examples For A Production Company

A production company has contracts to deliver electricity to its customers. The contracted production level must be kept, to avoid penalties to the system operator. On the other hand, if it is possible to generate more energy than what is contracted, it might be possible to sell it on the day ahead or ancillary services markets.

Suppose the production company stores the actual generated power from each power plant, along with other parameters such as, for instance, market price variations, load in the transmission network, contracts and weather variables. Then it might be possible to evaluate the contracts made, in relation to the actual earnings, and to see patterns in how this varies over the years. For example, to study how well the company has succeeded with the strategies they have had. Have they missed the peaks on the market price?

For hydro power plants an important application might be tracking and reporting of reservoir levels, as often required by authorities.

Examples For A Distribution Company

The distribution companies have a situation for which decision support using a data warehouse could provide many opportunities.

Areas that might be subject to improvement are for instance load studies, quality assessments and customer management.

Load Studies. The long-term load development in the distribution network is valuable to store in a data warehouse. Using this information, paired with weather and economical information, it is possible to make long-term load forecasts. If this information is compared with information from the asset management system, decisions on optimal maintenance, extension and strengthening of the distribution network can be made. Also, it makes it possible to find and learn from similar operating situations in the past.

Quality Assessment. In a deregulated market it is also needed to be able to report statistics to the authorities, and to customers with special service agreements, about how the quality has been in the network in terms of e.g. interruptions and voltage levels. Such information can also be used for solving disputes. If the appropriate local devices are connected to the Utility Data Warehouse, it is possible to store overtone measurements and disturbance recordings from the relay protection units, and by this increase the knowledge of how the network is run.

Customer Loyalty. Another case for data warehousing is to track customer loyalty. Suppose a distribution company is loosing many customers to competitors. Then the data warehouse can be used to find patterns that points to common characteristics of the customers that have left, such as where they live, what kind of service agreement they have etc. By using the information from the data warehouse, the company can develop a profile of the type of customer likely to leave the service of the company.

EXPERIENCES FROM INSTALLATIONS THROUGHOUT THE WORLD

Although the usage of data warehouses within the power industry is still in its initial stage, it is interesting to see how different power utilities are reasoning today around the usage of such tools in their business.

One transmission operator has chosen to store all measured and calculated values together with status, events and operator actions. Further, maintenance schedules and work orders are stored. The information is stored on-line for one year after which it is archived.
The system operators especially appreciate the tools that make it easy to define the database set up and to add new values to be logged using default patterns. Data used in on-line operation must be properly indexed in order to meet response time requirements. Ad-hoc SQL-queries to support different types of analyses may of course have a longer response time.

The information is used in a number of ways:

- Information may be retrieved either in the ordinary operational displays like one-line diagrams and event lists. Another way is to extract data by SQL-statements written in Microsoft Query and bring the information into a Microsoft Excel spreadsheet for further analyses like statistical calculations and reports.
- As all information is stored, later analyses can be made to improve control system set up or for operator training.
- Disputes about what has taken place are easier to resolve and preventive action may be taken.
- Real transmission costs are analyzed in order to improve the tariff setting for various energy transfers
- Stored data is used to estimate compensation to producers due to lost production or resolve disputes.
- Data could even be used as backup for the metering and settlement system if data is lost or questioned.
- Equipment statistics is gathered and used as input for maintenance planning.
- Data from executed work orders may be analyzed to find out if the maintenance could be performed more efficiently.
- Follow up of service quality by calculation of key performance indices to measure against set goals or benchmarking with other transmission operators.
- The company is convinced that they will realize further possibilities as they gain more experience

Discussions with other users show that they realize the potential in data warehousing but the development of applications combining data from different sources has not advanced so far. Typically they store data such as reservoir levels, water flows and weather data, and production plans are exchanged with an off-line system. Further, all switch positions, voltages and powers are stored. Reporting to authorities on water flows and reservoir levels is a very important application.

A recent study by Newton Evans Research Co shows that 24% of the utilities claimed that they have some form of data warehousing technology, which for many, was represented by the Data Base Management System in use at the utility. The most important requirements were speed of access for data warehousing and availability of archiving products.
Importantly, 56% of the American investor-owned subgroup indicated having already implemented some form of data warehousing technology, compared with one quarter of other U.S. utilities, but only 13% of the international community of respondents. However half of the respondents had plans to bring this technology to bear.

CONCLUSION

Changing requirements due to privatization and deregulation have created needs for analyzing information from different sources within the utility industry. These needs require new high performance solutions represented by the new Utility Data Warehouse and its characteristics outlined in the paper. Utilities have started to take advantage of this new technique and many other plan to follow. As the industry gains experience from this new tool new applications will evolve on the market.

BIOGRAPHIES

Ulf Hermansson is responsible for Product Management, Network Management Systems at ABB Utilities, Västerås, Sweden. He received an MSEE from Chalmers University of Technology in Gothenburg, Sweden. He has more than 25 years of experience from system development, sales, projects and management within the field of network management systems.

Mats Werner is responsible for Data Warehousing, Network Management Systems at ABB Utilities, Västerås, Sweden. He received an MSAP&EE from Linköping University of Technology, Sweden. He has several years of experience from managing development within the SCADA/EMS area. He has also a background from Network Management in the telecommunications industry.