

Data Quality and Management

Prepared by Dr. Amjad Qtaish

DS601
Data Mining

Lecture 05



Lecture: #05

Data Quality and Management



Dr. Amjad Qtaish

Faculty of Informatics
Midocean University



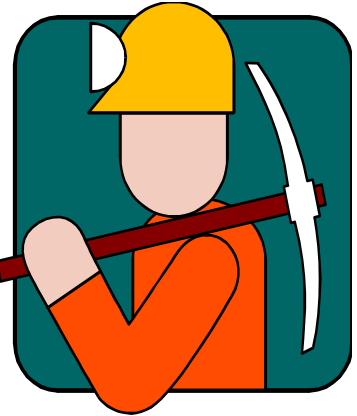
Data Mining (DS601)



Data Quality and Management

- Data, Information, and knowledge
- What is Data Quality ?
- Data Quality Dimensions, Rules, Process, Roles
- Data Quality Example
- Data Quality Technology Support
- Summary





- **Data**

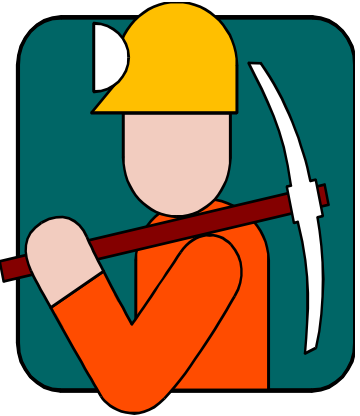
- Items that are the most elementary descriptions of things, events, activities, and transactions
- May be internal or external

- **Information**

- Organized data that has meaning and value

- **Knowledge**

- Processed data or information that conveys understanding or learning applicable to a problem or activity



- Raw data collected manually or by instruments
- Quality is critical
 - Quality determines usefulness
 - Contextual data quality
 - Intrinsic data quality
 - Accessibility data quality
 - Representation data quality
 - Often neglected or casually handled
 - Problems exposed when data is summarized



TABLE 5.1 **Data Problems**

<i>Problem</i>	<i>Typical Cause</i>	<i>Possible Solutions</i>
Data are not correct.	Data were generated carelessly. Raw data were entered inaccurately. Data were tampered with.	Develop a systematic way to enter data. Automate data entry. Introduce quality controls on data generation. Establish appropriate security programs.
Data are not timely.	The method for generating data is not rapid enough to meet the need for data.	Modify the system for generating data. Use the Web to get fresh data.
Data are not measured or indexed properly.	Raw data are gathered inconsistently with the purposes of the analysis. Use of complex models.	Develop a system for rescaling or recombining improperly indexed data. Use a data warehouse. Use appropriate search engines. Develop simpler or more highly aggregated models.
Needed data simply do not exist.	No one ever stored data needed now. Required data never existed.	Predict what data may be needed in the future. Use a data warehouse. Generate new data or estimate them.

Source: Based on Alter (1980), p. 130. Alter, S. L. (1980). *Decision Support Systems: Current Practices and Continuing Challenges*. Reading, MA: Addison-Wesley.





Data Quality and Management

- Data, Information, and knowledge
- What is Data Quality ? 
- Data Quality Dimensions, Rules, Process, Roles
- Data Quality Example
- Data Quality Technology Support
- Summary

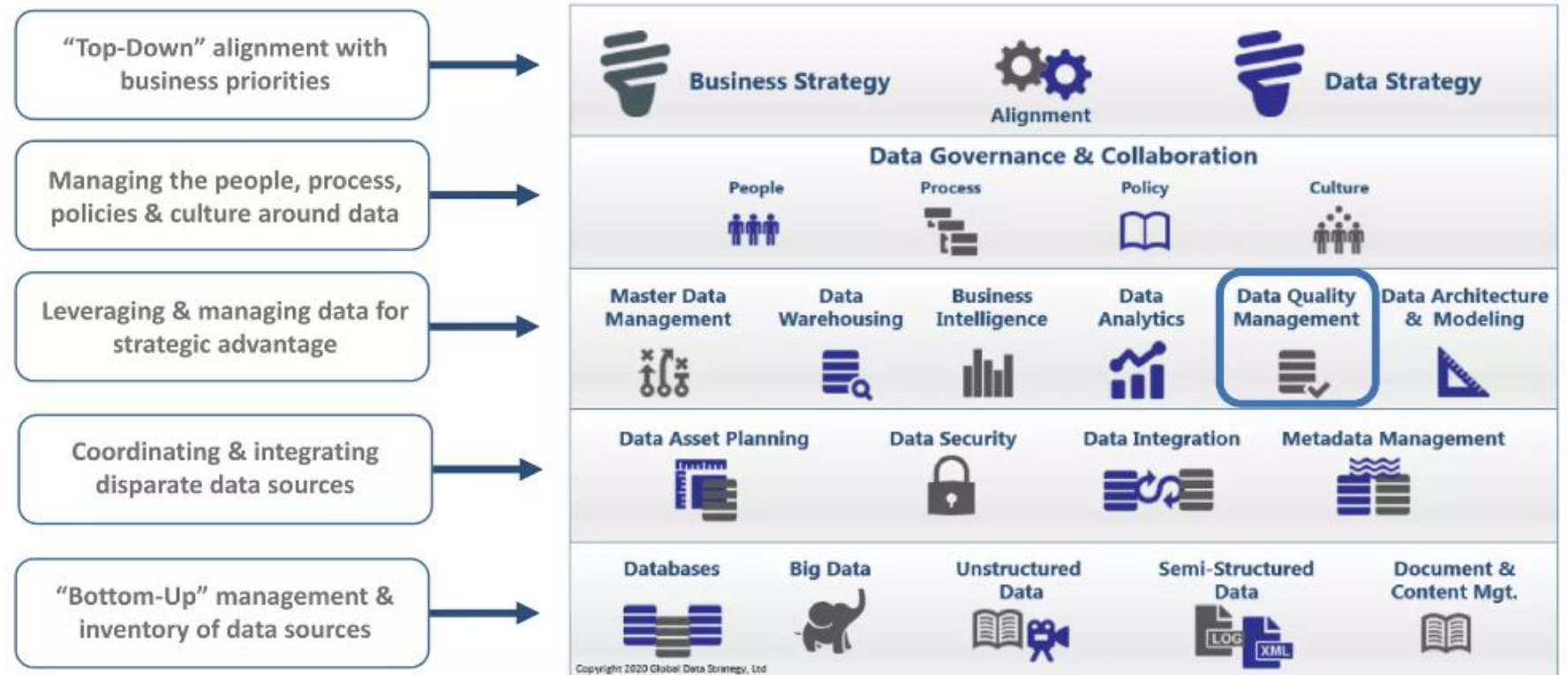




What is Data Quality ?

Data Quality is Part of a Wider Data Strategy

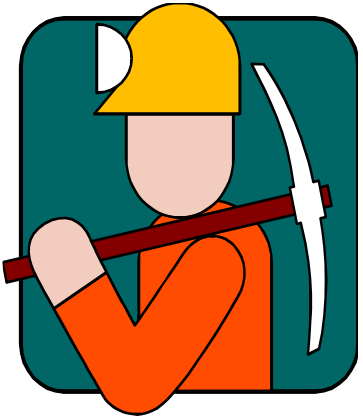
A Successful Data Strategy links Business Goals with Technology Solutions



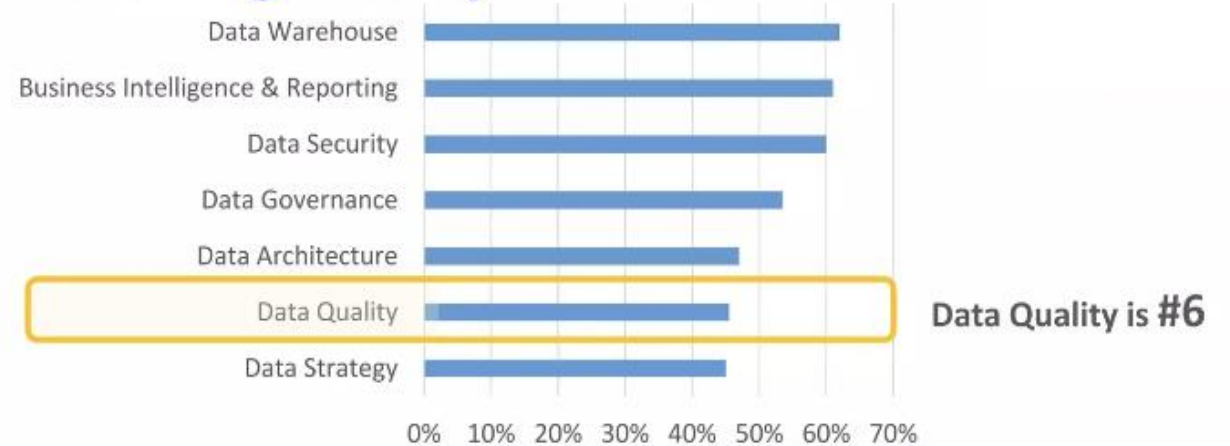


What is Data Quality ?

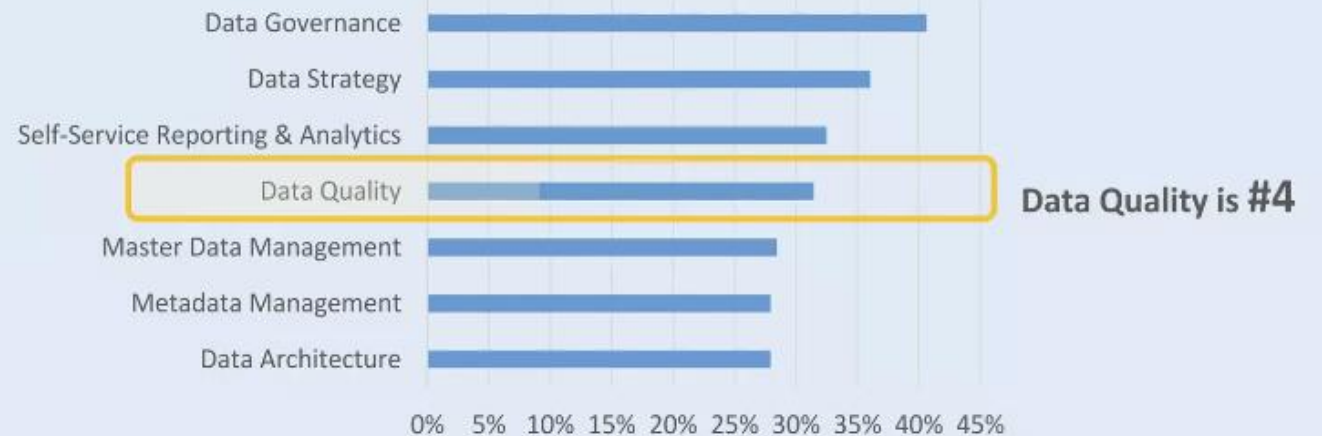
Data Quality Remains a Growing Priority



Data Priorities in Place for 2021



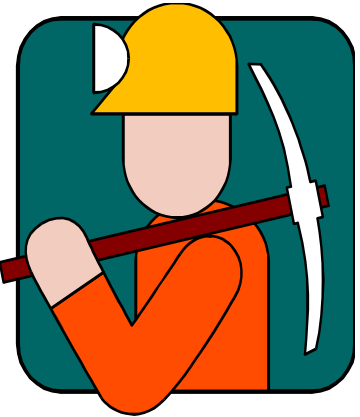
Future Data Priorities in 2022-2023





What is Data Quality ?

Data Quality refers to the **methodical approach, policies and processes** by which an organization manages the **accuracy, validity, timeliness, completeness, uniqueness, and consistency** of its data in systems and data flows.



The subject of Data Quality
is Critical Data Element



CDE Definition:
"The day on which a person is born"

Is it **accurate**?

Is it **valid**?

Is it on **time**?

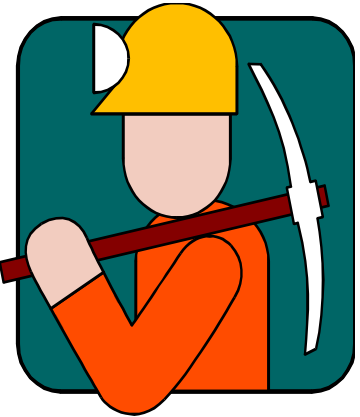
Is it **complete**?

Is it **unique**?

Is it **consistent**?

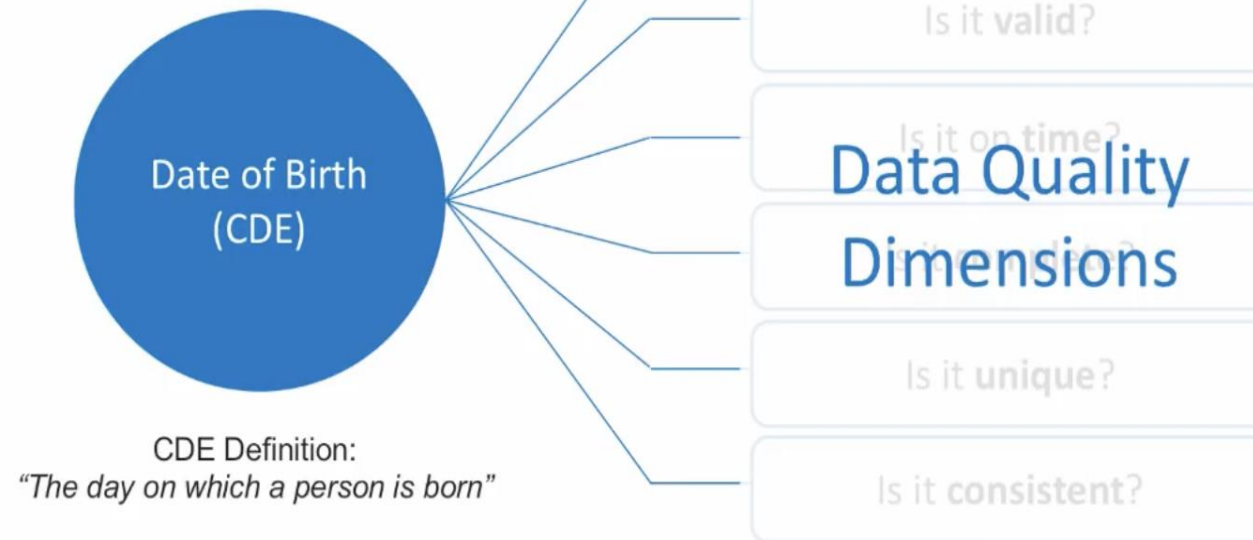


What is Data Quality ?



Data Quality refers to the **methodical approach, policies and processes** by which an organization manages the **accuracy, validity, timeliness, completeness, uniqueness, and consistency** of its data in systems and data flows.

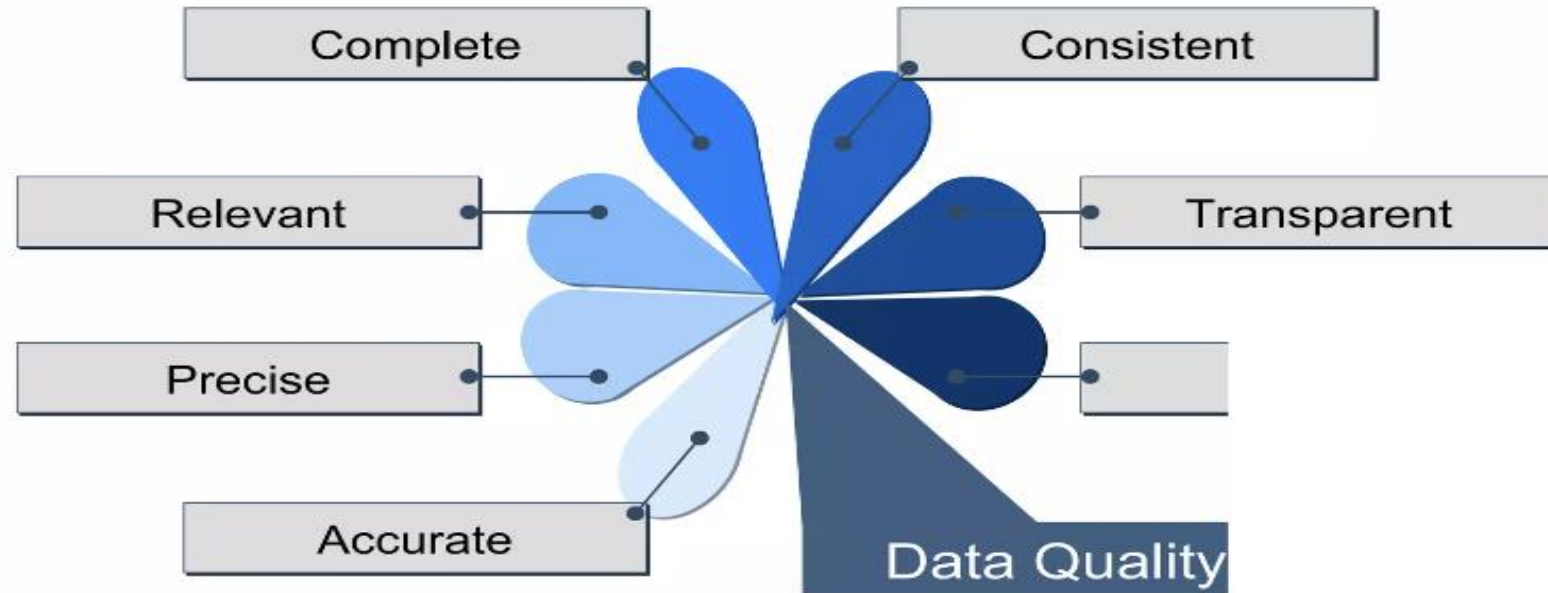
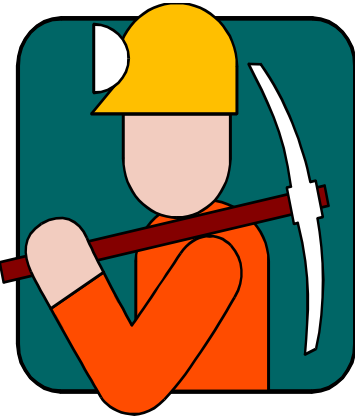
The subject of Data Quality
is Critical Data Element



Not all Data Quality
Dimensions are
applicable on
particular CDE (e.g.
Date of Birth can be
assessed against
Validity and
Completeness only)



What is Data Quality ?



Data Quality stands for:



- 1 Harmonized information need and provision
- 2 Mutual understanding of data capability
- 3 Trustworthy and credible information



What is Data Quality ?

Accuracy stands for:



- Good fit between the data and reality
- The ability to draw correct conclusions from data
- Business processes that match reality



Examples for Data Accuracy issues:

- Frank Meyer is recorded as "Fritz Meier" in the Database.
- An incident is reported with €23m when the loss was €12k.
- The amount invoiced does not represent the customer's usage.



Accurate

Data Accuracy is the degree at which a data object overlaps with the real world object or event described.

Data accuracy is measured as *reciprocal maximum gap* between data and reality. [high is good]



What is Data Quality ?

Precision stands for:



- A close link between desired and offered information
- The ability to pinpoint decisions based on data.
- Lean Business processes.



Examples of Data Precision issues:

- Frank Meyer lives in Bonn - or Cologne? Or was that Jon Myers?
- This Billing incident was caused by Mediation... I think...
- Why do we charge the customer 2 minutes for a 59sec call?



Precise

Data Precision is the closeness between
all possible interpretations of a data object.

Data precision is measured as *reciprocal maximum distance*
between all applicable data interpretations. [high is good]



What is Data Quality ?

Relevance stands for:



- Data that helps you know what you want.
- The ability to use data with maximum efficiency.
- Not having to sort through information you don't need.

Examples of Data Relevance Issues:

- The Revenue Assurance report also tells you about the weather!
- A CSR asks the cell phone customer if they have a microwave.
- You need to fill in a 7-page form to apply for a tariff change.



Relevant

Data Relevance is the closeness between data consumer need and data provider output.

Data relevance is measured as *percentage* of all data required divided by all data provided. [100% is best]



What is Data Quality ?

Completeness stands for:



- Data that does not leave any open questions.
- The ability to make a good decision based on available data.
- Closeness between “need to know” and what the data tells you.



Examples of Data Completeness issues:

- We can not tell how many cell phone contracts Egon Huber has.
- The CC application does not provide a “Call back wanted” field.
- A summary report includes projects that did not report status!



Complete

Data Completeness is the extent by which the data consumer's need is met.

Data completeness is measured as *percentage* of data available divided by the data required. [100% is best]

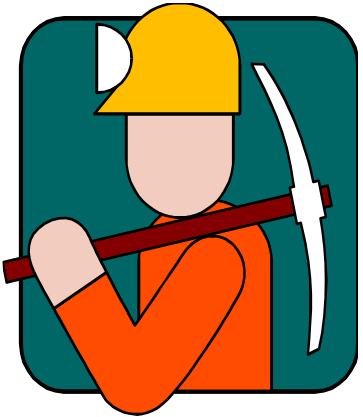


What is Data Quality ?

Consistency stands for:



- Data in harmony across the company.
- The ability to trust in data regardless of source.
- Identical information available to all processes and units.



Examples of Data Consistency Issues:

- We send Mr. Smith's invoices to "Smith" and ads to "Schmitz".
- Asking DWH or SAP for revenue yields different numbers.
- Mr. Kim defines "churn" as *cancel/total* and Mr. Jones as *cancel/new*.



Consistent

Data Consistency is the synchronization of data objects across the company.

Data consistency is measured as *reciprocal ratio* of distinct data objects per described object or event. [100% is best]

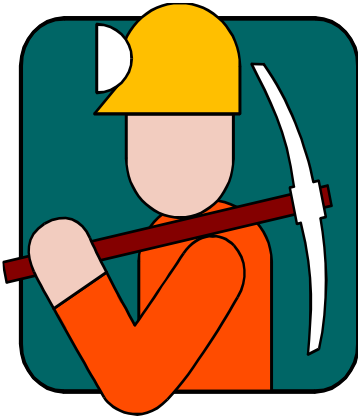


What is Data Quality ?

Transparency stands for:



- Trustworthy data in the entire data supply chain.
- The ability to connect data with its real meaning.
- Real accountability for data objects.



Examples of Data Transparency issues:

- We can't tell why Frank Müller is now "Udo Huber" in the DB!
- A report contains a figure which nobody can explain.
- Project leaders get away with reporting "green" when it's "red"!



Transparent

Data Transparency is the ability to trace back data to its origin and find out its real world meaning.

Data transparency is measured as *percentage* of maximum traceable distance by total processing steps. [100% is best]

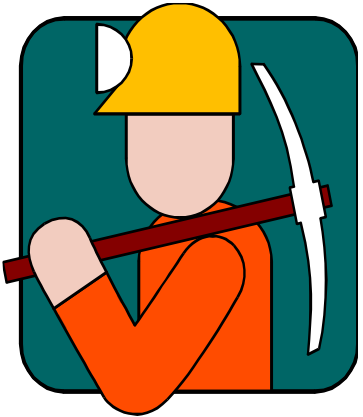


What is Data Quality ?

Timeliness stands for:



- Data that is available without delay.
- The ability to know what you need, when you need.
- Smooth Information Flow: 'Data Delayed' is 'Data Denied'!



Examples of Data Timeliness Issues:

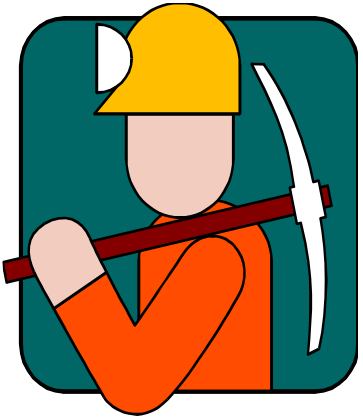
- The agenda is distributed during the Telco!
- Customers decide for a competitor before credit is approved!
- Receiving a "budget exceeded" SMS *after* you went over the limit!



Timely

Data Timeliness is the availability of data at the time it needs to be utilized.

Data timeliness is measured as *percentage* of processing time attributed to waiting for data. [0% is best]



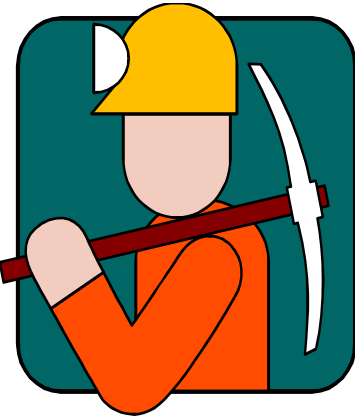
Data Quality Dimension refers to the an aspect or feature of information that can be assessed and used to determine quality of data.

6 Key Data Quality Dimensions

Accuracy	Validity	Timeliness	Completeness	Uniqueness	Consistency
Data accurately represent the “real-world” values.	Data conforms to the syntax (format, type, range) of its definition.	Data represents reality from the required point of time.	Data are complete in terms of required potential of data.	Data are properly identified and recorded only once.	Data are represented consistently across the data set.
Example: Incorrect spellings of product or person names, addresses.	Example: Incorrect classification values for gender or customer type	Example: Customer address change which is effective on July 1 st is entered into the system in July 15 th	Example: Customer address missing a ZIP code	Example: Single customer is recorder twice in the database, with different identifiers	Example: Customer account is closed, but there is a new order associated to that account



Data Quality Rules refers to business rules intended to ensure quality of data in terms of **accuracy, validity, timeliness, completeness, uniqueness, and consistency.**



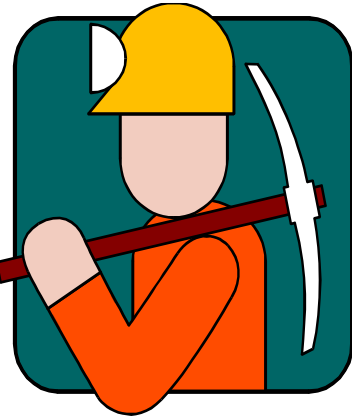
Validity: Date of Birth must be valid date in the range from 1900- Current Date

Completeness: Date Of Birth must be entered for each individual. Empty fields are not allowed.

CDE Definition:
"The day on which a person is born"



Data Quality Rules refers to business rules intended to ensure quality of data in terms of **accuracy, validity, timeliness, completeness, uniqueness, and consistency.**



CDE Definition:
"The day on which a person is born"

Validity: Date of Birth must be valid date in the range from 1900- Current Date

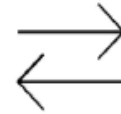
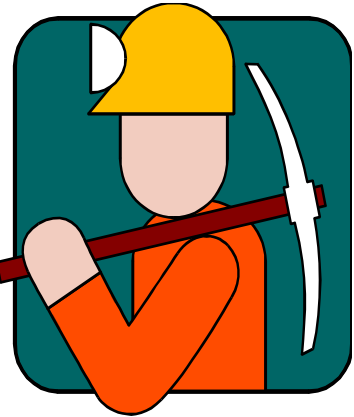
Data Quality Rules

Completeness: Date Of Birth must be entered for each individual. Empty fields are not allowed.

Each Data Quality Rule is associated to particular Data Quality Dimension. Multiple Data Quality Rules can be associated to one Data Quality Dimension



Data Quality Rules



Verify data types
and lengths



How should I
handle NULLs?



Domain value
constraints (ie. US
states)



Single source of
truth (master
data)



Late arriving
dimensions



Reference data
lookups



**De-Duplicate
Data**



**Descriptive data
statistics** (Length,
type, mean, median,
average, stddev ...)



**Frequency
distribution**



**How do I handle
missing values?**



Enumerations
(turn codes into
categorical data)



**Value
replacement**



Replacing Values

- Replace empty strings with nulls
 - `iif (length(title) == 0, toString(null()), title)`
- Replace nulls with values
 - `iif (isNull(customername), 'NA', customername)`
- Replace missing measurements with aggregates

All Functions Input schema Parameters

```

101 stddev(<numeric_value>)
102 stddevIf(<numeric_value> condition <numeric_value>)
103 stddevPopulation(<numeric_value>)
104 stddevPopulationIf(<numeric_value> condition <numeric_value>)
105 stddevSample(<numeric_value>)
106 stddevSampleIf(<numeric_value> condition <numeric_value>)
  
```

```

107 avg(<numeric_value>)
108 avgIf(<numeric_value> condition <numeric_value>)
  
```

Group by Aggregates

Grouped by: year

AverageRate toDecimal(avg(Rating), 10, 2)

```

109 mean(<numeric_value>)
110 meanIf(<numeric_value> condition <numeric_value>)
  
```



Multiple inputs

New Branch

Join

Conditional Split

Exists

Union

Lookup

Schema modifier

Derived Column

Select

Aggregate

Surrogate Key

Pivot

Unpivot

Window

Row modifier

Filter

Sort

Alter Row

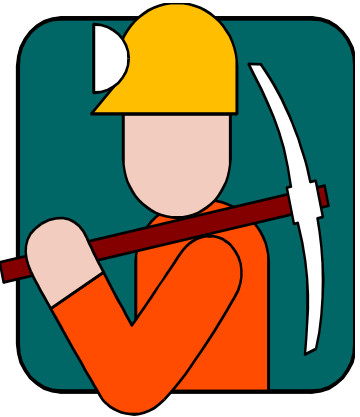
Destination

Sink



Splitting Data Based on Values

- Create separate paths to separate training data from testing data
- Separate rules can be applied to streams based on value conditions
- Create a full copy of data at any point
- Useful for both data value and metadata conditions



Conditional split settings Optimize Inspect Data preview

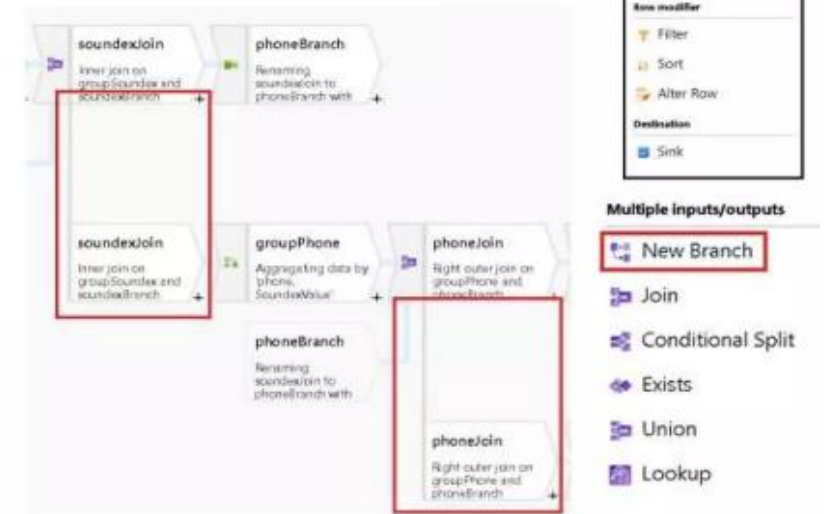
Output stream name * [Learn more](#)

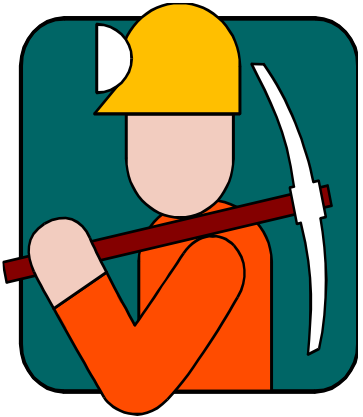
Incoming stream *

Split on: ☒ First matching condition ☐ All matching conditions

Split condition:

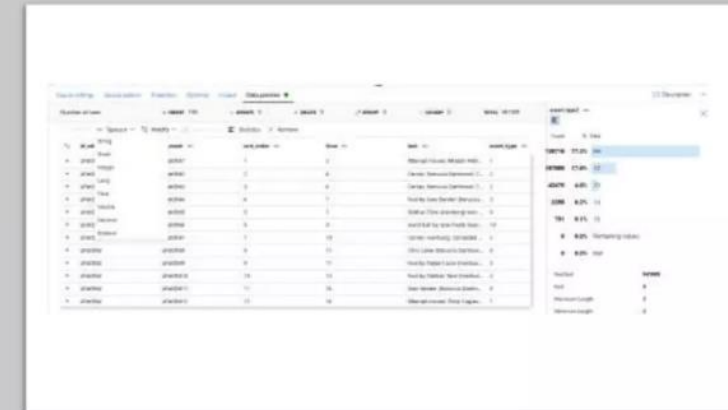
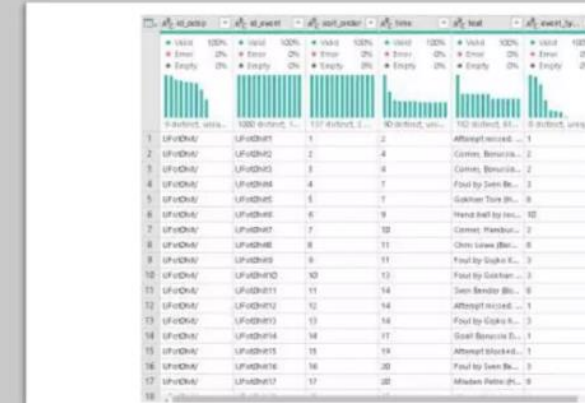
STREAM NAME	CONDITION
badSchema	isNotNullByValue('ProductID')
goodSchema	Rows that do not meet any condition will use this output stream

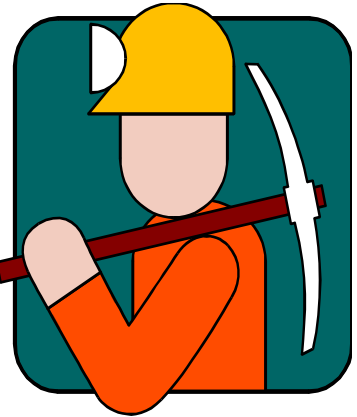




Data Profiling

- Summary statistics describing the shape, size, content of your data
- Helps you to understand what is inside your data
- As a Data Engineer, you have the responsibility to provide proper data for analytics and models
- For big data sets, use sampling / good enough approach
- View data statistics at any step in your data transformation
- Here is how to persist your data profile stats:
<https://techcommunity.microsoft.com/t5/azure-data-factory/how-to-save-your-data-profiler-summary-stats-in-adf-data-flows/ba-p/1243251>





Pattern Matching

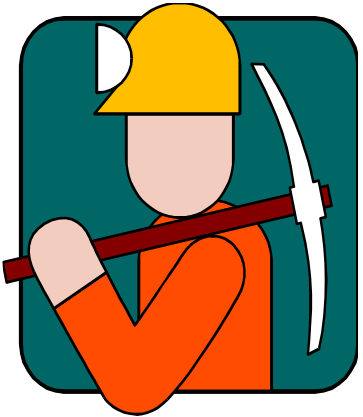
With very wide datasets or datasets where you cannot anticipate all column names, generate pattern-matching rules to apply data quality rules en masse without the need to name each column and describe each rule individually.



The screenshot displays a configuration window for a data quality rule, divided into two panels. The left panel shows a rule configuration for a stream named 'FishNames'. It has two tabs: 'Derived Column's Settings' and 'Optimizer'. Under 'Derived Column's Settings', the 'Output stream name' is 'RenameColumns' and the 'Incoming stream' is 'FishNames'. There are two rule entries under 'Columns':

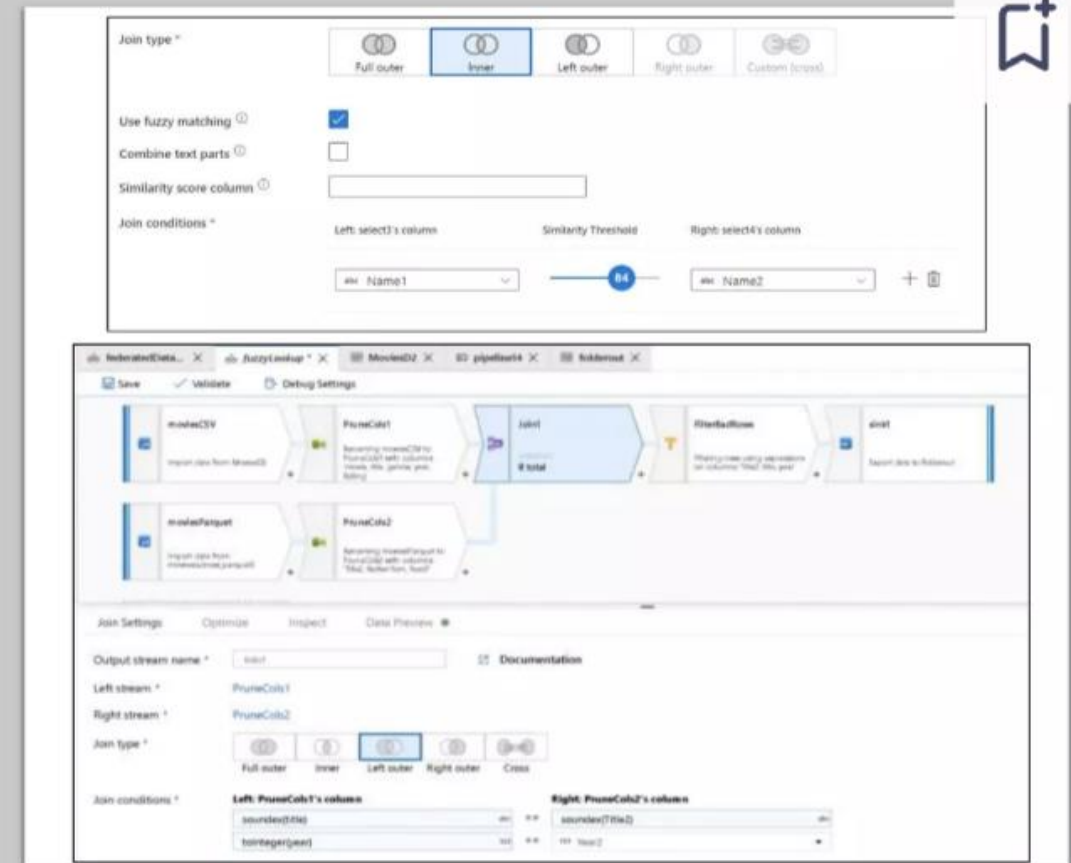
- Rule 1: 'Each column that matches: type = "string"' creates 1 column(s). The transformation is '\$\$ = trimmed' followed by 'trim(toString(\$\$))'.
- Rule 2: 'Each column that matches: type = "integer"' creates 1 column(s). The transformation is 'int_ = \$\$' followed by 'toInteger(\$\$)'.

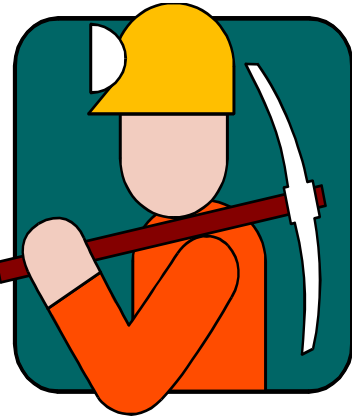
The right panel shows another rule configuration. It has a single rule entry: 'Each column that matches: position > 1 && position < 25' creates 1 column(s). The transformation is '\$\$' followed by 'avg(toInteger(\$\$))' with a value of '1.2'.



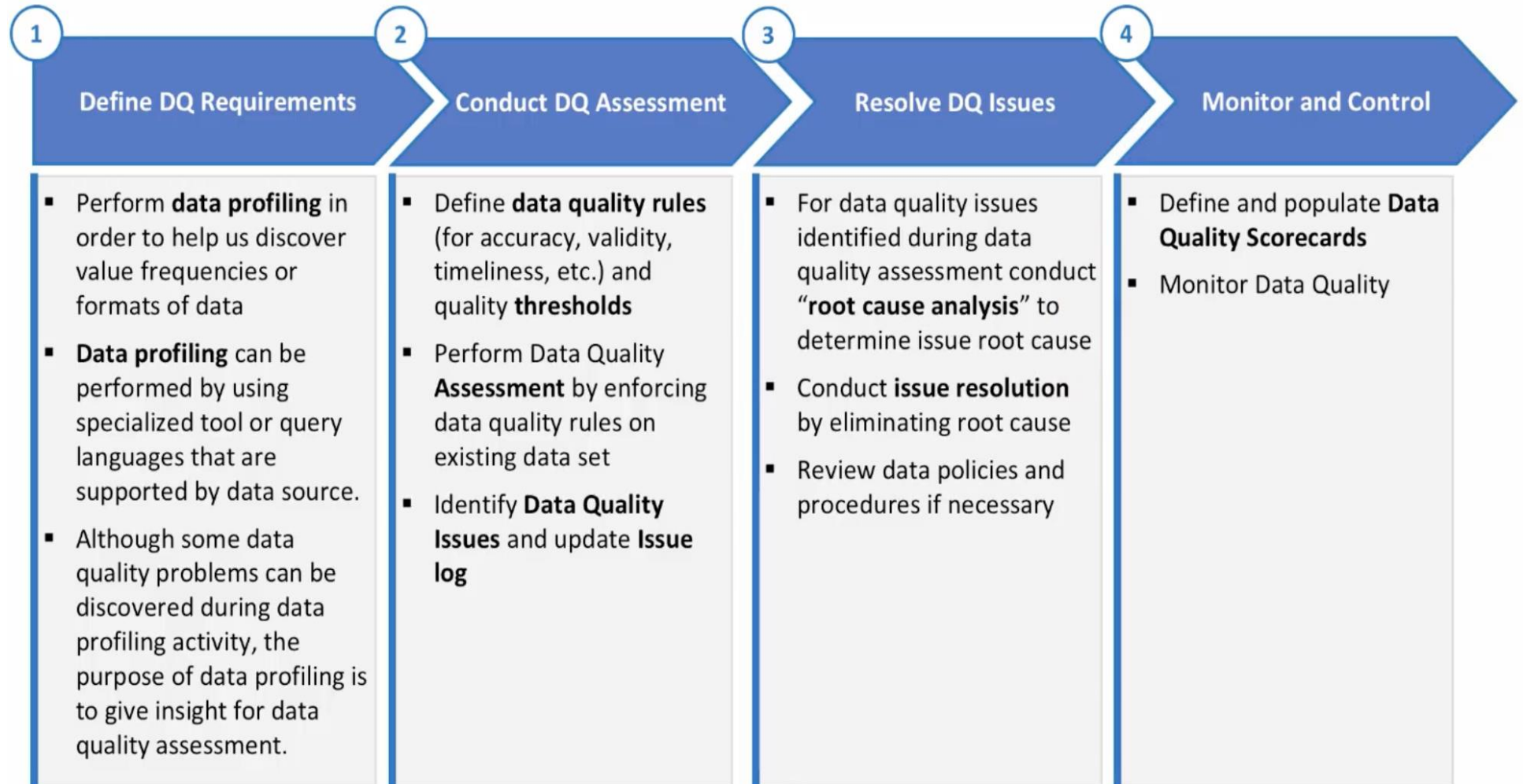
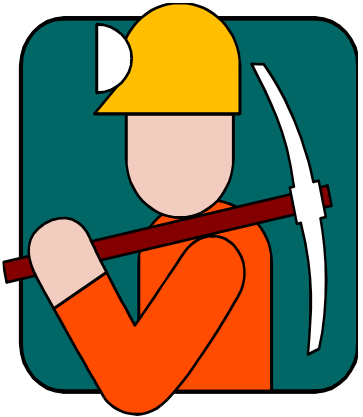
Fuzzy Lookups & Joins

- Sometimes when performing inline lookups, you don't have exact matches when looking for references
- Fuzzy Lookups with Soundex helps find matches based on phonetic algorithms
- Very useful in data lake scenarios where joins and lookups are against data that is not normalized or cleaned



**TABLE 7-2** Fuzzy Matching Options

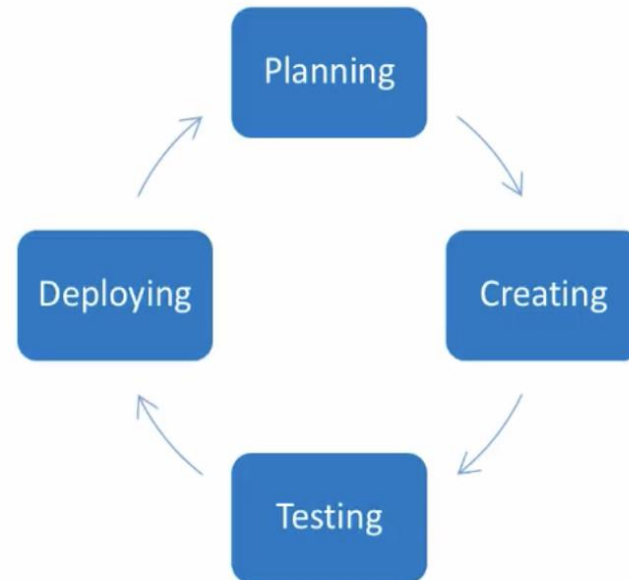
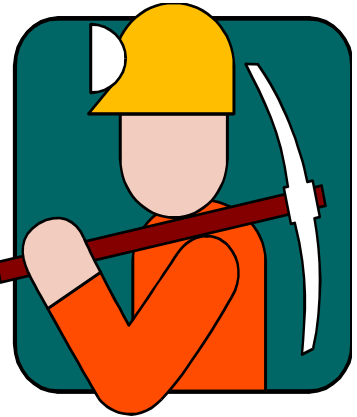
Fuzzy Matching Option	Description
Similarity threshold	Values are from 0 to 1. When values are 0, values are said to match each other, no matter how far apart they are. With 1, you get a match only when the match is exact.
Ignore case	Treats upper- and lowercase the same.
Maximum number of matches	Limits the number of rows from the second table that matches the first, which is helpful when the result set produces multiple matches.
Match by combining text parts	Attempts to combine separate words into a single entity, looking to find matches between keys.
Transformation table	Equivalent of a to-and-from, which means there must be at least two columns.





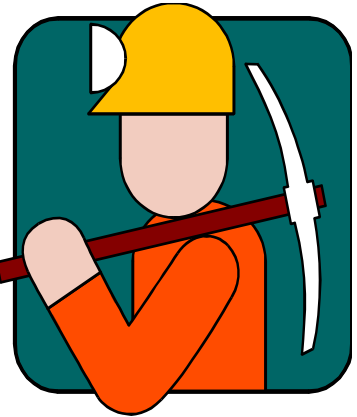
Data Quality process should be part of SDLC (System Development Live Cycle).

SDLC (System Development Life Cycle) refers to the process of process for planning, creating, testing, and deploying an information system.

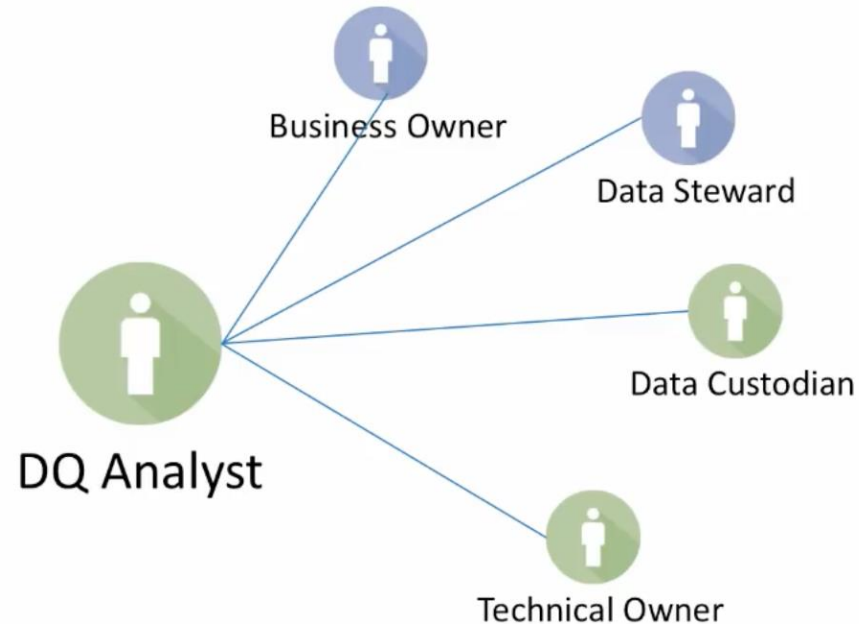




Data Quality Roles



Data Quality Analyst is the key role responsible to perform activities associated with the data quality process. Although this is the only role specific to Data Quality, Data Quality Analyst will closely work with Business Owners, Data Stewards, Technical Owners and Data Custodians. That includes, but not limited to: definition of data quality rules, analyze results of data quality profiling and assessments, investigating root causes for data quality issues, etc.



Business roles

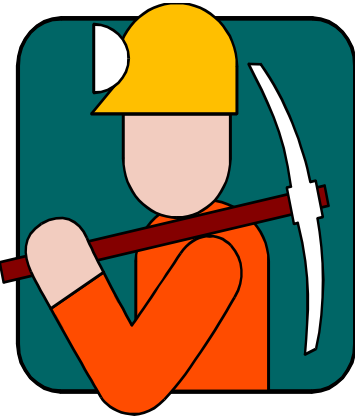


Technical roles



Use Case Definition

Conduct the Data Quality Assessment for Employee Data, specifically to the data element "Date of Birth". The threshold for quality of data is 99%, for each data quality dimension.



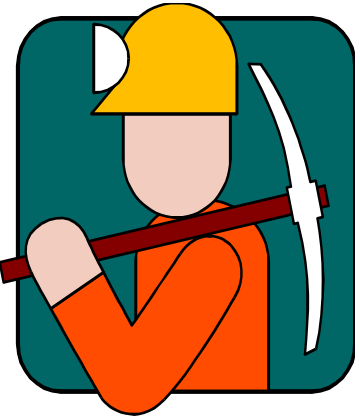
Validity: Date of Birth must be valid date in range from 1900-Current Date

Completeness: Date of Birth must be entered for each individual. No blanks allowed.

CDE Definition:
"The day on which a person is born"



Employee Data Profiling



Employee Data - Actual Data Set

ID	Employee Full Name	Date Of Birth
1	Michael Stanton	10/1/1968
2	Joe Irvine	12/3/1990
3	Jenifer Cipriani	7/24/1973
4	Salvatore Mendini	2/30/1968
5	Eva Carlos	
6	Courtney O'Brien	12/12/1981
7	Frank Damon	9/18/1983
8	Katherine La Sal	1/22/1991

Employee Data – Data Profiling Results

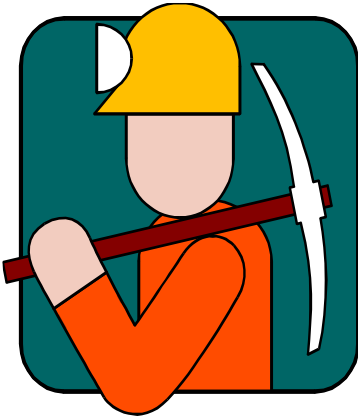
Category	Result
Number of Records	8
Number of Unique Values	7
Number of Blanks	1



Employee Data DQ Assessment

Data Quality Rules

- **Validity:** Date of Birth date must be valid date in range from 1900- Current Date
- **Completeness:** Date Of Birth must be entered for each individual. No blanks allowed.



Employee Data - Actual Data Set

ID	Employee Full Name	Date Of Birth
1	Michael Stanton	10/1/1968
2	Joe Irvine	12/3/1990
3	Jenifer Cipriani	7/24/1973
4	Salvatore Mendini	2/30/1968
5	Eva Carlos	
6	Courtney O'Brien	12/12/1981
7	Frank Damon	9/18/1983
8	Katherine La Sal	1/22/1991

Employee Data – Data Assessment Results

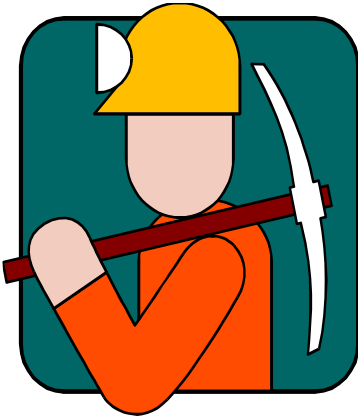
DQ Dimension	Rule	Passed	Failed	Score
Validity	Birth date must be valid date in range from 1900- Current Date.	6	2	75%
Completeness	Date Of Birth must be entered for each individual. No blanks allowed.	7	1	88%



Employee Data DQ Assessment

Data Quality Rules

- **Validity:** Date of Birth date must be valid date in range from 1900- Current Date
- **Completeness:** Date Of Birth must be entered for each individual. No blanks allowed.



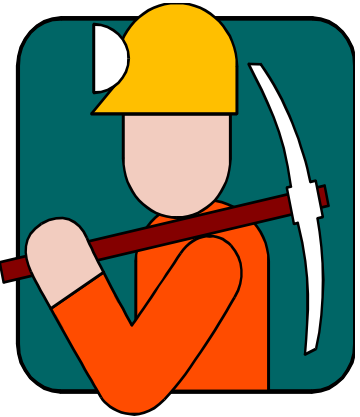
Employee Data - Actual Data Set

ID	Employee Full Name	Date Of Birth
1	Michael Stanton	10/1/1968
2	Joe Irvine	12/3/1990
3	Jenifer Cipriani	7/24/1973
4	Salvatore Mendini	2/30/1968
5	Eva Carlos	
6	Courtney O'Brien	12/12/1981
7	Frank Damon	9/18/1983
8	Katherine La Sal	1/22/1991

Employee Data – Data Assessment Results

DQ Dimension	Rule	Passed	Failed	Score
Validity	Birth date must be valid date in range from 1900- Current Date	6	2	75%
Completeness	Date Of Birth must be entered for each individual. No blanks allowed.	7	1	87.5%

Data Quality Scorecard



Employee Data Issue Resolution

Root Cause of DQ Issues:

Validity: Data entered in this field are not validated against the valid date format

Completeness: Date of Birth is not set mandatory field in database

Employee Data - Actual Data Set

ID	Employee Full Name	Date Of Birth
1	Michael Stanton	10/1/1968
2	Joe Irvine	12/3/1990
3	Jenifer Cipriani	7/24/1973
4	Salvatore Mendini	2/30/1968
5	Eva Carlos	
6	Courtney O'Brien	12/12/1981
7	Frank Damon	9/18/1983
8	Katherine La Sal	1/22/1991



Employee Data Issue Resolution

Root Cause of DQ Issues:

Validity: Data entered in this field are not validated against the valid date format

Completeness: Date of Birth is not set mandatory field in database

Employee Data - Actual Data Set

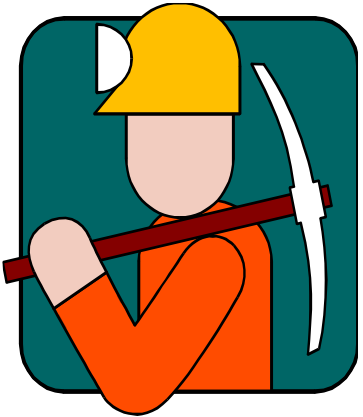
ID	Employee Full Name	Date Of Birth
1	Michael Stanton	10/1/1968
2	Joe Irvine	12/3/1990
3	Jenifer Cipriani	7/24/1973
4	Salvatore Mendini	2/23/1968
5	Eva Carlos	4/14/1971
6	Courtney O'Brien	12/12/1981
7	Frank Damon	9/18/1983
8	Katherine La Sal	1/22/1991

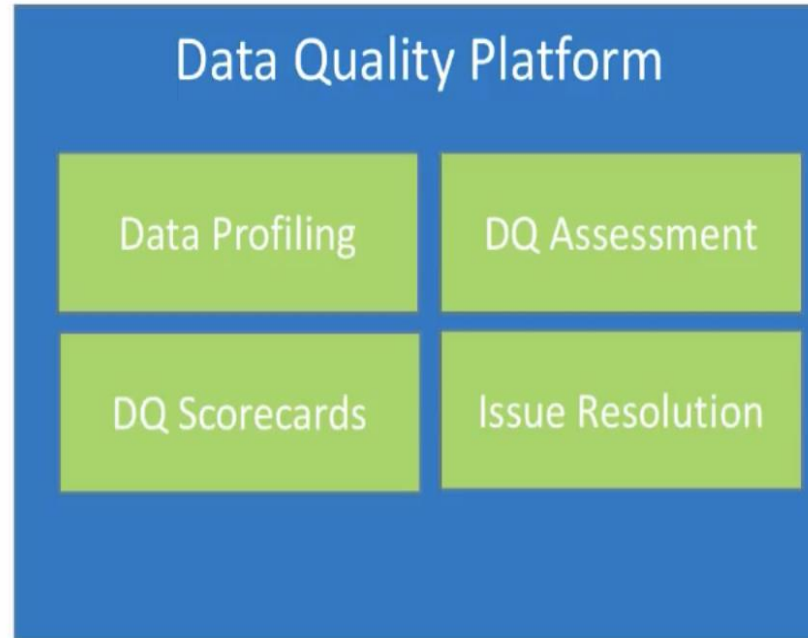
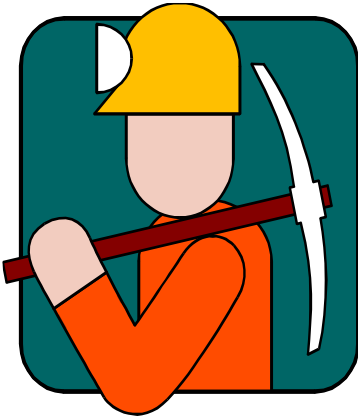
Employee Data – Data Assessment Results

DQ Dimension	Rule	Passed	Failed	Score
Validity	Birth date must be valid date in range from 1900- Current Date.	8	0	100%
Completeness	Date Of Birth must be entered for each individual. No blanks allowed.	8	0	100%

Corrective measures (to prevent further issues):

1. Implement validation control for “Date of Birth” field against the valid dates within defined range
2. Set “Date of Birth” field as mandatory in database



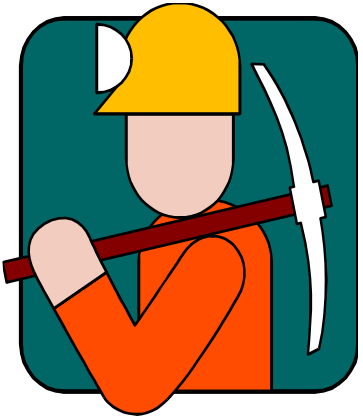


Key Metadata Tool Requirements

- Ability to conduct data profiling, including statistical analysis of data sets
- Ability to define and execute data quality rules for critical data elements which are subject of data quality check
- Ability to store data quality profiling and assessment results
- Ability to conduct issue resolution process and discover issue patterns
- Ability to create and visualize data quality scorecards



Summary



The subject of **Data Quality** is **Critical Data Element (CDE)**

Data Quality is about different aspects or features of CDEs. These aspects or features are also called **Data Quality Dimensions**

6 Key Data Quality Dimensions are:
accuracy, validity, timeliness, completeness, uniqueness, and consistency

Data Quality Process consist of 4 major steps: Define DQ Requirements, Conduct DQ Assessment, Resolve DQ Issues and Monitor and Control. All activities are performed by **Data Quality Analyst** in cooperation with business owner, data steward, technical owner and data custodian.

Data Quality Profiling is about getting statistical information about data

Data Quality Assessment is about to assess data quality and identify data quality issues. It's based on data **quality rules and thresholds**. Data Quality rules are associated with data quality dimensions. Data Quality Scorecards visualizes results of the assessment

Root Cause Analysis is about to find root cause of particular data quality issue

Data Quality Tools provide technology support to Data Quality process to ensure the quality of data in accordance to business requirements



Thank You