JAMES COOK UNIVERSITY
AUSTRALIA

Week 3

MA5831 – Advanced Data Management and Analysis using SAS

Dr Mostafa Shaikh
mostafa.shaikh@jcu.edu.au

online.jcu.edu.au

Cairns
Singapore
Townsville

JAMES COOK UNIVERSITY AUSTRALIA

Agenda

- Week 2 quick review
- Week 3 Parsing, matching and standardising data
  - The standard data-cleansing process steps
  - Entity resolution with multiple inputs and outputs
  - SAS Quality Knowledge Base and customisation
  - Data types and definitions
- Assessment 2 – SAS DataFlux to answer 13 questions

JAMES COOK UNIVERSITY AUSTRALIA

Week 2 Recap

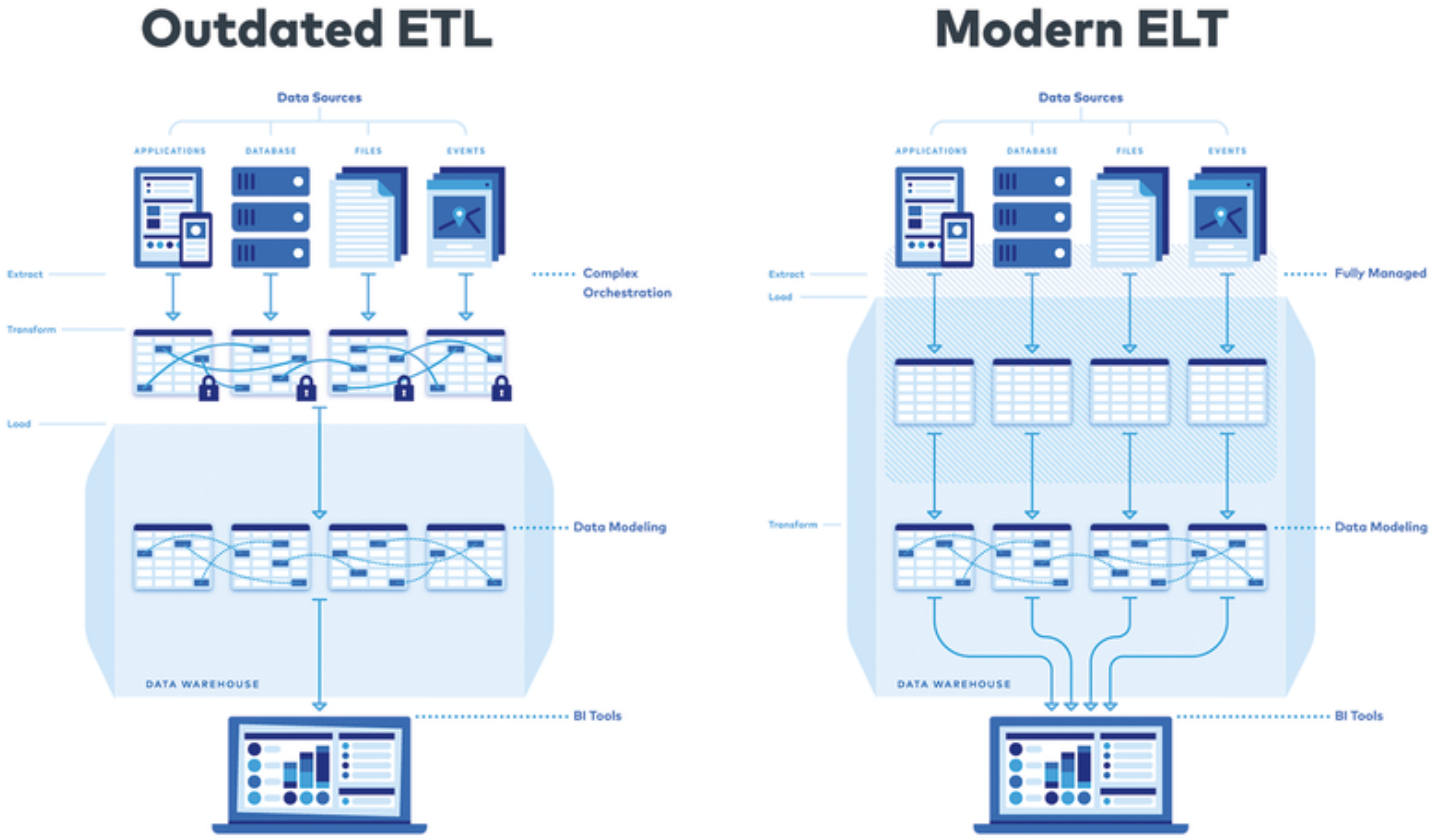
Treatment of Missing Data

JAMES COOK UNIVERSITY AUSTRALIA

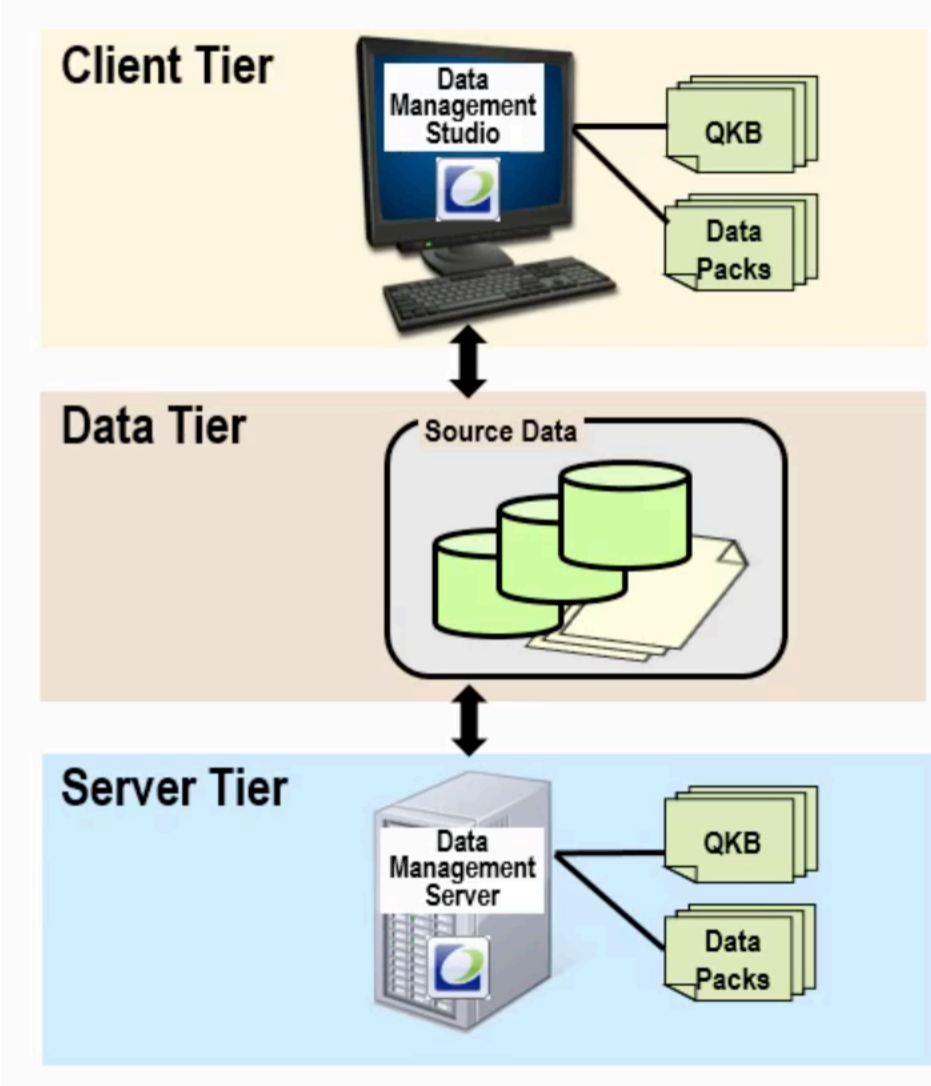
- The nature of missing data
  - Missing completely at random (MCAR): completely random
  - Missing at random (MAR): missing conditionally at random

- 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

Differences between ETL and ELT



Data Management Platform Architecture

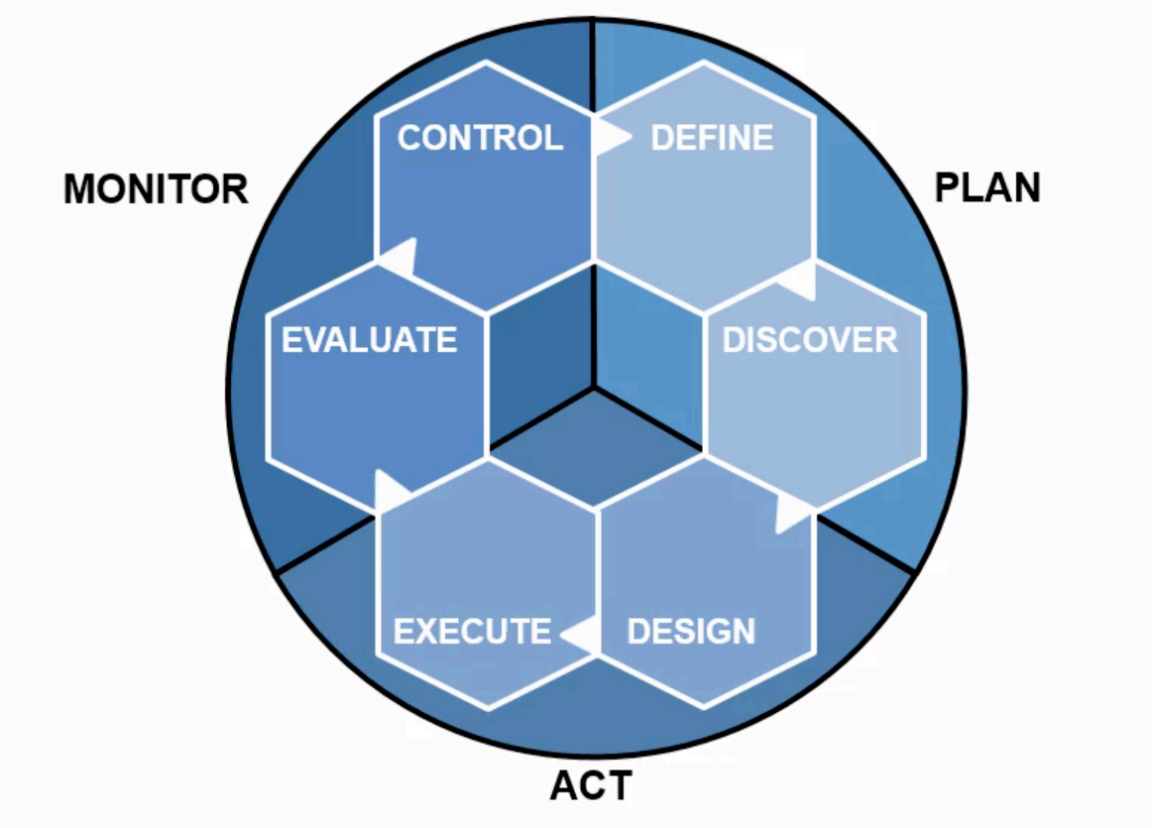


Data Management Studio is used to

- 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



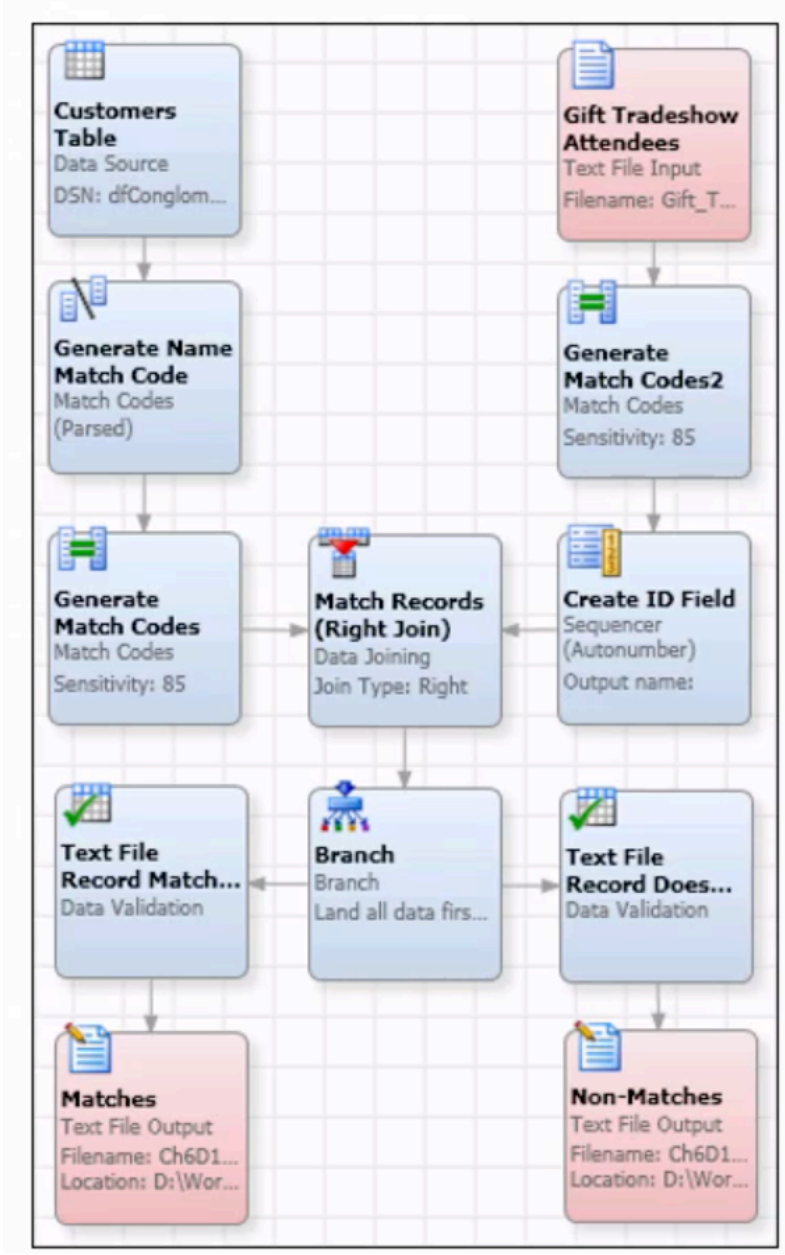
•**Plan** - Identify patterns and problems in your data.
•**Act** - Create processes to improve data quality and data integration.
•**Monitor** - Monitor your processes for data quality and data integration.

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

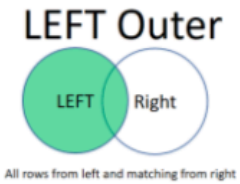
Customer Matches Example

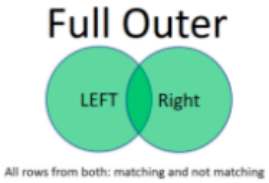


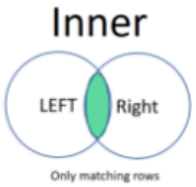
Joining by matched definition

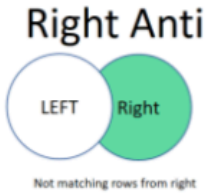
Different Join Kinds in Merge returns different result set.

- Left Outer: Rows from left table and matching with the right
- Right Outer: Rows from right table and matching with the left
- Full Outer: Rows from both tables (matching or not matching)
- Inner: Only matching rows from both tables
- Left Anti: Not matching rows from left table
- Right Anti: Not matching rows from right table









SAS DataFlux

JAMES COOK UNIVERSITY AUSTRALIA

Step 1: Data Repository

Step 2: QKB

Step 3: Data Connection

Step 4: Data filtering, Data Investigation, Data Preview

Step 5: Data Collection, Standardisation Scheme

Step 6: Data Exploration

Step 7: Data Profile

Step 8: Standardisation Scheme

Step 9: DMS option setup for Jobs

Step 10: Data Job

Step 11: Data Quality Job

Step 12: Data Enrichment Job

Step 13: Entity Resolution Jobs

Step 14: Data Profiling with Business Rules and Alerts

Step 15: Data Jobs with Business Rules

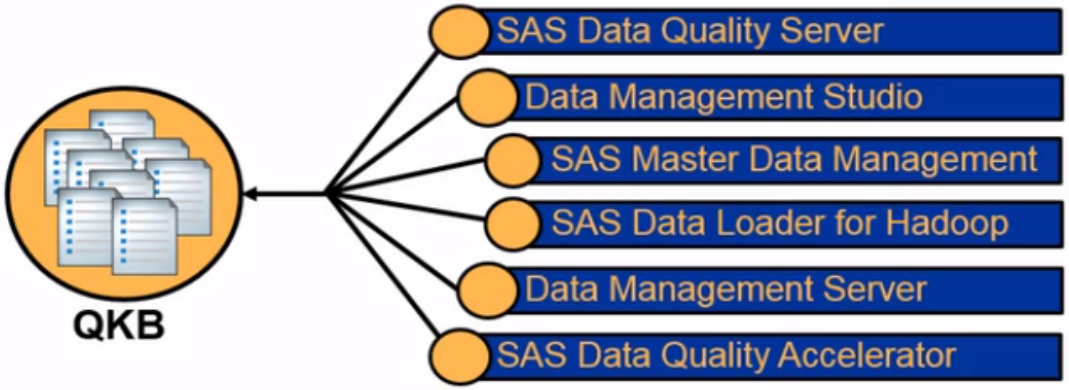
Step 16: Data Jobs with Monitoring Tasks

JAMES COOK UNIVERSITY AUSTRALIA

Week 3: QKB

SAS Quality Knowledge Base (QKB)

The *SAS Quality Knowledge Base* (QKB) is a collection of files that store data and logic that define data management operations.

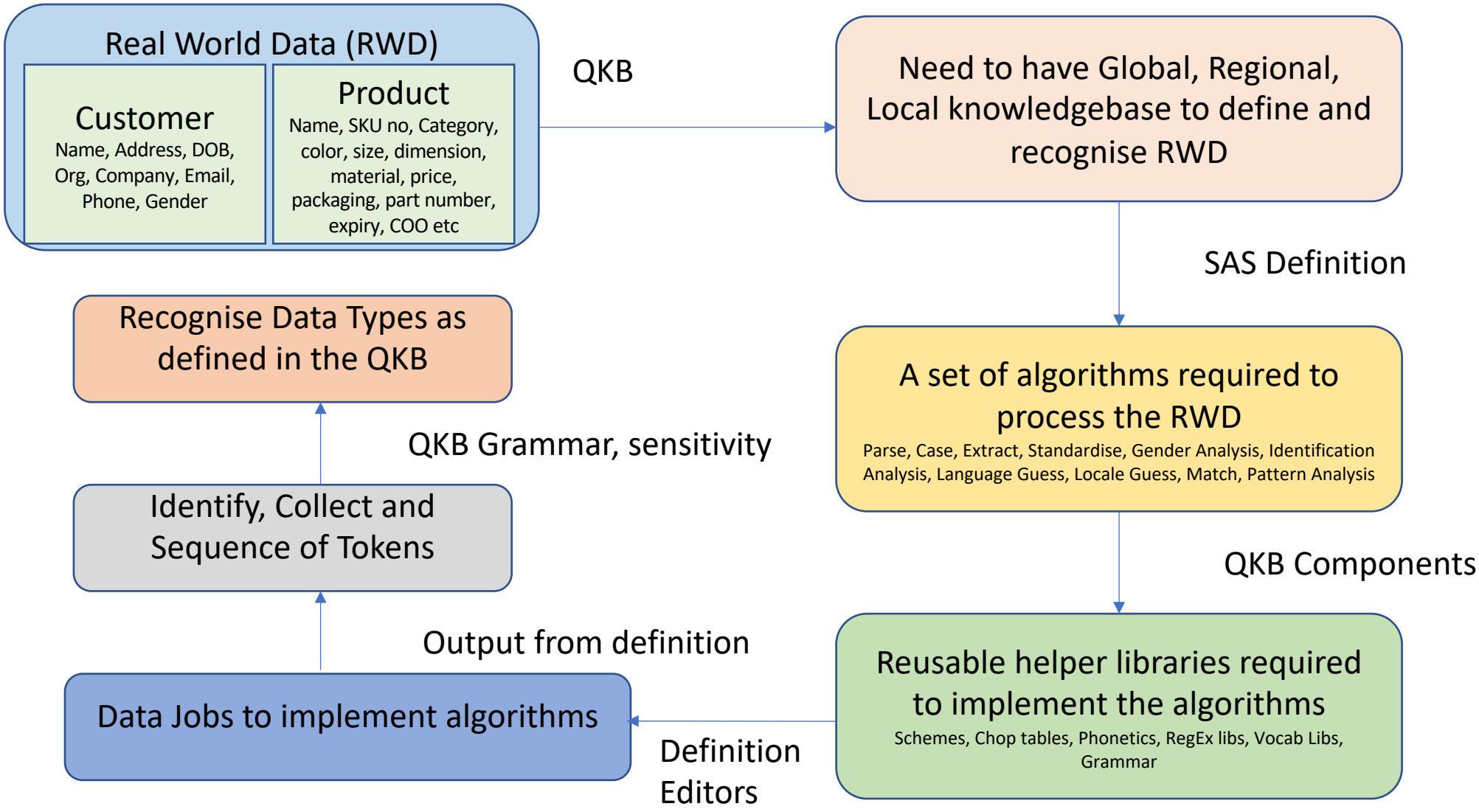


Two Types of QKBs

There are two types of QKBs available for licensing.

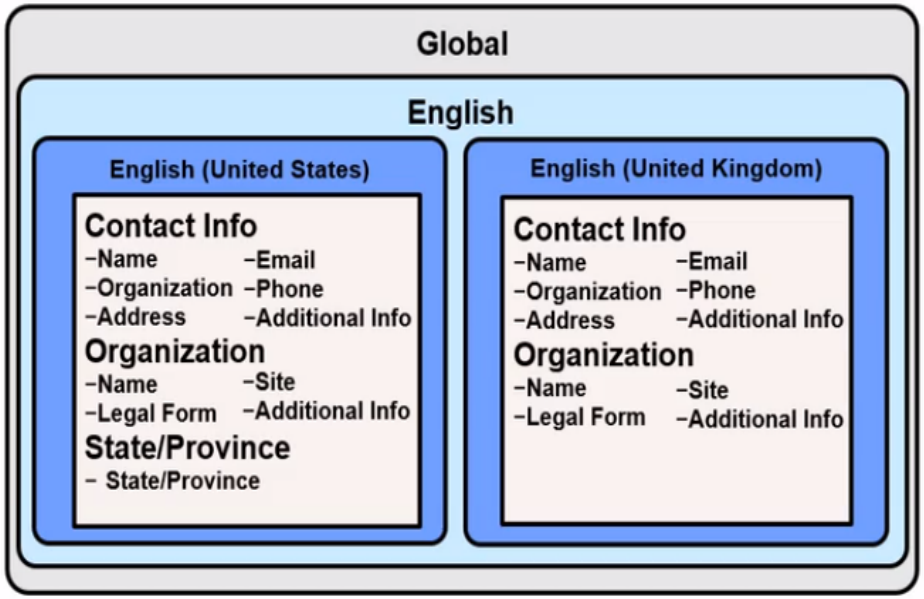
SAS Quality Knowledge Base for Contact Information (CI)	Supports management of commonly used contact information for individuals and organizations, such as names, addresses, company names, and phone numbers
SAS Quality Knowledge Base for Product Data (PD)	Contains extraction, parsing, standardization, and pattern analysis definitions to handle the following attributes in generic product data: • brands/manufacturers • dimensions • part numbers • packaging terms and units of measurement • colors • sizes • materials

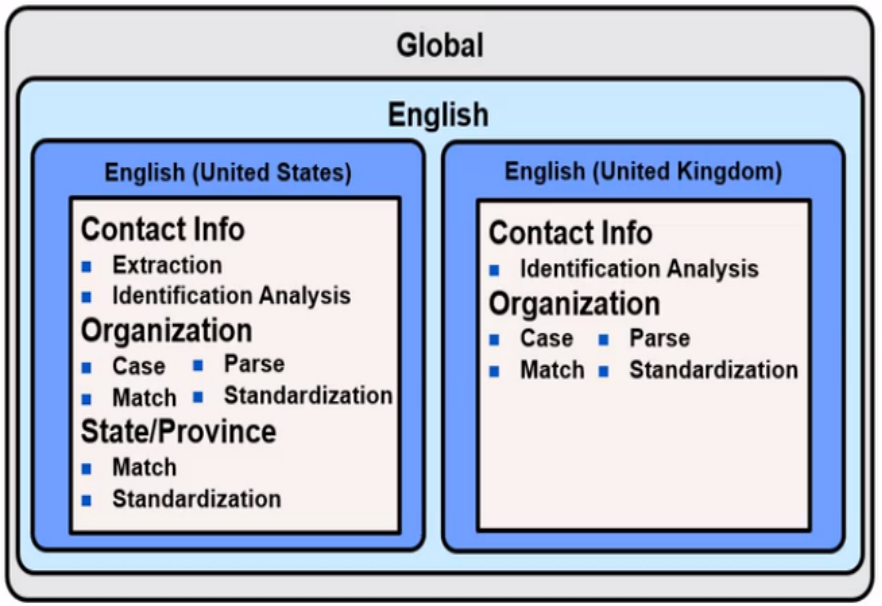
What problems QKB solves?



Structure of QKB

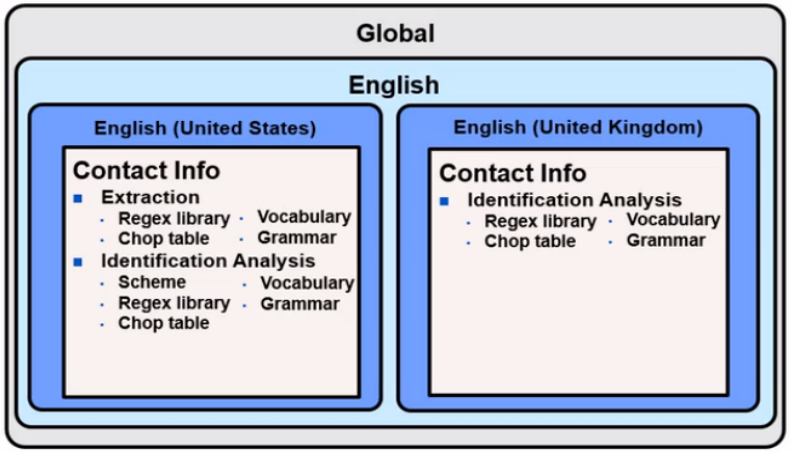
Variant datatypes and definitions





Organization of the QKB: Component Files

Each definition consists of various component files.



Component Files

Each definition in the QKB consists of one or more of the following types of component files:

- Schemes
- Chop Tables
- Phonetics Libraries
- Regex Libraries
- Vocabulary Libraries
- Grammars

JAMES COOK UNIVERSITY AUSTRALIA

What Is a Chop Table?

A *chop table* is a collection of character-level rules that are used to create an ordered word list from an input string.

- A classification and an operation are specified for each character in the chop table.
- Chop tables are used to perform element analysis and by parse definitions to parse a text string into tokens.

Classification of Characters

The following classifications can be used for any character in a chop table:

Classification	Description
LETTER/SYMBOL	A letter or non-separating symbol
NUMBER	A numeric digit (0-9)
LEAD SEPARATOR	A delimiter attached to the beginning of a word (example: open parenthesis)
TRAIL SEPARATOR	A delimiter attached to the end of a word (example: period)
FULL SEPARATOR	A delimiting character (examples: space, hyphen, comma)

Operation for Characters

The following operations can be specified for any character in a chop table:

Operation	Description
USE	Use the character as is in the word list and output tokens
TRIM	Omit the character from the word list; trim leading and trailing characters in output tokens
SUPPRESS	Omit the character from the word list and output tokens

JAMES COOK UNIVERSITY
AUSTRALIA

What Is a Vocabulary?

A *vocabulary* is a collection of words.

- A list of words can be entered manually or imported from text files.
- Each word in a vocabulary is associated with one or more categories.
- A likelihood is assigned to each category associated with a word.

What Is a Grammar?

A *grammar* is a set of rules that represent expected patterns of words in a given context.

- A grammar consists of two category types: basic and derived.
- Basic categories in a grammar correspond to categories associated with words in vocabularies.
- Derived categories in a grammar consist of ordered patterns of basic or derived categories.

Example: EN Name Grammar

Recursive rules are used to resolve variable numbers of categories into a single derived category.

Text String:	**Bob Brauer, Justice of the Peace**
Name Rule:	**N > GNW FNW COMMA NA where NA > NAW NAW NAW NAW**

A recursive rule can be defined that can shorten this:

NA > NAW
NA > NAW NA

Category Abbreviations	
Category	**Abbreviation**
Name	N
Name Prefix	NP
Given Name	GN
Middle Name	MN
Family Name	FN
Name Suffix	NS
Name Appendage	NA
Name Prefix Word	NPW
Given Name Word	GNW
Family Name Word	FNW
Name Suffix Word	NSW
Name Appendage Word	NAW
Comma or Semicolon	COMMA

What Is a Regular Expression?

A *regular expression* consists of a pattern matched against a text string from left to right.

Patterns are commonly found in data.

- Dates contain many different types of patterns:
  - MM/DD/YYYY
  - DD/MM/YYYY
  - Mon DD, YYYY
- IP addresses consist of a similar pattern:
  - number.number.number.number
- Names can match a specific pattern:
  - LastName, FirstName MiddleInitial

What Is Phonetics?

Phonetic analysis (*phonetics*) is used to generate match codes. During match code generation, phonetic rules are applied to reduce an input string. The goal is to create phonetic rules that produce the same output string for input strings with similar pronunciations or spellings.

Example: Using phonetic rules, SCHMIDT and SHMITT are reduced to SHMIT.

What Is a Standardization Scheme?

A *standardization scheme* is a lookup table that is used to transform data values to a standard representation.

- Schemes can be applied to an element of a string (element analysis) or to the entire string (phrase analysis).
- Schemes are used in many types of definitions and are often applied at the token level.

Data	Standard
ADV	ADVOCATE
BR	BR
BROTHER	BR
BRIGADIER GENERAL	BRIGADIER-GENERAL
BRIGADIER-GENERAL	BRIGADIER-GENERAL
CAPT	CAPT
CAPTAIN	CAPT
CPT	CAPT
CDR	COMMANDER
COUNCILLOR	COUNCILOR
DOCT	DR
DOCTOR	DR
DR	DR
FATHER	FR

What Is a Data Type?

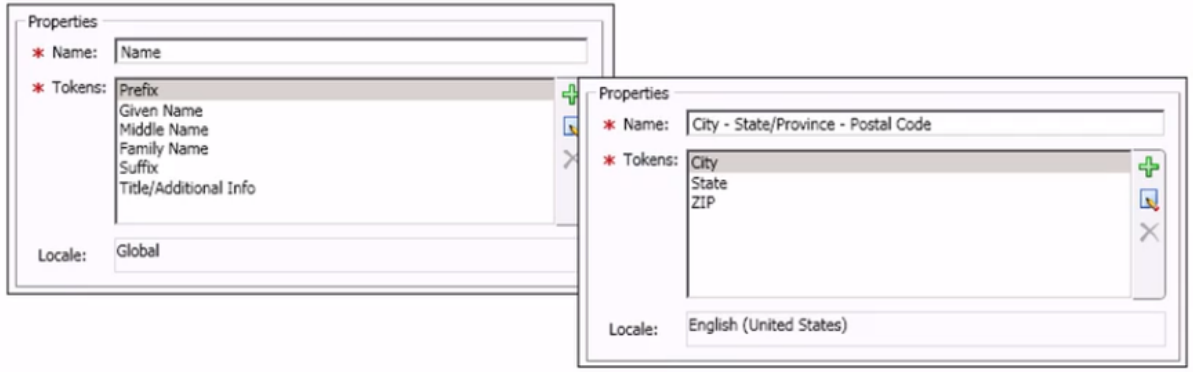
In the QKB, a *data type* has the following characteristics:

- is an object that represents the semantic nature of some data value (for example, address data, name data, organization data)
- serves as a placeholder (or grouping) for storing data quality algorithms called *definitions*
- consists of one or more tokens

Data Type Tokens

A *data type token* is

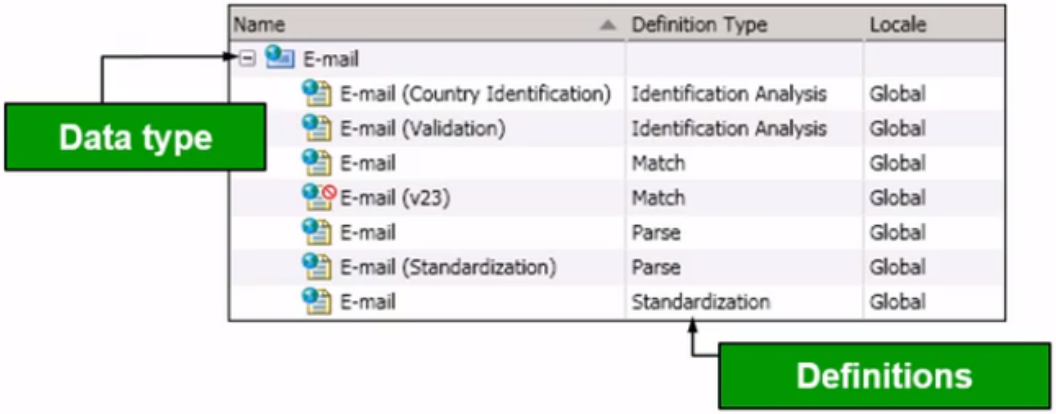
- an "atomically semantic" component of a data value
- the smallest meaningful part of a data value
- used as output for parse definitions
- often used as input for other types of definitions.



What Is a Definition?

In the QKB, a *definition*

- is a collection of metadata
- is associated with a data type
- defines a single context-sensitive data management (data-cleansing) operation.



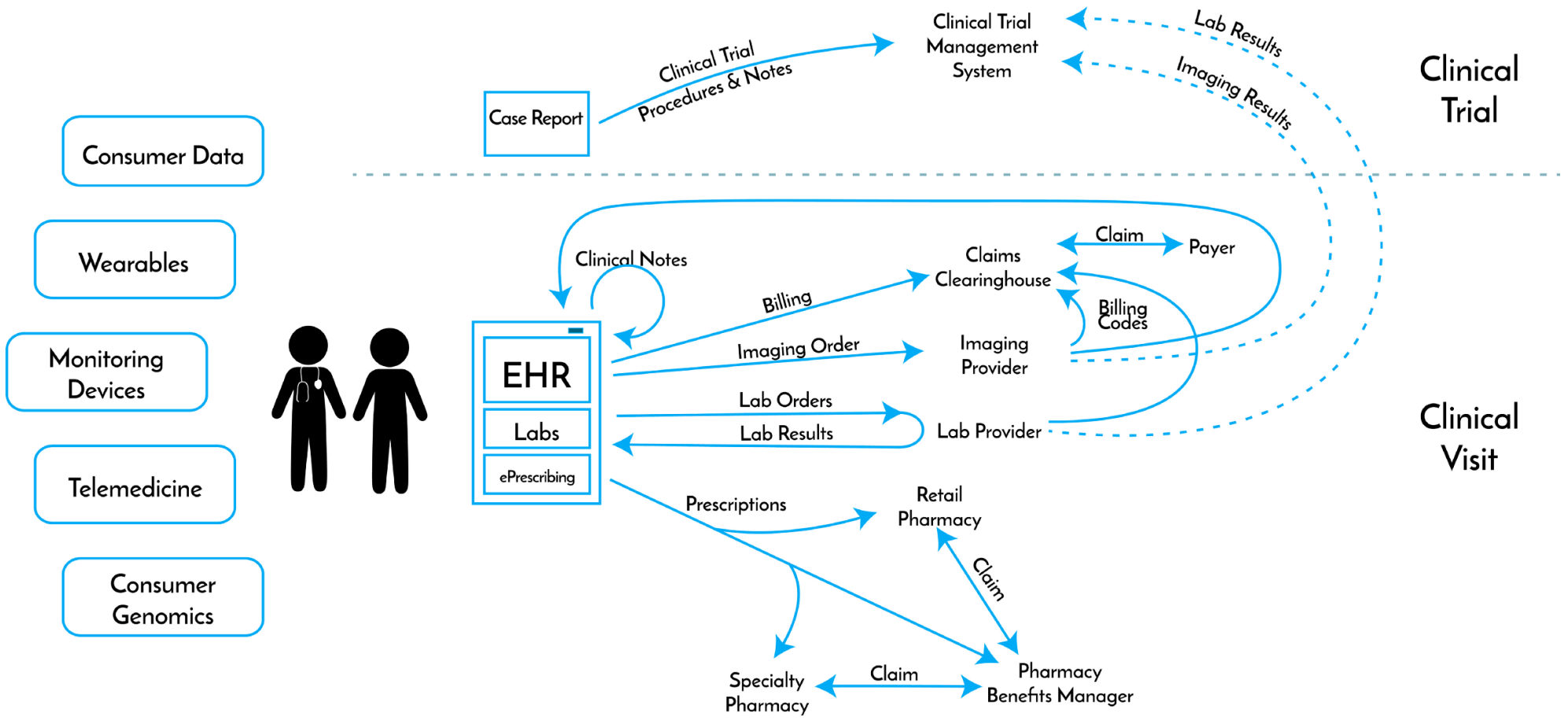
A definition is a set of steps for processing a piece of data. Here is a list of definition types:

- Case
- Extraction
- Gender Analysis
- Identification Analysis
- Language Guess
- Locale Guess
- Match
- Parse
- Pattern Analysis
- Standardization

The standard data-cleansing process steps

- **Parsing** – locates and identifies individual data elements in the source files and then isolates these data elements in the target files [chop, vocabulary, regex, grammar]
- **Correcting –** compares individual parsed data components using basic data matching algorithms and independent secondary data sources (eg, dictionaries) [vocabulary, regex]
- **Standardising –** transforms data into its preferred (and consistent) format using both standard and custom business rules [schemes, grammar]
- **Matching –** scores how exact the pattern matching is between two different records [phonetics, vocabulary, regex]
- **Consolidating –** analysing and identifying relationships between matched records is to consolidate them into ONE representation or single record

Single customer/patient view

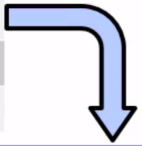


Parsing

Input text: John.Smith@DataFlux.com

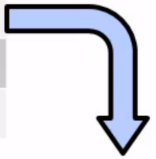
Token	Value
Mailbox	John.Smith
Sub-Domain	DataFlux
Top-Level Domain	com

Original Addresses
4001 Weston Parkway Suite 300
Ste 300, 4001 Weston Pkwy
3645 NW John Maynard Rd.



Tokenized Addresses					
Street Number	Pre-direction	Street Name	Street Type	Address Extension	Address Extension Number
4001		Weston	Parkway	Suite	300
4001		Weston	Pkwy	Ste	300
3645	NW	John Maynard	Rd.		

Original Names
Dr. James Goodnight, President and CEO
Mr. Warren Allan Richardson III, MBA
Jodi Drapeau



Tokenized Names					
Prefix	Given Name	Middle Name	Family Name	Suffix	Title/Additional Info
Dr.	James		Goodnight		President and CEO
Mr.	Warren	Allan	Richardson	III	MBA
	Jodi		Drapeau		

JAMES COOK UNIVERSITY
AUSTRALIA

Layout of a Parse Definition

The template for any parse definition contains the following types of items:

- preprocessing regular expression(s)
- chop table
- morph analysis containing the following:
  - casing
  - normalization regular expression(s)
  - vocabulary
  - categorization regular expression(s)
  - number check
  - specification of default categories
- grammar
- token mappings

Data Parsing Overview

1	**Pre-process Text**	If necessary, apply cleansing rules to the data before it is chopped into words.
2	**Chop Text**	Chop the text string into individual components or words.
3	**Categorize Words**	After a text string is chopped into words, assign the words to certain categories.
4	**Identify Patterns**	After categorizing the words, identify patterns in the categories.
5	**Score Solutions**	After the patterns are recognized in the grammars, a score is calculated.
6	**Select "Best" Solution**	The solution with the highest score is selected as the best solution for the parse.
7	**Map Words to Tokens**	For best solution, map words to appropriate tokens and create a tokenized word list.

Standardising

JAMES COOK UNIVERSITY AUSTRALIA

Original String	Standardized Value
411 elm street	411 Elm St
DALLAS TEXAS 75202	Dallas, TX 75202
united STATES	USA
Todd B. Donovan, PHD	Todd B Donovan, PhD
9196778000	(919) 677 8000

What Is a Standardization Definition?

A *standardization definition* is a context-specific transformation algorithm.

Standardization definitions consist of the following:

- Parse definition (for multi-token data strings)
- Regular expression libraries
- Transformation schemes
- Casing algorithms or definitions

Standardising

JAMES COOK UNIVERSITY AUSTRALIA

Example: Standardizing Names

The Name standardization definition for the Name data type standardizes names as shown.

Original Names	Standardized Names
John Jones Junior	John Jones, Jr
Junior Jones	Junior Jones

Example: Standardizing City-State-Zip (CSZ)

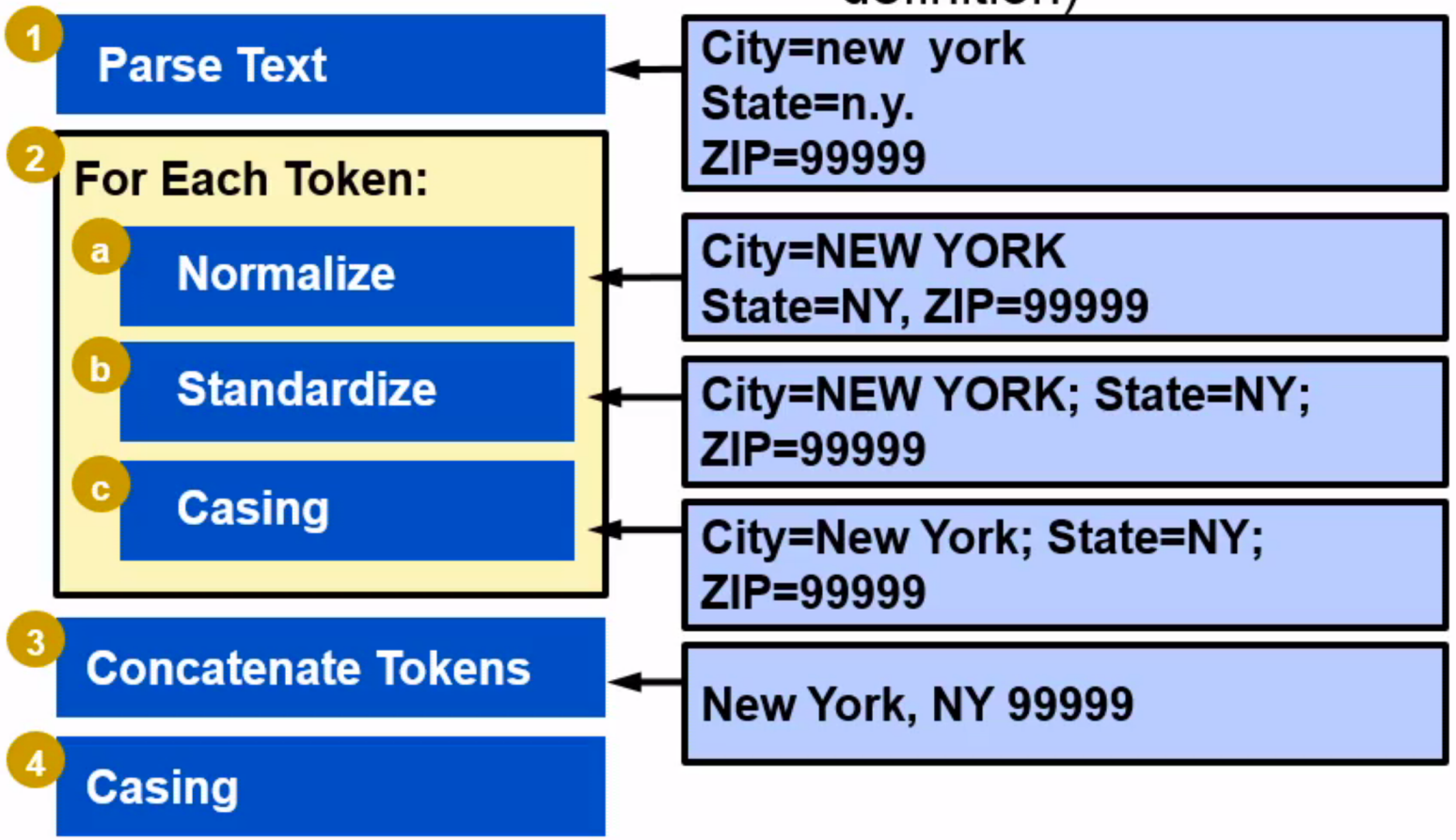
The City - State/Province - Postal Code standardization definition for the City - State/Province - Postal Code data type standardizes values as shown.

Original CSZ	Standardized CSZ
Richmond Virginia	Richmond, VA
New York New York 10019	New York, NY 10019
KANSAS CITY, KANSAS 66202	Kansas City, KS 66202
St. Louis, Mo. 63011	Saint Louis, MO 63011
washington dc 20005	Washington, DC 20005

Standardising

Example: Standardizing Addresses

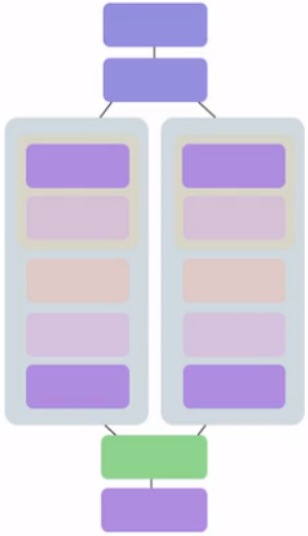
Example String: new york, n.y. 99999 (City-State/Province-Postal Code definition)



Layout of a Standardization Definition

The template for any standardization definition contains the following types of items:

- tokenization (parse definition)
- token nodes containing
  - trimming
  - casing
  - pre-scheme regular expression(s)
  - transformation schemes
  - post-scheme regular expression(s)
  - casing
- concatenation
- casing



1	**Parse Text**	Identify the tokens of the text (parse definitions).
2	**For Each Token:**	
a	**Normalize**	Normalize text (regular expressions).
b	**Standardize**	Standardize text (schemes).
c	**Casing**	Change case (casing definition).
3	**Concatenate Tokens**	After each token is processed, resultant tokens are concatenated.
4	**Casing**	Casing is applied to the concatenated modified tokens.

What Is a Match Definition?

Match definitions are a combination of the following:

- Parsing
- Normalization
- Standardization
- Phonetic reduction

The result of applying a match definition is a match code.

A *match code* is a fuzzy value that represents portions of a data string that are considered to be significant.

Example: Matching Names and Addresses

Match definitions generate identical match codes for records that represent the same entity even if the records themselves are not identical.

CONTACT	ADDRESS	CONTACT_MatchCode	ADDRESS_MatchCode	Cluster_Num
James E. Briggs	19 East Broad Street	MY&F$$$$$$$$$$C&B_4$$$$$$$	Z-$$$MY~$$$$$$$	0
Mr James Brigs	19 E Broad St	MY&F$$$$$$$$$$C&B_4$$$$$$$	Z-$$$MY~$$$$$$$	0
Jim Briggs	19 E. BROAD ST.	MY&F$$$$$$$$$$C&B_4$$$$$$$	Z-$$$MY~$$$$$$$	0
Bob Brauer	6512 Six Forks Road - 404B	MY&L&Y$$$$$$$$$M@M$$$$$$$$$	65ZH$6GY3S0SM$$	1
Bob Brauer	6512 Six Forks #404B	MY&L&Y$$$$$$$$$M@M$$$$$$$$$	65ZH$6GY3S0SM$$	1
Robert Brauer	6512 Six Frks Road Ste 404B	MY&L&Y$$$$$$$$$M@M$$$$$$$$$	65ZH$6GY3S0SM$$	1

The match codes from one or more fields can be used to identify duplicate records within a table or to join records between tables.

JAMES COOK UNIVERSITY

Match Code Generation Overview

1 **Pre-process Text**	Apply standardization schemes to the text before processing.
2 **Parse Text**	Identify the tokens of the text (parse or extraction definitions).
3 **Morph Analysis**	New and experimental steps meant for specialized use.
4 **Token Combination Rules**	
5 **For Each Token:**	
a **Suggestions**	Identify spelling variants (scheme).
b **Normalize**	Normalize text (regular expressions).
c **Remove Noise**	Remove noise words (vocabulary).
d **Standardize**	Standardize text (standardization schemes).
e **Phonetic Reduction**	Reduce text phonetically (phonetics library).
f **Matchcode Layout**	Specify ranges at various sensitivities.
6 **Concatenate Tokens**	After processing each token, resultant tokens are concatenated.
7 **Generate Match Code**	Encoded match codes generated for concatenated tokens.

Assessment 2

JAMES COOK UNIVERSITY AUSTRALIA

1.Register **data sources** for the two data tables in DataFlux Data Management Studio

2.Create an exploration to identify the types of data contained in the table, as well as **create any collections** that might be useful in the future.

3.Create a profile that analyses the demographic and high school information for **at-risk first- year students**, to determine whether it is appropriate for analysis:

1. You might want to create a **filter on only the necessary data** for the profile prior to processing, because the profile could fill up the memory on your Virtual Lab machine and fail.
2. Pay close attention to the **metrics** that you calculate for each column, because some are very process intensive and some metrics are not applicable to certain data elements.

4.Create **charts and graphs** from the profile report that **outline the overall cleanliness** of the data.

5.If time permits, **build jobs to improve** the quality and consistency of the data. Be sure to save the target tables for later use.

Assignment 2

JAMES COOK UNIVERSITY AUSTRALIA

Your report needs to answer the following questions based on the outcomes of your data analysis:

1. What is the overall quality of the available data?
2. What data elements need to be cleansed before you proceed with analysis and reporting?
3. What outliers exist in the numeric fields?
4. Do the columns of demographic data contain the expected data values?
5. Are the fields containing demographic data suitable for analysis and reporting? If not, what changes need to be made to the data?
6. How many students are identified as at-risk?
7. Are there any notable demographic patterns that exist among the at-risk students?
8. What processes should be used to improve the quality of the data to be used in reports?
9. Which columns require data standardisation?
10. Which columns need the case changed for consistency?
11. Are there any records where entity resolution might help?
12. Are there any applications where parsing can add to the value of the data?
13. Do you have full data for the relevant variables for analysis and reporting? If not, what additional information might be helpful for analysing the data?