A Data Quality Practical Approach

Maria del Pilar Angeles
Facultad de Ingeniería
Universidad Nacional Autónoma de México
México, D.F
pilarang@unam.mx

Francisco García-Ugalde
Facultad de Ingeniería
Universidad Nacional Autónoma de México
México, D.F
fgarciau@servidor.unam.mx

Abstract **This paper describes a Data Quality Framework and its application within a Data Quality Project for heterogeneous multi-database environments. The quality assessment of derived data was performed by considering data provenance and conflict resolution functions. A Data Quality Assessment tool provides information regarding the elements of derived non-atomic data values. The assessment and ranking of non-atomic data is possible by the specification of quality properties and priorities from users at any level of experience. Consequently, users are able to make effective decisions by trusting data according to the description of the conflict resolution function that was utilized for fusing data along with the quality properties of data ancestors.**

Keywords- data quality; quality assessment; derived data; cleansing; data integration

I. INTRODUCTION

Multi-database systems provide integrated access to autonomous, distributed, and heterogeneous database systems. The process of data integration requires fusing conflicting data through the use of conflict resolution functions. Therefore, when users retrieve data from disparate data sources, they have no information about the corresponding components and how they were integrated.

This paper is based on previous work regarding the assessment of derived data by considering conflict resolution functions shown in [1], as part of a Data Quality Manager (DQM), which is a prototype to assess data quality and inform users about qualitative characteristics of integrated data, the elements it comes from and how it was fused in order to trust data according to its quality. The aim of this document is to propose a Data Quality Framework (DQF) within a heterogeneous multi-database context, and to present its implementation within a data quality project.

The Data Quality Manager implementation corresponds to the Data Quality Assessment element of the Data Quality Framework, but it could be part of any Data Quality Project life cycle. The DQM provides qualitative information that can be used to determine the current state of data, the business impact of erroneous data and the possible root causes of poor data quality.

We have already identified generic and usable quality criteria to measure and assess data quality of primary data sources, and integrated data at multiple levels of granularity in [2] and [3].

During the data integration process, data administrators require developing conflict resolution functions in order to solve data discrepancies. We enhanced the data lineage algorithm we developed in [4] to trace back the conflict resolution functions in order to provide further quality information to users.

The DQM implementation was based on a Framework for Data Quality Assessment developed in [2][3][4] composed by the identification of quality properties, its corresponding metrics, the process of assessment by data provenance, analysis of data quality, and ranking of data sources.

The implementation of our Data Quality Framework allowed users to determine causes of data quality problems and refine the data quality through data cleansing, monitoring, ensuring data quality during data production process, improvement, etc.

The outline of this paper is organized as follows. We briefly present a data quality overview in Section 2. Related work is described in Section 3. Section 4 describes a framework for conducting data quality projects. Section 5 explains the Data Quality Assessment Process as an element of the previous framework. Section 6 presents a practical approach by following the Data Quality Framework proposed. The last section concludes with relevant and novel features of the research and outlines future work.

II. DATA QUALITY OVERVIEW

This section presents a generic overview of data quality, starting from commonly causes of data quality degradation, the impact of low information quality, the cost of data cleansing and our perspective for addressing data quality issues.

A. Data Quality Definitions

The subjective nature of the term Data Quality (DQ) has allowed the existence of general definitions such as “fitness for use” in [18], which implies that quality depends on customer requirements.

The definition established by Redman et al in [33], suggests that data quality can be obtained by comparing two data sources “A datum or collection of data X is of higher or (better) quality than a datum or collection of data Y if X meets customer needs better than Y”.

Another definition is “The distance between data views presented by an Information System and the same data in the Real World” in [34], which means that quality depends on the capacity of an information system to represent facts of the real world. Consequently, careful handling of data shall be done during its life cycle.

Recently, data quality has been defined as “the capability of data to be used effectively economically and rapidly to inform and evaluate decisions” [32].

However, these definitions are not very useful when data quality requires to be evaluated. Consequently, data quality rather than being defined has been characterized by multi attributes or dimensions according to specific application domains, types of assessment or customer requirements for instance, that shall be accomplished in order to be suitable for use.

As the determination of data quality is by comparing its corresponding attributes [30], [33], this collection of attributes must be defined, classified, measured and compared in order to determine an overall quality.

However, quality properties are often of a quantitative or qualitative nature, the former being easy to measure, but not the latter, which are subject to personal expertise.

Furthermore, “..What may be considered good quality information in one case (for a specific application or user) may not be sufficient in another case” [31], which means that even defining the quality attributes, and identifying their corresponding measures and assessment methods, the overall quality will depend on the specific priorities given by data consumers.

From our point of view, data quality is a multidisciplinary area, which involves management, statistics and computer sciences. We consider data quality not as the end but the means for making informed decisions.

The relevant data quality properties, its priorities, and the level of expected data quality depend not only on the data consumer experience, but also on the underlying type of information system.

B. Causes of Data Quality Degradation

Data are being deteriorated by processes bringing data from outside; incoming data may be incorrect and simply migrate from one place to another such as data conversion, batch feeds or real-time interfaces.

High volumes of data degradation are also introduced by wrong designed Extraction, Transformation and Loading (ETL) processes.

Data errors arise due to processes that manipulate the already existing data in the database such as periodic system updates with improper integrity constraints implementation.

Data are impacted by changes that for any reason are not captured, and wrong designed processes changing data from within.

There are some other processes that cause accurate data to become inaccurate because time related data changes over time and those changes are not reflected in the system.

C. Impact of low Data Quality

Poor data quality might affect every sector of industry such as finance [24], where an error attributed to the New York Stock Exchange resulted in several inaccurate stock quotes being picked up and posted at a number of news and investment organizations; within the medicine sector [25] a woman underwent a double mastectomy after being advised that she had breast cancer. After the surgery she was informed that the laboratory had switched her lab results with another patient and that she never had cancer; in the Academy sector [26], a University emailed 1,700 applicants to announce their acceptance into the class of 2007.

Unfortunately, 550 of the applicants received the letter in error they had already received rejection notices. The error was attributed to a "systems coding error". However, there is a possibility that the acceptance status of the 550 students was updated by mistake after sending the rejection notice.

Users should be aware of the quality of data they are accessing along with the cause of its degradation. For instance, identifying which data are time-related becoming obsolete as time goes by; quality of data might be application-related due to missing or wrong designed constraints; integrated data have been passing from one application to other or from one data source to other through data fusion or transformation; etc.

D. The cost of data cleansing

According to T. Anderson in [22] the cost of poor data quality is the sum of the cost to prevent errors and the cost to correct them and the cost to make them good for the customer. Pragmatically speaking, the cost of poor data quality extends far beyond the cost to fix it.

The Data Warehousing Institute estimates that data quality problems currently cost U.S. businesses over $600 billion annually. Errors are very hard to repair, especially when systems extend far across the enterprise, and the final impact is very unpredictable.

The first reaction at cleansing personal details would be determining if a single record is "correct" by calling the corresponding telephone number, and ask the person whose name shares the record with the telephone number. If the person comes to the phone, ask if all the values are accurate, and correct those that are not. If there is no one there by that name, the record is incorrect. The next step in data cleansing requires additional information, and if none is available, then the algorithm ends. This is a simple and accurate algorithm. However, commonly is neither cost effective nor scalable because depends on the number of records, staff members and telephones. Automated solutions may be more scalable, more costly, less accurate, more complex, require more expertise, etc.

D. Loshin in [23] states that the cost of cleansing data requires to analyze which is the size of data in number of records and columns, which would be the criteria in order to define data “clean”, if the relevant data are in a single table or scattered across many data sources, and the number and level of experience of customers. The level of reasonable

effort for spending on data cleansing must be less than the value of the accrued business benefits, and this provides an upper limit to what could be budgeted for the process.

The subject of this work is concerned with the specification and implementation of a Data Quality Framework for the identification, measurement, and assessment of data quality of derived data, and data sources at any level of granularity to provide ranking of data sources based on the user specified context. After the data quality diagnosis, feasible data cleansing within a monitoring process shall be possible, according to the business requirements and the level of data quality pre-established.

As low data quality impacts on business, and the process of assessing and cleansing data is not trivial, important research has been done recently. Section 3 presents recent developed frameworks for data quality projects, how previous approaches have dealt with data inconsistencies during data integration and how the assessment of data quality has been addressed in particular.

III. RELATED WORK

A. *Data Quality Frameworks*

The Massachusetts Institute of Technology (MIT) and the Cambridge Research Group, among other institutions, have co-founded the MIT Total Data Quality Management program (TDQM) [28]. The aim of TDQM is to create a theory of data quality based on disciplines such as Computer Science, Statistics, and the Total Quality Management field, and is focused on the definition and measurement of data quality, the identification and analysis of data quality impact, and the redesign of business practices and implementation of new technologies to improve information quality.

In Total Data Quality Management the concepts, principles and procedures are presented as a methodology, which defines the following continuous life cycle: define, measure, analyze and improve data as essential activities to ensure high quality, managing information as a product.

There are more detailed approaches such as the one proposed by D. McGilvray in [19] who proposes ten steps for executing data quality projects. The main objective of data quality projects is to achieve a reasonable level of quality that brings success to companies. Therefore, the project starts by the identification of business needs. After an analysis of information environment it is possible to identify the essential data and information corresponding to those business needs. During the assessment of data quality as a third step, the design and implementation of an assessment plan for relevant data is a key in order to evaluate the current state of data. As the following step, the assessment results should be analyzed and documented to determine the business impact of poor quality of relevant data. Step 5 corresponds to the identification of root causes of data issues and initial recommendations. The sixth step is the development of improvement plans. The implementation of the improvement plan will correct current data errors, and prevent future data errors (steps 7 and 8). Step 9 is concerned with monitoring if the improvement plan is providing the expected results through implementing controls allows finishing the cycle and starts it over again. However, communicating actions and results along the whole process is a key for success.

David Loshin in [23] identifies 17 steps required for data quality management.

The first step is to recognize the problem, if there are some issues that are affecting the business then there is evidence that poor data quality is having an impact in order to determine whether such evidence points to any particular problems with data quality or not.

The second step is to obtain the management support by showing them how the business is affected or can be affected by poor data quality, and at the same time their support and enforcement of a data ownership policy document for guiding the roles associated with information and the responsibilities accorded those roles. The third step is to spread the word by a data quality education program. The forth step is mapping the information chain in order to understand how information flows through the organization, which is a chart that describes processing stages and the channels of communication between them. Data Quality Scorecard is the fifth step, which is concerned with the overall cost associated with low data quality and can be used as a tool to help determine where the best opportunities are for improvement. The sixth step is to perform a current state assessment to obtain information regarding the causes of data quality issues, this step requires identifying which data quality dimensions will be relevant and identifying points within the information chain and for measuring for understanding the scope and magnitude of data quality problems. The seventh step is requirements assessment, which is in charge of problems prioritization, assigning responsibility and creating data quality requirements for identifying the location in the information chain where the requirement is applied, a description of the measurement rule, the minimum threshold for acceptance among others. Step eight is choosing the first problem to address. Therefore such problem should have a noticeable impact in order to ensure the continued operation of the data quality program. The next step is regarding to build the team to solve the problem. The step ten is related to the identification of proper data quality tools in order to support data cleansing, data standardization, etc. The eleventh step is to define a metadata model to store enterprise reference data. The next step is the definition of data quality rules. Step 13 is related to the Archaeology/Data mining to look for data domains, mappings, and data quality rules that are embedded in data. The fourteenth step is for managing suppliers, a corresponding program will be required to impose requirements on external data suppliers to specify the rules that are being asserted about expectation of the data along with penalties for nonconformance. Step fifteen is concerned with actually executing the data improvement. The next step is related to measuring the improvement in order to demonstrate success at improving data quality by performing the same measurements from current state assessment. The last step is to build on each success. Each small success

should be used as leverage with the senior level sponsors to gain access to bigger and better problems.

For the above mentioned frameworks we can say that there is no consideration of data quality within heterogeneous multi-database environments or enterprise information integration contexts, where data come from a number of data sources facing semantic and syntactic heterogeneities and derived data are product of integration processes.

B. Previous Approaches of Data Quality Assessment

A particularly important element within data quality projects is the data quality assessment. Therefore, this section presents previous approaches of data quality assessment.

Gertz developed some data integration techniques in [9], based on data quality aspects within an object oriented data model, and data quality information stored in a metadata. In the case of data conflicts between semantically equivalent objects, the object with the best data quality must be chosen. However, the quality goals specification limits the possibility of more combinations of priorities from the user, because they are not given in weights or percentages, just the "the most accurate" or "the most up to date". Consequently, not just one or two combinations of quality priorities will satisfy users. One result might be good enough for one user under a specific situation, but of poor quality for other.

The project Multiplex directed by Motro and Rakov [11] was based on accuracy and completeness as quality criteria. A voting scheme, using probabilistic arguments, identifies the best set of records to provide a set of ranked tuples to the user, but no further information about their associated quality. Therefore, users are neither able to establish their quality preferences or priorities nor to take part in the resolution process.

The project Quality-driven Integration of Heterogeneous Information Systems was developed by F. Naumann in [12]. The aim was to identify and to rank high quality plans, which produce high quality results. There is a classification of specific quality criteria according to the level of granularity (in this approach data sources, queries and attributes). However, there is no further specification of how to assess quality at different levels of granularity. Data sources are ranked using the DEA method. Therefore, there is no consideration of user priorities for this process. Besides, subjective criteria are used for discarding data sources such as reputation and understandability.

The aim of the Data Quality in Cooperative Information Systems (DaQuinCIS) project [15] was to define an integrated framework to improve data quality in cooperative environments. Such a framework started from the Total Data Quality Management methodology which was extended to suit the cooperative information systems requirements, and supporting data quality monitoring and improvement. The use of a metadata was required to store the quality score, the meaning of the quality value, and how the measurements were carried out. This approach takes into account the specification of data granularity as the combination of elementary data items that are subject to quality metrics. There is also a difference between computing the quality of aggregated data and computing an aggregate indicator over a set of items. However, the measurement is not only subjective but also different methods are utilized to measure quality, yielding different results. Furthermore, data derived from multiple data sources is not considered.

A Generic Framework of Information Quality was developed by Burgess in [8] with around 60 information quality properties classified hierarchically according to time, utility and cost. Nevertheless, this approach was focused on information search not on measurement and assessment of quality at data value level.

A. Maydanchik proposes a methodology in [10] for data quality assessment to identify all data errors. In order to do so the project shall involve business users, IT specialists, data quality experts to a project team.

The data quality project plan which in turn consists of four steps a) planning for identifying project scope and objectives; b) preparation for gathering relevant data and metadata; c) implementation concerned with designing the data quality rules, and d) fine tuning, where data experts validate error reports in order to enhanced data quality rules.

It is desirable to monitor data quality on an ongoing basis, in order to see data quality trends, identify new data problems, and check the progress of data quality improvements initiatives.

Within the implementation phase of the data quality assessment, data quality rules can be executed automatically in order to find such data errors, the first step is design, cataloguing, and coding data quality rules. The second step for data quality assessment is the process to identify and eliminate rule imperfections by manual verification of the sample data by data experts, the analysis of sample verification findings and the search for patterns; and to enhance the rules to eliminate as many flaws as possible; and repeat until obtain the expected results. The third step is concerned with storing information about all the identified data errors in an error catalogue in order to identify and analyze error patterns and enhance data quality rules and identify how to correct data errors. The next step is to identify and tabulate aggregate data quality scores. Accurate data quality scores help to translate data quality assessment results into cost of bad data, return of investment from data quality improvement and expectations from the projects. The fifth step is to identify the content and functionality of the data quality metadata warehouse which contains tools for organization and analysis of all meta data relevant to or produced by the data quality initiatives, contains aggregate meta data, rule metadata atomic metadata and general meta data. The last step is the recurrent data quality assessment for an ongoing data quality monitoring.

When data quality assessment is done on a regular basis and if the target database contains large volumes of data, running the rules directly against the production database might be a better solution than replicating it to the staging are data quality assessment is technically and technologically challenging, the best solution depends on the dynamics of the data.

C. Important Remarks

Within the previous approaches, there is no consideration of derived data. Data in all of these approaches has been considered as a product of a primary source. However, due to the explosion of information over the last decade, we cannot assume that any data source is necessarily the point of origin of the data users require. Hence, the fundamental presumption of current data management practice, the "Presumption of Primary Authorship" must be challenged.

Users should be provided with information regarding data as an atomic value, or if it is composed data, what the atomic values were and the quality generated from. This challenges the "Presumption of Atomicity".

The assessment methodologies presented until now do not consider data provenance as part of root causes of poor data quality. Cleansing derived data with no consideration of data fusion or conflict resolution functions is not an effective solution for assessing data within heterogeneous multi-database environments.

The next three sections present our Data Quality Framework, a Data Quality Manager prototype as an implementation of the Assessment of data quality and a practical application of both of them.

IV. THE DATA QUALITY FRAMEWORK

We propose a framework for Data Quality composed by seven phases. The first phase is the identification of data quality problems by their impact on the business, considering data quality experts, data domain experts and end users of any level of experience. The second phase is the identification of relevant data that has direct impact on the business for an estimation of poor data quality cost. The third phase is the creation, identification, or modification of relevant business rules. Commonly, some business rules have not been considered during the application development or they might exist but require enhancement. The fourth phase is the Data Quality Assessment Model for the analysis of data quality at different levels of granularity considering data provenance. The analysis of data quality assessment enables expert users to establish different priorities to quality properties and different levels of granularity for assessment. The fifth phase corresponds to the determination of the business impact through data quality comparison. The difference between the expected data quality and the actual data quality scores will establish the feasibility of the data quality project for cleansing and continuous assessment and the business impact in terms of operational efficiency, or increased revenue, money saved, etc. The sixth phase corresponds to the cleansing of data by enforcing the business rules, data standardization, and data matching. The last phase corresponds to monitoring data quality and executing the assessment phase on regular basis.

The proposed Data Quality Framework is simple enough to be suitable to any size of data quality project, and at the same time its data quality assessment element considers data provenance, data fusion and conflict resolution functions for comparing and resolving extensional inconsistencies within virtual or materialized data integration.

Fig. 1 shows the elements of the Data Quality Framework.

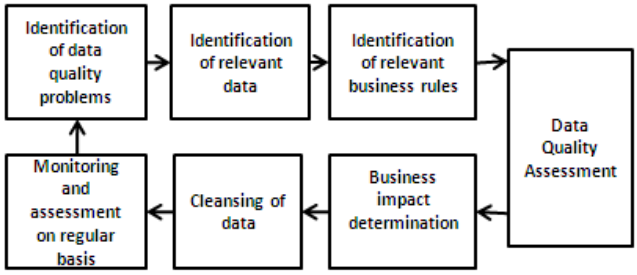

Figure. 1 The Data Quality Framework

In Section 5 we explain in more detail the Data Quality Assessment element of the Data Quality Framework.

V. THE DATA QUALITY ASSESSMENT PROCESS AND ITS IMPLEMENTATION

A. The Data Quality Assessment Process

The first step corresponds to the identification of useful data quality properties for the measurement, and assessment of data quality of derived data, and data sources at multiple levels of granularity, to provide data consumers with qualitative information directly related to the relevant data and business rules identified during the first three steps of the Data Quality Framework. The outcome of this step is called a Data Quality Reference Model, which contains an objective and effectively set of quality criteria to provide an unbiased measure of quality to users at any level of experience they might have. A generic set of data quality properties has been classified and summarized according to different user perspectives such as internal and external focuses or representation, value, and context in [3].

As we are addressing any level of experience user, the aim of the second step is to discuss which existing metrics are suitable for an unbiased, and user independent estimation of data quality scores to provide a more objective quality metadata. The development of new metrics is not relevant for this research, but to extend existing metrics to assess data quality at different levels of granularity. Therefore, the outcome of this step is called a Measurement Model [4], which assembles and extends the already existing data quality metrics [6] [11] [14] for the measurement at database, relation, tuple, and attribute levels of granularity.

The third step is concerned with the identification of methods required to represent, to interpret, and to assess data quality indicators. The assessment methods utilised should provide meaningful and useful scores. Therefore, objective criteria, and process criteria should be included in the Assessment Model which are user independent, rather than subjective criteria, which can only be determined by individual users based on their experience and background.

The Assessment Model provides a mechanism for tracking data lineage for the assessment of quality of derived

data. Previous approaches work from the presumption of primary authorship and the presumption of atomicity. Therefore, the utilization of data lineage as a mechanism for assessing data sources at different levels of granularity, challenging the presumptions of primary authorship and atomicity are novel.

The forth step corresponds to the estimation of the quality scores of primary data sources, which will be stored in a Quality Metadata.

The fifth step is the assessment of derived data, which requires the definition and population of a provenance metadata. The assessment is based on the quality scores of their corresponding ancestors, and the computed scores are also stored in a Quality Metadata.

The sixth step presents two options for the analysis of data quality, according to user requirements and business information stored in the organizational metadata

a) The selection of the best data sources before the query execution on the bases of its quality scores. Therefore, the consideration of data quality scores helps the query planning by finding the best combination of data sources for the execution plan.

b) The comparison of data quality aggregated scores corresponding to different query plans for the same business question.

The seventh step is the ranking of data sources, where the data quality scores previously stored in the metadata are used as a whole with their corresponding priorities stated by the user. Fig. 2 shows the Data Quality Assessment Process.

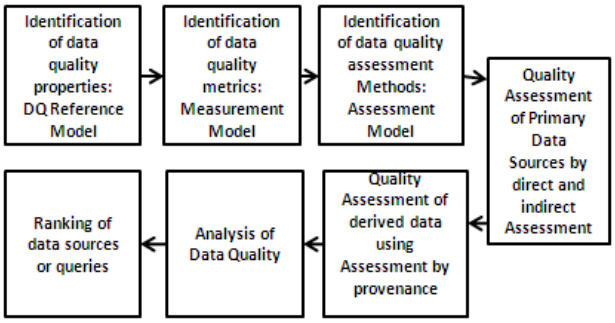

Figure. 2 The Data Quality Assessment

B. *The Data Quality Manager*

The process of assessment of data quality has been developed within the Data Quality Manager through the implementation of the already mentioned models and a quality metadata, a provenance metadata, and an organizational metadata.

The Quality Metadata is a repository to contain the quality scores per each data source obtained during the data quality assessment process, and reloaded to assess at lower levels of granularity.

The provenance metadata is a repository to contain ancestors' information for the tracking of provenance of the participant data sources.

The Organizational Metadata is a repository to contain the information required to map from the global schema to the local schema in order to resolve intensional inconsistencies (semantic differences) within the multidatabase environment, for further information regarding intensional inconsistencies, please refer to [29]. The organizational metadata will also contain business rules and relevant information for business understanding.

The DQM is part of a diagnostic pre-process for data cleansing, or after data cleansing to evaluate data quality improvement.

The DQM represents the data quality assessment component of the Data Quality Framework. The DQM is designed to utilise data quality measures to provide qualitative information. As we have explained, such information could be further used within the data integration processes.

The Data Quality Manager (DQM) is a system designed specifically for centralized processing of multiple interfaces between multiple databases; it allows maintaining detailed data provenance and data quality metadata for future reference.

The architecture of the DQM is shown in Fig. 3.

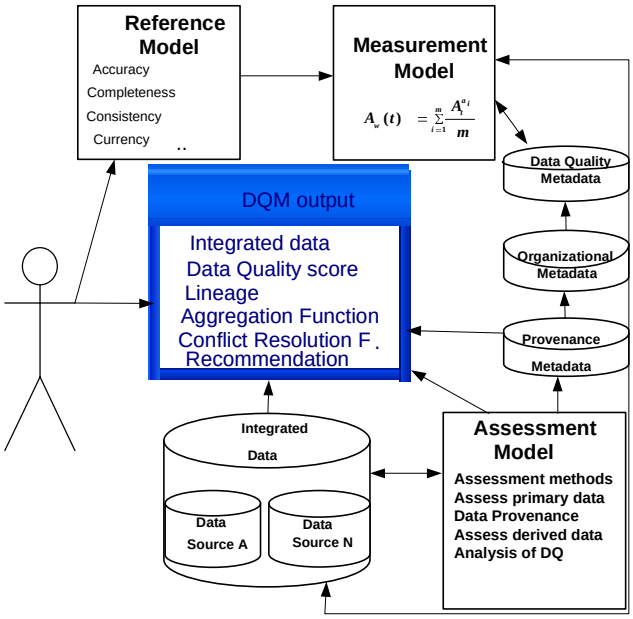

Figure. 3 Components and outcomes of the Data Quality Manager

The DQM provides qualitative information to any level of experience users to extend the scope and range of information available relative to the integrated data within the quality properties and priorities they state.

The DQM in the case of naive users provides an appropriate combination of scaling with ranking methods. In the case of expert users, they will have the ability to define scaling, ranking, quality properties and the priorities for a higher level of analysis. Users should be able to select the quality priorities. The specification of Multi-attribute

Decision and Scaling Criteria methods is also possible by experienced users.

The functionality and capability of the Data Quality Manager prototype has been validated against the specifications based on a testing plan detailed in [5]. We have also demonstrated that the DQM provides appropriate scores according with the expected outcomes based on the actual quality of data and information relative to the conflict resolution function utilized during the integration process.

VI. A PRACTICAL APPROACH FOR DATA QUALITY

This section is aimed to explain the implementation of our Data Quality Framework within a data quality project, and is intentionally more focused on the results presented by the Data Quality Manager for the assessment of derived data.

As the Data Quality Manager (DQM) tool is aimed to work within a multi-database environment, the conducted tests are based on a TPC Benchmark™H (TPC-H) [17].

TPC-H is a decision support benchmark. It consists of a suite of business oriented ad-hoc queries and concurrent data modifications. The queries and the data populating the database have been chosen to have broad industry-wide relevance. This benchmark illustrates decision support systems that examine large volumes of data, execute queries with a high degree of complexity, and give answers to critical business questions. The names of the implemented databases are TPCH, TPCHA and TPCHB.

A. Data Quality Issues

Users are unable to make informed decisions because they are retrieving different results for the same query. The problem is also called extensional inconsistencies, and it refers to the data value differences between the participating data sources during data integration. The cause of extensional inconsistencies is that queries can be executed on different data sources semantically equal. For further information regarding extensional inconsistencies, please refer to [29].

In order to determine business needs we require a list of the most important business queries, after the identification of such queries, executive users prioritize the queries according to their impact on business. Focusing in what is relevant and appropriate is critical for finding relevant data.

At this point the analyzed data, processes, technology, and people allows a better understand of all these components and their impact on information quality.

B. Relevant Data

The identification of relevant data affecting business questions was performed by the identification of such conflictive business queries. This paper will present just three queries corresponding to one possible option. However, similar analysis shall be done for each semantically equal business question executed on different data sources. The important business questions identified are Customer Distribution, Product Type Profit Measure, and National Market Share.

The business query called Customer Distribution seeks relationships between customers and the size of their orders. It determines the distribution of customers by the number of orders they have made, including customers who have no record of orders, past or present. It counts and reports how many customers have no orders, how many have 1, 2, 3, etc.

A check is made to ensure that the orders counted do not fall into one of several special categories of orders. Special categories are identified in the order comment column by looking for a particular pattern. Please refer to [17] for further detail. The query `Cus_Distribution` has been integrated by the outer join of two tables `CUSTOMER` and `ORDERS`, and the relevant data columns are `C_CUSTKEY`, `O_ORDERKEY` and `O_COMMENT`. The SQL Text of the `Cus_Distribution` query is presented as follows.

```
SELECT C_CUSTKEY AS C_COUNT,
       COUNT (O_ORDERKEY) AS HOW_MANY
FROM
CUSTOMER LEFT OUTER JOIN ORDERS ON
         C_CUSTKEY = O_CUSTKEY
         AND O_COMMENT NOT LIKE
          '%UNUSUAL%DEPOSITS%'
GROUP BY C_CUSTKEY
```

The Product Type Profit Measure business question finds for each nation and each year, the profit for all parts ordered in that year which contain a specific substring in their part names and which were filled by the Supplier in that nation.
The corresponding instantiation of the business question is called `pt_profit` and it contains relevant data such as `PART.P_PARTKEY, PART.P_NAME, SUPPLIER.S_SUPKEY, LINEITEM.L_SUPPKEY, L_PARTKEY, L_ORDERKEY, PARTSUPP.ORDERS and NATION.NATIONKEY.`
The SQL text code of the query `pt_profit` is presented below.

```
SELECT      N_NAME AS NATION,
 EXTRACT(YEAR FROM O_ORDERDATE) AS YEAR,
 L_EXTENDEDPRICE * (1 - L_DISCOUNT) -
PS_SUPPLYCOST * L_QUANTITY AS AMOUNT
FROM PART, SUPPLIER, LINEITEM, PARTSUPP,
ORDERS, NATION
WHERE S_SUPPKEY = L_SUPPKEY
AND PS_SUPPKEY = L_SUPPKEY
AND PS_PARTKEY = L_PARTKEY
AND P_PARTKEY = L_PARTKEY
AND O_ORDERKEY = L_ORDERKEY
AND S_NATIONKEY = N_NATIONKEY
AND P_NAME LIKE '%MINT%'
```

The National Market Share business question shows the market share for a given Nation within a given Region. It is defined as the fraction of the revenue from the products of a specified type in that Region that was supplied by Suppliers from the given Nation. The query determines this for two years. The relevant data are PART.P_PARTKEY, PART.P_TYPE, SUPPLIER.S_SUPPKEY, LINEITEM.L_PARTKEY, LINEITEM.L_SUPKEY, ORDERS.O_ORDERKEY, ORDERS.O_ORDERDATE, ORDERS.O_CUSTKEY, CUSTOMER. CUSTKEY, NATION.N_NATIONKEY AND REGION.R_NAME. The SQL text code for the corresponding query C_Market_Share is shown as follows.

```
SELECT      EXTRACT(YEAR FROM
O_ORDERDATE) AS O_YEAR,L_EXTENDEDPRICE *
(1 - L_DISCOUNT) AS VOLUME,
      N2.N_NAME AS NATION
FROM  PART,SUPPLIER,    LINEITEM,ORDERS,
      CUSTOMER,NATION N1,NATION N2,
      REGION
WHERE P_PARTKEY = L_PARTKEY
      AND S_SUPPKEY = L_SUPPKEY
      AND L_ORDERKEY = O_ORDERKEY
      AND O_CUSTKEY = C_CUSTKEY
AND C_NATIONKEY = N1.N_NATIONKEY
AND N1.N_REGIONKEY = R_REGIONKEY
AND R_NAME = 'AMERICA'
AND S_NATIONKEY = N2.N_NATIONKEY
AND O_ORDERDATE BETWEEN DATE 'date' AND
DATE 'date'
AND P_TYPE = 'LARGE PLATED NICKEL'
```

C. Business Rules

Once obtained the relevant data, the next step is to identify their corresponding business rules. They shall be enforced within the relevant data in order to detect data errors and correct them.

In the case of the business questions National Market Share and Product Type Profit, the corresponding trigger that inserts a new tuple into REGION whenever a new tuple is inserted into NATION, and the trigger that inserts a new tuple into NATION whenever a new tuple is inserted into REGION were enforced.

D. Assessment of Data Quality

Data quality assessment tells us about existing data problems and their impact on various business processes. When done recurrently, it also shows data quality trends.

The elements of the Data Quality Assessment Process produced during the practical approach will be explained in detail in the following subsections.

Data Quality Properties

Considering the relevant data and business rules, the identification of which quality properties are relevant for assessment is required. However, according with Lee and Strong in [21], the responses from data collectors, data custodian, and data consumers within the data production process determine data quality because of their knowledge.

Data collectors are associated to the quality properties accuracy, accessibility, relevance, completeness and timeliness. Data Consumers are more interested in the accuracy of and uniqueness of data in order to use them for making decisions. Their research was oriented to determine the causes of poor data quality during the data life cycle and how the knowledge of the participant users reflects the quality of data. Therefore, the identification of the relevant quality properties is also directly related to the knowledge of the data according to the experience of users.

In this Data Warehouse context, the quality criteria vary depending on the data source, for example for look up tables there will be low volatility, but accessibility is important. In case of Fact tables, as they provide the sales detail, accuracy, uniqueness, and completeness are important because they would be directly reflected in the generation of aggregate data in the summarize tables.

The integration of data sources that contain duplicated tuples could result in extensional inconsistencies. Therefore, the quality property called uniqueness should be included as a relevant quality criterion for the assessment of data quality to help in the resolution of extensional inconsistencies.

A Generic Data Quality Reference Model has been discussed in [2]; it is suitable to any application domain and supports the full range from the internal focus to the external focus.

After an analysis of the proper quality properties according to the expert users, the type of information system and the relevant data identified, we have reduced the number of quality properties from the Generic Reference Model to those corresponding to the data value level in order to obtain results fast for a rapid return on investment (ROI). Therefore, the quality properties or data quality dimensions used for this assessment are accuracy, completeness, consistency, currency, timeliness, uniqueness and volatility.

Data Quality Metrics

Designing the right metrics is the most challenging task during the process of data quality assessment. However, the challenge is to design them and make sure that they indeed identify all or most errors, avoiding metrics that reflect the same error in many different ways and produce comprehensive error reports.

Once identified the relevant quality properties the next step is to assess them through the measurement model, and synthesize the results from the assessments.

The Measurement Model corresponds to the metrics for data quality properties identified in previous step, and to the business rules already identified.

Accuracy is the measure of the degree of agreement between a data value or collection of data values and a source agreed to be correct. [27].

Completeness is the extent to which data is not missing [14] and is divided by two quality dimensions: coverage and density in [12].

Consistency is the extent to which the values are the same for overlapping entities and attributes. Data are consistent with respect to a set of constraints if they satisfy all constraints in the set [11]. Often referred as integrity constraints state the proper relationships among different data elements" [14]

The following SQL text code shows the measurement of referential integrity between `LINEITEM` and `PART` and `LINEITEM` and `SUPPLIER` as one of the requirements for the query `C_Market_Share`. Finally, the data quality score is stored in the quality metadata through an insert-select sentence.

```
/* lineitem with part */
begin
declare @part real
declare @supplier real
select @part=
case
when convert(real,count(L_PARTKEY))=0
then 1
when  convert(real,count(L_PARTKEY))> 0
then convert(real,count(L_PARTKEY))
end
from lineitem
where not exists (select * from part
                  where
P_PARTKEY=TPCHA.dbo.lineitem.L_PARTKEY)
/* lineitem with supplier*/
select @supplier=
case
when convert(real,count(L_SUPPKEY))= 0
then 1
when  convert(real,count(L_SUPPKEY))> 0
then convert(real,count(L_SUPPKEY))
end
from lineitem
where not exists (select * from supplier
                  where
S_SUPPKEY=TPCHA.dbo.lineitem.L_SUPPKEY)
select
object_id,12,@part,@supplier,mrows,"1-
inconsistent/total rows"
from Metadata.dbo.numrows
where object="TPCHA.dbo.lineitem"
group by object_id,mrows
insert Metadata.dbo.Scores
select object_id,12,1-
((@part/convert(real,mrows))*(@supplier/
convert(real,mrows))),"1-
inconsistent/total rows"
from Metadata.dbo.numrows
where object="TPCHA.dbo.lineitem"
group by object_id,mrows
end
```

Currency is the time interval between the latest update of a data value and the time it is used [11].

Timeliness is the extent to which the age of data is appropriate for the task at hand [6], and is computed in terms of currency and volatility. Timeliness has also been presented as context related dimension.

Uniqueness is the extent to where an entity from the real world is represented once. The below SQL code computes the ratio between the number of non-unique rows and the total number of rows in the nation relation.

```
insert into Scores select 301,2,
convert(real,count(distinct
N_NATIONKEY))/convert(real,count(*))
,"non-duplicated/total values"
from TPCHA.dbo.nation
```

Volatility is the interval of time where data remains valid on the system and is related to the update frequency [6].

Assessment Methods

Most metrics proposed until now are just at one level of granularity. Particularly, completeness has been deeply approached in [12] and [20] with the coverage and density concepts in the former, and at different levels of granularity in the latter. However, we have taken into account not only attribute, and relation levels of granularity following the completeness example given in [20] but also the database level. We are considering the cardinality of a relation when measuring its quality. Therefore, the estimation of quality at database level is taken from the average score of its relations as a representative aggregation function.

The strictness of data quality assessment is a weak or strong characterization depending on evaluating the quality property as a percentage or as a Boolean function respectively [20]. The strong characterization of the quality metrics is useful in applications in which it is not possible to admit errors at the corresponding level of granularity. For instance, in the case of accuracy at tuple level, it would be useful if and only if all the instances of its attributes are accurate. The remainder of this section presents 16 formulas corresponding to the relevant quality properties already identified, for further information regarding such formulas please refer to [5].

In this practical approach the assessment of data quality considers the weak strictness to make possible the comparison of data sources for a number of data quality properties. However, as there might be alternatives where strictness could depend on the level of quality required, according to specific applications we present both characterizations.

During the assessment of data quality, we identified a granularity-based assessment classification according to the level of granularity in which the quality assessment is done a) Direct assessment; b) Indirect Assessment; and Assessment by provenance.

Assessment of primary data sources

Direct assessment is the process of assessment that relates directly to the level of granularity. For instance, the uniqueness dimension $U(t_j)$, which relates at the tuple level.

$U(t_j) = 1$ *if tuple j is represented once in a relation*

$U(t_j) = 0$ *otherwise*.....(1)

Accuracy at value level corresponds to the presence of the correct data value within a specific attribute a_i in a tuple t, and is set by the following notation:

$A_t^{a_i} = 1$ *if value in* a_i *is correct*

$A_t^{a_i} = 0$ *otherwise* ....(2)

b) Indirect assessment: The score is calculated based on other scores at other levels of granularity of the same source. For example,

Weak accuracy at attribute level $A_w(a_i)$ is the number of tuples with correct values for a specific attribute a_i divided by the cardinality of the relation S.

$$A_w(a_i) = \sum_{j=1}^{n} \frac{A_{t_j}^{a_i}}{n} \quad \text{..... (3)}$$

The accuracy of an attribute $A_s(a_i)$ is strong if all instances t_j of the attribute a_i in the relation S are correct.

$A_s(a_i) = 1$ *if* $A_{t_j}^{a_i} = 1 \quad \forall j \in [1..n]$

$A_s(a_i) = 0$ *otherwise* ..... (4)

Weak relation accuracy $A_w(S)$ is the number of tuples where every attribute is correct divided by the total number of rows.

$$A_w(S) = \sum_{j=1}^{n} \frac{A_s(t_j)}{n} \quad \text{..... (5)}$$

Strong relation accuracy $A_s(S)$ is that when all the tuples contain correct values in every attribute, or when a relation contains strong tuple accuracy, and strong attribute accuracy.

$A_s(S) = 1$ *if* $A_s(t_j) = 1 \quad , \quad \forall j \in [1..n]$

$A_s(S) = 0$ *otherwise* ..... (6)

Then accuracy at database level $A(D)$ can be derived from the average of all accuracy scores at relation level.

$$A(D) = \frac{\sum_{k=1}^{w} A(S_k)}{w} \quad \text{..... (7)}$$

Consistency at the relation level depends on consistency at the row level. The weak consistency at the relation level $Cn_w(S)$ is the percentage of tuples t_j with all instances of the attributes consistent.

$$Cn_w(S) = \sum_{j=1}^{n} Cn_s(t_j) \Big/ n \quad \text{..... (8)}$$

Direct and indirect assessments are performed on the ancestors' data sources.

In the case of the data quality assessment cannot be computed directly for performance issues then if it is possible, the assessment by provenance is applied.

The following subsection is concerned with the quality estimation of integrated data as part of the Assessment of Data Quality.

Assessment of derived data

Assessment by provenance is the process of assessment when the score of an object is computed based on the quality indicators of its ancestors.

In order to explain how quality of derived data might be assessed through data provenance, consider a query or a source *s* that comes from *n* ancestors α_j.

For instance, accuracy of derived data $A(s)$ is computed by the average of the scores of its ancestors.

$$A(s) = \frac{\sum_{j=1}^{n} A(\alpha_j)}{n} \quad \text{..... (9)}$$

Completeness of derived data $C(s)$ is determined by the average value of the completeness of its ancestors.

$$C(s) = \frac{\sum_{j=1}^{n} C(\alpha_j)}{n} \quad \text{..... (10)}$$

Consistency of derived data $Cn(s)$ is determined by the average of the consistency of its ancestors. The consistency of its foreign keys is checked with its corresponding primary keys in each ancestor.

$$Cn(s) = \frac{\sum_{j=1}^{n} Cn(\alpha_j)}{n} \quad \text{..... (11)}$$

The currency of derived data $Cu(s)$ is the greatest value of the corresponding currency measures from the different ancestors.

$$Cu(s) = \max\left(Cu(\alpha_j)\right) \quad , \quad \forall j \quad \in [1 \cdots n] \quad \text{..... (12)}$$

Volatility is the update frequency. When there are a number of data sources with different volatilities, the volatility of derived data $Vo(s)$ is the greatest value of the corresponding volatility measure from its different ancestors.

$$Vo(s) = \max\left(Vo(\alpha_j)\right) \quad , \quad \forall j \quad \in [1 \cdots n] \quad \text{..... (13)}$$

The following SQL code shows the implementation of the measurement of volatility considering the maximum volatility value from its ancestors.

```
 insert Scores
select 1210,8,max(score),"max(volatility
of ancestors)"
from Scores
where object_id in
 (select ancestor_id
  from Ancestors
  where object_id = 1210)
and criterionID=8
```

Uniqueness of derived data $U(s)$ is obtained from the average of its ancestor's uniqueness.

$$U(s) = \frac{\sum_{j=1}^{n} U(\alpha_j)}{n} \quad \text{.....(14)}$$

Timeliness of derived data $T(s)$ is estimated in terms of its maximum currency and volatility.

$$T(s) = \max\left(0{,}1 - \frac{Cu(\alpha_j)}{Vo(\alpha_j)}\right) \quad , \quad \forall j \quad \in [1 \cdots n] \quad \text{..... (15)}$$

Consistency of derived data is determined by the average of the consistency of its ancestors. The consistency of its foreign keys is checked with its corresponding primary keys in each ancestor.

$$Cn(s) = \sum_{j=1}^{n} Cn(\alpha_j) \Big/ n \quad \text{.....(16)}$$

During the assessment of data quality, the Data Quality Manager tool obtains information about the quality of the ancestors from which derived data was produced. Assessing the quality of the available primary data sources from which the integrated data has been obtained is addressed in case there is no possibility of computing data quality from the data itself.

Once obtained the quality properties of the ancestors, the Data Quality Manager is able to assign quality scores to derived data by the aggregation of the quality properties of its ancestors. This assessment requires that all the quality scores of the corresponding ancestors are available. A quality aggregation function combines components of quality into an overall quality specification.

The DQM can show quality scores of the ancestors or derived data by selecting them from the provenance tree, and a brief formula is shown in the Unit column in order to provide the metric from which it was computed.

Fig. 4 shows qualitative information based on data provenance of a query `Cus_Distribution`.

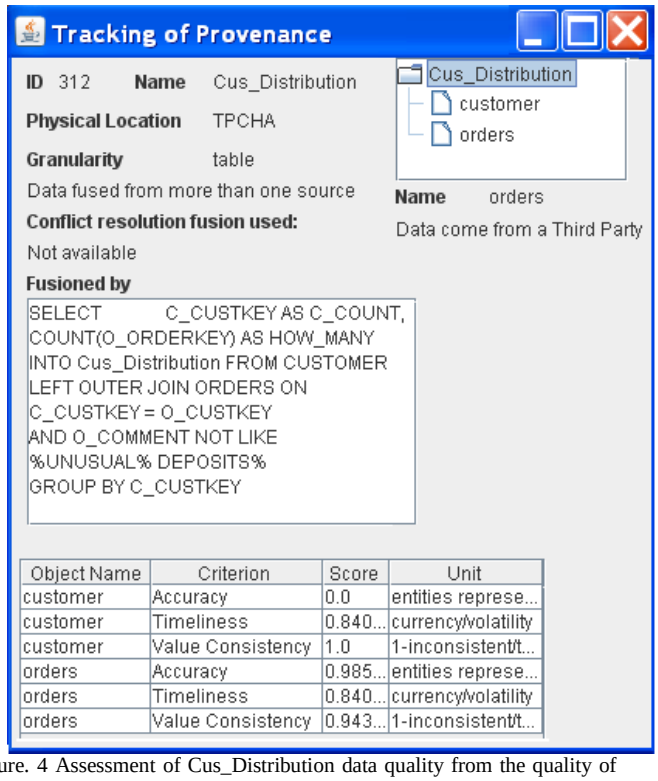

Figure. 4 Assessment of Cus_Distribution data quality from the quality of customer and orders, its ancestors

A statistically sound aggregation is when the quality property was obtained by mean values with given sample size n and one of standard deviation or standard error. If statistically soundness is to be preserved, a mean value can only be calculated for numeric values with an underlying normal distribution.

We have considered average as a default conservative aggregation function for accuracy, completeness, consistency, and uniqueness and a default pessimistic aggregation function for time related quality properties.

There might be different criteria for the aggregation of the qualitative measures. However, the DQM is able to ask expert users which aggregation function would they like to apply for the quality estimation.

Fig. 5 shows the quality estimation for `Cus_Distribution` given by the average of the scores from the ancestors in the case of accuracy, the maximum value as pessimistic approach for the assessment of timeliness.

Users are able to obtain their quality scores in order to decide whether `Cus_Distribution` is suitable for use or not.

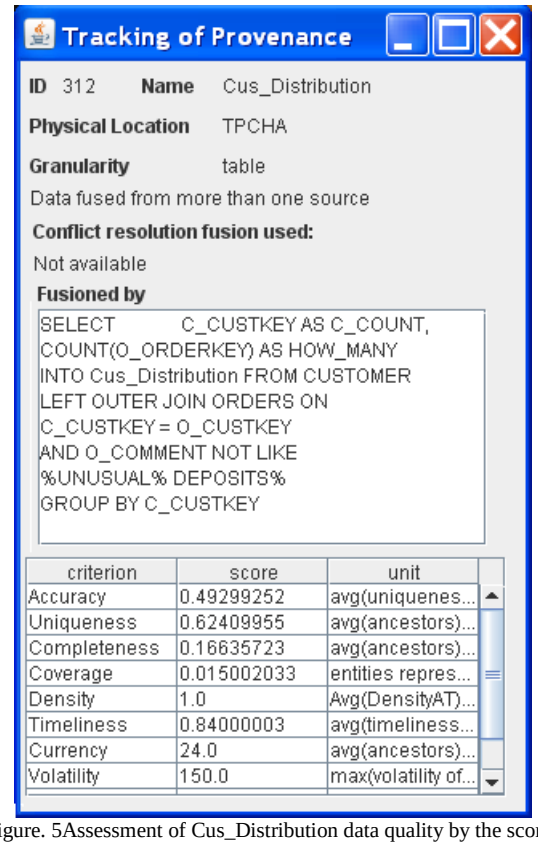

Figure. 5Assessment of Cus_Distribution data quality by the scores aggregation of customer and orders

The following subsection presents some commonly used conflict resolution functions found within the data integration process, and how presenting such information can help users to understand retrieved data.

Enhancing Qualitative Information with the Conflict Resolution Strategies

Previous approaches have developed a number of strategies to resolve conflicts within data fusion [7] [13] [16]. Such information should be taken into account for relying on a data source. Some conflict resolution functions are presented as follows:

Most recent data value: When quality of data is time-related, choosing the most recent value is an option for the solution of conflicting data. When time related data quality dimensions are a priority, then recent value would be preferred.

Most complete data value: Returning the value from the source that contains the fewest NULL values in the attribute in question is recommended if users prefer completeness among other quality properties.

Expert users select data value: The data source has been identified as the best option according with expert users. Therefore, users should take into account that the information retrieved was integrated by a quality dimension called believability, which is particularly relevant in the context of Web.

Selection of the most active data value: In case usability, usefulness, or both are quality properties relevant to the user, this conflict resolution function shall be a good option.

Selection of data value based on the highest quality: The DQM recommends the use of this data value if the quality measure is according to the quality preferences of the data consumer.

Selection of data based on standard aggregation function: The function returns the average, sum, or median value. The DQM recommends this data value as an unbiased and reliable conflict resolution function.

We enhanced the data lineage algorithm we developed in [3] to trace back the conflict resolution functions in order to provide further quality information to users as shown in [1].

During the assessment of data quality by the Data Quality Manager tool, such strategies can be trace back and presented to the user in order to have a better idea what information is being accessed.

The Data Quality Manager prototype provides the physical location, the granularity, the query code or the formula utilized for the data fusion in case of non-atomic data, the provenance tree, and the quality scores of data sources at different levels of granularity.

As we mentioned before, the `pt_profit` query determines how much profit is made on a given line of parts, broken out by supplier nation and year.

The profit is defined as the sum of `[(l_extendedprice*(1-l_discount)) - (ps_supplycost * l_quantity)]` for all line items describing parts in the specified line. Refer to [17] for further detail. Figure 5 presents `pt_profit` as an example of the above mentioned query.

The strategy by which `pt_profit` was selected among other possibilities was because its ancestors where the most active elements within the application of interest. Therefore, the conflict resolution function is presented as “Chosen the most often used data”.

Fig. 6 also presents the scores of the quality properties as a result of assessment by provenance. As we can observe this query is taken information from data sources, which are correct in 82% but not complete (20%), is timely data but very volatile.

The main intension of providing such information is to help users retrieve proper data for operational efficiency and sound decision making.

In the case that a conflict resolution function has been utilized for integrating data, the DQM presents a proper recommendation to users.

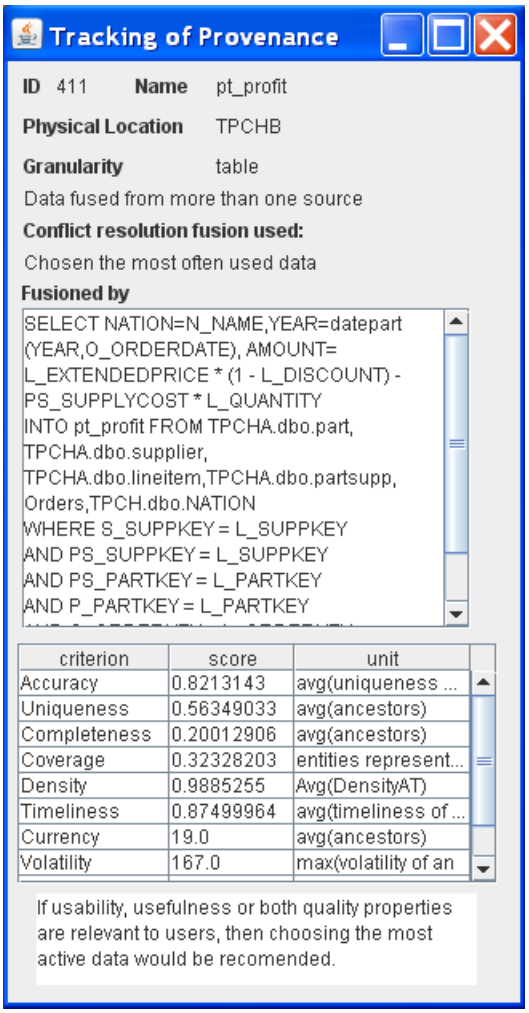

Figure. 6 Assessment of Cus_Distribution data quality by the scores aggregation of customer and orders

Analysis and Ranking of Data Sources

Once the assessment of data has been achieved, the DQM provides the facility to compare data quality of integrated views in order to select the best option. A data quality comparison is presented as follows:

Consider the business question called Market Share, as it was mentioned before, this query determines how the market share of a given nation within a given region has changed over two years for a given part type. There are three possible alternatives to answer the query, called `C_Market_Share`, `D_Market_Share` and `E_Market_Share`. A comparison of such alternatives is possible by the specification of the quality properties of interest. Figure 7 presents accuracy, completeness, and uniqueness as the desired quality properties with their corresponding scores for options C, D, and E.

By default, the DQM is able to apply the proper combination of such methods in order to rank the possible alternatives for the desired global query.

Figure 7 shows assessment and ranking of integrated data, which correspond to the expected outcomes by changing the priority values of chosen quality criteria stated by the user.

We have already explained that the DQM can estimate an overall quality score by providing qualitative information at different levels of granularity, which can vary according to the context specification given by data consumers.

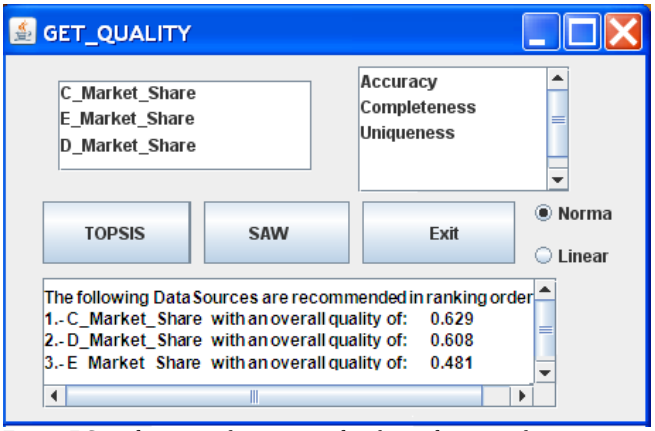

Figure. 7 Specification and execution of ranking of integrated views

As the process of data quality assessment uses a provenance metadata and creates a data quality metadata, in order to analyze data quality changes, the access to these metadata for an ongoing assessment process is required. If data quality assessment is done on a regular basis, users would be able to describe the state of data, to understand problematic data sources, and estimate the cost of data problems to the business.

Assessment of data helps to plan and prioritize data cleansing for improvement, to understand implications of the data quality on newly planned data uses and data driven process before they are put in place [10].

The assessment of data allows the understanding of the current state of data along with the business impact and finding the root causes will lead to a number of activities aimed to prevent data quality problems in addition to correction of current data errors which will be verified by periodic assessments.

E. Business Impact

The enforcement of business rules, the assessment of data quality and the ranking of queries or data sources, let users to identify how root causes affects business.

The data quality scores obtained from the Data Quality Manager inform users which relevant data sources require data cleansing.

The business impact determination varies according to the characteristics of the project, resources, time, and complexity. There are a number of useful techniques such as anecdotes, usage, ranking and prioritization, cost benefit analysis.

In this practical approach we identified ranking of business questions or data sources at different levels of

granularity as a very helpful mechanism, for determining the impact of poor data quality on the business.

Ranking of data sources according to its quality allows users to identify which sources of information should be cleaned and which local applications should enhance its business constraints.

F. Data Cleansing

Data Cleansing determines causes of errors and possible treatments. It also creates an audit trail of corrections.

The process of data cleansing requires on the first place identifying the types of errors reducing the data quality then on the second place choosing appropriate methods to automatically detect and remove such anomalies, applying the corresponding methods to the data sources and as a final step examining the results and perform exception handling for the tuples not corrected.

The correct use of metadata has been very useful in order to detect data failing and to establish data profiling and cleansing mechanics. Data consolidation specifications are now built with deep understanding of the actual structure, content, and quality of the data in each source.

Comprehensive data profiling and quality assessment has been a key for success. We started with a comprehensive set of tests, comparing the data between all sources, then we analyzed the discrepancies and look for patterns, if for instance, some time values in two data sources coincide we can trust them and make some corrections on the third one.

Data matching is a very common mechanism to merge and eliminate duplicated rows and keep correct data. At this point we are in enhancing the data matching program. For instance, in the case of text, we have executed a data quality pattern analyzer [35] in the following SQL code:

```
SELECT generate_mask(LINEITEM.ORDERKEY)
AS ORDERKEY_Pattern,
FROM LINEITEM;
```

Table 1 presents the corresponding patterns identified for the O_COMMENT text column.

Table 1 Patterm for O_COMMENT teext column.

O_COMMENT_Pattern
UUUUNNNNNNUUUUUUNN
LLLLNNNNNNWLLLLUNN
UUUUUUNUNUUUNUUNNUU
UUUUNNNNNNUUUUUUUN

After executing the data cleansing processes a certain acceptable level of data quality has been achieved. Therefore, data consumers are able to make effective and informed decisions on the basis of cleansed data at the level of data quality expected. However, as we mentioned before, what is correct today may be completely erroneous tomorrow. In order to maintain the data quality status by preventing new errors from being introduced into the data we require monitoring data integration interfaces and ensuring quality of data conversion and consolidation.

G. Continuous Monitoring and Assessment

After the initial data quality assessment and cleansing, the next step is to ensure that improvements are assigned and implemented. Therefore, we need to plan and implement controls, monitor improvements, and document the results. The successful improvements should be standardized.

Assessing data quality on a regular basis on large volumes of data of a production database is not always viable and technically challenging [10]. The assessment frequency and the level of granularity to assess depend mainly on the objectives stated for the project. A certain level of quality shall be achieved and in the case of that level is inappropriate then assessment and cleansing will be required.

VII. CONCLUSION AND FUTURE WORK

From the existing Data Quality Frameworks, data have always been considered as the product of a primary data source. Therefore, no consideration of derived data has been approached until now. The qualitative information provided to the user contains measures of quality, the original data sources where data come from, and the components of integrated data by considering the process of data integration (i.e. data fusion, data replication, or data transformation) during data quality measurement and assessment. In other words, measuring quality of derived data as part of a Data Quality Framework for multi-database environments has not been addressed before. Very few approaches have considered quality properties at different levels of granularity on databases [12] [14]. Not to mention levels of granularity within derived data.

In the present document, we have shown a practical approach for a proposed Data Quality Framework, where the Data Quality Assessment tool is able to assign quality scores to derived data by considering them as primary data sources, by comparing the available quality scores of its ancestors, or by the aggregation of the quality properties of all its ancestors. Therefore, we presented a new granularity-based assessment classification. Furthermore, qualitative information has been enhanced by including the conflict resolution function and the code or formula utilized for integrating data, depending on the granularity of data along with a brief recommendation to users for trusting data according to the conflict resolution function utilized.

As we mentioned before, data quality degrades during the data integration process [2]. The objective of monitoring these data integration processes is to prevent these errors from getting into the target database. The solution is to design and develop tools between the source and the target data before it is loaded and processed such as the Data Quality Manager for the assessment and ranking of non-atomic data and therefore allow users to be able to make

effective decisions by trusting according to the description of qualitative information such as the quality scores, the conflict resolution function, and the quality properties of their ancestors.

The process of determination of cost of data quality by computing the cost to prevent errors, and the cost to correct them is part of our future work.

The process that applies conversion routines to transform data into its preferred and consistent format using both standard and custom business rules stills on development.

We also are planning to extend the presented Data Quality Assessment process to consider semi-structured data.

ACKNOWLEDGMENT

This work was supported by a grant from Dirección General de Asuntos del Personal Académico, UNAM.

REFERENCES

[1] P. Angeles and F. Garcia-Ugalde “Assessing Quality of Derived Non Atomic Data by considering conflict resolution function”, First International Conference on Advances in Databases, Knowledge, and Data Applications. 978-0-7695-3550-0/09 © 2009 IEEE DOI 10.1109/DBKDA.2009.10, pp. 81-86, Cancun, Mexico, 2009.

[2] P. Angeles and L. MacKinnon, “Detection and Resolution of Data Inconsistencies, and Data Integration using Data Quality Criteria”, Quality in Information and Communications Tech., pp. 87-94, Porto, Portugal, 2004.

[3] P. Angeles and L. MacKinnon, “Tracking Data Provenance with a Shared Metadata”, Postgraduate. Research Conference in Electronics, Phot., Comm. and Networks, and Computing Science, pp. 120-121, Lancaster England, 2005.

[4] P. Angeles and L.MacKinnon, “Quality Measurement and Assessment Models Including Data Provenance to Grade Data Sources”, Int. Conference on Computer Science and Information Systems”, pp. 101-118, Greece, 2005.

[5] P. Angeles, “Management of Data Quality when Integrating Data with Known Provenance”, PhD Thesis, Heriot-Watt University, Edinburgh, UK, April 2007.

[6] D. Ballou, G. Tayi, "Examining Data Quality", Communications of the ACM, vol. 41,no.2, pp.54-57, 1998.

[7] J. Bleiholder. Declarative Data Fusion, Syntax, Semantics, and Implementation. Advances in DB and I S, Estonia, 2005, pp. 58-73, 2005

[8] M. Burgess, W. Gray, and N. Fiddian, "A Flexible Quality Framework For Use Within Information Retrieval", Int. Conference on IQ,Cambridge, MA, USA, 2003.

[9] M. Gertz and I. Schmitt, "Data Integration Techniques Based on Data Quality Aspects", 3rd National Workshop on Federal Databases, 1998.

[10] A. Maydanchik, Data Quality Asseessment, Data Quality for Practitioners Series, Technics Publications New Jersey ISBN 978-0-9771400-2-2, 2007.

[11] A. Motro and I. Rakov I, "Estimating the Quality of DB", Int. Conference on Flexible Query Answering Systems, pp. 298-307,Springer-Verlag, Germany, 1998.

[12] F. Naumann, "Quality-Driven Query Answering for Integrated IS", Lecture Notes in Computer Sciences LNCS 2261, Springer Verlag, 2002.

[13] F. Naumann, A. Bilke, J. Bleiholder, M. Weis ”Data Fusion in Three Steps: Resolving Inconsistencies at Schema, Tuple and Value-level, IEEE Data Engineering Bulletin 29(2):21-31, 2006.

[14] L. Pipino, W.L. Yang and R. Wang, "Data Quality Assessment", Communications of the ACM, Vol. 44 no. 4e, pp.211-218, 2002.

[15] M. Scannapieco, A. Virgillito, et.al. "The DaQuinCIS Architecture: a Platform for Exchanging and Improving DQ in Cooperative IS", Information Systems, Elsevier, pp. 551-582, 2004.

[16] Schallehn E., Sattler Kai-Uwe, Saake G., Efficient similarity-based operations for data integration Data & Knowledge Engineering, Vol. 48, 3, 2004, Pages 361-387

[17] TPCH, TPC Benchmark ™ H, Standard Specification Revision 2.3.0 Transaction Processing Performance Council, http://www.tpc.org , 2006, (date information as accessed by the author citing the references, e.g. 23 Sept. 2009.)

[18] R. Wang, "A Product Perspective on Total Data Quality Management", Communications of the ACM, vol. 41, no. 2, pp.58-65, 1998.

[19] D. McGilvray Executing Data Quality Projects Ten Steps to Quality Data and Trusted Information, ISBN 978-0-12-374369-5, Morgan Kaufman, Publishers, 2008.

[20] M. Scannapieco, C. Batini, "Completeness in the Relational Model: A Comprehensive Framework”, Research Paper, in Proceedings of the 9h International Conference on Information Quality (ICIQ-04, Cambridge, MA, USA, November 2004.

[21] L. Young and D. Strong “Knowing-Why about Data Processes and Data Quality”, Journal of Management Information Systems, Vol. 20, No. 3, pp. 13 – 39. 2004.

[22] T. Anderson , The Penalties of Poor Data, Immedia smart targeted solutions., http://www.goimmedia.com/ArticlesWhitepapers/ThePenaltiesofPoorData.aspx , (date information as accessed by the author citing the references, e.g. 23 Sept. 2009.)

[23] D. Loshin, Enterprise Knowledge Management, The Data Quality Approach, 2007.

[24] New York Times, December 1, 2002, by Jennifer Bayot.

[25] New York Times, January 19, 2003, by The Associated Press.

[26] New York Times, February 28, 2003, by Karen W. Arenson.

[27] Y.Lee, D. Strong, “Knowing-Why about Data Processes and Data Quality”, Journal of Management Information Systems, Vol. 20, No. 3, pp. 13 – 39. 2004.

[28] The MIT Total Data Quality Management web site, http://web.mit.edu/tdqm/, (date information as accessed by the author citing the references, e.g. 23 Sept. 2009.)

[29] P. Anokhin, A. Motro, "Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources", Technical Report ISE-TR-03-06, Information and Software Engineering Dept., George Mason Univ., Fairfax, Virginia, 2003.

[30] J. Cavano, “A Framewok for the Measurement of Sotware Quality”,Rome Air Development Center, James A. McCall, General Electric Company (1978),pp.133-139.

[31] K.T. Huang, Y.W. Lee, R.Y. Wang, Quality Information and Knowledge Management,Prentice Hall PTR Upper Saddle River, NJ, USA, ISBN:0-13-010141-9.

[32] A.F. Karr, A.OP. Sanil, D.L.Banks ,” Data Quality: A Statistical Perspective”, Technical Report 151, March 2005, National Institute of Statistical Sciences.

[33] T. C. Redman, “Data Quality for the Information Age”, Boston, MA., London : Artech House, 1996, ISBN:0890068836.

[34] Wang R. Y., Strong D.M. “Beyond accuracy: What data quality means to Data Consumers”, Journal of Management of Information Systems, vol. 12, no 4 1996, pp. 5 -33.

[35] Data Quality Pro Forum, http://www.dataqualitypro.com/(date information as accessed by the author citing the references, e.g. 23 Sept. 2009.)