Chapter-4

Enforcing Data Quality, Extending SQL Server Integration Services

❖Introduction To Data Quality

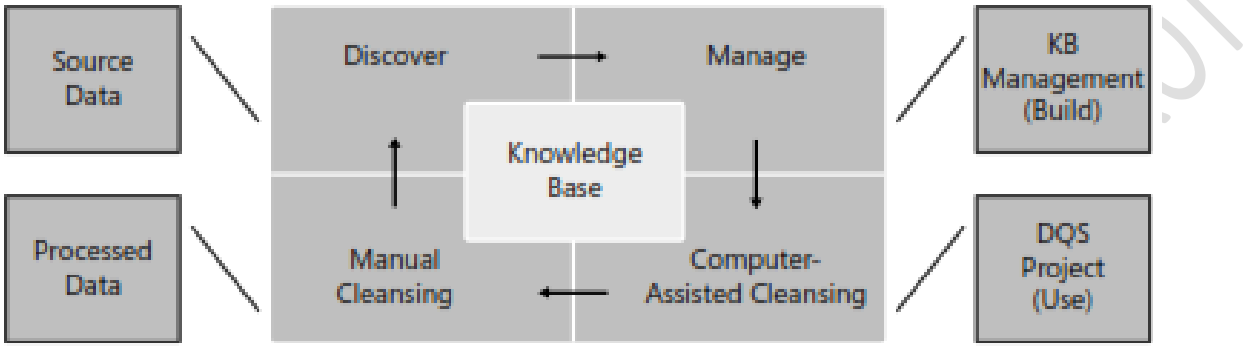
- Data quality refers to the condition of a set of values quantitative or qualitative variables.
- Data is generally considered high quality if it is used in operations, decision making & planning.
- Data quality is way of measuring data properties from different perspectives it is a comprehensive of the application efficiency & reliability & fines of data an especially data residing in a data warehouse.
- Data Quality Assurance (DQA) is a procedure intend to verify the influence & reliability of data.
- Data quality refer to the overall utility of a data set its ability to be easily processed & analyzed for other used usually by a database, data warehouse or data analytics system.
- Data quality has the following dimension:
  1) **Accuracy.**
  2) Completeness.
  3) Consistency.
  4) Integrity.
  5) Reasonability.
  6) Uniqueness.
  7) Validity.
  8) Accessibility.
  9) Timeness.

- The repeat on data quality of data warehouse estimates that data quality problems cause business more than 600 million dollars per year (US).
- The finding of the report based on the interview with industry expert, many customer & survey data from responses.
- Data Quality Management (DQM) community there is a generally held view that quality of a data set is depend on whether it meets defined requirements.
- SQL Server 2012 is used to Data Quality Services (DQS) & Master Data Service (MDS).
- You can solve data quality problems in a proactive way.

❖Using Data Quality Service To Cleanse Data

- One of the most powerful option in the MDS Master Data Service for excel is the ability to duplicate data by using data quality services.
- You can use this option if your data quality services instance is installed on the same SQL Server instance as master data services.
- Data cleansing is performed on your source data using a bowleg base that has been built in DQS assigns a high quality data set.
- A SQL Server data quality service is a bowleg driven data quality product aim at the data if professional to improve the quality of their business data.
- The data cleansing process are using the data quality tools.
- DQS allows creating a knowledge base by discovering & managing the information about the data we will use the bowleg base for cleansing data.
- The data cleansing is that incorrect value should be corrected & incomplete values should be made complete.
- A DQS knowledge base must be available on data quality server again which you want to compare & cleansing your source data.
- DQS uses knowledge base automatic & computer assisted data cleansing.

- After the automatic process is done you can manually review & additionally edit the process data.



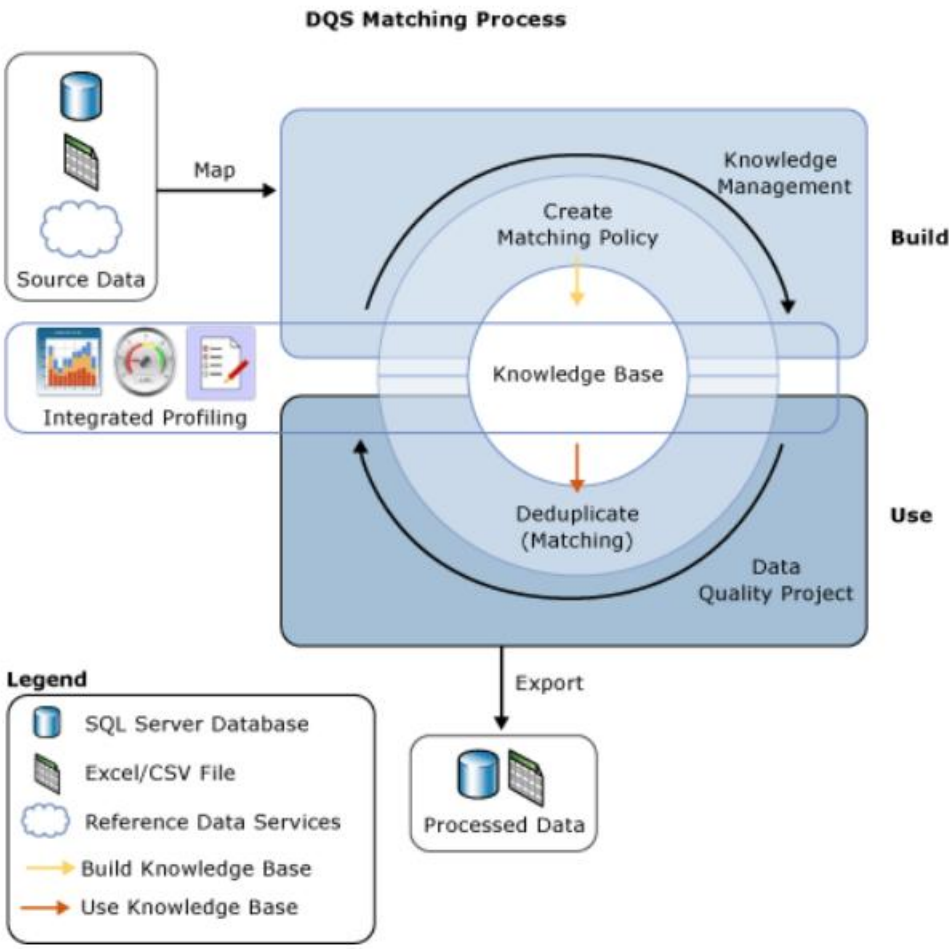
- You can use SQL Server or excel data source if the source data is in the source data is in an excel file, then excel must be installed on the same computer data quality client.
- A DQS knowledge base must exist before you can start a DQS cleansing or matching project.
- **For Example:**
  - ✓ If you are cleansing company name the knowledge base you use should have high quality data about company name.
  - ✓ A KB used for cleansing company name could have synonyms & turn based relations define.
  - ✓ A DQS project uses a single KB multiple projects can use the same KB (Knowledge Base).
- A cleansing process has the following stage:

1) Mapping.
2) Computer Assisted Cleansing.
3) Interactive Cleansing.
4) Export.

- After you have a knowledge base you can use DQS process to validate & cleanse your data.

❖Using Data Quality Service To Match Data

- Data Quality Service Data Matching Process enables you to reduce duplication & improve data accuracy in a data source.
- Matching analyzes the degree of duplication in all records of a single data source, returning weighted probabilities of a match between each set of records compared.
- Matching enables you to eliminate differences between data values that should be equal, determining the correct value and reducing the errors that data differences can cause.
- For example, names and addresses are often the identifying data for a data source, particularly customer data, but the data can become dirty and deteriorate over time.
- Performing matching to identify and correct these errors can make data use and maintenance much easier.
- DQS enables you to create a matching policy using a computer-assisted process, modify it interactively based upon matching results, and add it to a knowledge base that is reusable.



- A matching policy consists of one or more matching rules that identify which domains will be used when DQS assesses.
- The matching policy activity in the Knowledge Base Management wizard analyzes sample data by applying each matching rule to compare two records at a time throughout the range of records.
- The matching policy is run on domains mapped to the sample data.
- A knowledge base is available for use in a data quality project only when it is published.
- The matching rules in it cannot be changed by a user other than the person who created it.

- DQS performs data de-duplication by comparing each row in the source data to every other row.
- Using the matching policy defined in the knowledge base, and producing a probability that the rows are a match.
- A data matching project consists of a computer-assisted process and an interactive process.
- DQS performs the matching analysis, it creates clusters of records that DQS considers matches.
- You can export the results of the matching process either to a SQL Server table or a .csv file.

❖Using Script In SSIS

- The script task is ability to create a variable value from the SSIS package into the script & other than write a message out of the script task.
- The script task can interact with SSIS variable. You can use .net code to manipulate & respond to variable values.
- The SSIS script allows you to add functionality to your SSIS package that does not already exists with the other pre-defined task.
- SSIS script task one of the most interesting tools to increase SSIS capability.
- The script task you can program new functionality using C# & VB.
- The script task & script command have to design in time mode.
- The script task provides the entire required infrastructure for the custom code for you,letting you focus exclusively on the code.
- The Script Task Editor exposes property expressions on the Expressions page as other taskeditors do.
- That you need to list the SSIS variables you want to use in your script in theReadOnlyVariablesand ReadWriteVariables properties of the script task.

- The script task supports Microsoft Visual Basic & C# language.
- The script task provides code to perform function that are not available in the built in task& transformation that SQL Server Integration Service provides.
- The script task uses Microsoft Visual Studio tools for Applications (VSTA).
- As the environment in which you write the script & the engine that run those scripts.
- The script task & script component have greatly increase your possibility when it comes to script base ETL development in SSIS.
- ETL developers have a creative way of handling logic in their package.

❖Using Custom Component In SSIS

- Script task & component you can implement custom programmatic logic in SSIS package quality & efficiency.
- The definition of a script task or component is embedded in the definition of the SSIS package itself.
- In custom task & components can be developed, deployed & maintained independently in the SSIS package.

❖ Planning Of Custom Components

- SSIS data flow components, and after you have determined that due tocomplexity, dependency, or reusability requirements.
- The following guidelines to plan the design of thecustom component:

  1) Role.

  2) Usage.

3) Access To External Data.
4) Behavior.
5) Configuration.

- **Role:**

- Acustom source would be needed if none of the existing sources support thespecific connection manager that you are using, or if an appropriate connection manageris not available.
- If the source data is extracted from an incompatiblesource or is stored in an incompatible format, you could develop a custom data source.

- **Usage:**

- Source or transformation component going to use multiple outputs& a transformation or destination component going to use multiple inputs.
- A source component accessing a compositedata set could be programmed to produce multiple, normalized row sets, instead of asingle de-normalized one.

- **Access To External Data:**

- If the component will perform lookup operations or will need to access data that isnot available in the current data flow, it will require access to external data sources.
- Toaccess data stored in variables or parameters, the component will also need access tothose variables and parameters.

- **Behavior:**

- If the component is going to pass rows to the destination without having to retainthem, such as to calculate running totals (partially blocking), or to sort them (blocking),the component will not block the data flow.
- New rows cannot be added to a synchronous output and cannot be removed from it.

➢ Configuration:

- To improve the reusability of a custom component, specific settings used to control itsoperation should be exposed, allowing the developer to set them at design time, oreven expose them to the environment.
- Typically, custom components are developed separately from SSIS packages.
- If copy or paste the code from a package to another package you can create a custom component.
- There are two group of methods to customize & design custom components:
  1) **Design-Time Methods.**
  2) **Run-Time Methods.**

➢ Design-Time Methods

- Design-time methods facilitate the interaction between the SSIS developer and the dataflow component, allowing the component to be placed in the data flow and configuredappropriately.

➢ Run-Time Methods

- Run-time methods represent the component's operational programmatic logic, which isexecuted when the component is used at run time.

- These methods provide the completeoperational capabilities of the component; without them, the component performs no actionsother than being validated and returning a validation result.
- The run-time methods in this section are listed in the order of operation—from validation,through row processing, to cleanup.