Data Quality

Introduction

Today is world of heterogeneity.
We have different technologies.
We operate on different platforms.
We have large amount of data being generated everyday in all sorts of organizations and Enterprises.
And we do have problems with data.

Problems

Duplicated , inconsistent
, ambiguous, incomplete.

So there is a need to collect data in one place and clean up the data.

Why data quality matters?

Good data is your most valuable asset, and bad data can seriously harm your business and credibility…

1. What have you missed?
2. When things go wrong.
3. Making confident decisions.

What is data quality?

Data quality is a perception or an assessment of data's fitness to serve its purpose in a given context.
It is described by several dimensions like

•Correctness / Accuracy : Accuracy of data is the degree to which the captured data correctly describes the real world entity.

•Consistency: This is about the single version of truth. Consistency means data throughout the enterprise should be sync with each other.

Contd…

- Completeness: It is the extent to which the expected attributes of data are provided.

- Timeliness: Right data to the right person at the right time is important for business.

- Metadata: Data about data.

Maintenance of data quality

Data quality results from the process of going through the data and scrubbing it, standardizing it, and de duplicating records, as well as doing some of the data enrichment.

1. Maintain complete data.
2. Clean up your data by standardizing it using rules.
3. Use fancy algorithms to detect duplicates. Eg: ICS and Informatics Computer System.
4. Avoid entry of duplicate leads and contacts.
5. Merge existing duplicate records.
6. Use roles for security.

Inconsistent data before cleaning up

Invoice 1

Bill no	CustomerName	SocialSecurityNumber
101	Mr. Aleck Stevenson	ADWPS10017

Invoice 2

Bill no	CustomerName	SocialSecurityNumber
205	Mr. S Aleck	ADWPS10017

Invoice 3

Bill no	CustomerName	SocialSecurityNumber
314	Mr. Stevenson Aleck	ADWPS10017

Invoice 4

Bill no	CustomerName	SocialSecurityNumber
316	Mr. Alec Stevenson	ADWPS10017

Consistent data after cleaning up

Invoice 1

Bill no	CustomerName	SocialSecurityNumber
101	Mr. Aleck Stevenson	ADWPS10017

Invoice 2

Bill no	CustomerName	SocialSecurityNumber
205	Mr. Aleck Stevenson	ADWPS10017

Invoice 3

Bill no	CustomerName	SocialSecurityNumber
314	Mr. Aleck Stevenson	ADWPS10017

Invoice 4

Bill no	CustomerName	SocialSecurityNumber
316	Mr. Aleck Stevenson	ADWPS10017

Data Profiling

Context

In process of data warehouse design, many database professionals face situations like:

1. Several data inconsistencies in source, like missing records or NULL values.
2. Or, column they chose to be the primary key column is not unique throughout the table.
3. Or, schema design is not coherent to the end user requirement.
4. Or, any other concern with the data, that must have been fixed right at the beginning.

To fix such data quality issues would mean making changes in ETL data flow packages., cleaning the identified inconsistencies etc.

This in turn will lead to a lot of re-work to be done.

Re-work will mean added costs to the company, both in terms of time and effort.

So, what one would do in such a case?

Solution

Instead of a solution to the problem, it would be better to catch it right at the start before it becomes a problem.

After all “**PREVENTION IS BETTER THAN CURE**”.

Hence data profiling software came to the rescue.

What is data profiling ?

It is the process of statistically examining and analyzing the content in a data source, and hence collecting information about the data. It consists of techniques used to analyze the data we have for accuracy and completeness.

1. Data profiling helps us make a thorough assessment of data quality.
2. It assists the discovery of anomalies in data.
3. It helps us understand
   content, structure, relationships, etc. about the data in the data source we are analyzing.

Contd…

4. It helps us know whether the existing data can be applied to other areas or purposes.
5. It helps us understand the various issues/challenges we may face in a database project much before the actual work begins. This enables us to make early decisions and act accordingly.
6. It is also used to assess and validate metadata.

When and how to conduct data profiling?

Generally, data profiling is conducted in two ways:

1. Writing SQL queries on sample data extracts put into a database.
2. Using data profiling tools.

When to conduct Data Profiling?

-> At the discovery/requirements gathering phase

-> Just before the dimensional modeling process

-> During ETL package design.

How to conduct Data Profiling?

Data profiling involves statistical analysis of the data at source and the data being loaded, as well as analysis of metadata. These statistics may be used for various analysis purposes. Common examples of analyses to be done are:

Data quality: Analyze the quality of data at the data source.

NULL values: Look out for the number of NULL values in an attribute.

Candidate keys: Analysis of the extent to which certain columns are distinct will give developer useful information w. r. t. selection of candidate keys.

Primary key selection: To check whether the candidate key column does not violate the basic requirements of not having NULL values or duplicate values.

Empty string values: A string column may contain NULL or even empty sting values that may create problems later.

String length: An analysis of largest and shortest possible length as well as the average string length of a sting-type column can help us decide what data type would be most suitable for the said column.

Identification of cardinality: The cardinality relationships are important for inner and outer join considerations with regard to several BI tools.

Data format: Sometimes, the format in which certain data is written in some columns may or may not be user-friendly.

Common Data Profiling Software

Most of the data-integration/analysis soft-wares have data profiling built into them. Alternatively, various independent data profiling tools are also available. Some popular ones are:

- Trillium Enterprise Data quality
- Datiris Profiler
- Talend Data Profiler
- IBM Infosphere Information Analyzer
- SSIS Data Profiling Task
- Oracle Warehouse Builder

Thanks…