Taking the Integrated Data Warehouse Global:

Part 1 – The IDW Architecture

3.16 EB9305 ANALYTICS

TERADATA.

What happens when the CEO says he wants a global view of his business all in one place, complete with drill down and analytical capabilities? A typical data warehouse addresses basic concerns; however, issues like currency conversions, date and time synchronization, privacy laws, customer data crossing international borders, workload management, and security become exponentially harder for a domestic data warehouse to address.

This is the first in a series of three whitepapers that will take you through the journey of how an Integrated Data Warehouse (IDW) addressed one CEO's many concerns for his global business. Having the capacity to handle only the data from their own source systems and users in the United States, we grew their IDW to effectively operate globally, for users across the enterprise. In this paper, I will discuss the architecture that was implemented, as well as the several other architectures that were available for us to deploy. Ultimately, it is important to analyze all possibilities and choose the one that is best for your situation and company needs. I will address the IDW Global Standards and Governance in the two subsequent white papers.

The Architecture

In this case study, the Teradata customer had a new Teradata IDW in the U.S., but the customer also had data warehouses (of other vendors) in five other countries around the globe. Each country ran their own profit and loss, and each was in the same business (television and online retail sales). The source systems that fed each data warehouse were basically the same, but each had its own unique data sets. Additionally, there were no reference data standards among the six data warehouses regarding SKUs, product codes, customer profiles, units of measure, or vendors.

To address the disorganization, we started by creating a data architecture in the key data groups and building data profiles to determine where there were differences, gaps, and similarities. This profiling proved to be very beneficial in designing the architecture, because our team recognized where data needed to be brought together for transformation and standardization. Also, we knew where we had differences and how we could design around them. Next, we turned our attention to the typical globalization issues that global enterprises face:

- Currency conversions
- Date and time synchronization
- Privacy laws
- Customer data crossing international borders
- Workload management
- Security

The architecture we built provided a clean transformation that could be reversed and audited.

Data Profiling

One of the biggest mistakes that data warehouse developers make is assuming the data coming from the sources is clean and standardized. This is very rarely the case, especially when dealing with sources from multiple countries that have all been developed separately. Many times the developers will try to code around the data problems, and this leads to nothing but a lot of extra code and follow-up maintenance in the following years.

Data profiling identifies these data problems early in the development process. The earlier they can be identified, the faster they can be eliminated—again saving time and costs later on.

The profiling work is analysis of the data from both a business and technical level. It starts with a set of basic statistics on the populated data (taken on a full production set of the source data) and utilizes that information in conjunction with source and core IDW models (logical and physical), data dictionaries and definitions of the data attributes and values, data flows (from operational systems), data lineage, and redundancy. This is done across systems and business requirements for data. It involves:

- Evaluation and identification of all relationships (structural and semantic)
- Detection of gaps, defaults, and missing metadata (definitions, documentation, etc.)
- Cross-country reconciliation and integration (semantic and structural)
- Review of discovery work sessions with appropriate subject matter experts, as needed (local and IDW global)

Data profiling is used to determine suitability of the source data, as well as to identify data management requirements for the physical design. It can also be used to set expectations about the quality of the data. In the case of multi-country integration into a pre-established target, this will provide the information to prioritize semantic integration and set expectations for the business on the level of integration that is both feasible and sustainable.

Profiling does require a considerable amount of time in the initial development process, but the time and cost savings found later in the process far exceeds the cost of profiling.

Data profiling also provides better requirements to the data modelers resulting in a cleaner data model for the integrated and semantic layers.

Logical and Physical Architecture

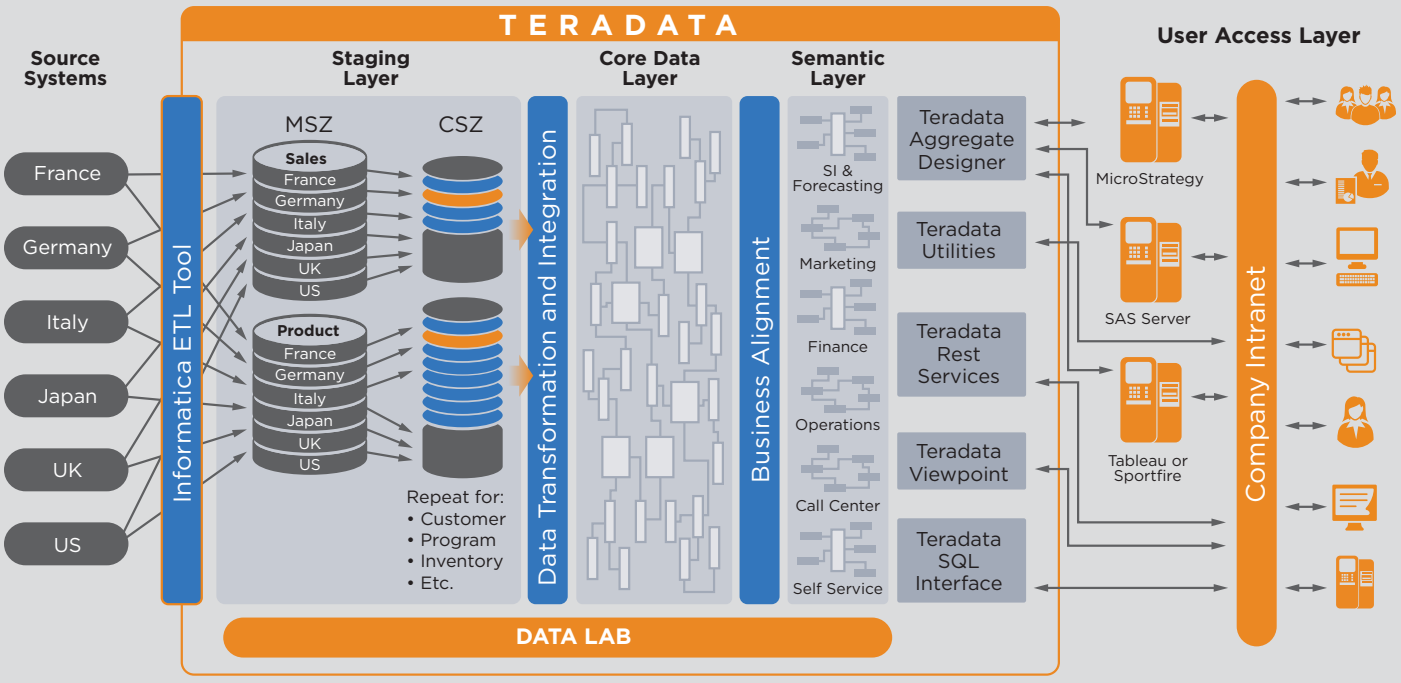
The Global IDW architecture has many parts working in the process of gathering, organizing, and analyzing data from virtually every part of the company. To do this, it takes a well-defined architecture that provides an "industrial-strength" environment, while at the same time is flexible enough to handle new business requirements as the business evolves.

A well-defined architecture provides an "industrial-strength" environment, while at the same time is flexible enough to handle new business requirements as the business evolves.

Central to the IDW architectural approach are reusable components, which include both best practices and intellectual property, reducing the time required to achieve value from an organization's analytic ecosystem and improving end results. Ideally, an analytic architecture should:

- Be driven by business requirements rather than being system-focused
- Use appropriate technologies which best meet the business requirements
- Support a strategic vision and business goals which drive clear direction for the analytic ecosystem

Depiction of the Global IDW Architecture



TERADATA

- Link operational, informational, and analytical data needs to drive actionable insights
- Address customer needs for timely and accurate business information
- Recognize and leverage current capabilities and cultural alignment
- Develop a business-centric (or analytic) roadmap that plans for data reuse and usage across the organization

Overall, setting up a good logical and physical data architecture becomes the cornerstone of a solid foundation for the IDW. Remember, the whole purpose of the IDW is to provide an analytical system that the general user can access and provide good data for critical business decisions.

Explanation of the IDW Architecture Layers

As someone who has been building data warehouses for 30 years and business intelligence platforms for 40 years, one thing I have found to be consistent in a successful IDW is a sound architecture used from the beginning. Often, companies will start throwing data into a database and a year or two into the life of the data warehouse, things start falling apart. Jobs start running longer than needed; computer code to build queries become very complex; extra space is needed to handle intermediate data builds. The list of issues is seemingly never ending. The architecture that I have found to be the most successful is the one that we implemented for this customer. It is comprised of seven main layers, sometimes with separate stages within the layer, which we call *zones*:

Source Layer

The source layer includes data from transactional systems where data is created on a routine basis. A source system is considered to be the data or record.

Stage Layer

The stage layer is the place in the IDW where data from the source is landed with little or no transformation. In a Global IDW, this layer will have two zones: a Market Staging Zone (MSZ) and Common Staging Zone (CSZ).

- **Market Staging Zone (MSZ)** The MSZ is used for getting the data from various countries ready for processing into the Common Staging Zone (CSZ). No major transformations are done in this zone; it is virtually a copy of what the data looks like in the source system.

- **Common Staging Zone (CSZ)** This layer is used to bring data from all countries together before putting the data through the transformation process and into the Core Data Layer (CDL). Moving data from the MSZ into the CSZ has some minor transformations dealing with currency conversions, numbers, and date formatting. This stage may be skipped if the data coming from the MSZ has no transformation or reformatting required. Tables in this circumstance may be loaded directly into the Core Data Layer.

Transformation Layer

The transformation layer is the process of moving and organizing data from the structure of the source to the structure in the Core Data Layer (CDL). This layer is mostly computer code and is performed using Informatica (unless special needs exist where Informatica cannot handle them).

Core Data Layer (CDL)

The core data layer houses all the integrated data and historical data. It is the main workhorse of the IDW environment and must be maintained with governance and methodology. This layer is modeled from a collection of industry models supplied by Teradata. It is as close to third normal (relational) form as possible. It is also the first layer that can become cluttered and unmanageable if not governed and maintained. No users should have direct access to this layer.

Business Alignment Layer

The business alignment layer is where data is taken out of the CDL (sometimes virtually, sometimes physically) and put into a view or semantic structure that the business users can easily understand and use for their reports and analytics. This layer is mostly computer code.

Semantic Layer

The semantic layer is where data structures reside for the users. It is how the business understands the data and can access it with their analytical tools. No data should be extracted from this layer except what the business intelligence group (BI) may need for their tools in processing.

Access Layer

The access layer is what the end user will know best. In this architecture, the main BI tool is MicroStrategy, but Tableau and SAS are other tools that can be used at this layer instead. SAS processing can be done within the IDW, so no data should be moved into a separate SAS environment.

> The whole purpose of the IDW is to provide an analytical system that the general user can access for critical business decisions.

Data Lab

The data lab is the extension of what many customers term as an *analytic sandbox*. It enables agility in business analytics through rapid self-service loading and analysis of new data combined with existing production warehouse data.

Agile analytics using an integrated data lab provides both the DBA and the user community a solution to enable agility for new data. It helps quickly assimilate untested data into a separate "non-production" portion of the data warehouse and provides a self-provisioning, self-service environment for swift prototyping and analysis on new, external, unclean data, which is temporary in nature.

Businesses need agile analytics to quickly test theories and new ideas to drive innovation and react to competitive pressures. They require flexibility in terms of speed and agility to explore new, unrefined data and experiment on new theories without long planning periods. They also must determine that the analysis was a success, and then swiftly shift it into a production environment—or, they need to fail fast and move on.

Lessons Learned

After any project, it is critical to analyze the work so that successes can be replicated and missteps can be avoided in the future. In this deployment, we took away three key lessons:

Data Profiling was the Greatest Challenge

Data profiling will always be a difficult task—not because of the tedious work it requires, but because of the international laws on data crossing borders. The profilers in the United States can only see data that was not deemed 'private,' but this policy varies from country to country—and, to make matters more confusing was the retraction of the Safe Harbor Act by a European court in the midst of our profiling work! Files had to literally be erased overnight because of the retraction, which left each individual European country to enforce their own laws.

Businesses need agile analytics to quickly test theories and new ideas to drive innovation and react to competitive pressures.

Some countries had relatively new laws in place, while others have virtually no laws in place. Thus, we had to become familiar with each country on a one-on-one basis. For example, Japan has the toughest laws on customer privacy. In fact, customer data cannot physically be moved out of Japan.

While in some cases encryption can be used, it is not always the case. A new method known as *tokenization* seems to be catching on, which would be a relief in some ways, yet it requires specialized tools and an added cost to the IDW development.

The Creation of the Market Staging Zone

The Market Staging Zone was created to bring in data from each country, so the currency conversion and data formatting could be done within the Teradata environment. By doing this, data meets international standards and common formatting before it goes in the Common Staging Zone (CSZ) and, finally, into the Core Data Layer.

Though this required extra development time, it proved to be a valuable step taken and provided a cleaner audit trail.

Standardization of the ETL Tool Set

By working to standardize the ETL tool set ahead of development, the development process was able to flow with little to no incidents. Using this with the Teradata Mapping Manager (TMM) tool made the development process faster and easier to change in testing.

TMM is a tool used by the Teradata Professional Services which provides a great place to maintain all the mappings from one layer to the next. Though this takes some discipline, the added benefits in the long run far outweigh the challenge in development. It also provides great input into a metadata tool for use in answering questions about data lineage.

Bibliography

Data Profiling for the IDW—Barb Johnson, Business Consultant, Teradata

Global Solutions Architecture for the Integrated Data Warehouse—Bob Bender, Principal Industry Consultant, Teradata

About the Author

As a Principal Industry Consultant and Global IDW Architect for Teradata's Business Consulting Group, **Robert (Bob) Bender** has leveraged his 30 years of data warehouse experience to lead teams through data mart consolidations, data warehouse road maps, business value assessments, and data warehouse development projects. He has also directed business executives through the process of building successful analytical platforms and has expertise in developing business intelligence platforms.

10000 Innovation Drive, Dayton, OH 45342 **Teradata.com**

3.16 EB9305