INTERNATIONAL JOURNAL OF RESEARCH CULTURE SOCIETY ISSN: 2456-6683 Volume - 4, Issue - 6, June – 2020
Monthly, Peer-Reviewed, Refereed, Indexed Journal Scientific Journal Impact Factor: 5.245
Received on : 01/06/2020 Accepted on : 15/06/2020 Publication Date: 30/06/2020

DATA QUALITY ARCHITECTURE FOR DATA WAREHOUSES

[1]Madhusudhan Reddy Sureddy, [2]Prathyusha Yallamula

[1]Associate Director, [2]Vice President
[1] Application Development, Santander Bank
[2] Risk Treasury, Natixis CIB Americas
Email - [1]sureddy21@gmail.com, [2]prathyusha13y@gmail.com

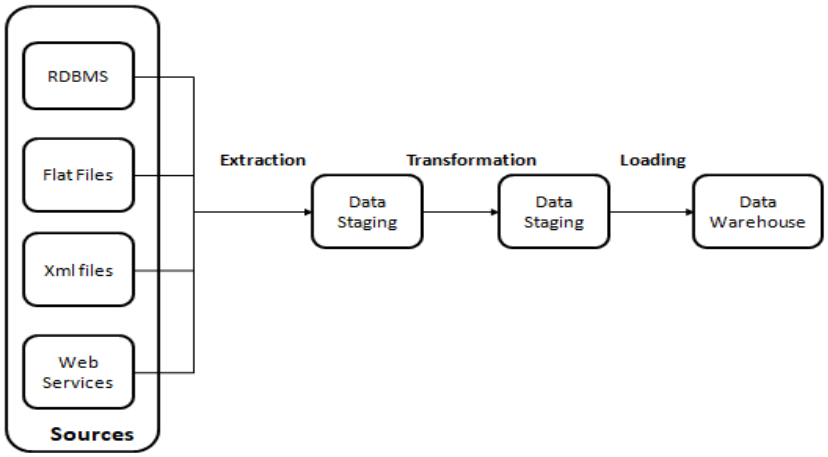
Abstract: *Understanding the Quality of data loaded into the data warehouses is one of the key criteria's for any data warehouse success. Bad quality data warehouses do not have longevity in organizations and decommissioning/replacing those leads to millions of dollar of additional spend. Data quality is needed at every data in and data out points and hence organizations should invest in data quality during the design phase of data warehouse itself and ensure it's integrated with the data integration approach. This article provides a comprehensive data quality architecture which satisfies all the key dimensions of data quality – completeness, consistency, uniqueness, Accuracy, validity, Timeliness.*

Key Words: *Business Intelligence (BI), Data warehousing (DW), Data Quality (DQ), Architecture, Framework, Data Integration (DI), extraction-transformation-load (ETL), Data Model, Data Quality Rules.*

1. INTRODUCTION:

Growth of any organization is always directly proportional to the depth of organizational data understanding. Business Intelligence (BI) and Data Warehousing (DW) together provide the keys to facilitate this data analysis and present the actionable information for senior executives to make informed decisions. Common functions of BI include reporting, analytics, data mining thus making it the front face of the decision making process. Data warehousing on the other hand stores the organizations data and provides the data needed for BI to function. Data Warehouses integrates data from business transactional systems and a wide range of other data sources - relational databases, flat file sources, xml sources and Web services into a structured reportable format. Genesis of data warehousing started in late 1970s with its concept fully formalizing in 1988 by IBM researchers Barry Devlin and Paul Murphy through their journal "An Architecture for a business and information system". Most Organizations always have a well defined data integration strategy in their data warehouse development life cycle and are able to build an operational data warehouse in a short span. Life of these quickly rising data warehouses is short lived as organizations tend to fall short with their data quality strategies.

Data warehouses normally follow two main approaches for data integration - extraction-transformation-load (ETL) and extraction-load-transformation (ELT). ETL based loading is preferred by organizations using ETL tools like Informatica, Talend, SSIS and ELT based approach is preferred by organization using Massively Parallel processing (MPP) Databases like Teradata, Netezza. Below illustration provides the data flow architecture of ETL based data warehouse which do not follow any kind of data quality architecture.



During the loading of a data warehouse there are multiple points of data entry, data modification and data exit. Data quality needs to be measured at each of these data movement points for building a successful data warehouse.

- Data is received from various heterogeneous sources - Data entry point for source files
- Data is Extracted from source files and loaded into the data staging database - Data exit point for source files and Data entry point for staging database

- Data in Staging from various sources is integrated and transformed to meet the data warehouse database model - Data Modification point for staging database
- Data from staging is loaded into the data warehouse - Data exit point for staging and Data entry point for data warehouse

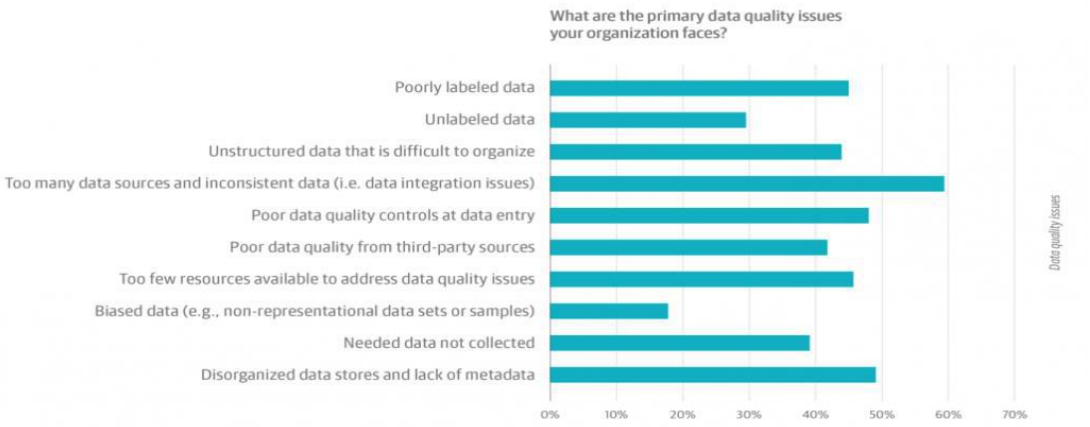
Data warehouse in the above illustration is prone to failure as the data movement points have gone unchecked for data quality. Understanding the Quality of data in the data warehouses is one of the key criteria's for any data warehouse success. Data Quality in data warehouses can be measured using six core dimensions - Completeness, Uniqueness, Timeliness, Validity, Accuracy, and Consistency.

- Completeness indicates the proportion of the stored data for all the data sets and data items against the potential of 100% complete.
- Uniqueness indicates the duplicity of the same information based on how that is identified.
- Timeliness indicates whether the data meets availability and accessibility requirements.
- Validity indicates if the data conforms to the syntax of its definition.
- Accuracy indicates the accuracy of data by validating it against third party data sources.
- Consistency indicates if inconsistencies exist within the data or between similar data obtained from different data sources.

In this paper author provides a comprehensive data quality architecture which satisfies all these key dimensions of data quality.

2. LITERATURE REVIEW:

Data quality has played a critical role since 1865 when Professor Richard Millar Devens first coined the term business intelligence in the Cyclopaedia of Commercial and Business Anecdotes to describe how a banker named Sir Henry Furness had gained profit by gathering data and acting upon it. Concept of measuring data quality using attributes called as "data quality dimensions" happened only in 1995 after emergence of data warehouses when Total Data Quality Management group of MIT University led by Professor Richard Y. Wang used a two-stage survey to identify four categories containing fifteen data quality dimensions. Even after 25 years of introduction of data quality measures organizations are still struggling with variety of data quality challenges as noted in the study "The state of data quality in 2020 ". This study pointed out that many top executives and IT professionals felt the need for organizations to reach more comprehensive levels of Data Quality. Based on the study even now the most common data quality problem is "data integration issues"



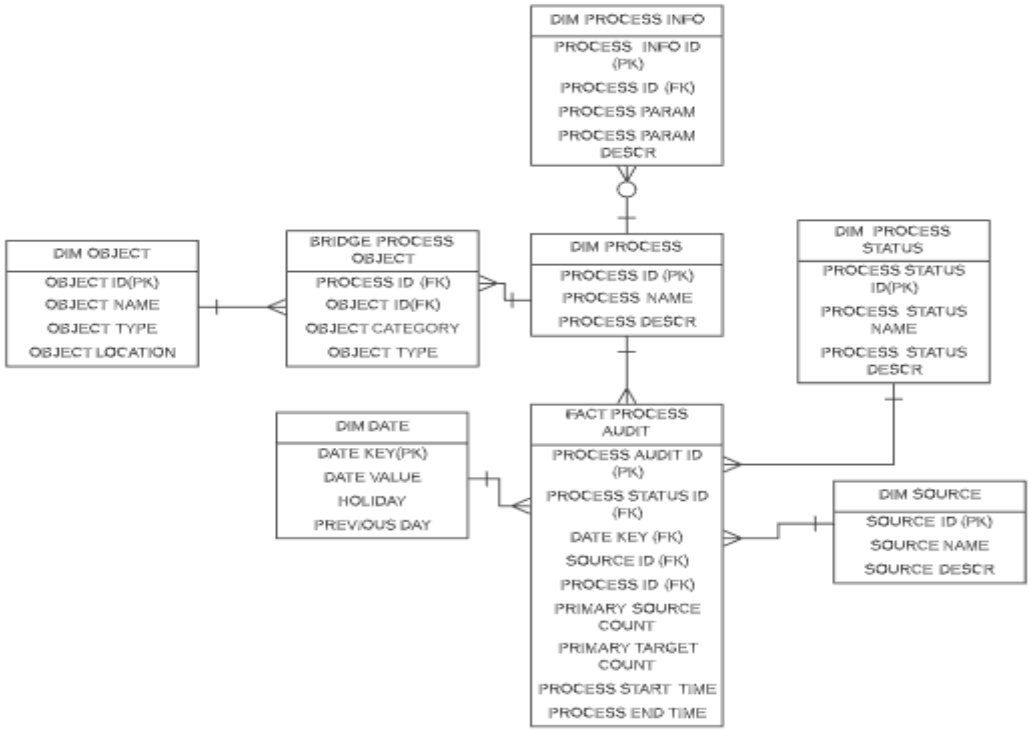
Above discussion indicates that data quality issues are still persistent in organizations and need to be addressed to be successful with business intelligence.

3. PROBLEM STATEMENT:

Bad quality data warehouses do not have longevity in organizations and decommissioning/replacing those leads to millions of dollar of additional spend to businesses. In many cases poor Data Quality has caused loss of billions of dollars to organizations through litigations and incorrect business decisions, as described by Nord in his 2005 journal "due to the fraud in food stamps wherein some of the recipients received the benefits even after they were long dead" and "due to inaccurate and outdated information in data systems, organizations have experienced financial losses". To eliminate such issues a comprehensive data quality architecture which satisfies all the key dimensions of data quality is required. Data quality is also required to be included into the data warehouse development life cycle, with its initiation in the early stages of data warehouse build and data quality architecture being integrated with the data integration architecture.

4. DATA QUALITY ARCHITECTURE FOR DATA WAREHOUSES:

Data warehouses have strict service level agreement on the time refreshed data needs to be made available for business reporting. In most data warehouses refreshed data as of previous business day closure is expected to be available in data warehouse before next business day start. Due to this reduced time available for data loading not all data quality checks can be done before data is loaded into the data warehouses and data quality checks have to be broken into critical and non-critical. Critical data quality checks happen as part of the data integration flow and a failure in these rules causes the data load to stop and hence needs to be addressed immediately. Non-critical data quality checks can run as part of the data quality flow and there is no impact to the data loading even if these rules fail. Though these are named non-critical failures these data quality checks also need to be solved before the next day's load to ensure bad data does not get accumulated in the data warehouses. Proposed data quality architecture in this paper needs two data marts – audit data mart and data quality data mart for storing the data integration batch load data and data quality result data. Audit data mart stores the static and run metadata needed for data integration execution and provides more control of the data loading. Below is a sample audit data model with entity names and only critical data elements. This model can be expanded based on the organization needs.



DIM PROCESS:
This table is used to store the details of the data integration jobs executing in the data warehouse.

DIM PROCESS INFO:
This table contains the configuration details of the data integration jobs and used for standardizing the data integration loads

DIM PROCESS STATUS:
This table contains the details of the process statuses which can occur for a data integration process.

DIM SOURCE:
This table is used to store the details of the various source systems which send data to the data warehouse

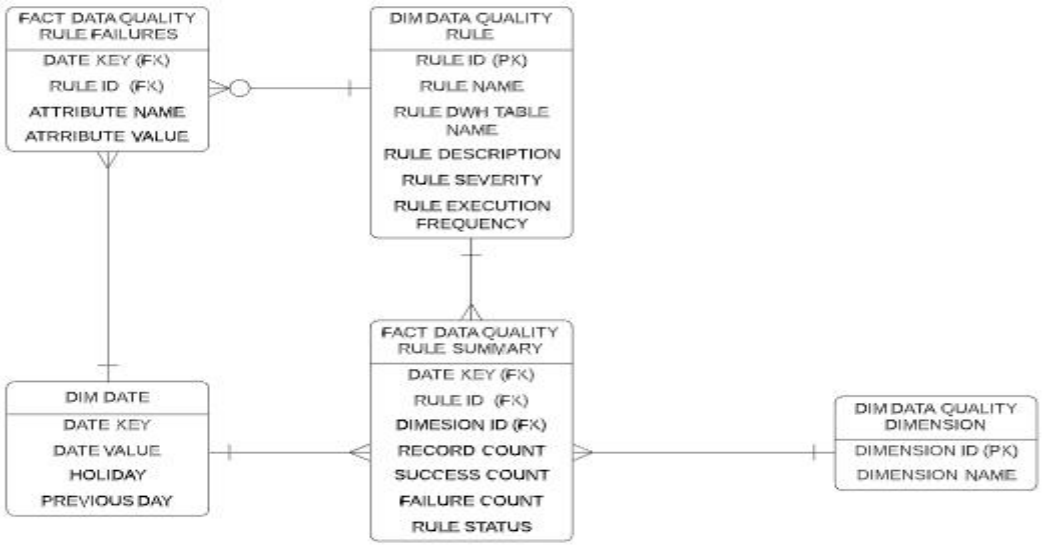
DIM OBJECT:
This table stores the details of every object used during the loading of the data warehouse. It includes the actual source files used, source tables used, data base staging objects and data warehouse objects.

BRIDGE PROCESS OBJECT:
This table acts a bridge between the DIM OBJECT and DIM PROCESS table and stores the information of all the source and target objects used in a data integration process. This also stores the information of the primary source and primary target of the data integration process.

DIM DATE:
This is the standard date dimension used in data modelling and stores the information of all the calendar dates.

FACT PROCESS AUDIT:
This is the main table of the audit data mart and stores all the measures needed for auditing the data integration loads. This stores the information of when the process started, process ended, primary source object count, primary target object count and the status of the process.

Data quality data mart stores the static and run data needed for data quality rule execution. Below illustration provides a sample data quality data model with entity names and only critical data elements. This model can be expanded based on the organization needs.



DIM DATA QUALITY RULE:

This table stores the metadata information of the data quality rule being executed. It also stores the data warehouse table for which data quality is being measured

DIM DATA QUALITY DIMENSION:

This table contains the details of the data quality dimensions which are being used for measuring the organizations data quality

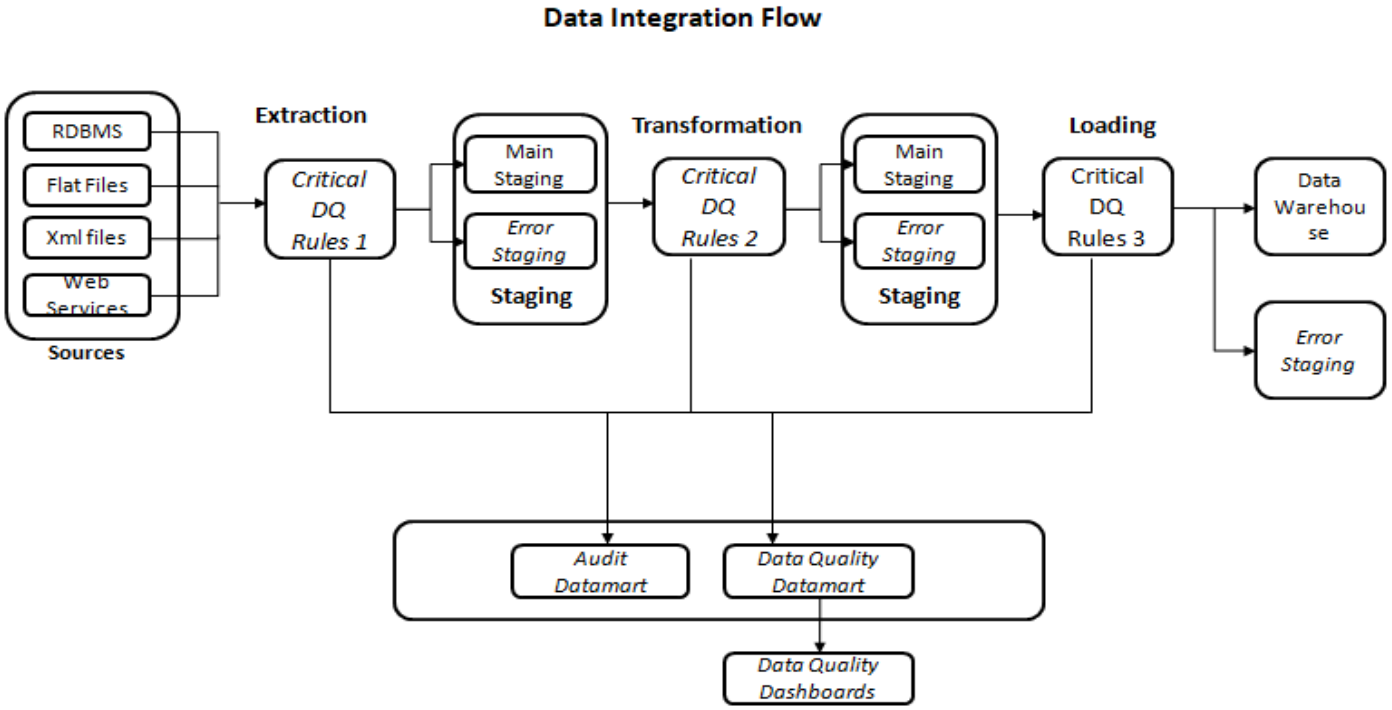
FACT DATA QUALITY RULE FAILURES:

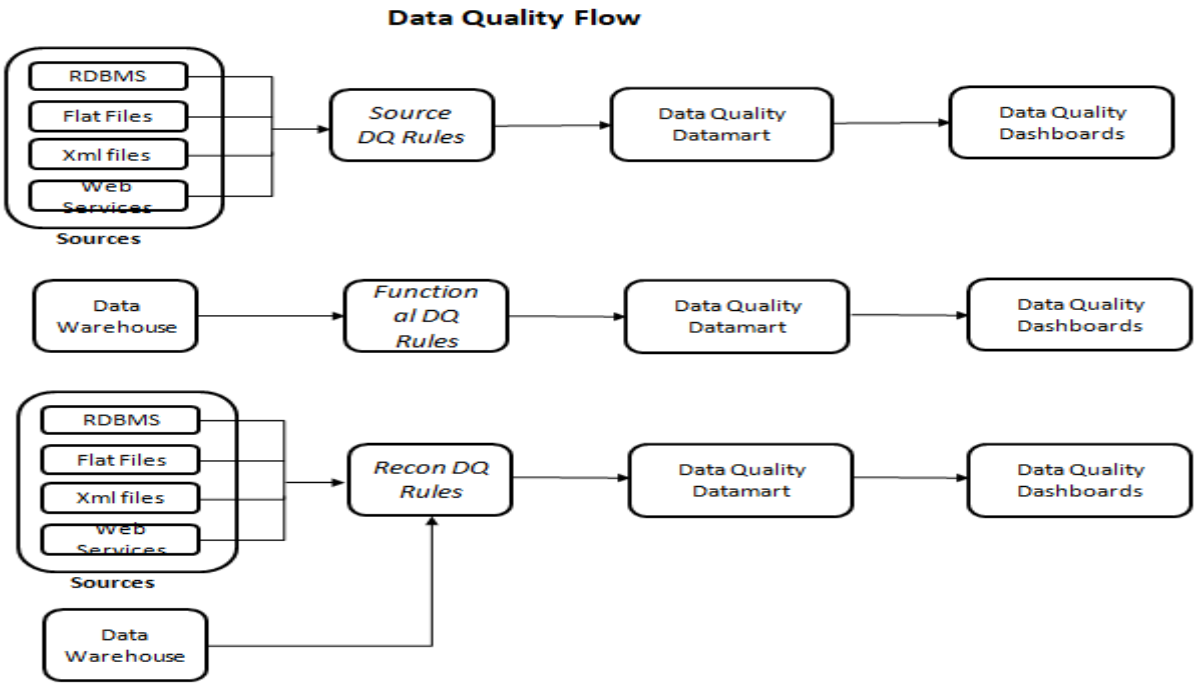
This table stores the actual data records for the data quality rules which were not successful. This in conjunction with fact data quality rule summary helps the data quality support resource to analyze the data issues.

FACT DATA QUALITY RULE SUMMARY:

This is the main table in the data quality data mart and stores all the measures which include the record count, success count, failure count and the rule status.

Below illustration provides the details of the data integration and data quality architectures to be followed for a successful data warehouse.





In the proposed data integration flow data quality check points are introduced at every data out, data in and data modification points during the data integration flow. Before the data extraction step begins audit data mart captures the information of record counts, source file availability and critical DQ rules1 are executed on the source files and audit data mart to identify the bad files and bad records. Bad Records which fail any critical DQ rules are loaded into a similar structured error staging table and FACT DATA QUALITY RULE FAILURES table. Only successful data records are loaded into the main staging tables. Error data is duplicated into error staging tables to allow skipping of these records if the business users deem these records as bad data and not fit for loading. After the data extraction step completion audit data mart captures the information of record counts and critical DQ rules2 are executed on the audit data mart and new staging tables to identify the bad records. After the data transformation step completion audit data mart again captures the information of record counts and critical DQ rules3 are executed on the audit data mart and new staging tables to identify the bad records. Any failures in critical data quality rules causes the data integration flow to stop immediately and load stays on hold until the data quality support resource and data governance resource solve the issue. Data quality rules in the data integration flow can be broadly classified into below:

Critical DQ Rules1:

These rules execute on the source data objects and during the extraction step of the ETL process and check for below:

- NULL values or badly formatted data in the critical source attributes. All source columns which eventually turn into a key column in the data warehouse or utilized in join conditions during the data integration loading are considered as critical source attributes.
- Differences in count and size of the source objects between source data transferred from the source system and the source data received on the data warehouse loading server.
- Differences in count and size of the previous day and current day source objects and if the changes in source files are under the thresholds defined by business.
- Delays in the source data availability

Critical DQ Rules2:

These rules execute during the transformation step of the ETL process and check for below:

- Differences in count between primary source object count and each data integration process primary target object count.
- Critical data quality rules defined by the business which allow checking of the data validity during the data transformation step itself.

Critical DQ Rules 3:

These rules execute during the loading step of the ETL process and check for below:

- Referential data integrity issues between the facts and dimensions of the data warehouse tables

A parallel data quality flow allows the execution of the non-critical data quality rules to be executed with the data integration flow. Data quality rules in this flow can be broadly classified into below:

Source DQ Rules:

These rules execute on the source data objects and run as part of the parallel data quality flow and check for below:

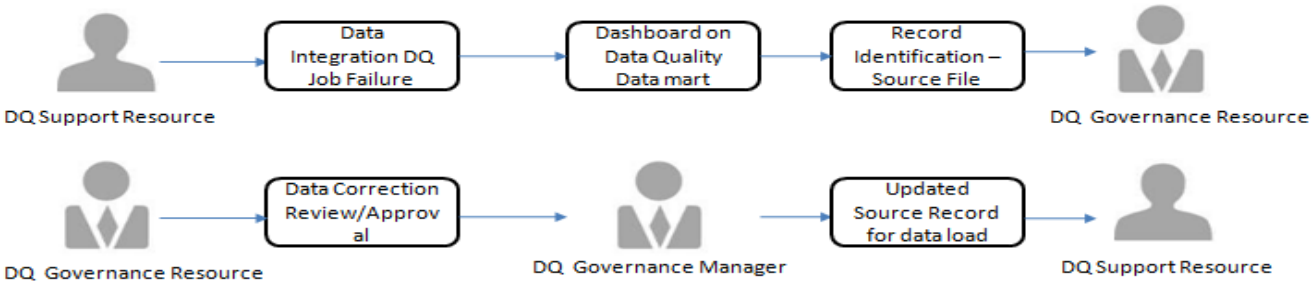
- NULL values or badly formatted data in all source file attributes and if the data quality of each attribute is under the threshold defined by business
- Data quality rules defined by the business which allow checking of the data validity in source files. These rules are defined during the initial source file analysis in the analysis phase of the project.

Functional DQ Rules/Recon DQ Rules:

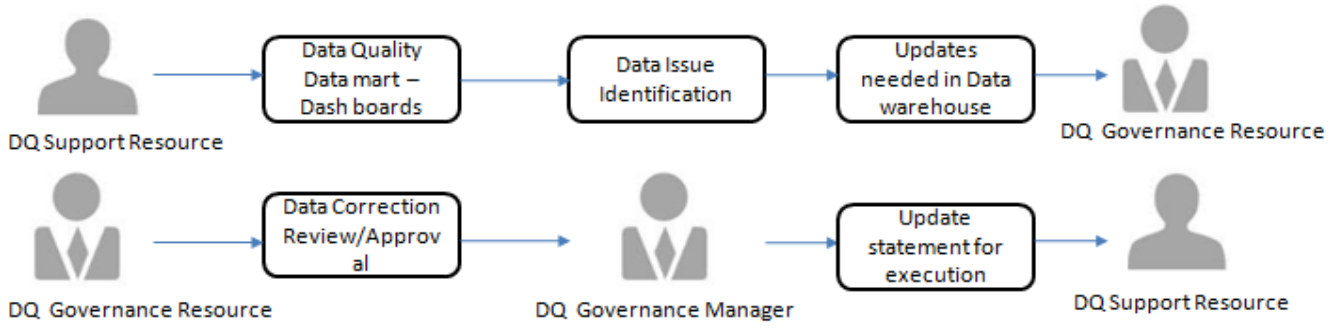
These rules are defined by the business during the data warehouse development phase and are written over the loaded data warehouse tables. Functional DQ rules are defined based on the source system functionality and vary for source system. Recon DQ rules do reconciliation between the source files and the data warehouse tables to validate the data quality on critical data measures.

Data quality of a data warehouse starts diminishing overtime if the support and governance processes for data quality are not well defined. Due to this reason Data quality maintenance team should be separate to that of the data integration team. DQ Maintenance team should be part of the business team which uses the data warehouse for reporting and should comprise of DQ support resources and DQ governance resources. DQ support resource responsibility is to monitor the data quality dashboards on a 24*7 basis and catch the abnormalities in the data. DQ support resource also provides the fix for the data issues and works with Data governance resource to apply the fix. Data Quality governance resource has an ability to directly modify data in the data warehouses and hence his actions are monitored through a workflow process, where in the DQ governance manager approves each change to the data warehouse. Below illustrations depicts the data correction flow in the case of data integration job failure and data issue in non critical business rules.

Data Correction Flow – Data Integration Job Failure



Data Correction Flow – Data Issues



Data quality architecture proposed in this paper can be implemented using any technology and can also be easily be expanded to the big data environments which do not follow the traditional data warehousing approach. Below table provides the information on how various aspects of the data quality architecture meet each of the core data quality dimensions:

DQ Architecture Aspect	DQ Dimension Satisfied
Critical DQ Rules1/ Critical DQ Rules2/ Critical DQ Rules3/Source DQ Rule	Completeness
Functional DQ Rules	Uniqueness
Critical DQ Rules1	Timeliness
Critical DQ Rules1/Source DQ Rule	Validity
Recon DQ Rules	Accuracy
Functional DQ Rules	Consistency

5. CONCLUSION:

Goal of any data warehouse is to provide the management with high quality data which allows them to make better decisions in improving the organizations performance. Exodus of data due to the emergence of big data solutions has made it more important now than ever to manage the data quality of data. This paper provides comprehensive data quality architecture for data warehouses to better handle the data quality issues which are still persistent in most of the organizations. DQ architecture proposed in this paper provides detailed data models for the data quality result storage and introduces data quality at every data movement point. DQ Architecture provided also satisfies all the key dimensions of data quality and is very well integrated with data integration data flow.

REFERENCES:

Journal Papers

1. B. A. Devlin and P. T. Murphy, An *architecture for a business and information system* in IBM Systems Journal, vol. 27, no. 1, pp. 60-80, 1988, doi: 10.1147/sj.271.0060
2. Wang, R., & Storey, V. (1995) *Framework for Analysis of Quality Research. IEEE Transactions on Knowledge and Data Engineering* 1(4), pp 623–637.
3. Ballou, D., & Tayi, G. (1999, January). *Enhancing data quality in data warehouse environments*, *Communications of the ACM*, 42(1), pp. 73-78.
4. Nord, G. D. (2005), *An Investigation of the Impact of Organization Size on Data Quality Issues*, Journal of Database Management, Vol. 16, No. 3, pp. 58-71.
5. Nemani, Rao & Konda, Ramesh. (2019). *A Framework for Data Quality in Data Warehousing*.
6. Richard H. Thayer, and Barry W. Boehm, *software engineering project management*, Computer Society Press of the IEEE, pp.130, 1986.
7. Rahul Gupta *Maintaining Data Quality in Data Warehouse* on International Journal of Scientific and Research Publications, Volume 6, Issue 12, December 2016

Proceedings Papers

1. Cai, L and Zhu, Y 2015 *The Challenges of Data Quality and Data Quality Assessment in the Big Data Era*. Data Science Journal, 14: 2, pp. 1-10, DOI: http://dx.doi.org/10.5334/dsj-2015-002

Web References

1. https://www.dataversity.net/a-short-history-of-data-warehousing/
2. https://www.dataversity.net/brief-history-data-warehouse
3. https://www.investopedia.com/terms/d/data-warehousing.asp
4. https://www.dataversity.net/a-brief-history-of-data-quality
5. https://www.investopedia.com/terms/b/business-intelligence-bi.asp
6. https://searchbusinessanalytics.techtarget.com/definition/business-intelligence-BI
7. https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf
8. https://resources.paxata.com/reports-white-papers/the-state-of-data-quality-in-the-enterprise-2018
9. https://www.oreilly.com/radar/the-state-of-data-quality-in-2020/

Author Biography

I am Madhusudhan Reddy Sureddy with qualifications of Bachelors of engineering from Osmania University. I have 14 years of Information Technology experience and worked with fortune 500 companies - GE Capital Americas, American Red Cross, CVS Caremark, Fannie Mae and Santander Bank and in the business domains BFSI (Banking, Financial Services and Insurance), Mortgage and Healthcare. My research fields include data warehousing, data integration, data quality, data virtualization, master data management, reference data management, metadata management, big data solutions, business rule engines, data modelling, data analytics, Operational and Support Model, server management and Capacity planning.