# Chapter 10 – Data Quality and Integration

1. Define each of the following key terms:

a. *Static extract.* A method of capturing a snapshot of the required source data at a point in time

b. *Incremental extract.* A method of capturing only the changes that have occurred in the source data since the last capture

c. *Data steward.* A person assigned the responsibility of ensuring that organizational applications properly support the organization’s enterprise goals

d. *Master data management.* The disciplines, technologies, and methods to ensure currency, meaning, and quality of reference data within and across various subject areas

e. *Refresh mode.* An approach to filling the data warehouse that employs bulk rewriting of the target data at periodic intervals.

2. Match the following terms and definitions:

a data transformation

b data scrubbing

c selection

d data steward

e changed data capture

3. Contrast the following terms:

a. *Static extract; incremental extract.* A static extract provides a capture of the source data at a point in time, while an incremental extract provides the changes that have occurred in the source data since the last capture.

b. *Data scrubbing; data transformation.* Data scrubbing uses pattern recognition and other artificial intelligence to upgrade the quality of raw data before they are transformed and moved to the data warehouse. Data transformation converts data from the format of the source operational systems to the format of the enterprise data warehouse.

c*. Consolidation; federation.* Consolidation is a technique for bringing together many disparate data sources into a single, authoritative source for data that support decision making. Consolidation is typified by the extract-transform-load (ETL) process used by most data warehouse technologies. Federation is a technique for data integration that provides a virtual view of integrated data without actually creating one centralized database.

d. *ETL; master data management.* ETL is the extract-transform-load process used by most data warehouse technologies to consolidate many data sources into a single, consolidated data warehouse. Master data management (MDM) refers to the disciplines, technologies, and methods to ensure the currency, meaning, and quality of *reference data* within and across various subject areas. MDM determines the best source for each piece of *reference data*, and then affords access to this “golden record” to the applications that need to use it. MDM uses a subset of integrated data while ETL usually provides much more integrated data in a data warehouse.

4. Key components of a data governance program:

a. sponsorship from both senior management and business units;

b. a data steward manager to support, train, and coordinate the data stewards

c. data stewards for different business units, data subjects, source systems, or combinations of these elements.

d. a governance committee, headed by one person, but composed of data steward managers, executives, and senior vice presidents, IT leadership, and other business leaders to set strategic goals, coordinate activities, and provide guidelines and standards for all data management activities

5. How data stewardship relates to data governance:

Data governance refers to high-level organizational groups and processes that oversee data stewardship across the organization. Data governance guides data quality initiatives, data architecture, data integration and master data management, data warehousing and business intelligence, and other data-related matters. Data stewardship refers to the role of individuals in the organization who are responsible for ensuring that organizational applications properly support the organization’s enterprise goals for data quality. Data stewards are held accountable for the quality of the data for which they are responsible. Thus, data stewardship is central to the success of data governance within an organization as data stewards are the actual “front-line” people who see that the data governance processes are undertaken on a daily basis.

**Alternate Solution**

Data governance is a comprehensive program in organizations that usually guides data quality initiatives, data architecture, data integration and master data management, data warehousing, business intelligence, and other data-related matters. Data governance oversees data stewardship in an organization. Data stewardship is a specialized task, whereas data governance exerts guidance over activities to develop a unified management of data across the enterprise and participated in by enterprise decision makers.

6. Importance of data quality:

According to Informatica (2005), data quality is important to:

a. Minimize IT project risk

b. Make timely business decisions

c. Ensure regulatory compliance

d. Expand the customer base

7. The effect of Sarbanes-Oxley Act on the need for organizations to improve data quality:

In a nutshell, the Sarbanes-Oxley Act (SOX) requires organizations to be accountable to their stakeholders regarding their operations and financial decisions; at the present time, this applies mainly to publicly-traded organizations. As some experts have indicated, information technologies and systems provide a means for organizations to comply with the various regulatory sections of SOX. Data are the heart and lifeblood of most information systems. In particular, various sections of the SOX act yield requirements for organizations to measure and improve metadata quality; ensure data security; measure and improve data accessibility and ease of use; measure and improve data availability, timeliness, and relevance; measure and improve accuracy, completeness, and understandability of general ledger data; and identify and eliminate duplicates and data inconsistencies. The SOX act could provide motivation for many organization executives to treat data quality as an important strategic and operational goal.

1. Characteristics of quality data:

*Uniqueness.* Each entity exists no more than once within the database and there is a key that can be used to uniquely access each entity.

*Accuracy.* The degree to which any datum correctly represents the real-life object it models.

*Consistency.* Values for data in one data set (database) are in agreement with the values for related data in another data set (database).

*Completeness.* All data that must have a value do have an assigned value.

*Timeliness* Meeting the expectation for the time between when data are expected and when they are readily available for use.

*Currency.* The degree to which data is recent enough to be useful.

*Conformance.* Whether the data is stored, exchanged, or presented in a format that is as specified by its metadata.

*Referential integrity.* Data referring to other data are unique and satisfy requirements to exist (that is, satisfies any mandatory one or optional one cardinalities).

9. Explain four reasons for poor data quality in organizations:

*External data sources.* Organizations continue to rely on many outside sources for data—Web forms, XML channels from business-to-business sources, and databases from external organizations (e.g., mailing lists, census data, etc.). Organizations have little to no control over the completeness, accuracy, timeliness, or compatibility of this data with their internal data. Thus, organizations experience issues with data quality in their internal systems when this external data is brought into the organization without some form of data quality audit or control.

*Redundant data storage and inconsistent metadata.* Often, organizations have data stored in many different formats across the organization: spreadsheets, desktop databases, legacy databases, data marts, data warehouses, and other data repositories. This varied data may be redundant, inconsistent, and incompatible. If the metadata are wrong (For example, a bad algorithm or formula in a spreadsheet), the data will be wrong as well.

*Data Entry*. Organizations have not always taken advantage of placing integrity controls, valid value controls, and other data quality controls within the database definitions. Thus, the control for data integrity has typically fallen to the user interface of various applications, which may or may not take advantage of automatically filling in stored data or using drop-down selection boxes.

*Lack of organizational commitment*. Organizations have failed to recognize the importance of data quality or doubt that there will be a positive return on investment if efforts to improve data quality are undertaken. Thus, organizations have not made the commitment or invested resources to address data quality in organizational systems.

10. Key steps to improve data quality:

a. Get the business buy-in

b. Conduct a data quality audit

c. Establish a data stewardship program

d. Improve data capture processes

e. Apply modern data management principles and technology

f. Apply TQM principles and practices