4.1 Introduction

In the classes of applications that heavily depend on enterprise data quality—business intelligence, finance reporting, market-trend analysis, and so on—a typical approach to the data-quality problem usually starts and ends with the activities scoped to the physical data-storage layer. Not surprisingly, according to the trade publications, most of these efforts have minimal success. Given that in business applications data always exists within the context of a business process, all the attempts to solve the "data-quality problem" at the purely physical data level are doomed to fail. This failure is because the physical level does not capture the requisite semantics to accurately communicate data across business processes. As a result, most of the semantic data issues exist at the process and organizational boundaries. The top level is the focal point with the highest probability for discrepancy. Most companies do not have domain models. If models exist at all, the majority of them exist at the project level as logical data models. Enterprise-level business and information models are practically absent, and therefore
there is no way to effectively communicate the data across organizational, department, project, or service boundaries.

Data Quality is affected by factors other than the system itself, such as whether it reflects real world conditions, and can be easily used and understood by the data user. If the data is not interpretable and accessible by the user, even accurate data is of little value [1]. Therefore, a methodology for designing and representing corporate data models is needed. The use of scenarios, subject areas and design rationale has been found to be effective in enhancing the understanding of corporate data models [2].

4.1.1 Data Quality

Major business initiatives in a broad spectrum of industries, private and public sectors alike, have been delayed and even cancelled, citing poor data quality as the main reason. The lack of quality data often leads to decisions being made more on the basis of personal judgment rather than being data driven [3]. Gartner Research finds that poor Data Quality hamstrings organizations by limiting their agility and growth, causing waste and wrong decisions, and adding significant business risk [4]. Informatica indicates that defective data leads to breakdowns in the supply chain, and hampers efforts to meet regulatory compliance responsibilities [17].

The problem of poor data quality has become so severe that it has moved to the top tier among the reasons for business customer’s dissatisfaction with their IT counterparts. While it is hardly an argument that poor data quality is probably the most noticeable issue,
in a vast majority of cases it will be accompanied by the equally poor quality of systems engineering in general, including requirements elicitation and management, application design, configuration management, change control, and overall project management.

The popular belief that "if we just get data under control, the rest (usability, scalability, maintainability, modifiability, and so forth) will follow" has proven to be consistently wrong. If business applications in general are failing to meet realistic expectations of the business users, data is just one of the reasons cited, albeit the most frequent one. Therefore, what is commonly called the "poor data-quality problem" should be more appropriately called the "data-quality deficiency syndrome." It is indeed just a symptom of a larger and more complex phenomenon that we can call "poor quality of systems engineering in general." Data quality is just the most tangible and obvious symptom that our business partners observe, not the root cause.

### 4.1.2 Importance of Achieving Data Quality

Achieving quality data at the early stage of database design is a critical issue for both the database researchers and practitioners. Conceptual design focuses on application issues such as entities and relations. As data increasingly outlives the application for which it was initially designed, is processed along with other data, and is used over time by users unfamiliar with the data, more explicit attention must be given to
data quality. Organizations are becoming more and more dependant on data; virtually everything the modern organization does both depends upon and creates enormous quantities of data. To meet the needs of the organization, a comprehensive data management program is essential [5].

Further, Levitin and Redman saw the need for management science for data, as data are different from other resources and require different management techniques [6]. Data quality is a multidimensional concept as data itself is multidimensional [7][8]. Modern definitions of data quality have a wider frame of reference and many more attributes than the obvious characteristics of accuracy. Strong take a consumer (people or groups who have experience in using organizational data to make business decisions) focused view, that quality data is ‘data that is fit for use’, and this view is widely adopted by the literature [9][10][11][12]. Redman comes to the following definition based on Joseph Juran [8];

“Data are of high quality if they are fit for their intended uses in operations, decision-making, and planning. Data are fit for use if they are free of defects and possess desired features [9].”

### 4.1.3 Initiatives for Data Quality

Everybody wants better quality of data. Some organizations hope to improve data quality by moving data from legacy systems to enterprise resource planning (ERP) and customer relationship management (CRM) packages. Other organizations use data profiling or data
cleansing tools to unearth dirty data, and then cleanse it with an extract/transform/load (ETL) tool for data warehouse (DW) applications. All of these technology-oriented data quality improvement efforts are commendable—and definitely a step in the right direction. However, technology solutions alone cannot eradicate the root causes of poor quality data because poor quality data is not as much an IT problem as it is a business problem. Major business initiatives in a broad spectrum of industries, private and public sectors alike, have been delayed and even cancelled, citing poor data quality as the main reason.

According to Joseph M. Juran, a well-known authority in the quality-control area and the author of the Pareto Principle, data are of high quality "if they are fit for their intended uses in operations, decision making, and planning. Alternatively, data are deemed of high quality if they correctly represent the real-world construct to which they refer" [8]. Again, this definition points to the notion that data quality is dependent on our ability to understand data correctly and use it appropriately. As an example, consider U.S. postal address data. Postal addresses are one of the very few data areas that have well-defined and universally accepted standards. Even though an address can be validated against commercially available data banks to ensure its validity, this validation is not enough for every purpose. In this example, if a business uses a shipping address for billing and vice versa, or uses a borrower’s correspondence address for appraisal, the results will obviously be wrong.
In general, data may be of poor quality because it does not reflect real world conditions or because it is not easily used and understood by the data user. The cost of poor data quality must be measured in terms of user requirements [13]. Even accurate data, if not interpretable and accessible by the user, is of little value.

4.2 Factors Affecting Data Quality
Maintaining the quality of data is often acknowledged as problematic, but is also seen as critical to effective decision-making. Examples of the many factors that can impede Data Quality are identified within various elements of the Data Quality literature [14]. These include: inadequate management structures for ensuring complete, timely and accurate reporting of data; inadequate rules, training, and procedural guidelines for those involved in data collection; fragmentation and inconsistencies among the services associated with data collection; and the requirement for new management methods which utilize accurate and relevant data to support the dynamic management environment. Clearly, personnel management, organizational factors, and effective technological mechanisms, affect the ability to maintain Data Quality.

4.2.1 Dirty data
Dirty data is a term used to refer to information/data that is misleading, incorrect or without generalized formatting, that has been collected by means of data capture forms. It could even be spelling mistakes or poor punctuation, incomplete or outdated data or even data that is
duplicated in the database. System data can degrade very rapidly, starting with customer information such as spelling of names, addresses, and missing information. A database that offers any kind of unsecured access can become unreliable-and ultimately worthless within two months. Even in a professionally designed and operated system, where the data is strictly controlled, errors exist.

### 4.2.1.1 Causes of dirty data

- Poor or "creative" data entry, including misspellings, typos, transpositions, and variations in spelling, naming or formatting.
- Missing data.
- Lack of company-wide or industry-wide data coding standards (e.g. a big problem in health care).
- Multiple databases scattered throughout different departments or organizations, with the data in each structured according to the idiosyncratic rules of each.
- Older systems that contain poorly documented or obsolete data.

### 4.2.1.2 Types of Dirty Data

Many of the dirty data examples described in the following list can be found in relational databases as often as they can be found in files:
• **Incorrect data**— incorrect data can be the most intractable, because much incorrect data will not be detected by validation, completeness, or consistency checking. Even though it is valid, complete, and consistent, it's just wrong. Not only is it harder for you to find, it can be maddening for the customer. For data to be correct (valid), its values must adhere to its domain (valid values). For example, a month must be in the range of 1–12, or a person’s age must be less than 130. Correctness of data values can usually be programmatically enforced with edit checks and by using lookup tables.

• **Inaccurate data**—a data value can be correct without being accurate. For example, the state code “CA” and the city name “Boston” are both correct, but when used together (such as Boston, CA), the state code is wrong because the city of Boston is in the state of Massachusetts, and the accurate state code for Massachusetts is “MA.” Accuracy of dependent data values is difficult to programmatically enforce with simple edit checks or lookup tables. Sometimes it is possible to check against other fields or other files to determine if a data value is accurate in the context in which it is used.

• **Inconsistent data**— inconsistent data can be even harder to find, because its detection requires even more inside knowledge. Even if the web of data for a customer is complete, it can still be inconsistent. Uncontrolled data redundancy results in inconsistencies. Every organization is plagued with redundant and inconsistent data. This is especially prevalent with customer
data. For example, a customer name on the order database might be “Mary Karlinsky,” the same name on the customer database might be “Maria Louise Karlinsky,” and on a downstream customer-relationship, decision-support system the same name might be spelled “Mary L. Karlynski.”

• **Incomplete data**— incomplete data is more difficult to find. Its detection requires a model of data relationships, and may involve multiple applications. During system requirements definition, we rarely bother to gather the data requirements from down-stream information consumers, such as the marketing department. For example, if we build a system for the lending department of a financial institution, the users of that department will most likely list Initial Loan Amount, Monthly Payment Amount, and Loan Interest Rate as some of the most critical data elements. However, the most important data elements for users of the marketing department are probably Gender Code, Customer Age, or Zip Code of the borrower. Thus, in a system built for the lending department, data elements, such as Gender Code, Customer Age, and Zip Code might not be captured at all, or only haphazardly. This often is the reason why so many data elements in operational systems have missing values or default values.

• **Nonintegrated data**— most organizations store data redundantly and inconsistently across many systems, which were never designed with integration in mind. Primary keys often don’t match or are not unique, and in some cases, they
don’t even exist. More and more frequently, the development or maintenance of systems is outsourced and even off-shored, which puts data consistency and data quality at risk. For example, customer data can exist on two or more outsourced systems under different customer numbers with different spellings of the customer name and even different phone numbers or addresses. Integrating data from such systems is a challenge.

4.2.3 Problems due to poor data quality

A common example of data quality problems is when trying to speak with a customer service representative (CSR) of a bank, credit card Company, or Telephone Company [15]. An automated voice response system prompts you to key in your account number before passing your call to a CSR. When a person finally answers the call, you are asked to repeat your account number because the system did not pass it along. Where did the keyed-in data go?

Another more serious data quality problem involves a report in 2003 about the federal General Accounting Office (GAO) not being able to tell how many H-1B visa holders worked in the U.S. The GAO was missing key data and its systems were not integrated. This presented a major challenge to the Department of Homeland Security, which tried to track all visa holders in the U.S.

According to Gartner, Inc., Fortune 1000 enterprises may lose more money in operational inefficiency due to data quality issues than
they spend on data warehouse and CRM initiatives. In 2003, the Data Warehouse Institute (TDWI) estimated that data quality problems cost U.S. businesses $600 billion each year.

At an Information Quality Conference in 2002, a telecom company revealed that it recovered over $100 million in “scrap and rework” costs, a bank claimed to have recovered $60 million, and a government agency recovered $28.8 million on an initial investment of $3.75 million. Clearly, organizations and government are slowly realizing that data quality is not optional.

Many companies realize that they did not pay sufficient attention to data while developing systems during the last few decades. While delivery schedules have been shrinking, project scopes have been increasing, and companies have been struggling to implement applications in a timeframe that is acceptable to their business community. Because a day has only 24 hours, something has to give, and what usually gives is quality, especially data quality.

4.3 Data Quality Characteristics

The idea of quality characteristics arises from the fact that there is more than one workable solution to most modeling problems, and hence we need some means of comparing alternative solutions. Understanding the key data quality dimensions is the first step to data quality improvement. Being able to segregate data flaws by dimension or classification allows analysts and developers to apply improvement techniques using data quality tools to improve both your information and the processes that create and manipulate
that information. Analyzing the parameters of quality helps in identifying how well the data supports the application.

Accuracy: Data accuracy means that the data are the correct values and are valid. To accurately analyze data, ensure that the algorithms, formulas and translation systems are correct. Also, each entry or record within the database should be correct. To store the data, appropriate edits should be in place to ensure accuracy. Exceptions or error reports should be generated and corrections should be made. Accuracy is checked where numeric data like amounts, counts, and quantities are involved. Ensuring accuracy involves appropriate education, training and appropriate communication of data definition to those who collect data. Data should be sufficiently accurate for the intended use and should be captured only once, although it may have multiple uses. Accuracy states that a piece of information is what it should be. In our definition, a telephone is a telephone if it has the right number of valid numbers and the right format for the right country. For example, this is a valid number for the Netherlands +31104065777, however if the field is supposed to contain mobile numbers then it would be invalid.

Consistency: Consistency means the value of the data should be reliable and the same across applications. Data should be analyzed under reproducible circumstances by using standard formulas, scientific equations, variance calculations and other methods to achieve consistency. Data is consistent when the value of the data is the same across the applications and systems. To ensure consistency, edits or conversions of tables should be employed
while storing data. Consistency is looked for in dates; like you cannot be born in the future.

**Definition:** Clear definitions should be provided so that current and future data users will know what the data means. Each data item should have clear meaning and acceptable values. For appropriate analysis, display data needs to reflect the purpose for which the data were collected. Appropriate comparisons, relationships linkages need to be shown. Clear, concise data definitions facilitate accurate data collection. The application’s purpose, the question to be answered or the aim for collecting the data element must be clarified to ensure appropriate and complete data definitions.

**Timeliness:** Timeliness is determined by how the data are being used and their context. Timely data analysis allows for the initiation of action to avoid adverse impacts. For some applications, time may be seconds and for some it may be years. To ensure timeliness while storing data means that data are available as per information management policy and retention schedules. Timeliness is looked for in data extractions, transactional details. Data should be captured as quickly as possible after the event or activity and must be available for the intended use within a reasonable time period. Data must be available quickly and frequently enough to support information needs and to influence service or management decisions.

**Relevance:** Data captured should be relevant to the purposes for which it is to be used. This will require a periodic review of requirements to reflect changing needs. The data are meaningful to the performance of the process
or application for which they are collected. The applications purpose, the question to be answered, or aim for collecting data elements must be clarified to ensure relevant data. For appropriate analysis, display data to reflect the purpose for which the data were collected. This is defined by the application. To ensure availability of relevant data, appropriate retention schedules should be established.

**Accessibility**: Accessibility means data items should be easily obtainable and legal to collect. To ensure accurate analysis, access to complete current data is better, otherwise results and conclusions may be inaccurate or inappropriate. Data ownership and guidelines for who may access data or system should be established.

### 4.4 Framework for achieving quality data

Almost every activity in which organizations engage involves data. Data provides the foundation for operational, tactical, and strategic decisions [16]. The quality of any business decision is only as good as the data at its foundation. Through data, managers plan, organise, and control an organization’s resources. They combine data in almost unlimited ways to search for new opportunities, market niches, process improvements, and innovative products and services. Data enables the better business decisions that drive performance, and builds the competitive advantage that ensures organizational success.

Data Quality doesn’t happen by chance. Organizations must establish certain guidelines for all personnel to follow to ensure that data quality is
addressed during the entire lifecycle of a system. Good quality data means that all master data is complete, consistent, accurate, time-stamped and industry standards-based. Data degrades day-by-day, so a consistent approach to entering the right data, cleaning data and importing good data will ensure that quality remains the same. The proposed framework helps to convert the dirty data into quality data. The above defined parameters help the system to maintain the quality of data while application development. The quality data used in the development process will help to improve the performance and outcome of the system.

The figure 4.1 displaying the structured framework is given below:

Figure 4.1: Quality Data Framework
In the proposed framework, data from different sources are collected and stored at one place after extraction and cleaning. Data cleansing is a labor-intensive, time-consuming, and expensive process, and cleansing all the data is usually neither cost-justified nor practical. On the other hand, cleansing none of the data is equally unacceptable. It is therefore important to carefully analyze the source data and to classify the data elements as critical, important, or insignificant to the business. Then, concentrate on cleansing all the critical data elements, and as time permits, cleanse as many of the important data elements as practical, leaving the insignificant data elements unchanged.

In order to confirm data quality, organizations must deal with both the subjective perceptions involved with the data and the objective measurements based on the datasets. Subjective analysis reflects the needs and experiences of the stakeholders. A follow up investigation into the root causes of differing assessments provides valuable insight on areas needing improvement. Objective analysis can be task-dependent or task-independent. Task-independent analysis reflects the states of the data without the contextual knowledge of the application and can be applied to any data set, regardless of the task in hand. Task-dependent analysis includes the organization’s business rules, company regulations, constraints provided by the database administrator, all are developed in specific application context.

After conducting subjective and objective analysis of the data, a comparative analysis of all the dimensions of the data quality is done. The important dimensions considered for the analysis are accuracy, consistency, timeliness, relevance, accessibility and definition.
Now, since all the dimensions of the quality of data are thoroughly analyzed in the above step, it becomes easier to identify the discrepancies or errors in the data or any scope for the improvement of data quality. The important step here is to optimize the data to attain a high quality data to be finally stored in the repository. Optimization may involve two steps: 1) Creation of quality schema 2) Evaluation.

Here some data quality improvement programs are implemented for preventing any kind of data defects. Quality modeling is done by using various quality dimensions mentioned above. The outcome of the data quality modeling, the quality schema, documents both application data requirements and data quality issues considered important by the design team. From creating quality schemas, few terms have to be understood. These are: data quality parameter, data quality attribute, data quality indicator. A data quality parameter is a qualitative or subjective dimension by which a user evaluates data quality. A data quality indicator is a data dimension that provides objective information about the data like source, creation time etc. A data quality attribute is a collective term including both quality parameters and quality indicators. A data quality parameter value is the value determined for a quality parameter based on underlying quality indicator values. Following are the steps to create a quality schema.

i) First step is to create an application view. The objective is to elicit and document application of the database.

ii) Second step is to add quality parameters to the application view i.e. to determine quality parameters that are subjective in nature. The main aim here is to elicit data quality needs. For each component of the application view, the design team determines those parameters
needed to support data quality requirements. This is called parametric view.

iii) Next step is to determine quality indicators that are objective in nature. The goal here is to operationalize the subjective quality parameters into measurable characteristics. This is called quality view.

iv) Last step is to integrate the quality view. When the design is large and more than one set of application requirements is involved, multiple quality views may result. Thus to avoid redundancy and inconsistency, these views may be consolidated into a single global view so that a variety of data requirements are met. This involves integration of quality indicators.

The final outcome i.e. the quality schema, documents both application data requirements and data quality issues. This is an effort similar to the traditional data modeling, but focusing on the quality aspects of data. The data quality modeling approach provides a foundation for the development of a quality perspective in database design.

Evaluation includes constant monitoring of the data and evaluation of the new quality schemas generated to maintain the quality of data. Finally, after optimization, the quality data is stored in the repository for further use. This data quality framework gives high quality data as output stored in repository for the use in the other phases of the development lifecycle.
4.4.1 Analysis of Data Quality Parameters:
Organizing data quality rules within defined data quality dimensions not only simplifies the specification and measurement of the levels of data quality, it also provides the underlying structure that supports how the expression of data quality expectations can be transformed into a set of actionable assertions that can be measured and reported. Defining data quality rules segregated within the dimensions enables the governance of data quality management. Dimensions of data quality are often categorized according to the contexts in which metrics associated with the business processes are to be measured, such as measuring the quality of data associated with data values, data models, data presentation, and conformance with governance policies. The dimensions associated with data models and data governance require continuous management review and oversight. However, the dimensions associated with data values and data presentation in many cases lend themselves handily to system automation, and are the best ones suited for defining rules used for continuous data quality monitoring.

Data accuracy refers to the degree with which data correctly represents the “real-life” objects that are intended to model. Accuracy of data is analyzed by identifying the correct data values for an attribute in terms of attribute’s dependency and its state in the real world. In many cases, accuracy is measured by how the values agree with an identified source of correct information. There are different sources of correct information: a database of
record, a similar collaborative set of data values from another table, dynamically computed values or perhaps the result of a manual process. An example of an accuracy rule might specify that for healthcare providers, the Registration Status attribute must have a value that is accurate according to the regional accreditation board. If that data is available as a reference data set, and automated process can be put in place to verify the accuracy, but if not, a manual process may be instituted to contact that regional board to verify the accuracy of that attribute.

Consistency refers to the data values in one data set being consistent with values on another data set. Another definition of consistency specifies that two data values drawn from separate data sets must not conflict with each other, although consistency does not necessarily imply correctness. Even more complicated is the notion of consistency with a set of predefined constraints. More formal consistency constraints can be encapsulated as a set of rules that specify consistency relationships between values of attributes, either across a record or message, or along all values of a single attribute. Consistency may be defined within different contexts:

- Between one set of attribute values and another attribute set within the same record (record level consistency)
- Between one set of attribute values and another attribute set in different records (cross-record consistency)
• Between one set of attribute values and the same attribute set within the same record at different points in time (temporal consistency)

• Consistency may also take into account the concept of “reasonableness,” in which some range of acceptability is imposed on the values of a set of attributes.

The data values for an attribute must be consistent when the attribute is duplicated for performance reasons and when it is stored redundantly for any other reason. The duplicate data values of an attribute must be based on the same domain and on same data quality rules.

   Definition specifies that all the data values of an attribute must be defined precisely as required by attribute’s requirements, rules, intended meaning, intended usage etc. Timeliness refers to the time expectation for accessibility and availability of information. Timeliness can be measured as the time between when information is expected and when it is readily available for use. For example, in the financial industry, investment product pricing data is often provided by third-party vendors. As the success of the business depends on accessibility to that pricing data, service levels specifying how quickly the data must be provided can be defined and compliance with those timeliness constraints can be measured.

   Relevance means that the purpose of the application must be clearly stated specifying the aim of collecting the data elements. It must be specified whether the instances of data are present in a format that is consistent with the domain of values as well as consistent with other similar attribute values.
4.5 Empirical validation

Software Engineering should take into account data quality issues in order to prevent, detect and solve problems in the use of systems caused by bad quality of data. Some of these problems are human in nature and others can be addressed using standard software techniques. We can use the GQM (Goal Question Metric) methodology to measure quality using metrics it describes how to incorporate data quality metrics to evaluate, improve and maintain levels of quality in an organization.

GQM is based on the assumption that in order to measure in a useful way, an organization must:

- Specify goals
- Characterize them by means of questions pointing their relevant attributes.
- Give measurements that may answer these questions.

We have chosen this framework because it is a top down approach that provides guidelines to define metrics, without a priori knowledge of the specific measures. Following GQM, we first are able to state which dimensions characterize our notion of data quality.

Then, we can ask questions characterizing each dimension, without giving a precise definition. Finally, we give metrics to answer these questions, giving us a more precise valuation of the quality of our data. A goal is defined for an object, with a purpose, from a perspective, in an environment. A question in GQM tries to characterize the object of measurement with respect to a selected quality issue and determine its quality from the selected viewpoint. A metric in GQM is a set of data associated
with every question in order to answer it in a quantitative way. Data can be objective, if it depends only on the object being measured and not on the viewpoint, or subjective, if it depends on both.

Data collection forms (DCF) is a technique for collecting information from users, in order to compute some of the subjective metrics defined. DCF have questions to be answered by data users. One can choose GQM because of its simplicity, its adequacy to our problem and because it is well known and proven in software engineering applications.

While following GQM to define metrics for data quality, we have identified two main objects to be measured: the set of data and the data model. The set of data is the data actually stored in the database. The data model is how data is structured, from an implementation or a logical point of views. While defining subjective metrics, our perspective is always the user point of view. Each goal is defined in a particular environment and elements of each particular environment can be among the following:

- A fixed data set
- A fixed data model
- A fixed query set
- A fixed set of temporal attributes

In some cases there is no operational ways to measure a specific item or we did not identify one yet or the one identified is too complex to use. In such cases, the technique used must be refined and the one presented here is only a reference for a future implementation.
<table>
<thead>
<tr>
<th>GOAL</th>
<th>QUESTION</th>
<th>METRIC</th>
<th>TECHNIQUE</th>
</tr>
</thead>
</table>
| Object: Set of data  
Purpose: Evaluate  
Quality: Accuracy  
Perspective: User  
Environment:  
- Fixed data set  
- Fixed query set | Do the obtained answers conform the expected answers? | No. of answers that conform the expected answers/ Total no. of answers | “Functional Data Test”  
DCF |
| Object: Set of data  
Purpose: Evaluate  
Quality: Relevance  
Perspective: User  
Environment:  
- Fixed data set  
- Fixed query set | Is there data never queried? | Percentage of tuples never returned as answers | Query set |
| Object: Set of data  
Purpose: Evaluate  
Quality: Timeliness  
Perspective: User  
Environment:  
- Fixed data set  
- Fixed query set  
- T: Set of temporal attributes | How often is data updated?  
Which percentage of data is updated?  
How much data has passed its deadline? | No. of updated operations per unit of time.  
No. of records with attributes in T/ NO. of records in the database  
No. of records with at least one attribute in T not updated/ No. of records with at least one attribute in T | LOG of database activities  
LOG of database activities  
Temporal testing |

Table 4.1: Data Quality Dimension
It is important to notice that we are not trying to be exhaustive in giving all the possible goals, questions and metrics for every data quality dimension. In this work, we want to define data quality metrics and analyze them.

1. **Accuracy**: We propose to measure accuracy of data as the relation between the number of obtained answers that conforms the expected answers and the total number of obtained answers. Several techniques may be developed to characterize expected answers, for example using predicates, or estimating number of elements returned for each query or defining the answers in terms of the instance. This approach is called Functional Data Test.

2. **Relevance**: We propose to measure the percentage of elements from the total number of elements that are never returned as answers. We know the number of tuples involved and we can count the number of tuples returned. How many times we need to query the database to have an accurate estimate. This question can only be answered empirically, after conducting adequate experimentation on real systems. The technique “Query Sets” refers to this notion of running queries and performing a calculation based on the results obtained.

3. **Timeliness**: Here we are interested in measuring how accurate is our data with respect to time. If we know which attributes are time dependent, and how to know if specific values are outdated, we may define a test over the data to estimate the number of records that are outdated. We call this technique temporal test. These are sample metrics for some of the parameters of quality data. Once we have defined data quality metrics, we can use them. We can now take our
relational database, identify the dimensions, choose the appropriate metrics and techniques, apply them and analyze the results. Now following are the steps to follow for evaluation of quality of data:

- Choose the interesting dimensions.
- Define the questions that characterize the dimensions.
- Define the metrics and techniques to answer each question.
- For each metric, define values or ranges presenting good and bad quality data.
- If subjective metrics have been chosen, define appropriate DCF and data collection procedures.
- Apply the techniques to correspondent objects when possible.
- Collect the information using the DCF.
- For each metric, determine if data quality is acceptable or not and take the appropriate corrective actions.
- Store the obtained results.

Following this procedure we can decide whether or not our current data satisfies our quality expectations. This procedure presented above is not a data quality plan. It only deals with measuring the quality of data at certain point that can help in deciding which corrective and preventive actions to implement. Although data modeling represents only a small proportion of the total system’s development effort, its impact on the quality of the final system is probably greater than any other phase.
4.6 Conclusion

Quality control involves applying data quality attributes to datasets to determine their viability. The main purpose of the framework is to provide a structured approach to achieve quality data to be stored in the repository. Quality data stored in the repository minimizes the chances of any failure in the later stages of the development lifecycle. The proposed framework finally converts the dirty data into quality data for the enhancement of the quality of the final product. GQM (Goal Question Metric) methodology is used to measure quality using metrics which describes how to incorporate data quality metrics to evaluate, improve and maintain levels of quality in an organization.
4.7 References


