In the 21st century the majority of data managers and consumers understand the importance of accurate robust data. We know that our data warehouses, CRM systems, ERP systems, and BI reports are compromised if the data we feed them is suspect. The millions of dollars we invest in those systems are wasted if the users do not trust the data in them and will not subsequently use them. Those systems will then sit marginalized, under used, and destined to become shelf ware. As one analyst put it, “it’s all about the data, not the software system.” We need to keep firmly in mind that the application is just a means of collecting, managing, and delivering the data.

Put simply, to realize the full benefits of their investments in enterprise computing systems, organizations must have a plan of how to monitor, cleanse, and maintain their data in a quantified state. Having a strategy of approaching the data quality challenge is the first step to building a program plan. The purpose of this paper is to provide you with an overview of creating a data quality strategy. It can be used as a guide for the early discovery sessions when the IT or business manager is first grappling with the topic and struggles to frame the problem into a manageable work effort. The data manager will encounter the need to build a data quality strategy in two different scenarios. The first is when they are implementing a new system, perhaps an ERP system, and they want data quality to be built in from the ground up. This means developing a strategy that encompasses three operations:

1. Cleansing data as it is migrated from the legacy system into the ERP system.
2. Validating and cleansing transactional updates to the ERP system.
3. Once the data is loaded, how to maintain it at regular intervals in its cleansed state.

The second scenario the data manager will be faced with is remediation of an existing marginal system or process. Marginal in this case being the quality of the information. In the second scenario greater emphasis must be placed on evaluating existing data structures and interface points as these will constrain access to the data quality cleansing functionality. Ultimately the goals between the two scenarios are the same: establish a system with maintainable levels of data quality going forward. However, the approach to achieve the goals will be different.

In reading this paper you can expect to learn the importance of organizational goals and the driving force behind the strategy, and how to use those goals to direct the strategy effort. From there the five aspects used to build the strategy are explored, and lastly we close the paper with a section on the implementation of the strategy.
The organization’s goals drive strategy. These goals flow down to a data quality project through the business unit, department, or operation that uses the data. That is why the first section of a data quality strategy must list the corporate (organizational) goals the strategy is designed to support. Moreover, it is not enough to just list the corporate goals, but the strategy must clearly articulate how the improvement of the data directly contributes to the corporation’s attainment of its goals. Failure to do this, to create that value chain from corporate goal to clean data will result in failure to obtain approval for the project. There is no business value in cleaning data just because the data is defective. Senior management will want to know why they should spend scarce resources on a data quality project with unknown benefits when they can expend the same resources on a different project with tangible outcomes. It is purely a business decision, and it is one that is decided in the goals section of the strategy.

As an example, a financial services firm had the corporate goal of increasing revenues by cross selling products between customers in different product lines. To obtain that goal they determined that they needed a single master customer information file (MCIF) that held the customers from all seven of their lines of business. Building a data warehouse is no small undertaking. The data warehouse project for the firm was progressing smoothly until they came to the problem of trying to match customers across source systems. The problem resolved itself by first standardizing the customer names and addresses so they could be accurately matched, and then developing the matching logic. What at first, in isolation, appeared to be two separate events: standardizing names and addresses, and generating $80 million in cross selling products, became inextricably linked when one was shown to have a direct impact on the other. The MCIF manager who was faced with developing a strategy to fix the matching problem and thus enabled the $80 million revenue stream, met with their business counterparts. They gathered the reasons why the MCIF, and hence sound data, were needed for cross selling, and listed those reasons in the strategy document. The project was funded at the next senior management meeting. They had connected all the dots and depicted the value chain.

Connecting the links in the value chain: The drawing above shows the MCIF example of moving from an organizational goal to a data quality action.
If the team responsible for creating the strategy cannot identify the business value for cleansing and maintaining the data one of three things is at work:

1. The person or team building the strategy does not truly understand the business operation that uses the data, and are therefore the wrong entity to build the strategy.

2. There is no value for cleaning up the data, and therefore the value of the data itself is called into question.

3. The organization has no clear objectives.

Of these three it is usually the first that is the culprit and can be rectified by finding the business manager whose job depends on that data. Of anyone in the firm, they will be able to explain how and why the data is used.

**THE FIVE ASPECTS OF DATA QUALITY STRATEGY**

When decomposing the data quality problem into dimensions or perspectives that aid the practitioner in building a framework through which to solve the problem there are five aspects of an organization's operations that must be considered. Those five aspects are:

1. **Subject Area** — Identifies the type and usage of the data being cleansed
2. **Connectivity** — Lists options for connecting data quality functionality to the data
3. **Data Flow** — Shows how the data moves through the environment
4. **Governance** — People and processes responsible for managing the data
5. **Data Monitoring** — Processes for regularly validating the data

The drawing below depicts the five aspects as they are driven by the organization's goals. As the strategy is created the team composing the strategy needs to be cognizant of the program constraints that will impact the options and alternatives put forward in the strategy.
SUBJECT AREA

The subject area aspect defines the domain of data and how it is used. Ultimately the data domain determines the necessary types of cleansing algorithms and functions needed to raise the level of quality. Examples of subject areas and the entities and attributes found in each are:

- **Customer** — first name, last name, street address, e-mail address, customer number, account number
- **Equities** — code, symbol, vendor, exchange, date, description
- **Supply Chain** — part number, description, quantity, supplier code, SKU code
- **Sales** — product name, code, amount, date, transaction ID, POS ID, agent ID

Subject areas can be matched against their appropriate types of cleansing algorithms. Using customer and our MCIF example, an attribute of customer is first name. In order to match across the seven different input source systems differences in first names such as Bob versus Robert, or Bill versus William need to be reconciled. To do that a specialized data cleansing algorithm is employed that contains information on the domain of the subject area (typically as a lexicon or custom algorithm). This specialized function is used to standardize, in this case, Bob to Robert so they can be matched. The function used on first names probably cannot be used without modification on equity names, hence the need of the strategy to list the subject areas and entities, and if time permits the crucial attributes, i.e. fields.

CONNECTIVITY

In order to monitor and correct the data the data quality function must be able to connect to the data repository and access the data. The first step in resolving the connectivity aspect is identifying the potential data repositories (source systems) that hold the data. These repositories can be cataloged when doing the data flow aspect as that evaluation interrogates the systems architecture.

In the connectivity aspect what we consider is whether the data is distributed across many systems or centralized in one repository, and whether those environments are homogenous (all the same, such as in SAP® BW) or heterogeneous (in a combination of, for example, MS Excel, Oracle, and DB2). If the data resides in an enterprise application (CRM, ERP, DW, etc.), the vendor, and platform of the application will dictate connectivity options to the data. Connectivity options between the data and data quality function generally fall into the following three categories:

1. Data extraction
2. Embedded procedures
3. Integrated applications

**Data extraction** occurs when the data is copied from the host system. It is then cleansed, typically in a batch operation, and then reloaded back into the host. Extraction is used for a variety of reasons, but typically it is because direct access to the host system is impractical.

**Embedded procedures** are the opposite of extractions. Here, data quality functions are embedded, perhaps compiled, into the host system. Custom-coded, stored procedure programming calls invoke the data quality functions, typically in a transactional manner. Embedded procedures are used when the strategy dictates the utmost customization, control, and tightest integration into the operational environment.
Integrated applications lie between data extraction and embedded procedures. Through the use of specialized, vendor-supplied interfaces, data quality capabilities are integrated into enterprise information systems. The published interface allows for a quick, standardized integration with seamless operation, and can operate in either a transactional or batch mode. Owners of CRM, ERP, or other enterprise application software packages often choose this type of connectivity option.

The three approaches for connectivity are supported by an array of technologies offered by vendors:

- **Low-level Application Program Interfaces (APIs)**
  - They offer in depth control of many parameters, but demand custom programming by the end-user

- **High-level APIs**
  - They offer access to functionality through a summarized parameter set, but reward the end-user with lower programming requirements and faster implementation times

- **Web-enabled applications**
  - For real-time e-commerce implementations

- **Enterprise application plug-ins**
  - For predefined integration into ERP, CRM, and other applications

- **GUI (graphical user interface) interactive applications**
  - For data profiling and visual interrogation of the data

- **Batch applications**
  - Can run in automatic or manual start modes and take extract files or data streams as input, and then output the results in a similar fashion

- **Web services/ASP connections**
  - Provide access to external (in the cloud) or on-premise sourced functions

The deployment options are in addition to custom programming or “home grown” solutions using SQL, ETL scripts, or other languages. It is the purpose of the strategy to select the deployment option(s) that fit the operational environment. We will discuss this further in the final section, drafting the strategy. The drawing below depicts the three categories of connectivity options in addition to the different types of deployment technologies.
DATA FLOW

Each of the five strategy aspects builds a different view of the data environment. With subject area (type of data) and connectivity (access to data) identified, the next step in developing a data quality strategy is to focus on data flow (the movement of data).

Contrary to perception, data does not sit static in a data repository. Data flows through an organization like blood in the circulatory system, and each day, each hour there are a myriad of touches to that “static” data. To the modern business, data is the crucial fluid that carries nutrients (information) to those business functions that consume it. The movement of data imposes another dimension on a data quality strategy. Picture it as a moving target, like blood in the circulatory system. The question becomes where is the best place to intercept the data, while in transit, so that it can be cleansed and validated? The human body has its own answer, and that is the liver, but for those of us building data systems, we have many more options.

A data flow or system architecture diagram (shown above) is created as part of a data quality strategy and will indicate where the data is captured, manipulated, and stored. Knowing these locations provides the strategist a selection of the best locations to cleanse and monitor the data given the project objectives (goals). The effort of evaluating the data flow will allow the strategist to refine the results compiled in the connectivity and subject area aspects as both of those are examined when building a data flow diagram. The data flow diagram will depict access options to the data, and catalogs the locations in a networked environment where the data is staged and manipulated. These can be thought of as opportunities to cleanse the data. These opportunities fall into the following categories:

- Transactional Updates
- Third Party Data
- Regular Maintenance
- Operational Feeds
- Data Migrations

The movement of data is spawned by two general operations: automated processes such as a nightly ETL script, and manual processes such as the data entry by a sales person into a mobile CRM application. Data flow analysis must consider both automated and manual initiators of data movement. For evaluation of manual activities the data flow analysis turns into the task of work flow analysis.
Work flow and data flow are closely related. A work flow, such as entering a new product code, immediately spawns a data flow. It is important to inventory these work flow touch points as they represent points of capture, and are opportunities to validate and cleanse data as it is created. The highest incidence of data quality errors, other than data aging, occur in the manual entry of data and therefore merits significant attention in the strategy. User interface drop-down fields where a value is selected rather than having it entered free form is a common tactic used to ensure data integrity at the point of capture. There are numerous other tactics, such as back-end business rule checking. Just because an entry in a field may be valid against a given domain of possibilities does not mean the entry is valid in the context of all the other data. For example, a California ZIP code may be valid, but entering it for a Michigan address invalidates it. Back-end business rule checking can catch these types of work flow-generated errors.

GOVERNANCE

Data governance is perhaps the most crucial component in a data quality strategy. The rules and definitions that govern the usage of the data and therefore determine the necessary level of quality come from the governance function. While there may be no formal data governance function in an organization, let there be no doubt that data governance is being performed, however imperfect. Any decision made regarding the management of information stems from data governance.

Typically a governance program will entail two distinct roles, the governance council, and the data stewards. While the purpose of this paper is not to define a data governance program, it is necessary to know the roles so the data quality strategist knows who to look for. Most likely, the person building the strategy is a data steward or a member of the governance council, or both. Certainly the strategist will have a key role in the team that is resolving the data quality issue. For simple definitional purposes a data steward (formal or informal) is any person who creates, manages, or provides information. Their role may be part-time, and they may not be accounted as a steward, but if they manage data, they are a data steward.

When evaluating the data governance aspect of a data quality strategy the following points should be addressed:

1. Who are the stakeholders of the data? What are the predominant user groups, and who are their representatives. Who is responsible for the creation, capture, maintenance, reporting, distribution, and archiving of the data? A list of these people is needed as they will be involved in the project. At a minimum they should be interviewed for their input to the project requirements. They will provide definition and context around the data, and will hopefully provide documentation, such as a data dictionary that will describe the data. Certainly there will be people on this list that will be required to sign off on the strategy, and will potentially have budgetary approval to authorize the project on an ongoing basis.

2. Once a near-final set of requirements for the strategy are ready they need to be reviewed with the stakeholders.

3. Consider the ramifications of how an improved and changed system will impact existing roles and responsibilities. The strategy should outline these potential changes so that personnel managers can provide insight as how to implement.

4. Identify pending process changes. Seldom can improved data quality be accomplished without a process change, because it is often a process weakness that causes the defect. The process change may be simple like changing an input screen from free-form text entry to a drop-down selection list, but it results in a process revision. Between the data flow analysis and data governance evaluation potential process improvement actions will become apparent. At this juncture it is important to keep the scope of the strategy effort constrained to building the strategy and not investing in a detailed process redesign. That effort is held as a task in the detailed project plan. For the strategy, as an example, it would be suffice to say “the data entry process for new account entries needs to be redesigned.”
5. What pending decisions need to be made? What important issues surrounding the data have not been resolved? Often times this is because the stakeholders can’t agree on an approach or a clear delineation of authority over the data. This can be one of the biggest problems to gaining go-ahead approval. The data quality strategist must carefully list these outstanding decisions and offer a solution, perhaps “straw man” ideas to move forward. One example comes from a large retailer. The staff from the various marketing departments could not agree on the format of an account number field to be used for customer identification. The proposal was to use the credit card number, but using too many digits exposed too much information about the consumer, and too few digits made using the number ineffective. The deadlock was broken when a data steward recommended the decision be put to the corporate privacy and policy administrator. A simple decision in the end, but it held up a customer data quality project until it was resolved.

DATA MONITORING

The fifth aspect in a data quality strategy is data monitoring, which is the measuring, analyzing, and reporting on the data through a consistent and scheduled process. Data monitoring employs a process known as data profiling and assesses defined quality measurements and delivers them as early warning indicators to the data consumers. Data begins to age immediately after capture, and any process that touches the data can generate errors. Only a rigorous automated process will detect aging or newly generated defects before they impact operations.

An issue the strategy needs to address is how often to run the data profile or audit. The frequency of running the automated audit is determined by these factors:

- How often the data is accessed — hourly, daily, weekly, monthly, etc.
- The importance of the operation using the data, is it mission critical, is it for a direct mail campaign, or end of month financial reporting?
- The cost of monitoring the data. What are the human operator, process, software license, and equipment costs?
- Operational impact of monitoring the data. The strategist needs to consider the impact of assessing production data during daily operations, and the effect of the process on operations staff. The level of test automation and the need for the test results will be some of the parameters around monitoring operational data.

The benefit of data monitoring is it alerts managers and data consumers to deterioration in data quality early in the trend. It identifies which processes (depending on the granularity of testing) are functioning properly or have experienced an transient event. Moreover, the results of data monitoring can be used to quantify the effectiveness of data remediation activities. Finally, data monitoring provides regular information to end-users as to the usability of their data, which can increase their confidence, assuming the trend is positive, that the data is fit for use.
The best process for drafting the strategy is to begin with the organizational goals. Make sure relevant goals that are driving the project are documented, and then capture the link between those goals and the project objectives. Using the MCIF example, the organizational goal of increasing revenue through cross-selling was directly related to the data quality objective of standardizing and matching of customer names and addresses. There will be other objectives, each must be related to the goals.

The next steps are in order:

1. Define the subject area
2. Determine potential connectivity options to the repositories of the subject areas
3. Diagram the data flow(s) to evaluate candidate intersections where the connectivity options can be employed
4. Identify the program stakeholders and the communications to them. This will become visible when you know the repositories and work/data flows.
5. Plan how, when, and at what frequency the data will need to be monitored against the defined quality criteria.

Considerable license can be taken by the strategist when following these steps. Iteration and feedback from a later step into an early step is common as later steps will provide more information that is pertinent to the early ones. For example, if a decision is made to monitor and validate data at the point of entry into SAP® ECC, then a special emphasis would be placed on that connectivity option.

The dividing line between strategy development and detailed project planning is where the effort begins to quantify costs. The strategy process is dynamic, and is never truly finished. As the project planning effort delves into the cost of software licenses, resource allocations, data volumes, time constraints, operational impacts, and governance decisions, these implementation factors need to be fed back into the strategy in the form of parameters that will refine the strategy alternatives. Knowing that the strategy will be impacted by project planning constraints, creating strategy versions can be a useful exercise. As an example the strategy could be to implement cleansing across the system in a phased-in approach, allowing source systems to be added or delayed as the financial budget allows. A classic approach to scaling a data quality strategy is to plan a series of projects within the overall program where each project builds upon the previous, but where immediate value and benefit is derived from the project stand-alone.

Another factor to weigh when drafting the strategy is the organization's data quality maturity and corporate culture. If the concept of data quality is new to the organization a simple start to the strategy is best, and then build in comprehensiveness over time. The project-based approach suits this growth well as simple, quick win pilot projects can be invoked that build credibility and confidence in the team to deliver on the promised output. An example of a simple starter pilot that could be the first implementation of a larger strategy is to standardize and validate all product names in a material master. It involves one attribute, but has immediate pay-back in ordering efficiency. Success in this first project spawns the next and the next until one day the organization realizes that data quality and its benefits have become ingrained in the corporate culture. Herein lies a strategy of itself, divide and conquer. Successful small projects will drive future initiatives, and an initial strategy can be a planned part of a larger recurring cycle. The quality improvement process must be a repeatable process, because business and the data that drives it is dynamic. A challenge solved today just prepares for the data quality challenge of tomorrow.