JAMES COOK UNIVERSITY
AUSTRALIA

20-MA5831-ONL-EXT-SP85

Advanced Data Management and Analysis using SAS

Week-2

Presented by
- Ban Pradhan

Agenda

- Week 2 contents
  - Common causes of poor data quality
    - Data-Oriented reasons for poor data quality
    - System-Oriented reasons for poor data quality
  - Treatment of Missing Data
  - Key characteristics of transformed data
  - Standard data flow design
    - ETL
  - Data processing pipeline
  - SAS Data Flux – PLAN, ACT and MONITOR
- Assessment 2 – Data quality profiling and standardising

Data-Oriented reasons for poor data quality

JAMES COOK UNIVERSITY AUSTRALIA

- **Dummy values** – databases sometimes require that some value is entered, so the input isn't the right value its often a default dummy value
- **Absence of data** – some applications may not require that some data be entered, so some fields contain no values.
- **Multipurpose fields** – different business units use the fields for different purposes, causing inconsistencies in what data is stored in them
- **Cryptic data** – the documentation is weak or non-existent, and it is not clear what the field stores.
- **Contradicting data** – one source system stores one value and another source system stores another, such as address information.
- **Inappropriate use of address lines** – rather that line 1, 2, 3 and so on to break out the name and address, everything is in a single line.
- **Violation of business rules** – as an example, on a variable interest rate loan, the minimum interest rate is higher that the lowest.
- **Re-used primary keys** – for example, each bank branch has a number that is its primary key. A branch closes, but two years later, a new branch is opened, and the old number is re-used. Queries using the number will generate undesired results.
- **Non-unique identifiers** – different systems use different identifiers for the same person. This is a common problem with healthcare data.

System-Oriented reasons for poor data quality

JAMES COOK UNIVERSITY AUSTRALIA

- **Data integration problems** – Problems can occur when source systems are very diverse and disparate on multiple platforms and with different operating systems.
- **Poor adherence to a service level agreement (SLA) –** Some organisations have an SLA that defines the transformations and load expectations, but making an external party adhere to them is always a problem and missing data from source systems is one of the major factors
- **Legacy platforms –** Many source systems are older legacy applications running on obsolete database technologies that have evolved data models over time with poor documentation.
- **Replaying history –** Generally, historical data is not preserved in source operational systems. Historical information storage is, therefore, a critical function in a data warehouse or data lake to provide the record of change.
- **Source system change –** Data structures in source systems keep changing over time because of new business conditions. ETL functions must also be modified accordingly otherwise key information is missed.
- **Poor enterprise information architecture –** A gross lack of consistency among information models used in source systems is prevalent.
- **Inability to change source systems –** Even when inconsistent data is detected among disparate source systems, the lack of a means and the cost for resolving mismatches escalates the problem of inconsistency.

Treatment of Missing Data

- The nature of missing data
  - Missing completely at random (MCAR): completely random
  - Missing at random (MAR): missing conditionally at random

- Traditional treatments for missing data
  - listwise deletion: remove records completely
  - pairwise deletion: keep all records but omit missing fields
  - Mean substitution
  - kNN impute
  - Regression substitution

Key characteristics of transformed data

JAMES COOK UNIVERSITY AUSTRALIA

- Subject oriented

all the datasets relating to the same real-world business subject or event are grouped and catalogued together.

- Available

In active data warehouses and data lakes this availability needs to be 24/7 and the users need to be informed when a load or transform step fails, this is often achieved with a daily load report.

- Integrated

This integrated subject-oriented data is very valuable information for both data science and decision support, and often the only place to find it is the data warehouse. As a result data warehouses contain highly integrated data as opposed to data lakes, which are loosely integrated and require users to build the integration layers.

- Time stamped

Data is stored as snapshots over past and current periods, this is useful for auditing as the source data should not change once stored in the data warehouse. Changes to data are tracked and recorded so that, if necessary, reports can be produced to show changes over time.

- Reference data (metadata)

It is broadly true to say that data can carry higher levels of information when associated with metadata or after integration.

- Non-volatile

The data in data warehouses and in data lakes are not normally intended for the day-to-day running of the business.

ETL

JAMES COOK UNIVERSITY AUSTRALIA

Extraction

there are five main options for ongoing extraction of data. These options are:

- Option 1: Capture through transaction logs (known as Change Data Capture or CDC)
- Option 2: Capture through database triggers
- Option 3: Capture in source applications
- Option 4: Capture based on date and time stamp
- Option 5: Capture by comparing files

Transformation tasks

Clean the data extracted from each source

Standardise the data types and field lengths for the same data elements retrieved from the various sources.

Combining data from a single source record or related data elements

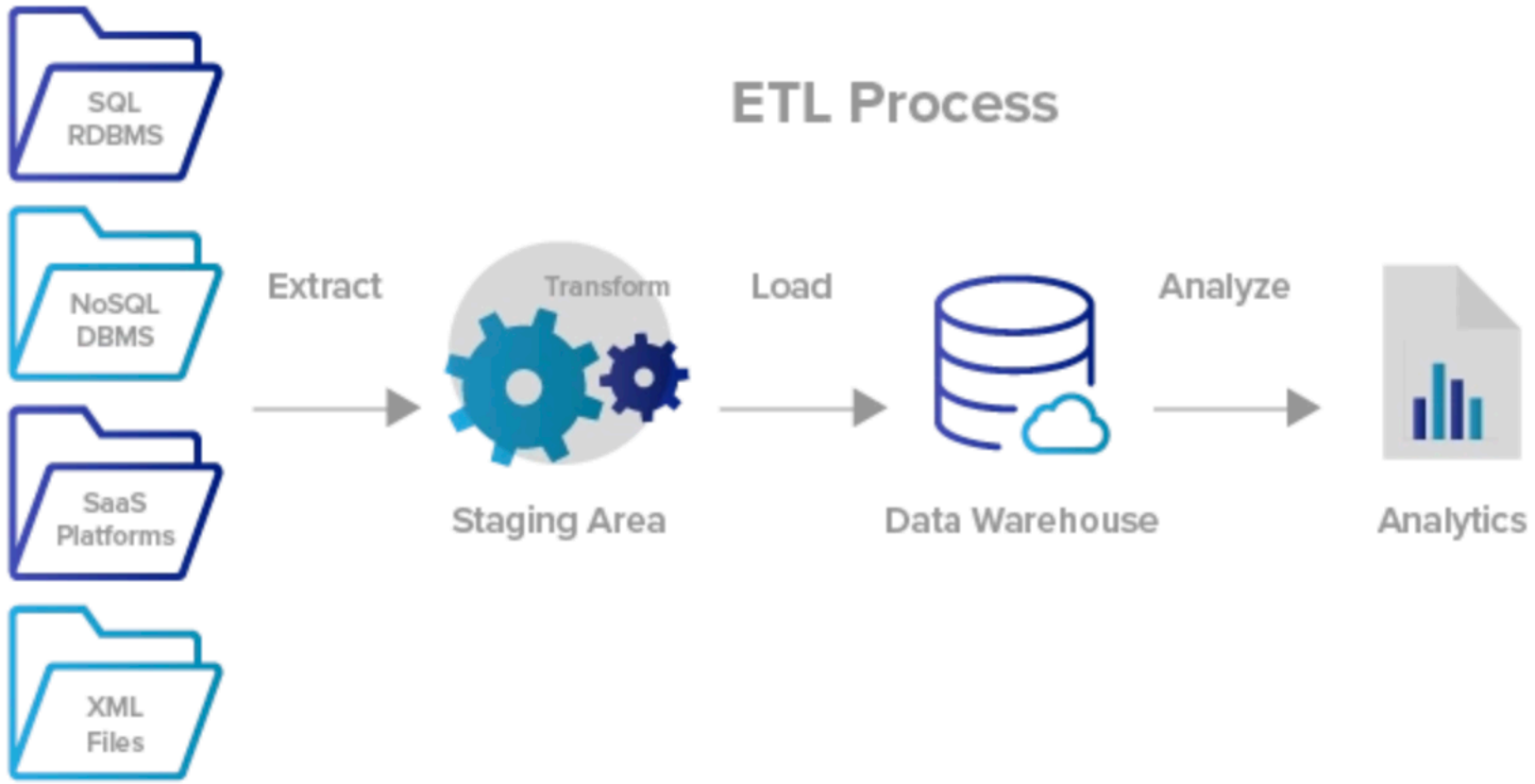
Sorting and merging of data takes place on a large scale in the data staging area.

Key Assignment often occurs in data transformation. This can include the assignment of surrogate keys derived from the source system primary keys.

Aggregation are commonly built in transformation steps for analytical applications that don't require all the details and to help improve response time for analytical applications.

Loading

Two distinct groups of tasks form the data loading function: Initial load and incremental data revisions.

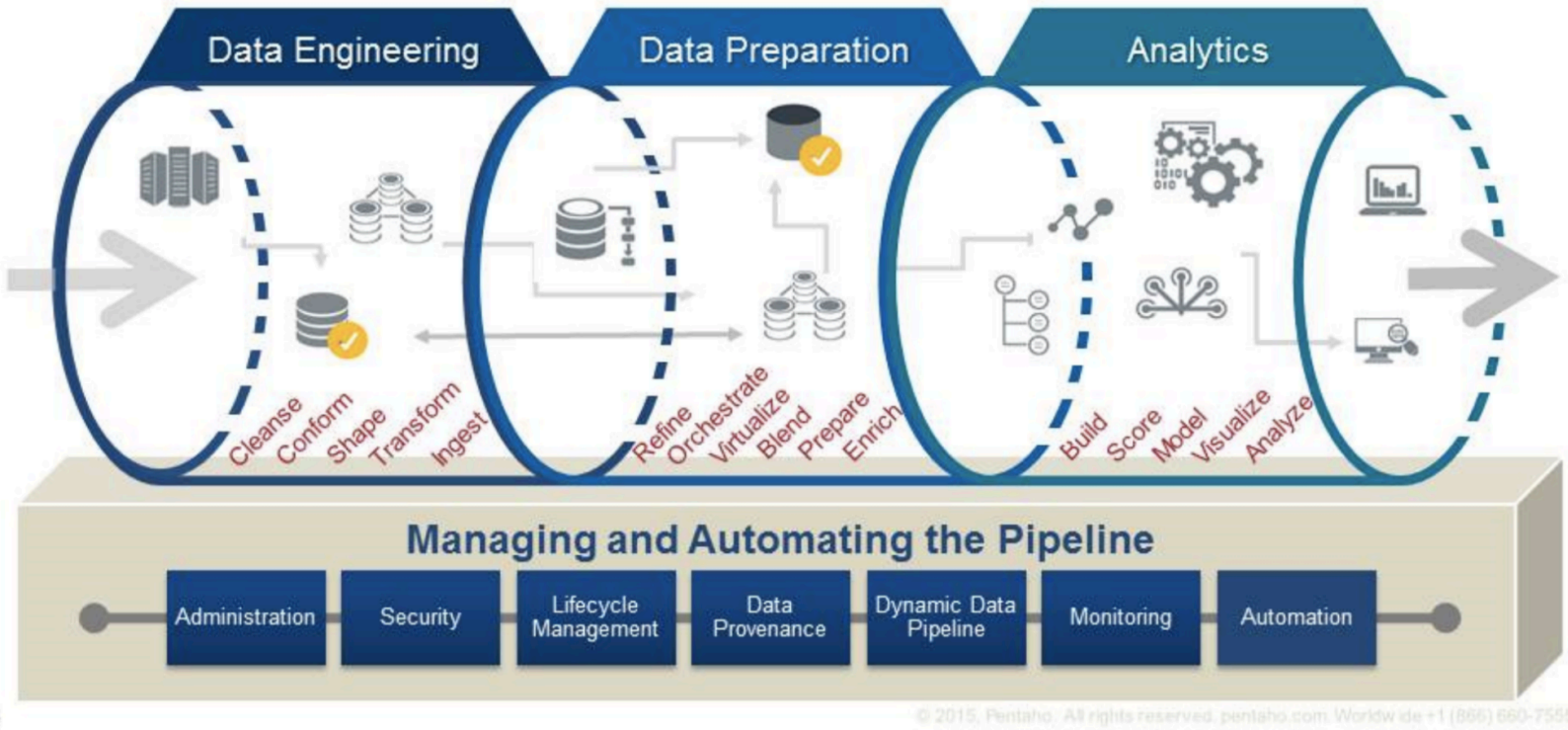

ETL Process
SQL RDBMS
NoSQL DBMS
SaaS Platforms
XML Files
Extract
Transform
Load
Analyze
Staging Area
Data Warehouse
Analytics

The five critical differences of ETL vs ELT:

1. ETL is the Extract, Transform, and Load process for data. ELT is Extract, Load, and Transform process for data.
2. In ETL, data moves from the data source to staging into the data warehouse.
3. ELT leverages the data warehouse to do basic transformations. There is no need for data staging.
4. ETL can help with data privacy and compliance by cleaning sensitive and secure data even before loading into the data warehouse.
5. ETL can perform sophisticated data transformations and can be more cost-effective than ELT.

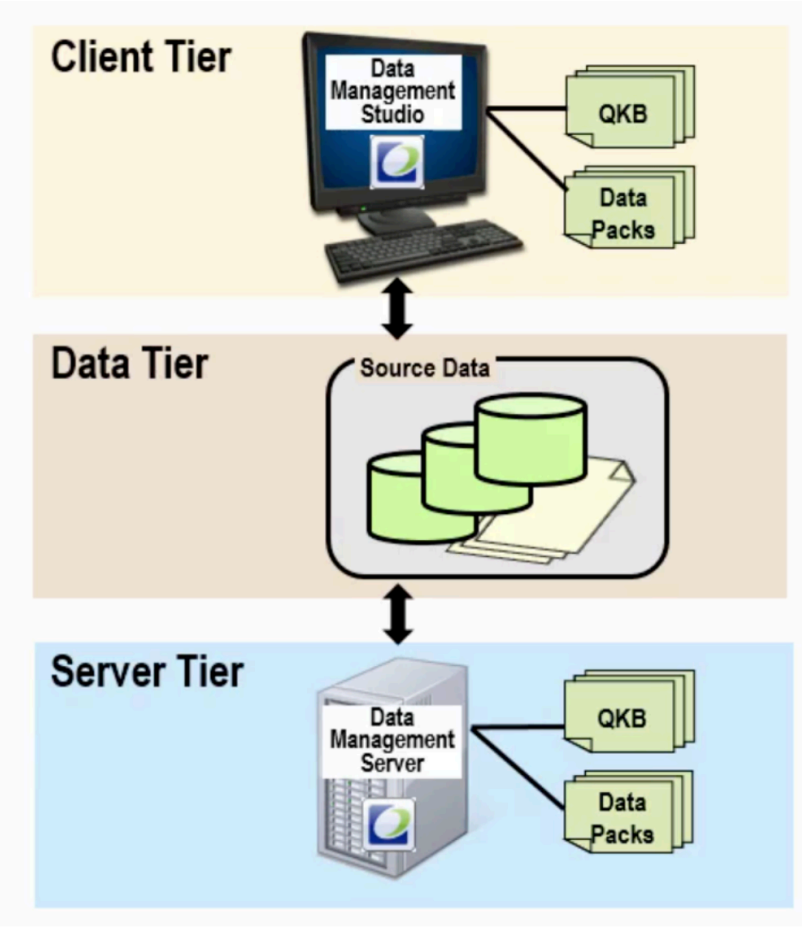
Data processing pipeline

JAMES COOK UNIVERSITY AUSTRALIA



Data Management Platform Architecture

JAMES COOK UNIVERSITY AUSTRALIA

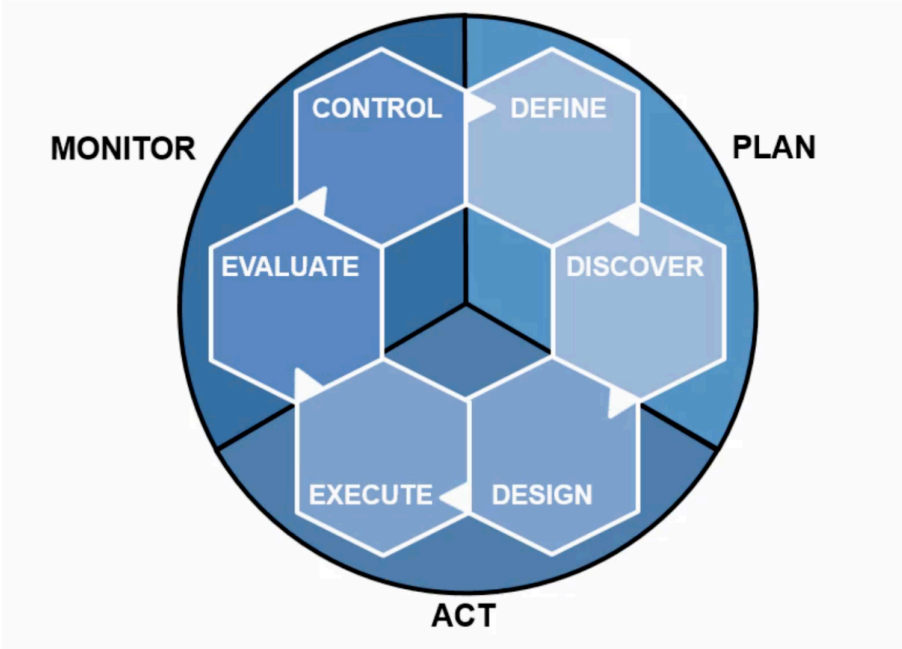


Data Management Studio is used to

- define data connections
- establish links to the QKB and data packs
- create repositories
- create
  - data explorations
  - profiles
  - data jobs
  - process jobs

- The Data Management Server needs access to a copy of the QKB and data packs that are used in the jobs and profiles.

DataFlux Data Management Methodology



Plan
- Data collection
- Data exploration
- Data profile
- Data standardisation

Act
- Data job
- Data quality job
- Data enrichment jobs
- Entity Resolution jobs

Monitor
- Business rules
- Data profiling with business rules
- Alerts
- Data jobs with business rules
- Monitoring tasks: log error, launch data flow job or run a local job etc

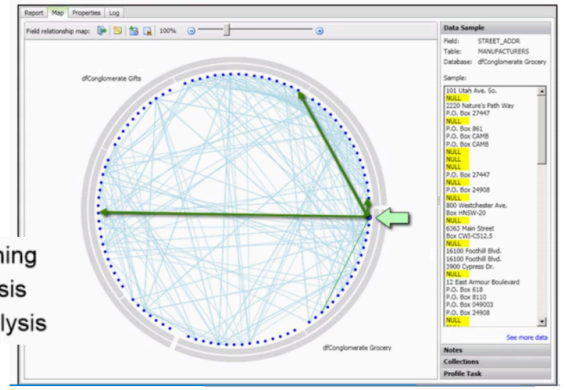
PLAN phase

JAMES COOK UNIVERSITY AUSTRALIA

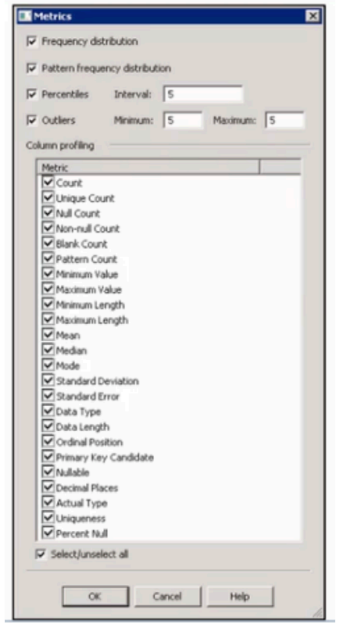
Data Collection A set of data fields from different tables in different data connections

Data Exploration Reads data from databases and categorizes the fields in the selected tables into categories.

- field name matching
- field name analysis
- sample data analysis



Data Profile Provides the ability to inspect data for errors, inconsistencies, redundancies, and incomplete information.

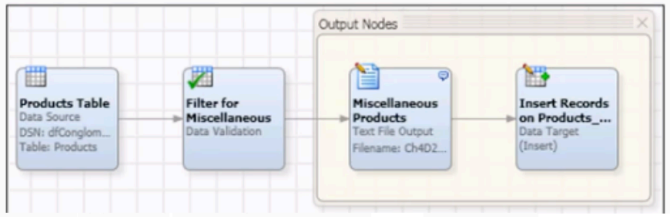


Data Standardisation A *standardization scheme* takes various spellings or representations of a data value and lists a standard way to consistently write this value.

ACT Phase

JAMES COOK UNIVERSITY AUSTRALIA

Data Job The main way to process data in Data Management Studio. Each data job specifies a set of data-processing operations that flow from source to target.



Data Prior to Standardization	Definition	Data After Standardization
Mister John Q. Smith, Junior	Name	Mr John Q Smith, Jr
dataflux corporation	Organization	DataFlux Corp
123 North Main Street, Suite 100	Address	123 N Main St, Ste 100
U.S.	Country	UNITED STATES
9194473000	Phone	(919) 447 3000

Name Information
Dr. Alan W. Richards, Jr., M.D.

Parsed Name	
Prefix	Dr.
Given Name	Alan
Middle Name	W.
Family Name	Richards
Suffix	Jr.
Title/ Additional Info	M.D.

Address Information
123 N MAIN ST APT 201

Parsed Address	
Street Number	123
Pre-direction	N
Street Name	MAIN
Street Type	ST
Post-direction	
Address Extension	APT
Address Extension Number	201

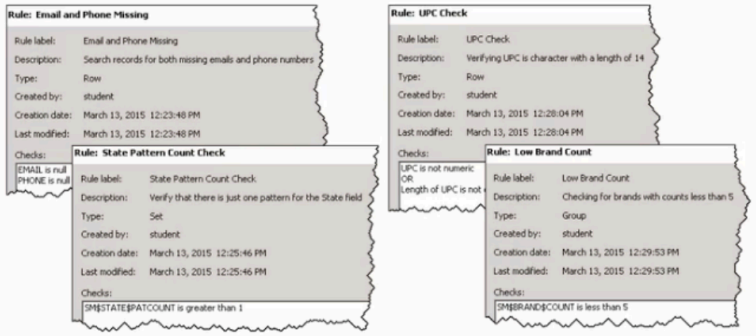
Match Codes

Name	Match Code @ 85 Sensitivity
John Q Smith	4B&~2$$$$$$$$$$C@P$$$$$$$$$
Johnny Smith	4B&~2$$$$$$$$$$C@P$$$$$$$$$
Jonathon Smythe	4B&~2$$$$$$$$$$C@P$$$$$$$$$

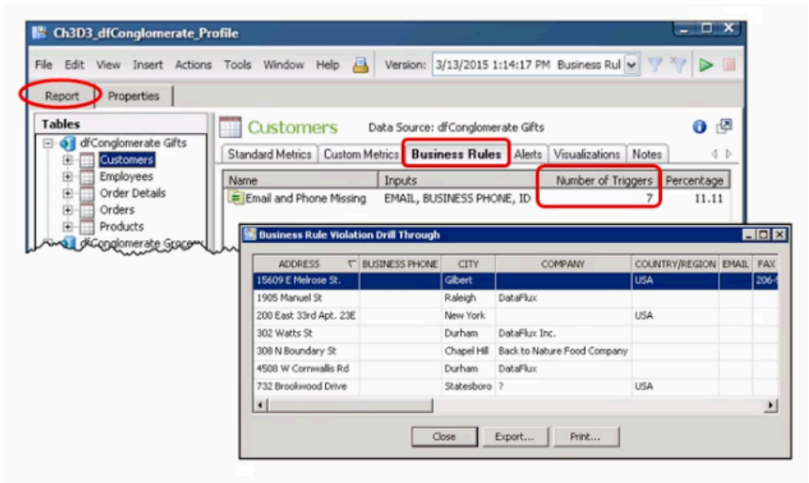
Monitor phase

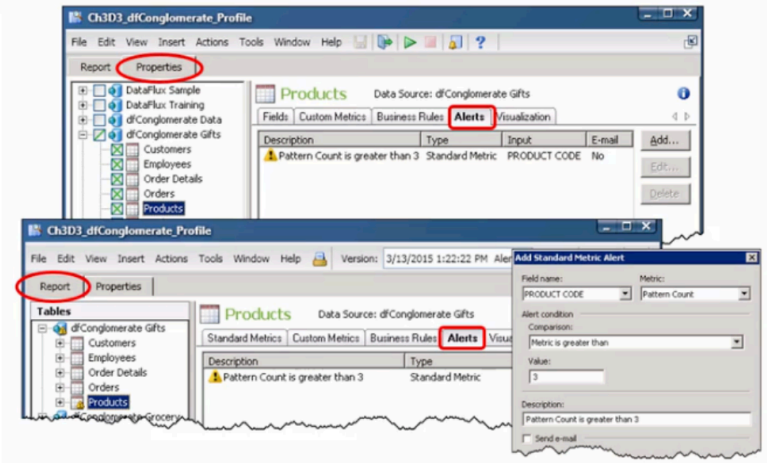
JAMES COOK UNIVERSITY AUSTRALIA

Business Rule A formula, validation, or comparison that can be applied to a given set of data. Data must either pass or fail the business rule.

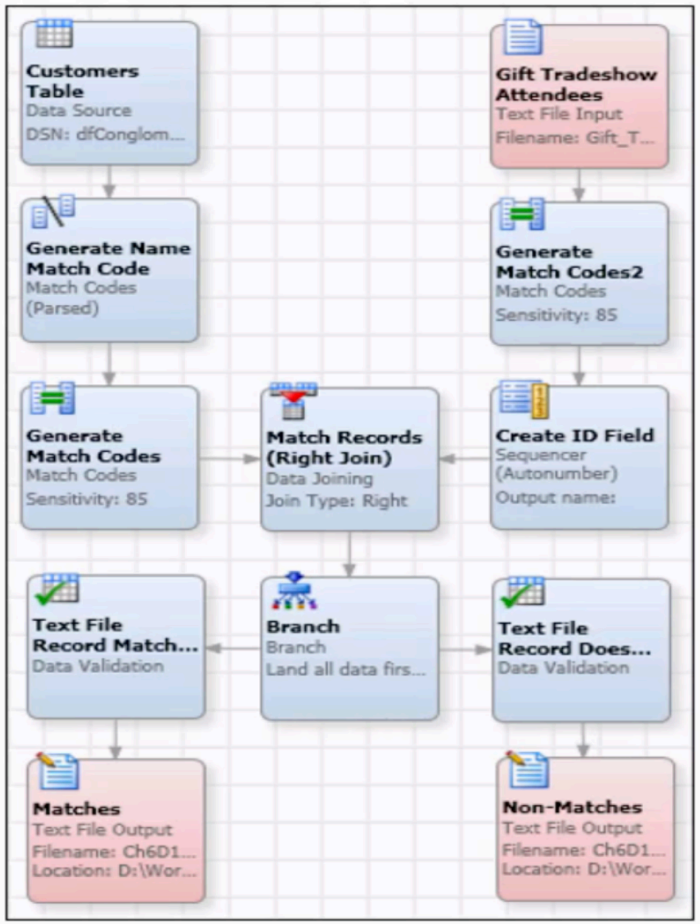


- Row based rule
- Set based rule
- Group based rule

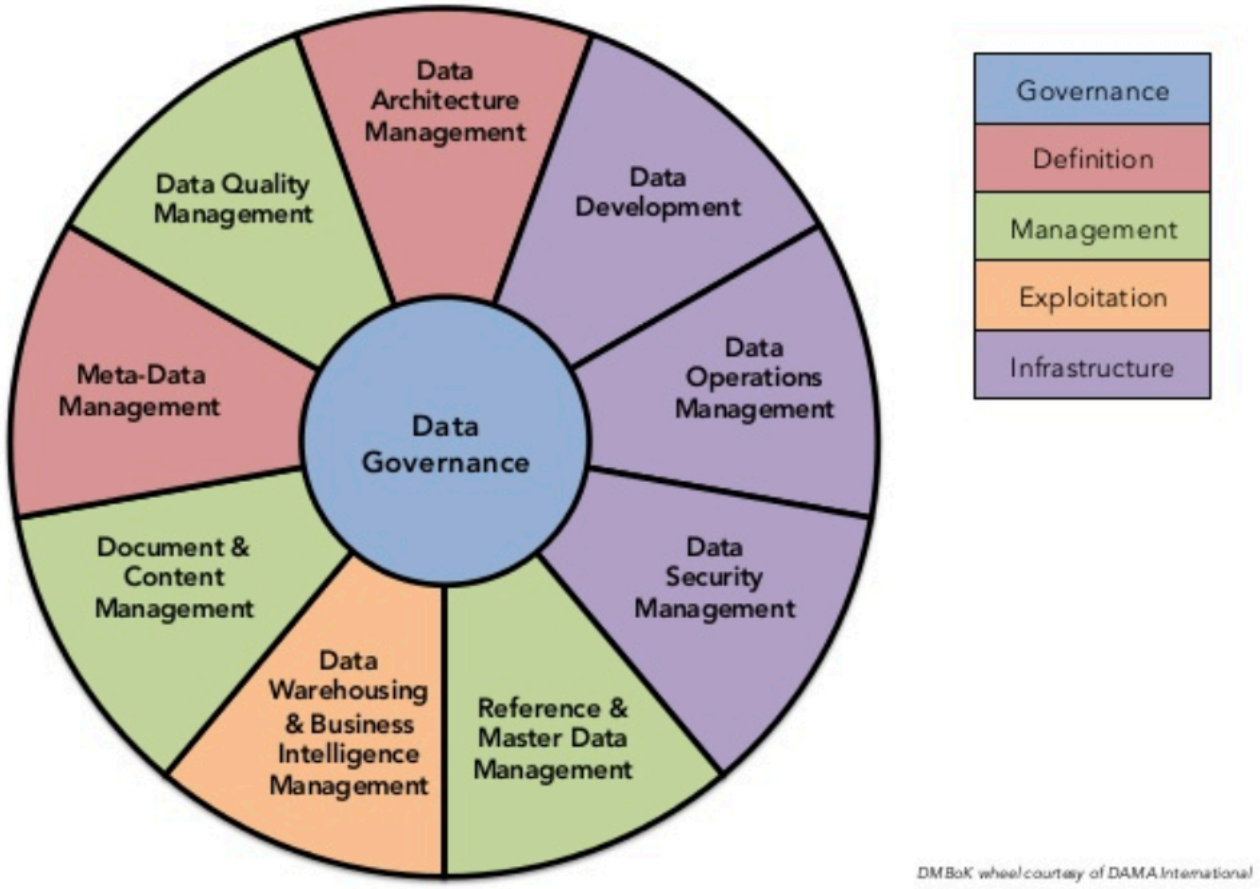




Customer Matches Example



DAMA Wheel



DMBoK wheel courtesy of DAMA International

QUIZ ANSWERS

Question 1 The time-variant nature of the data in a data warehouse is important for which of the following reasons? (Select all that apply.)

A. Allowing analysis of the past
B. Relating information to the present
C. Enabling forecasts for the future
D. Real-time updates on transactions

Correct answer: A, B and C.

Explanation: Real-time updates is a characteristic of OLTP systems not a data warehouse

Question 2 Which of the following statements are true? (Select all that apply.)

A. Organisations started creating data warehouses to improve their access and insight into historical information from transactional business support systems

B. Data lakes originated before data warehouses

C. Organisations started creating data lakes to improve their systems of record

D. Data warehouses need to handle distributed data across departments, servers and location

Correct answer: A and D

Explanation: Datalakes are a recent development within enterprises with a poorly defined architecture and are not recognised as record management platforms and do not comply with the principles and concepts set out in International Standards (ISO 15489 Records management).

Data warehouses were originally created to allow for more historical source system transactional data to be stored to provide insights across departments, servers and locations.

Question 3 Which of the following is not a good reason to use a data lake?

A. To store multiple datasets together in a common location

B. To enable scalable computing power

C. To build a business intelligence application for management reporting

D. To verify a customers details across multiple sources

Correct answer: C

Explanation: Building a business intelligence application requires a structured data schema organised into an normalised business view of the underlying data. Data lakes are not designed to do this as they keep their data as close to source structure as possible for experimentation

Question 4 Why is OLAP separated from OLTP? (Select all that apply.)

A. Due to performance problems
B. For archiving reasons; OLAP is more suitable for tape-archiving
C. Out of security concerns
D. Because OLAP systems tend to be larger

Correct answer: A and D

Explanation: The runtimes of analytical queries are significantly higher than those of transactional ones. Based on this characteristic, analytical processing negatively affected day-to-day business, that is, in terms of delayed sales processing. The separation of analytical and transactional queries to different machines was the inevitable consequence of the hardware and database prerequisites of these times.

Question 5 Consider a Formula 1 race car, with each race car having 512 sensors, each sensor records 32 events per second whereby each event is 64 byte in size.

How much data is produced by a F1 team, if a team has two cars in the race and the race takes two hours?

Please use the following unit conversions: 1 000 B = 1 KB, 1 000 KB = 1 MB, 1 000 MB = 1 GB.

A. 14 GB

B. 15.1 GB

C. 32 GB

D. 7.7 GB

Correct answer: B

Explanation:

Total time: 2h = 2 x 60 x 60 s = 7200 s

Total events per car: 7200 s x 512 sensors x 32 events/second/sensor

= 117 964 800 events

Total events per team; (2 x total events per car) = 235 929 600 events

Total amount of data per team: 64 bytes/event x 235 929 600

= 15 099 494 400 bytes = 15.1 GB

Question 6 For a column with few distinct values, how can dictionary encoding significantly reduce the required amount of memory without any loss of information?

A. By mapping values to integers using the smallest number of bits possible to represent the given number of distinct values.

B. By converting everything into full-text values. This allows for better compression techniques, because all values share the same data format.

C. By saving only every second value.

D. By saving consecutive occurrences of the same value only once.

Correct answer: A

Explanation: The correct answer describes the main principle of dictionary encoding, which automatically results in a lossless compression if values appear more often than once. Saving only every second value is clearly lossy. The same applies for saving consecutive occurrences of the same value only once, if the quantity of occurrences is not saved as well. Additionally, this does not describe dictionary encoding, but run-length encoding. Transforming numbers and other values into text values increases the data size, since each character value is at least 1 byte in order to allow the representation of the full alphabet. Number representations are usually limited to certain upper limits and achieve much smaller data sizes.

Question 7 Given a population table (50 million rows) with the following columns:

- Name (49 bytes, 20 000 distinct values)
- Surname (49 bytes, 100 000 distinct values)
- Age (1 byte, 128 distinct values)
- Gender (1 byte, 2 distinct values)

What is the compression factor (uncompressed size/compressed size) when applying dictionary encoding?

A. Approximately 20

B. Approximately 90

C. Approximately 10

D. Approximately 5

Correct answer: A

Explanation: Calculation without dictionary encoding: 5000 mb

Number of bits needed for the attributes:

Names: $\log_2(20\ 000) < 15$

Surnames: $\log_2(100\ 000) < 17$

Ages: $\log_2(128) \leq 7$

Genders: $\log_2(2) \leq 1$

Size of the attribute vectors:

50 million x(15 + 17 + 7 + 1) bit = 2000 million bit = 250 mb

Size of the dictionaries:
Names: 20 000 x 49 B = 980 Kb
Surnames: 100 000 x 49 B = 4.9 mb
Ages: 128 x 7 B = 896 B
Genders: 2 x 1 B=2 B

Total dictionary size: 4.9 mb + 980 kb + 896 B + 2 B = 5 mb
Overall size of attribute vectors + dictionary size = 250mb + 5mb = 255mb
Compression Rate 5000 mb / 255 mb ≈ 20

Question 8 What information is saved in a dictionary in the context of dictionary encoding?

A. Cardinality of a value
B. All distinct values
C. Hash of a value of all distinct values
D. Size of a value in bytes

Correct answer: B

Explanation: The dictionary is used for encoding the values of a column. Therefore, it consists of a list of all distinct values to be encoded and the resulting encoded values (in most cases ascending numbers). The distinct values are used to encode the attributes in user queries during look-ups and to decode the retrieved numbers from query results back to meaningful, human-readable values.

Question 9 Which of the following is an advantage of dictionary encoding?

A. Sequentially writing data to the database is sped up
B. Aggregate functions are sped up
C. Raw data transfer speed between application and database server is increased
D. INSERT operations are simplified

Correct answer: B

Explanation: Aggregate functions are sped up when using dictionary encoding because less data has to be transferred from main memory to CPU. The raw data transfer speed between application and database server is not increased; this is determined by the physical hardware and exploited to the maximum. INSERT operations suffer from dictionary encoding because new values that are not yet present in the dictionary have to be added to the dictionary and, if the dictionary is sorted, then the operation might require a resorting of the related attribute vector and sequentially writing data to the database is not sped up.

Question 10 Which of the following compression techniques can be used to decrease the size of a sorted dictionary?

A. Cluster encoding
B. Prefix encoding
C. Run-Length encoding
D. Delta encoding

Correct answer: D

Explanation: Cluster encoding, run-length encoding and prefix encoding cannot be used on dictionaries because each dictionary entry is unique.

Thank You!