Paper 163-28

"How Do I Love Thee? Let Me Count The Ways." SAS® Software as a Part of the Corporate Information Factory

John E. Bentley, Wachovia Bank, Charlotte, North Carolina

Abstract

Many organizations either have built or are trying to build a Corporate Information Factory (CIF), and SAS® Software is often an integral part of it. A limitation of the CIF framework, however, is that it suggests hardware and database architectures and shows the flow of data and information through the system but it doesn't provide guidance about the software needed to run the factory. As a result, it's common for a CIF to take a "best of breed" approach and use different software for each process and task: ETCL, data QA, metadata management, ad hoc querying, production analysis, and reporting, OLAP, data mining, *ad nausium*. This approach complicates the CIF system, increases costs, and reduces ROI. In this best-of-breed solution, SAS Software is often deployed only in an analytical role that clearly fails to take advantage of the power of software. If SAS is leveraged to the full extent of its capabilities, then the CIF can dramatically reduce its software complexity, thereby reducing costs and increasing flexibility and ROI. After reviewing the CIF concept, this paper will discuss how specific SAS products should be integral parts of the Corporate Information Factory. The paper is generic with regard to operating systems and appropriate for all skill levels.

Disclaimer: The views and opinions expressed here are those of the author and not his employer. Wachovia Bank does not endorse or necessarily use any of the models, frameworks, approaches, processes, architectures, hardware, or software discussed here.

Introduction

As a concept, the Corporate Information Factory (CIF) has been around since the early 1980s when W. H. Inmon used it to describe the information ecosystem. The CIF is well suited for this because its generic structure can be used to identify the ecosystem in totally different corporations. At the same time, it's flexible enough to handle the impacts of the numerous forces that affect each corporation—business and economic, culture and politics, technology—and therefore affect it.

Just as important as its descriptive usefulness, the CIF framework provides broad direction by telling us what we should be doing and what our goal should be in building our information systems. Organizations have increasingly complex business systems that include data warehouses, data marts, operational data stores (ODS) and multiple decision support systems (DSS) used by hundreds or thousands of information consumers and decision makers. Even smaller organizations' information systems are increasingly complex. For all organizations, then, the CIF concept provides a strategic view or picture of the integrated IT system that is needed.

In a nutshell, the Corporate Information Factory provides a proven way to organize our corporate information systems for maximum effectiveness. Without this strategic framework, investments in IT will produce a balkanized set of information systems that fail in terms of

- Systems integration
- Ability to change
- Infrastructure and support costs
- Performance and efficiency
- End-user satisfaction
- Contributions to organizational success

Where does SAS Software fit into the CIF framework? Almost everywhere! SAS provides a platform-neutral, integrated suite of data management, analysis, and presentation software to generate and deliver intelligence about customers, suppliers, IT systems, and the organization itself. In the sections that follow, we look at the individual parts of the CIF and then describe which SAS products and solutions can be used in each particular location.

An Overview of Corporate Information Factory

To gain competitive advantage in recent years, organizations have been implementing new technological capabilities that promise to deliver best-of-breed data management, business intelligence, and information delivery solutions. The result is that most organizations have built or implemented many of the following technologies:

- Data warehouse
- Data repository
- Operational data store
- Data marts
- Exploratory data processing
- Data mining
- Online Analytical Processing
- Multidimensional Online Analytical Processing
- Internet and intranet capability
- Multidimensional and relational databases
- Star schema and snowflake relational database designs
- Data extract, transform, clean, load processes
- High-performance computing platforms (SMP and MPP)
- Data warehouse administration
- Metadata management

Each of these technologies has great promise, but they're all "point systems" designed to address a specific problem or need. Implemented without a guiding strategic vision and plan, they will not completely deliver on their promise but instead the combined result will be a confusing, intimidating, and wasteful hodgepodge of systems that don't cooperate. In fact, in many cases the systems will compete with one another. Admittedly, each system may adequately perform the task for which it was designed, but without an overarching framework to guide implementation few synergies will be gained from the systems as a whole. At best, there will be a dysfunctional information ecosystem. At worst, there will be no ecosystem.

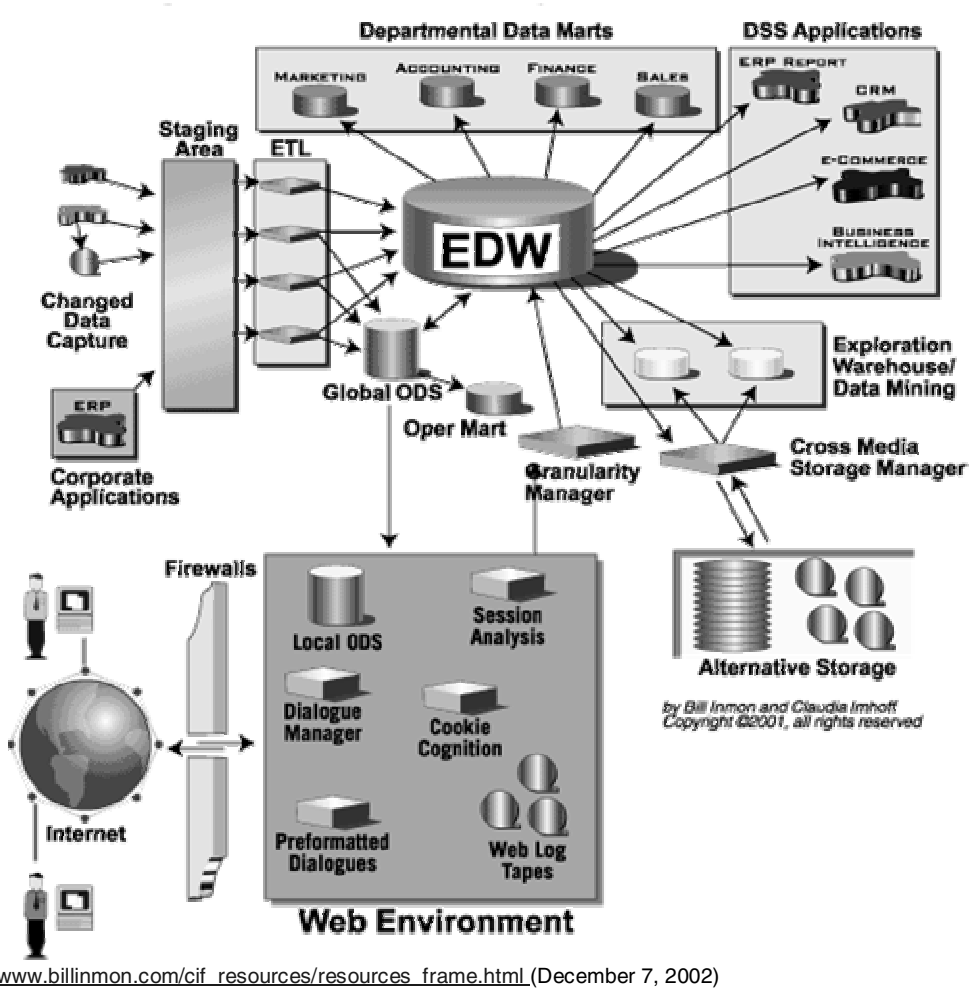
An information ecosystem is a system of different components cooperating to produce a cohesive, balanced information environment. (Inmon, 2001.) The ecosystem is adaptable, changing as the needs of the "inhabitants" change. As in nature, when needs change the ecosystem adapts, changes, and rebalances itself. For example, a data warehouse that feeds a series of data marts delivering business intelligence is an environment common in marketing IT. When the need emerges to manage customer interactions, an operational data store (ODS) is added to the system to provide near real-time access to current customer data. Data may no longer be loaded directly into the data warehouse but may instead flow into the warehouse through the ODS.

The Corporate Information Factory is the physical depiction of the corporate information ecosystem. In Corporate Information Factory, Bill Inmon identifies the following as components of the CIF:

- The external world
- Applications
- Integration and transformation layer
- Operational data store
- Data warehouse
- Data mart(s)
- Exploration and data mining warehouse
- Alternative storage
- Internet and intranet
- Metadata
- Decision support systems.

Although components such as decision support systems are specifically designed to provide data analysis and information delivery for decision-making, the flexibility of the CIF allows this to be accomplished from almost any point in the ecosystem. Figure 1 is Inmon's current published concept of the Corporate Information Factory, showing details of the Internet component.

Figure 1. The Corporate Information Factory



Source: http://www.billinmon.com/cif_resources/resources_frame.html (December 7, 2002)

Roles, workflows, methods, and tools are critical to the smooth functioning of the CIF, but they are perhaps the most difficult issues to address because of the corporate factors that impact decisions in these areas—culture, tradition, politics, personalities, budgets, and geography. These activities include

- Systems administration
- Configuration management
- Data quality assurance
- Request for information
- Data analysis
- Information delivery
- Customer communications

Because of the unique factors that impact decisions made in these areas, the CIF framework has the most difficulty in addressing them. For example, organizations that manage information systems at the line-of-business level may have difficulty giving up the local control needed to create a healthy responsive ecosystem. Alternatively, a highly centralized organization may resist creating the data marts needed by the lines of business and thereby unbalance the ecosystem. How does one build organizational pathologies into a model or methodology? The CIF doesn't really try to.

The CIF is like any other factory—raw materials go in, they are processed and changed, and a product goes out. The purpose of the CIF is to turn data into actionable information and then deliver it to those who need it, so the best way to understand the CIF is to follow the flow of data.

Data enters the CIF in a raw form, collected by functional applications and systems. The raw data is first refined by cleaning programs and then passed to a set of programs that transforms and integrates the functional data into corporate data. The data then passes into the operational data store and the data warehouse. From the data warehouse, the data can be processed further (or not) and passed to data marts and an exploratory data warehouse. From these sources the data is transformed, analyzed, and reported by decision support system (DSS) applications. The data in the warehouse itself is, of course, available for use directly by decision support applications. From the ODS, the data can pass to operational data marts for DSS usage.

Data in the CIF

Data is the raw material used by the CIF, and there are three types of it:

- External
- Reference
- Historical.

External data is that which originates outside the organization. An example is proprietary data provided by a credit-reporting agency such as Experian. External data usually requires modification or addition of a key structure to make it meaningful within the CIF.

Reference data is generated internally by the organization, such as that describing products and product lines. This data allows the organization to standardize on commonly used names and provides the basis for consistent interpretation of corporate data across departments. Unfortunately, reference data requires standardization and formalization and is therefore difficult to acquire and maintain. With reference data, we can be confident that three analysts all analyzing sales of a specific widget will all get the same answers. Without reference data, each analyst's report may be based on a different geographic roll-up of the organization.

In one sense, almost all data becomes historical data ten seconds after it is acquired. In reality, the time period after which data is classified as historical differs among CIF components. A data warehouse that loads monthly may consider data to be historical after 30 days but an ODS that loads daily will consider data more that 24 hours old to be historic. Alternative storage may contain only historic data at least twelve months old. Clearly, it's important to know how historic data is defined and where to look for it.

The External World

External data originates in the external world. This is where the activity—generically called "transactions"—that the organization is interested in takes place. Without transactions from the external world, the CIF would not exist. The external world generates the raw data transactions, the CIF processes them into information and delivers them to the organization's information consumers—managers, sales people, statisticians, vendors, suppliers, etc.

The interface (channel) to the external world is important to the CIF because this is how data gets in the door. The interface must be fast, efficient, inexpensive, simple, and scalable—rather like a loading dock. It can be high tech or low tech, as long as it meets these criteria. The CIF usually has numerous interfaces to the external world in the form of "Applications".

Applications

An application in the CIF is a data collection mechanism that uses a specific interface to the external world. An application's core function would likely not be described as "transaction data collection". In fact, its core function may instead described as automating a key business process, such as accepting payments, tracking accounts payable and receivable, or order processing. In doing this, the application may also audit and adjust the transaction data being collected.

A major challenge of applications is that they are generally not integrated due to a variety of reasons, which include (1) the application may have been built for a specific, narrow purpose or (2) it may have been acquired in a merger and never converted. Regardless of the reason, unintegrated applications cause problems by using and maintaining:

- Inconsistent key structures
- Inconsistent valid values
- Inconsistent data structures
- Inconsistent data definitions
- Inconsistent transformation and derivation code
- Inconsistent report generation
- Inconsistent programming languages

A specific application must be designated as the "source system of record" for specific data. Many applications will collect name and address, for example, but only one can be "the best" because it has the most complete, accurate, and current data conforming to corporate requirements. The source system of record probably won't provide perfect data, but will be the best available to the organization.

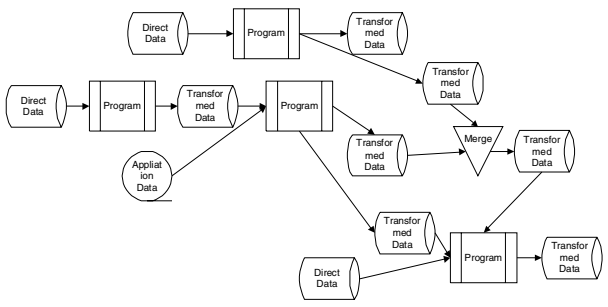
Integration and Transformation Layer

The integration and transformation layer (the I&T layer) is made up of programs that process the data provided by the applications. These programs capture, transform, and move data from the applications' environments to the ODS and data warehouse environments. This layer is also known as extract-transform-clean-load (ETCL) processing.

In the I&T layer, similar and related data from individual applications is combined, integrated, and transformed into corporate data. In this, a number of things can happen.

- The data simply can be passed from one platform or database to another, changing only the layout or not.
- Data can be aggregated or summarized.
- Transformations can standardize data elements using a set of common coding, derivation, formatting and substitution standards.
- Intermediate files for use as input to other programs can be produced.
- A single program can output multiple files.
- The data can be sorted for processing in a specific order or merging with other data.

Figure 2. Complex Integration and Transformation



Processing speed requirements often drive development of integration and transformation programs because feeds from the I&T layer go into both the operational data store and the data warehouse. A terabyte class data warehouse, for example, usually processes huge volumes of data in a narrow time window. Without an optimized I&T layer, the data quickly backs up and service level commitments are missed.

The integration and transformation programs themselves are often a combination of simple and complex. They constantly change over time, however, because the applications are constantly changing and the data requirements of the ODS and data warehouse are constantly evolving. Maintenance can be difficult because, as with the applications, it's not uncommon for organizations to use a number of different programming languages. Some programs could be Perl, some COBOL, or some in a vendor's proprietary language. As a result, significant staff resources may be needed for maintaining the I&T layer.

Operational Data Store

The Operational Data Store is a hybrid structure that performs both operational and decision support processing functions to provide tactical business information. It is not an essential part of the CIF, but where it's implemented it brings together data from related operational systems (e.g., consumer loan and home equity loan) to be used for integrated business management reporting.

The ODS may be structured like a data warehouse, but it contains only dynamic, currently valued detail data. The data in it is accurate only at the moment of presentation. Because the data is so volatile, summary data is not stored but instead is calculated on the fly.

Operational data stores are classified according to the speed with which the data moves from the application into the ODS. Usually, an ODS will consider multiple classes of data.

- Class I: Asynchronous—one to two second delay
- Class II: Store and forward—two to four hour delay
- Class III: Batch processing—overnight
- Class IV: Data from the warehouse

The primary feed into an ODS is directly from the I&T layer but data from the data warehouse may provide a secondary feed. An ODS, however, often feeds a data warehouse. Depending on the volume of data, load and update processing can be time consuming and require careful scheduling and efficient programming to avoid limiting user availability. This is also true for data marts.

User access to the ODS often entails queries returning only a few rows of data. Just as often, though, it may require DSS analytical processing and this requires a different database optimization than does other styles of processing. Optimizing the ODS for one purpose will necessarily be done at the expense of the others, so architecting an ODS is one of the most complicated tasks in building a CIF.

Data Warehouse

The standard description of a data warehouse is a data management structure that is subject-oriented, integrated, time-invariant, nonvolatile, and comprised of both detail and summary data. The data warehouse is the most prominent component of the Corporate Information Factory and is often the first place where data integration is achieved in the organization.

The data warehouse is the centerpiece of the CIF. In Figure 1, the data warehouse is drawn to scale. In terms of data contents and processing, it is much larger than any other component. Inmon (2001) suggests that as a general rule, a data warehouse is one to one and a half *orders of magnitude* larger than any other CIF architectural component.

The Integration and Transformation layer and the ODS both feed the data warehouse on a scheduled basis, and alternative storage is available to provide data on an as needed basis. The data warehouse in turn sends data to multiple recipients—data marts, an exploratory data warehouse, DSS applications, the ODS, and alternative storage. The integration and transformation layer moves massive amounts of data into the warehouse, and the load may be made by a combination of monthly, weekly, or daily processing. Records are never updated in the sense that existing data is lost, but instead an existing record is "expired" and a new current record is added.

In addition to storing data, the warehouse provides important data processing services.

- Data quality monitoring during the load process
- Post-load data aggregation and derivation, such as scoring and customer segmentation.
- Transformation and derivation processes prior to loading data marts
- Monitoring data usage
- Running background utilities needed for the day-to-day operation of the system
- Limited amounts of DSS query processing

Archiving old data is important for maintaining the data warehouse as a manageable entity. Budget-wise, there's no such thing as a database that scales forever, so at some point the data must be purged. A typical approach is to move the data to a siloed sequential storage medium. Archiving has its own set of complex issues, and many organizations do not spend much time thinking about them until forced to.

Data Mart(s)

The demands for information and analytical capabilities quickly exceed that which can be provided by the data warehouse. After all, the data warehouse is designed primarily for storing data, not using it. A data mart, on the other hand, is a subset of the data warehouse that is tailored to the information needs of a particular group of people or the needs of a particular analytical topic. That is, you can have a data mart for the marketing department and also a data mart used by anyone doing regulatory reporting. In contrast to a data warehouse, data in a data mart is denormalized, summarized, and optimized for query performance.

The data warehouse is the single-source feed to the data mart, and a data warehouse can and usually does support multiple data marts. There are several types of data mart.

1. A data mart can be a simple sample or subset of the warehouse data. The tables are identical to the source and no transformation processing is needed to prepare the data.
2. A Multidimensional OLAP (MOLAP) data mart contains data summarized by specific dimensions, most commonly product, time, and location, for online analytical processing. Multi-dimensional database technology allows fast data retrieval and the ability to drill down from one summarization level to another. Significant processing, however, is needed to prepare the data.
3. A Relational OLAP (ROLAP) data mart uses relational database technology to provide an OLAP database. A star schema database design is most common, and the database design is optimized to support the queries of a primary group of users. The amount of data transformation and preparation processing may be less than that needed for a MOLAP data mart, but query

response is usually slower due to the overhead from SQL processing and table joining.

It is important to note that data marts are "owned" by a business unit or group of users. The data mart exists to serve the needs of a specific user group and so those users must have control over the database design and contents as well as the underlying OLAP technology. Inmon (1996) points out, however, that responsibility and accountability come with ownership. Data Mart owners do not work in isolation, and they have obligations to participate as "active members of the corporate information community".

Exploration and Data Mining Warehouse

Large queries are generally banned from the data warehouse for the simple reason that they drain huge amounts of resources and cause overall system performance to suffer. "Explorers" who ask questions and look for patterns no one else has thought of are the ones most affected by this ban. One solution that may be tried is to allow explorers to use the data warehouse late at night, between 1 and 5 AM perhaps, but that requires that queries be broken into pieces that can be scheduled and processed in the time window allowed.

A better long-term solution is to create a data warehouse dedicated exclusively to exploration and data mining for handling long-running, complex statistical analyses. Data exploration uses detail data to find patterns and then formulates hypotheses about why those patterns exist. Data miners take those hypotheses and test their validity against historical data.

A major benefit of an exploration warehouse is that it relieves the data warehouse of a processing burden. But why can't a data mart be used as an exploration warehouse?

- An exploration warehouse needs granular data—transaction-level is best. A data mart contains summarized data.
- A data mart is built from specific requirements about the data it needs to contain. An exploration warehouse is a constant work in progress.
- Exploration warehouse users are statistically oriented and use statistical analytical software tools. Data mart users are business managers and knowledge workers who use OLAP software.
- A system that tries to do all things for all people will satisfy nobody.

Alternative Storage

A data warehouse grows quickly (usually faster than anticipated) because it contains detailed data and each load multiplies the amount of data it contains. For a number of reasons though, a data warehouse may contain a lot of dormant data—data that is either never used or is used rarely. By moving this dormant data to alternative storage, performance is improved, the cost of the warehouse drops, and the warehouse can be redesigned to contain an even finer level of granularity.

A database vendor-supplied activity monitor should be used by to log the queries that users submit. The log, in turn, can be analyzed to identify dormant data that is a candidate for alternative storage.

Two types of alternative storage exist: secondary storage and near-line storage. Secondary storage is disk-based, but it's slower and less expensive than the high-performance disks arrays that normally support a data warehouse. Near-line storage is tape-based and managed robotically for fast cartridge mounting when retrieval is needed. With each type, a contents directory and a cross media storage manager insure that data in alternative storage is available when needed.

When properly implement with a product such as FileTek's StorHouse, the cost for high-performance disk arrays can be dramatically reduced, better query performance can be achieved, and the depth and breadth of the warehouse can grow almost indefinitely. For the users, queries are executed in a transparent manner so that they see no difference.

Internet and Intranet

The Internet and organizational Intranets provide an increasingly important application channel for data entering the CIF and for communication and interaction between the components. The Internet and Intranets can be used to:

- Transport data
- Distribute processing
- Schedule and coordinate processing
- Report status
- Allow DSS software access to the data mart

Factors that inhibit or restrict Internet and Intranet usage in the CIF environment include:

- The volume of data to be moved
- Network capacity
- Bandwidth availability
- Cost for usage

Metadata

According to Inmon (2001) "[t] he most important yet most ambiguous, most amorphous component of the Corporate Information Factory (CIF) is the metadata." "It is the glue that holds the CIF together." This author whole-heartedly agrees.

The definition "data about data" does very little to describe what metadata is or does. For data warehouse technical staff, metadata "refers to anything that defines a data warehouse object, such as a table, query, report, business rule, or transformation algorithm. It also supplies a blueprint that shows how one piece of data is derived from another." For business users, "metadata shows non-technical users where to find information, where it came from and how it got there, describes its quality, and provides assistance on how to interpret it." (Bentley, 2001.)

Good arguments can be made for both a centralized metadata management system and an autonomous system in which each data mart/data base has its own facility. The best system, as usual, is a balance between centralization and autonomy. A centralized facility should insure completeness, reliability, and ease of access. Autonomy, though, allows customization—the valid values of a field in a data mart may not be the same as those allowed for the data warehouse source field.

A good solution is to establish a central metadata repository along with smaller metadata repositories for each CIF component that has a database. This provides control of standardized sharable metadata in the central repository while still allowing a customizable component-specific solution.

Very often, the difficulty of capturing metadata is what causes it to receive a low priority. Many of the CIF applications that feed the data warehouse date to the late 1980s and early 1990s, and the people who built them moved on long ago. Even after Y2K efforts, documentation is either sparse or simply doesn't exist. Updates and maintenance applied to those systems makes recreating documentation even more difficult.

Yet good accurate metadata is critical for DSS processing. Without a map of the database and field-specific metadata, an analyst, especially one new on the job, will have a very hard time getting started, much less doing an effective analysis and making sense of the output.

Entire books have been written about metadata, but an excellent overview of metadata as it relates specifically to the CIF is in Chapter 12 of Corporate Information Factory. A SAS-focused overview is "Metadata: Everyone Talks About It, But What Is It?" in the SUGI 26 Proceedings.

SAS Software provides a product for maintaining and updating a metadata repository and more will be said later about that. One custom SAS solution is presented in "Creating Multi-Purpose Metadata for Data Quality, Trending, and a Data Dictionary" in the Proceedings of the Southern SAS Users Group Conference.

Decision Support System Applications

The purpose of the CIF is not to store data—its purpose is to receive data, transform it into information, and then deliver that information to those who need it so that they can add context and turn it into knowledge. Decision support applications are used to solve strategic and tactical business problems, and they use data stored in data marts. The data warehouse's data is available, of course, when background or history is needed and the ODS is available when the freshest data is needed.

Many decision support applications use a data mart dedicated to a specific analytical process that cuts across departmental boundaries, such as risk management or marketing campaign analysis. This narrow focus and broad audience of decision support data marts contrasts with the generally broad focus and narrow audience of departmental data marts. Of course, departmental data marts are available to the DSS applications.

Table 1. Data Mart Features

	DSS Data Mart	Departmental Data Mart
Focus	Narrow	Broad, generic
Outputs	Standard analyses, repetitive reporting	General analyses and reports, ad hoc queries
Usage	Highly structured	Unstructured
Access Pattern	Predictable	Unpredictable
Audience	Corporate-wide	Single department or group

Many vendors have developed off-the-shelf DSS applications that address either strategic or tactical problems. These applications can be either vertical, e.g., for the financial services, energy utilities, or telecommunications industries, or horizontal for areas like risk management, CRM, or fraud identification. Many of these packages include a data model, database schemas, and even transformation rules. Among other difficulties, however, the trick is to determine the degree of fit between a packaged solution and your documented business requirements.

Buying a DSS application solution can have advantages over building one in-house:

- Shorter implementation time
- Expanded functionality and usage of an existing data mart
- Forced standardization of data and processes

On the other hand, a packaged solution:

- Can be expensive and may require consultants for implementation
- Provides specific limited functionality and might not be scalable
- May have functionalities and capabilities that you don't need but must pay for
- May allow only limited customization
- May require that the business users change their processes to fit what the application offers
- May not provide competitive advantage because your competitors are using the same solution

SAS in the Corporate Information Factory

So where can SAS Software be used in the CIF? As we said earlier, almost everywhere because SAS products offer a complete end-to end solution for almost all the processes needed for a CIF and most of the elements and components associated with it. The SAS Rapid Warehousing Methodology 4.1 is based on architecture very similar to the Corporate Information Factory. Not all the listed components must be present, but in general the SAS architecture contains:

- Operational source systems and external data
- An staging area to land data from the source systems prior to integration and transformation
- An enterprise data warehouse
- An operational data store
- Data marts at the departmental or small group level
- Application specific data extracts to support business applications or processes
- A business intelligence portal via a single web interface
- Metadata

To address these architectural issues, in broad terms SAS's integrated software products and solutions either provide or support:

- Applications for data capture and automated business processing
- Data warehouse, ODS, and data mart design, construction, and maintenance
- Metadata capture and the metadata repository
- Data integration and transformation
- Data exploration and mining
- Online analytical processing, both the relational and multidimensional flavors
- Generic and specific Decision Support System needs
- Web enablement for data access and information delivery via the Internet and Intranets

Overall CIF design, construction and maintenance

Integration Technologies® is in a category by itself because it provides the foundation for SAS to be integrated with the non-SAS enterprise applications and information systems that are part of the Corporate Information Factory. Integration Technologies supports industry standards such as the distributed object models COM/DCOM and CORBA, the Lightweight Directory Access Protocol (LDAP), and message-oriented middleware. By implementing these standards SAS extends access to and presentation from SAS-based decision support and business intelligence applications as well as SAS-format data warehousing/data mart data sets. Using Integration Technologies, building and deploying solutions in thin-client, Web, and n-tier distributed application topologies is possible, along with is development in non-SAS programming languages including Visual Basic, Java, and C++.

Applications for data capture and automated business processing

SAS/AF® and Frame provides a development tool set for building object-oriented, menu-driven on-line transaction processing applications (OLTP). Using SAS/AF without Frame, you can still build text-based OLTP applications. These applications are scalable and portable across all computing platforms. Frame's library of pre-built components supports rapid application development and allows easy modification and enhancement of existing Frame applications. When used

with SAS's web enablement tools, the Internet becomes a channel for data collection.

Base SAS®, the SAS Macro Language, SAS/CONNECT® and other "traditional" SAS products can combine to build powerful automated business processing programs and applications that run in batch to process data collected by non-SAS applications. Using a scheduling utility such as AutoSys, SAS programs can be scheduled to run at a specific time or they can be combined with a shell script so that initialization or execution is event driven.

Data warehouse, ODS, and data mart design, construction, and maintenance

As noted above, the SAS Institute has developed a detailed Rapid Warehousing Methodology that complements the CIF concept. The CIF provides a picture of what the end product should be and SAS's Rapid Warehousing Methodology provides detailed guidance on how to build that product.

The SAS/Warehouse Administrator® serves as a control center for the SAS software components that are needed to build and maintain a data warehouse, ODS, or data mart. Erwin models can be imported and registered, and the underlying databases can be SAS, any major database or another data source—or a combination of any of them. The Warehouse Administrator uses SAS/CONNECT so that it can reside on a local Windows desktop but control processing and databases on remote mainframe, UNIX, and other computing platform.

Warehouse Administrator integrates ETCL programs, captures and consolidates metadata, schedules processes, and facilitates usage of decision support tools. A graphical process editor uses flow diagrams to generate on-the-fly custom code or to run existing source code such as COBOL routines, not just SAS programs. The current version (2.1) has a number of optional add-in tools that provide more functionality on an as-needed basis, including a delimited file loader, a constraint generator, a surrogate key generator, and job dependency and impact analysis reporting.

The Scalable Performance Data Server® (SPD Server) is designed to manage massive SAS data sets containing millions of rows and can be a viable alternative to relational database software for a data mart. SPD Server data sets are stored in a format different from "normal" SAS data sets, and SPD Server input/output functions, such as WHERE processing and INDEX creation, SQL pass-through and GROUP BY processing are designed to run in parallel against those data sets. Because it is designed for parallel execution, SPD Server requires symmetrical multiprocessor (SMP) hardware. On an SMP box, it will use all resources available to the machine to achieve maximum scalability and performance.

Metadata capture and the metadata repository

The SAS Warehouse Administrator and other off-the-shelf solutions make extensive use of metadata. SAS Version 9 products include an enhanced Open Metadata Server (OMS). The OMS fully complies with the Common Warehouse Metamodel (CWM) standard and provides common metadata services to SAS and other applications.

As a key member of the Object Management Group, the SAS Institute designed and built the Open Metadata Server® to insure compatibility and data transfer capability between SAS and other diverse data warehouse/mart components and software applications. This saves development time and effort because developers no longer have to implement their own metadata facilities.

The OMS is based on the Open Metadata Model, which defines about 150 metadata types. The model is organized into sub-models to assist in navigation and usage. To make the Model more understandable, the types are hierarchically organized into over five dozen groups that include:

- AnalyticTable
- AnalyticColumn
- ResponsibleParty—associates a set of persons with the object
- AbstractTransformation—documents data transformation and has five sub-groups, some with sub-groups
- QueryClause—defines transformations that are performed and has six sub-groups.
- Variable—defines substitution strings and replacement values
- Classifier—defines the structural or behavioral features of the object
- Event—describes a condition that occurs that causes other things to happen
- ContentLocation—provides location information and has five sub-groups, some with sub-groups
- Key—identifies table key columns and has two sub-types of unique and foreign.

The OMS uses the XML transport format to make it easier to exchange metadata between applications or publish metadata to a web page or in another channel.

Data integration and transformation

One of the many things that SAS is famous for is its ability to manipulate, transform, and manage data. The DATA step and PROCs in Base SAS provide an amazingly broad programming toolset. Any way you can think of to change your data, you can do it with Base SAS.

dfPower Studio® from DataFlux (a SAS-owned company) provides a methodology and a data quality tool for provides data cleaning, conversion, and scrubbing during application data integration. Among other capabilities, it allows ad-hoc data quality analyses to be run against specific data sources.

SAS Data Quality-Cleanse® is a Release 8.2 product for data analysis, cleaning, and standardization. Data cleaning, scrubbing, and standardization are critical because only with solid data quality controls in place can enterprise data standards be enforced and the validity of data assured.

All SAS products can directly read and write to major data sources by using the appropriate SAS/ACCESS® Interface to Relational Database (DB2, Informix, Oracle, Teradata, etc.) Interface to Enterprise Resource Planning system (SAP R/3 and BW, Peoplesoft, Baan). SAS/ACCESS can also work with non-relational data sources such as ADABAS, IMS-DL/I, CA-IDMS, and others). The Interface to ODBC enables access to data sources that include AS/400 data and Microsoft SQL Server. Using the ODBC driver provided by the appropriate vendor, SAS can easily read any ODBC-compliant data source.

SAS Software is built with a Multi-Vendor Architecture that allows a program written for one platform to run on another. Depending on the source and target systems, 90 percent or more of the code will be portable and only code that deals the operating system—such as filenames and libnames—may have to be customized. This makes is easy to move a SAS application from a system like an IBM S-80 running AIX to a system like a Sun Fire 4800 running Solaris. When changing databases, SAS/ACCESS's SQL Pass-Through capability and LIBNAME Statements can eliminate the need to change one database's SQL extensions to another.

SAS/CONNECT provides an n-tier client-server programming solution that allows programmers to take advantage of all the networked computing resources available to the organization.

- Remote Compute Services allows a program to execute in the most efficient location. You can move any or all of the processing to a remote machine to take advantage of those resources.
- Remote Data Services allows access to data stored in a remote environment. Remote data transfer moves a copy of the data to another platform, and remote library services transfers data on an as-needed basis from and to a remote platform as the local execution requests it.

Data exploration and mining

Base SAS provides well-known tools for data transformation, manipulation, and statistical description. SAS/GRAPH® provides the capability to analyze and then present data as two- and three-dimensional graphs, charts, plots, and maps. The Output Delivery System® (ODS) makes it easy to report results in a variety of formats in addition to listings, including RTF, PDF, and HTML.

Base SAS includes SAS/INSIGHT®, a point-and-click tool for exploring and analyzing SAS data sets. In the data window, you can add new data points, sort, and create subsets. Then compute descriptive statistics, create plots of univariate and multivariate data, and fit models using regression, analysis of variance, and GLM. The data, analyses, and graphs are linked so that changes flow through dynamically. For example, if you exclude a set of observations from calculations, all analyses immediately recalculate and the plotted data is redrawn automatically.

Enterprise Guide® is a thin-client product with a Windows-like interface that uses a point-and-click, drag-and-drop approach to data access, manipulation, and reporting. No programming is required, but you can run previously written SAS routines. Enterprise Guide can capture the code that it generates in the background so it's easy to move manipulations and transformations developed during data exploration into production or into a batch routine. A powerful feature of Enterprise Guide is its ability to rotate graphical output in "real time", thereby delivering a dynamic data visualization capability.

SAS/STAT® is a statistical workhorse, offering analytical techniques ranting ranging from simple to incredibly complex. The procedures are grouped into categories that include regression, analysis-of-variance, multivariate, clustering, survival analysis, and others. A number of PROCs perform nonparametric analysis, and other procedures allow analysis of structural equations with latent variables

SAS/IML® is a higher-level advanced programming language for manipulation of two-dimensional (row × column) matrices of numeric or character values. IML has built-in operators and call routines that perform tasks such as matrix inversion or eigenvector generation and also allows users to define their own functions and subroutines. Operations can be performed on an entire data matrix or a single value.

Enterprise Miner® (EM) was designed specifically to sample, explore, and model the huge amounts of data contained in and exploration data warehouse. EM's graphical interface uses flow diagrams to select and work with the data and build models. After running the models, examining and comparing results is also made easier with a flow diagram. EM provides decision tree, neural network, association, and linear and logistic regression algorithms. Models that fit the data can easily be moved into production because code is automatically generated at all stages of model development.

Online analytical processing, both the relational and multidimensional flavors

"Online analytical processing is a hardware/software configuration that provides a fast way of viewing and analyzing data from any perspective without having to specify the perspective and the exact level of detail required in advance." (SAS Institute, 1996).

SAS supports both multidimensional OLAP (MOLAP) and relational OLAP (ROLAP). The key difference between the two is the ease and speed at which application interfaces to the underlying database can move from one view of the data to another and retrieve pre-summarized data when the view's dimension changes. Many arguments are made about the merits of MOLAP vs. ROLAP but they won't be reviewed here.

A dimension is a hierarchy that can be used for aggregating. A time dimension can be day, month, and year. Using dimensions to view the data is commonly called 'slicing and dicing'. For example, a location dimension can be store, city, region, and state and data such as total sales can be aggregated at each level. Combined with the time dimension, data can be sliced and diced to view total city sales by month, for example.

The ability to view progressively more detail is called 'drilling down'. We may start with city sales percent change by month but quickly decide that we want to investigate one specific city-month combination that appears to be an anomaly, so we drill down into data to examine daily sales for that city. Expanding and collapsing dimensions—moving from a dimension summary to the dimension's components and back—is a key OLAP capability. Obviously, for OLAP to be successful the data we want to see must be available quickly.

Base SAS or Enterprise Reporter® combined with SAS/ACCESS can easily retrieve data from a ROLAP database. True OLAP functionality such as drilling down and expanding and collapsing dimensions will not be possible, but interim tables that subset data by dimension or subtotal on a class variable can be constructed and used to give the appearance of these functions. Performance will probably be slower than MOLAP, however, because of the overhead from the SQL processing needed for data retrieval.

SAS has two products that deliver MOLAP capability. SAS/MDDB Server® enables creation and management of multidimensional databases (MDDBs) that can be used by MDDB viewers. When used in conjunction with a graphical interface such as one designed with SAS/EIS®, a browser front-end from web/EIS® or web/AF®, or the MDDB Report Viewer that comes with SAS/Intrnet®, the power of MOLAP can be seriously exploited. You get fast slice and dice access to large amounts of pre-summarized data.

SAS OLAP Server® includes the SAS/MDDB Server and also supports registering OLAP cubes, 'reach through' to the data source underlying the MDDB (called Hybrid OLAP or HOLAP) and advanced Access Control capabilities. For those who want to use a non-SAS front-end, the OLAP Server enables access to SAS MDDBs by external sources.

Generic and specific Decision Support System needs

As with data exploration and mining, SAS Institute provides so many tools to build or exploit decision support systems that we run the risk of presenting a laundry list of products. Here are only a few of them.

Much has been written about the benefits of "build versus buy" and vice-versa. For those who decide to build a decision support application for information retrieval and distribution,

SAS Institute provides an integrated suite of products and training that make it a realistic endeavor.

SAS/EIS provides a syntax-free environment for building user-friendly enterprise information systems. Using object-oriented development and over thirty pre-built objects included with the software, it is a straightforward process to assemble a solid application with little or no programming. The reporting objects include a multidimensional data viewer, a comparison report that combines tabular data with graphics, and an expanding report that allows drill-down into the data. Fill-in-the-blank screens are used to tell the system where to find the data, which can be SAS data sets, flat files, or any of the major databases.

SAS/EIS-built applications use point-and-click menus with pull-down windows, give access to native host applications such as e-mail, and allow drill down, what-if analyses, variance and exception reporting, and multidimensional data viewing and analysis. Graphical capabilities include showing critical success factors, grouped bar charts, and "hotspots" that allow linking to other graphs, reports, or the underlying data. Enterprise Reporter is a thin-client package that has a GUI with a Microsoft Office look and feel. It provides a "palette" for report design that allows users to create, preview, and publish reports by pointing-and-clicking and dragging-and-dropping while the software handles the data in the background. SAS Institute describes Enterprise Reporter as "the users' window to the warehouse, fulfilling their information needs for creating reports on paper or the Web."

Data manipulation and transformation requires no hands-on programming, and the output options include HTML for creating web pages, PDF files for printing or e-mail attachments, and (of course) the listings for hard copies. If it is used by the technical staff in conjunction with SAS/Warehouse Administrator, Enterprise Reporter can be an excellent a tool for documenting the contents of a warehouse or data mart.

SAS Institute has a broad offering of pre-developed "solutions" that provide almost off-the-shelf answers to organizations problems and questions.

- Used with a financial data warehouse, SAS Financial Management Solutions provide answers to questions about organizational planning, budgeting, allocations, consolidation, reporting and analysis and support analysis and reporting. Strategic performance management and financial statistical analysis is also enhanced with SAS Financial Management Solutions.
- SAS Risk Management delivers a suite of credit, market, and operational risk analysis techniques and reporting capabilities. The risk management solution has been specifically adapted to the energy, banking, and insurance industries.
- Many large organizations are developing Human Resources Data Warehouses. The web-enabled SAS Human Capital Management Solution is designed to provide insights needed for planning effective human capital strategies and measuring HR-related practices.
- In addition to an HR data warehouse, many large organizations are building a Purchasing Data Warehouse to help manage their supplier relationships. SAS has a Supplier Relationship Management (SRM) Solution that has procurement scorecard, spending analysis, supplier ranking, and other detailed analysis and reporting features.
- A Customer Relationship Management (CRM) data mart is a must-have in customer-centric organizations. To leverage the data in a CRM data mart, the Customer Relationship Management Solution provides analytical modeling, cross-selling and up-selling modeling, segmentation and profiling, E-channel analysis and reporting capabilities along with the standard SAS data analysis and information delivery capabilities. Other Solutions related to CRM include marketing automation, interaction management, customer retention, and credit scoring.

Web enablement for data access and information delivery via the Internet and Intranets

In the past couple of years, SAS Institute has developed some very robust and comparatively easy-to-use products for web enablement. These include SAS/IntrNet, AppDev Studio®, web/AF, and web/EIS.

SAS/IntrNet "integrates the SAS System and the World Wide Web." It provides both Common Gateway Interface (CGI) and Java technologies for building dynamic Web applications and allowing users to access and execute remote SAS programs for analysis and decision support via the Web. SAS/IntrNet allows a thin-client, either a web browser or Java applet, to do:

- Web publishing, in which passive content is delivered to a Web server.
- Report distribution, where simple queries are constructed on the client and passed through the Web server to a data repository from which data is retrieved. The information is then passed back to the client in a report format.
- Dynamic application processing that handles custom data and information requests initiated on the client and executed by an application server using macro driven programs. The results are then passed back to the client for viewing.

AppDev Studio, webAF, webEIS each allows development of web-based, thin client information delivery applications. Each has a visual development environment that simplifies building object oriented applications.

AppDev Studio provides everything you need to create:

- Java client applications and applets;
- Java server applications (servlets, JSP and EJB);
- CGI/HTML applications;
- Active Server Pages applications
- Traditional full-client applications.

webAF software is used for building Java-based applications and applets that connect to SAS software such as SAS/AF objects, SAS data sets, and SAS programs. webAF generates 100 percent pure Java code, has complete JDBC and JavaBean support, and also integrates with non-SAS Java-based applications. It has a drag-and-drop application builder interface and wizards that simplify the development process.

Web-enabled MOLAP "documents" are created with webEIS software. A webEIS application reaches across the Web to retrieve data from a multidimensional database and provides the dynamic drill-down, sub-setting, and exception highlighting, and other capabilities of SAS/EIS. This allows users to explore live data using ad hoc computations, charts, graphs, tables, and embedded hyperlinks. webEIS itself is built with webAF.

Summary

The Corporate Information Factory is a powerful framework for organizing corporate information systems to achieve maximum performance and ROI. It presents a complete picture of what an integrated corporate information system looks like. It is less useful, however, in providing specific guidance on the nuts and bolts of how to build it and what tools to use. For that reason (among others) the SAS Institute's Rapid Warehousing Methodology and SAS Software are essential to getting the job done and achieving that vision.

SAS in the Corporate Information Factory—How do I love thee? Let's count some of the ways.

1. Integration Technologies
2. Base SAS Software, the SAS Macro Language, and SAS/STAT
3. The Output Delivery System
4. SAS/CONNECT and Remote Library, Compute, and Data Services
5. SAS/Access Interfaces to Relational Databases
6. The Scalable Performance Data Server
7. The SAS OLAP Server and SAS/MDDB
8. SAS/Data Warehouse Administrator
9. The Open Metadata Server
10. SAS/AF and Frame
11. SAS/EIS
12. SAS/Intrnet
13. web/AF and web/EIS
14. Enterprise Miner
15. Enterprise Guide
16. Enterprise Reporter
17. SAS Solutions, including Financial and Risk Management, Human Resources Management, SRM and CRM

References and Resources

Bentley, John E. (2001) "Metadata: Everyone Talks About It, But What Is It?" Proceedings of the 26th Annual SAS Users Group International Conference.

Bentley, John E. (2001) "Creating Multi-Purpose Metadata for Data Quality, Trending, and a Data Dictionary." Proceedings of the Southern SAS Users Group Conference.

Ekerson, Wayne W. (2000) "Ignore Meta Data Strategy at Your Peril." Application Development Trends, March.

Inmon, W.H. (1996) Building the Data Warehouse, 2nd edition. Wiley Computer Publishing: New York.

Inmon, W.H., Claudia Imhoff, Ryan Sousa (2001) Corporate Information Factory, 2nd edition. Wiley Computer Publishing: New York.

Kimball, Ralph et al. (1998). The Data Warehouse Lifecycle Toolkit. John Wiley & Son: New York.

Mattison, Rob (1999) Web Warehousing and Knowledge Management. McGraw-Hill: New York.

SAS Institute (1996). A SAS Institute White Paper: A Formula for OLAP Success. SAS Institute: Cary, NC.

SAS Institute, (2002). SAS® Rapid Warehousing Methodology 4.1. SAS Institute: Cary NC.

Bill Inmon's web site: http://www.billinmon.com

Intelligent Solutions web site: http://www.intelsols.com

About the Author

John E. Bentley has used SAS Software since 1987 in the healthcare, insurance, and banking industries. For the past six years he has been with the Corporate Data Management Group of Wachovia Bank with responsibilities of supporting users of the bank's data warehouse and data marts and managing the development of SAS client-server applications to extract, manipulate, and present information from them. John teaches a short course in parallel processing with SAS Software and regularly presents SAS User Group Conferences. He is chairing the SUGI 28 Systems Architecture section and organized the Weekend Workshops for SESUG 2002.

Contact Information

John E. Bentley
Wachovia Bank
Corporate Data Management Group
201 S. College Street
Mailcode NC-1025
Charlotte NC 28288
704-383-2686
John.Bentley@wachovia.com