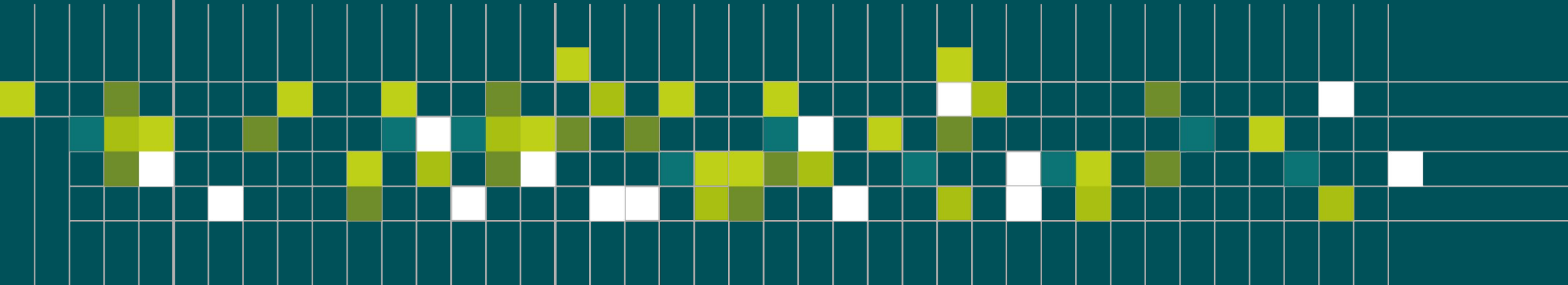




Tabular dataset analysis

Concepts and principles



Contents

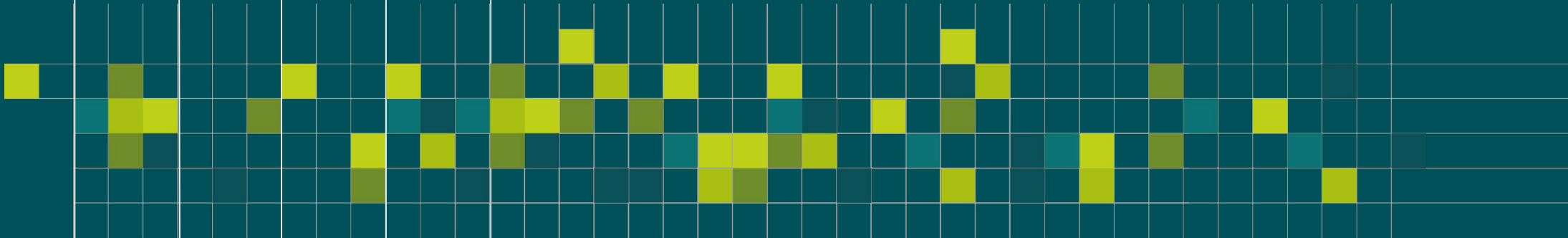
Introduction

Field

Relationship

Dataset

Example



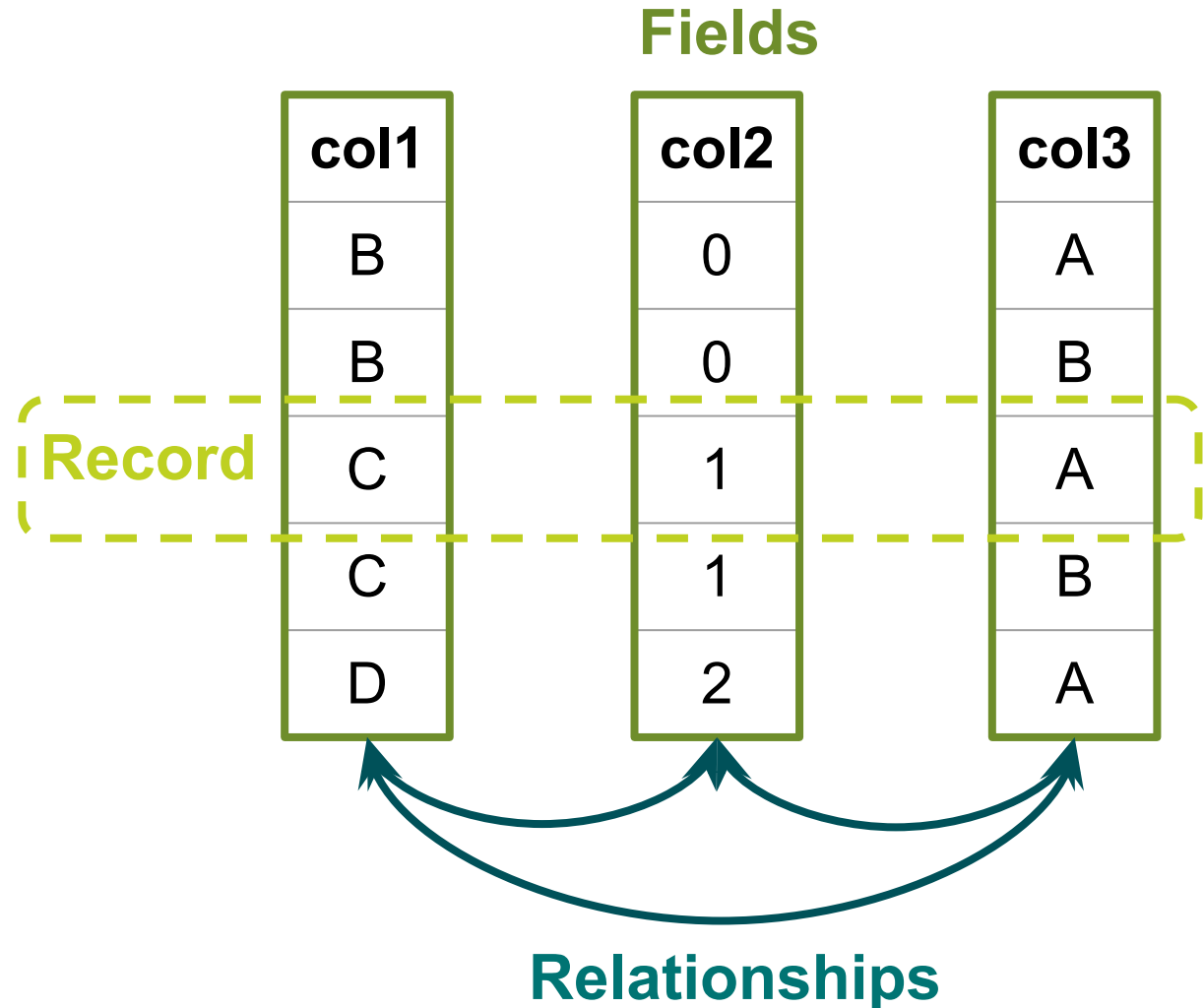
Dataset structure

Record oriented

- Dataset is a list of records
- A semantic entity is a record
- Length is variable

Field oriented

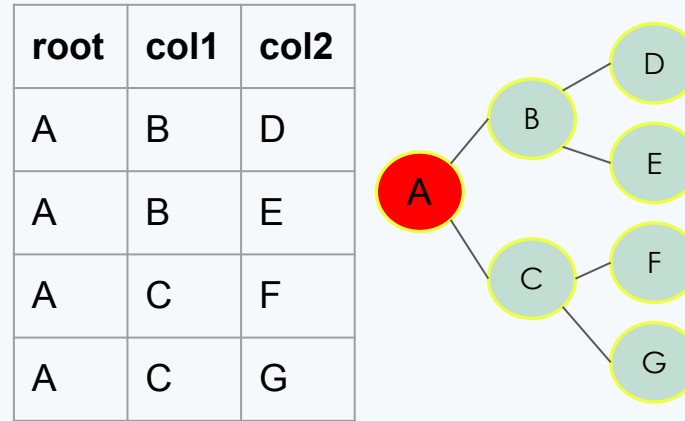
- Fields have semantics
- Fields are dependent
- A semantic entity is a set of record or the entire Dataset



Dataset structure analysis

Dataset structure

- Tree structure
- Matrix structure
- Mixed structure



Tree structure

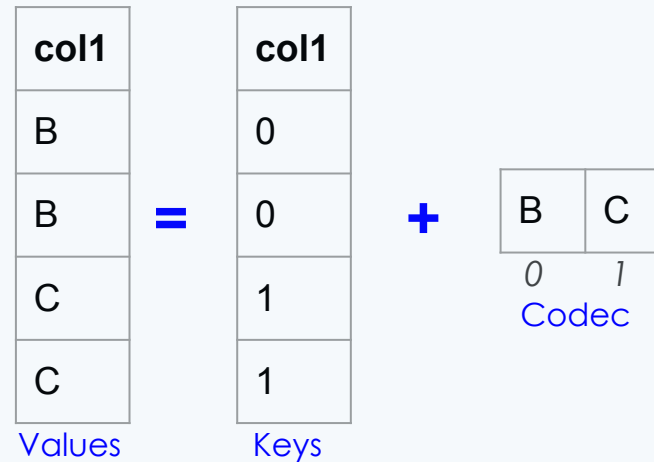
val	col3	col4
1	A	C
2	A	D
3	B	C
4	B	D

	A	B
C	1	3
D	2	4

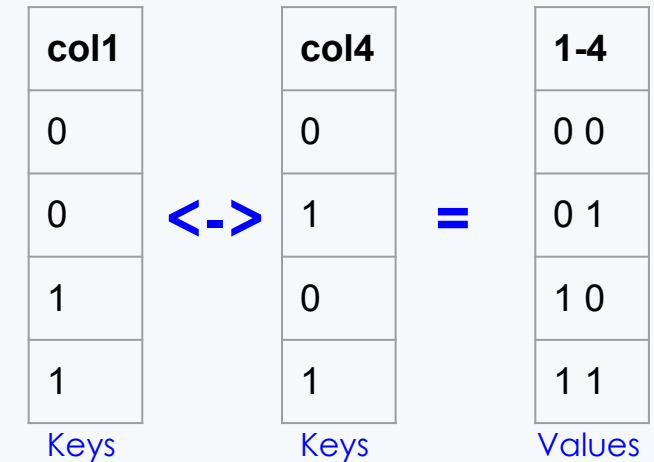
Matrix structure

Analysis structure

- Field
- Relationship



Field structure analysis



Relationship structure analysis

Definition

Values

[Anne, Paul, Anne, Lea, Lea]

Codec (row)

Anne	0
Paul	1
Anne	2
Lea	3
Lea	4

Anne	0
Paul	1
Lea	2
Anne	3

Anne	0
Paul	1
Lea	2

Keys

[0, 1, 2, 3, 4]

[0, 1, 3, 2, 2]

[0, 1, 0, 2, 2]

Full
codec : all values
keys : no duplicate

Default
codec : different values

A Codec defines the correspondence between values and keys (e.g.) :

- 1 : Anne
- 0 : Paul
- 2 : John

A Codec may not be bijective (e.g.) :

- 0 : Anne
- 1 : Paul
- 2 : Anne

A Keys is a list of integers where :

- the maximum is the length of Codec - 1
- Each integer is present in the list

- A Field is a representation of a list of Values
- A Field is defined by a Keys list and a Codec list
- A Field is canonical if the keys is ordered
- A Field where values are row number is the “root field”



Definition

Values	[Anne, Paul, Anne]	[Anne, Anne, Anne]	[Anne, Paul, Anne]
Keys	[0, 1, 2]	[0, 0, 0]	[0, 1, 0]
Length <small>(number of values)</small>	3	3	3
Codec <small>(row)</small>	<div>0</div> <div>1</div> <div>2</div>	<div>0</div>	<div>0</div> <div>1</div>
Type codec	full	unique	default
Property	R : 1 DM : 0	R : 0 Dm : 0	R : 0 Dm : 0
Representation	Codec : [Anne, Paul, Anne] Keys: implicit ([0,1,2])	Codec : [Anne] Keys: implicit ([0, 0, 0])	Codec : [Anne, Paul] Keys: [0, 1, 0]

Indicators :

M: len(values)maxcodec

m: len(set(values))mincodec

x: len(codec)lencodec

k: maxc(keys)maxkeys

* maxc : max(counter())

R: (M - x) / (M - m)ratecodec

Dm: x - mdmencodec

DM: M - xdmmaxcodec

Mm: M - mrancodec

Keys typology

M = 0	Keys empty
M = x (k = 1)	Keys without duplicate data
x = 1	Keys with unique 0 value
[0, 1, 2, ... , M]	identity Keys
k = M / x	Keys is distributed

Codec typology

M = 0 (m = x = 0)	null
x = 1 (m = 1)	unique
m = M = x (x > 1)	complete (rooted)
[0, 1, 2, ... , m]	Identity Codec
m < M = x	full (rooted)
x = m < M (x > 1)	default
m < x < M	mixed

The Root Field is the Field with Identity Keys and Identity Codec

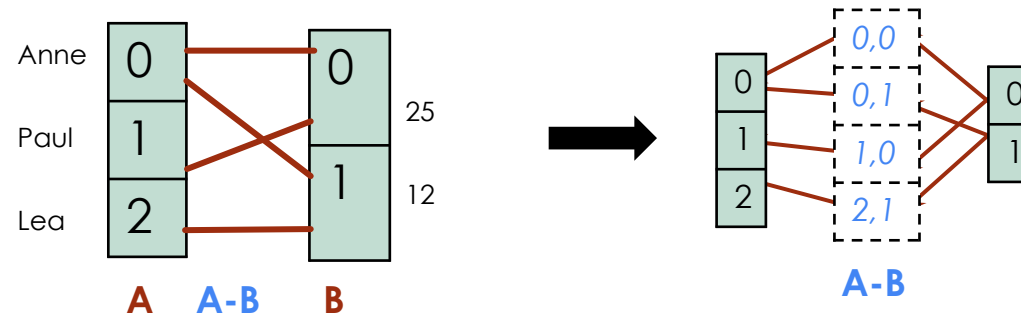
Properties

- Any Field have **default** Codec and **rooted** Codec
- **Default** Codec is the shortest Codec, **rooted** is the longest
- The only bijective Codec is the **default** (or **complete**) Codec
- The **Identity** Codec is **default** or **complete** Codec
- The maximum Keys value is the length of Codec plus one
- If Codec is the **Identity** Codec, Keys and Values are equals
- If Keys is the **Identity** Keys, Codec and Values are equals
- In a **distributed** Field all values are present with the same frequency
- **Full** and **unique** Field are **distributed**
- The **Root** Field is **complete**
- The **Root** Field has identical Keys, Values and Codec

Definition

	<i>values</i>	<i>codec</i>	<i>keys</i>
Field A	[Anne, Paul, Anne, Lea]	[Anne, Paul, Lea]	[0, 1, 0, 2]
Field B	[25, 25, 12, 12]	[25, 12]	[0, 0, 1, 1]
Relationship A - B	[0-0, 1-0, 0-1, 2-1]	[0-0, 1-0, 0-1, 2-1]	[0, 1, 2, 3]

Notation



- A Relationship is a virtual Field where
 - values are pairs of each keys
 - codec is the default codec
- The value of a Relationship is the lencodec of the virtual Field : X_{A-B}

Definition

Values	A [Anne, Paul, John, Paul] B [25, 26, 15, 26]	A [Anne, Paul, Anne, Lea] B [25, 25, 25, 12]	A [Anne, Paul, Anne, Lea] B [25, 25, 12, 12]	A [Anne, Anne, Paul, Paul, Lea, Lea] B [25, 12, 25, 12, 25, 12]
Codec				
Type	coupled	derived	linked	Crossed
	$\text{dist} = \text{dmin}$ $\text{diff} = 0$	$\text{dist} = \text{dmin}$ $0 < \text{diff} < \text{dmin}$	$\text{dmin} < \text{dist} < \text{dmax}$ $0 \leq \text{diff} \leq \text{dmin}$	$\text{dist} = \text{dmax}$ $0 \leq \text{diff} < \text{dmin}$
Keys	B Implicit (equal Keys A)	Relative to keys A	Absolute	Implicit (Matrix order)
Example B	Codec :[25, 26, 15, 35] Keys: implicit	Codec :[25, 12] Keys: relative [0,0,1]	Codec :[25, 12] Keys: absolute [0,0,1,1]	Codec :[25, 12] Keys: implicit ([0,1,0,1,0,1])

Indicators :

$$\begin{aligned}
 \text{dmax} &= x_A * x_B \\
 \text{dmin} &= \max(x_A, x_B) \\
 \text{dran} &= \text{dmax} - \text{dmin} \\
 \text{diff} &= \text{abs}(x_A - x_B) \\
 \text{dist} &= x_{(A, B)}
 \end{aligned}$$

Rules :

$$\begin{aligned}
 \text{A and B : derived} & \quad x_{(A, B)} = \text{dmin} \\
 \text{A and B : coupled} & \quad x_{(A, B)} = x_B = x_A \\
 \text{A and B : crossed} & \quad x_{(A, B)} = x_B * x_A
 \end{aligned}$$

Additional rules:

$$\begin{aligned}
 \text{A and B : distributed} & \quad k_{(A, B)} * x_{(A, B)} = M_A = M_B \\
 \text{A and B : full distributed} & \quad k_{(A, B)} = 1
 \end{aligned}$$

Relative keys :

Length:
 $\text{length}(\text{parent.codec})$
 Values:
 $\text{Keyder}(\text{parent.key}(i)) = \text{key}(i)$

Properties

- Type and Indicators are independent of Values (order or value)
- All Fields are **derived** or **coupled** with an **unique** Field
- All Fields are **derived** or **coupled** with a **rooted** Field
- If A is **derived** (**coupled**) with B ($x_B \leq x_A$) and B is **derived** (**coupled**) with C ($x_C \leq x_B$), A is **derived** (**coupled**) with C and $\text{diff}(A,C) = \text{diff}(A,B) + \text{diff}(B,C)$
- If A and B are **coupled**, all the relationships with other indexes are identical
- If A and B are **crossed**
 - if C is **derived** (**coupled**) with A ($x_C \leq x_A$): B and C are **crossed**
 - $x_A * x_B \leq M_A$ If $x_A * x_B = M_A$, A and B are **full distributed**
 - All combinations of values are present
 - If A and B are **distributed**, the relationship is **distributed**
- **Keys can be deduced with coupled relationship**
 - A and B are **coupled** $\Rightarrow \text{keys}(B) = \text{keys}(A)$
- **Keys can be reduced with derived relationship (relative keys)**
 - B is **derived** with A ($x_B \leq x_A$) $\Rightarrow \text{len}(\text{relative_keys}(B)) = \text{len}(\text{codec}(A))$

Distance

Distance (resp distomin): number of codec links to remove to be coupled (resp derived)

Distomax: number of codec links to add to be crossed

$$X_B \leq X_A: \quad dmin = X_A \quad diff = X_A - X_B \quad dmin - diff = X_B$$

Distance: $dist - dmin + diff$ $X_{(A, B)} - X_B$

Distomin: $dist - dmin$ $X_{(A, B)} - X_A$

Distomax: $dmax - dist$ $X_A * X_B - X_{(A, B)}$

RateCpl: $distance / (distance + distomax)$

1 - RateCpl: $distomax / (distance + distomax)$

RateDer: $distomin / (dmax - dmin)$

1 - RateDer: $distomax / (dmax - dmin)$

$X_B \leq X_A$	distance	distomin	distomax	rateCpl	rateDer
coupled	0	0	$X_A * X_A - X_A$	0	0
derived	$X_A - X_B$		$X_A * X_B - X_A$	$(X_A - X_B) / (X_A * X_B - X_B)$	
linked	$X_{(A, B)} - X_B$	$X_{(A, B)} - X_A$	$X_A * X_B - X_{(A, B)}$	$(X_{(A, B)} - X_B) / (X_A * X_B - X_B)$	$(X_{(A, B)} - X_A) / (X_A * X_B - X_A)$
crossed	$X_A * X_B - X_B$	$X_A * X_B - X_A$	0	1	1

Distance properties

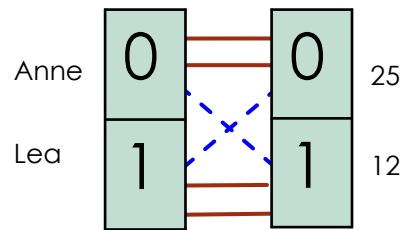
- If $\text{distance}(A,B) = 0$, A and B are coupled
- If A is derived from B and B is derived from C:
 $\text{distance}(A,C) = \text{distance}(A,B) + \text{distance}(B,C)$
- The maximal distance is between root Field and unique Field
- The distance of a Field to the Root Field is equal to the “dmaxcodec”

$X_B \leq X_A$	distance	distomin	distomax	rateCpl	rateDer
X_A and X_B unique	0	0	0	0	0
X_B unique	$X_A - 1$			1	
X_A root (len)	$\text{len} - X_B$		$\text{len} * X_B - \text{len}$	$(\text{len} - X_B) / (\text{len} * X_B - X_B)$	
X_A root X_B unique	$\text{len} - 1$		0	1	
$A = B$	0		$X_A * X_A - X_A$	0	
$X_A = X_B$	distance = distomin			rateCpl = rateDer	

Distance - example

coupled

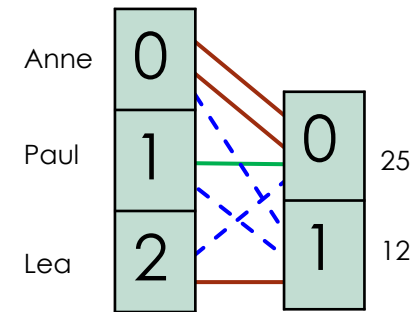
[Anne, Anne, Lea, Lea]
[25, 25, 12, 12]



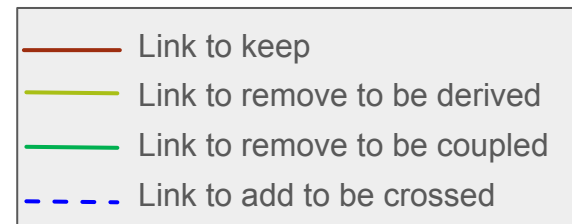
Dmax : 4
Dmin : 2
Diff : 0
Dist : 2
Distance : 0
Distomin: 0
Distomax: 2

derived

[Anne, Paul, Anne, Lea]
[25, 25, 25, 12]



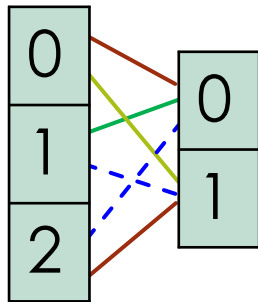
Dmax : 6
Dmin : 3
Diff : 1
Dist : 3
Distance: 1
Distomin: 0
Distomax: 3



Distance - Examples

linked

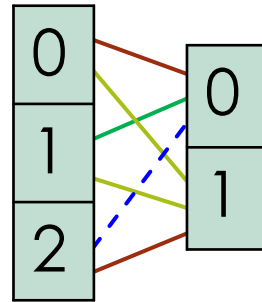
[Anne, Paul, Anne, Lea]
[25, 25, 12, 12]



Dmax : 6
Dmin : 3
Diff : 1
Dist : 4
Distance: 1 + 1
Distomin: 1
Distomax: 2

linked

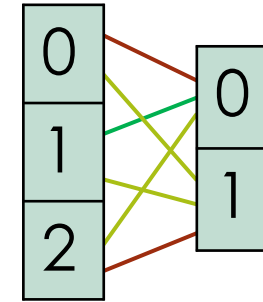
[Anne, Anne, Paul, Paul, Lea]
[25, 12, 25, 12, 12]



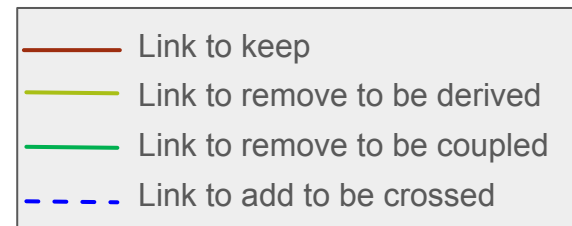
Dmax : 6
Dmin : 3
Diff : 1
Dist : 5
Distance : 1 + 2
Distomin: 2
Distomax: 1

crossed

[Anne, Anne, Paul, Paul, Lea, Lea]
[25, 12, 25, 12, 25, 12]



Dmax : 6
Dmin : 3
Diff : 1
Dist : 6
Distance : 1 + 3
Distomin: 3
Distomax: 0



Definition - properties

- **Dataset definition**

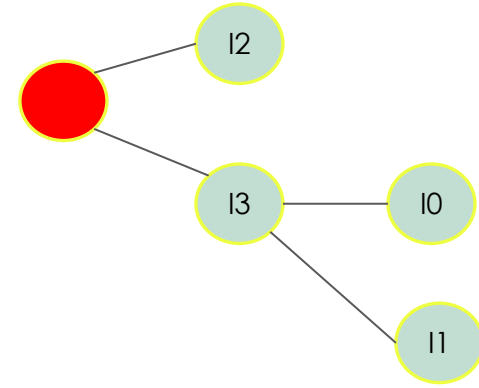
- Datasets have a hidden Field: **Root** Field
- Fields are ordered (**Root** Field is the first)
- Relationships are oriented :
 - **Parent** Field has the highest lencodec, **Child** Field has the lowest
 - If lencodec are equal, **Parent** Field has the lowest row (order of Fields)

- **Dataset properties**

- Each Field is **derived** (**coupled**) from at least one **parent** Field (the **root** Field)
- A dataset with only **derived** or **coupled** relationships is equivalent to a tree
- A dataset with only **crossed** relationships is equivalent to a multi-dimensional data
- With relationship adjustment, a dataset can be translated into a tree or a multi-dimensional data

Tree properties

- A dataset can be represented by a rooted tree where
 - **Nodes** are Field
 - **Root** is root Field
 - Each node has a single parent node (relationship parent)
- Different rules are available to define the parent node:
 - Option derived : First (row) derived Field with minimal distance
 - Option distance : First (row) Field with minimal distance
 - Option distomin : First (row) Field with minimal distomin



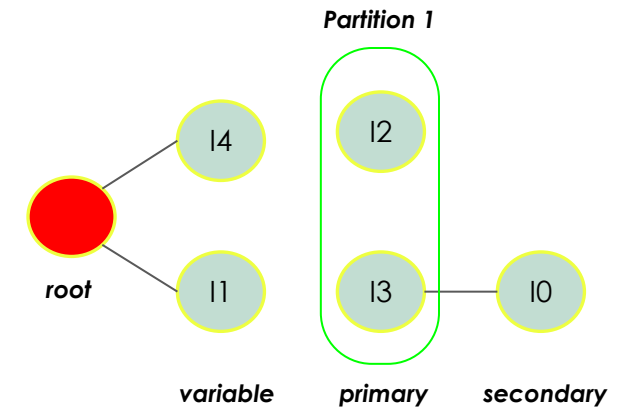
Fields tree typology

unique	Codec is Unique	leaf	Parent is root
rooted	Codec is complete or full (root coupled)	leaf	Parent is root
coupled	Field is coupled with a parent Field	leaf	Parent is the first previous coupled Field
derived	The Field is not derived with a Child Field	leaf	
mixed	Other Fields	Node	

Partition properties

- **Dataset partition**

- A **partition** is a set of Fields full distributed (i.e. each record is associated to a single combination of Field keys)
- Three categories of Fields are defined in a **partition** :
 - **Primary** Fields: Fields included in the **partition**
 - **Secondary** Fields: Fields **derived** / **coupled** with a **primary** Field
 - **Variable** Fields: other Fields (including **unique** Fields)
- The **default partition** is the **partition** with the largest size
- The **dimension** of a Dataset is the **default partition** size



- **Properties**

- A Dataset has at least one implicit **partition** (the **root partition**)
- A multi dimensional array is associated to each **partition**
- Keys data may be implicit for **primary** Fields
- **Dimension** can be reduced by codec extension
- **Dimension** can be increased by values extension
- In a **root partition**, all the Fields are **variable**, the **dimension** is 0 (the **primary** Field is the **root** Field)

Example

3 fields are derived

- First name
- Last name
- Group

1 field is coupled

- Surname

1 field is unique

- Year

3 fields are almost crossed

- Full name
- Course
- Examen

1 field is almost rooted

- Score

first name	last name	full name	surname	group	course	year	examen	score
Anne	White	Anne White	skyler	gr1	math	2021	t1	11
Anne	White	Anne White	skyler	gr1	math	2021	t2	13
Anne	White	Anne White	skyler	gr1	math	2021	t3	15
Anne	White	Anne White	skyler	gr1	english	2021	t2	10
Anne	White	Anne White	skyler	gr1	english	2021	t3	12
Philippe	White	Philippe White	heisenberg	gr2	math	2021	t1	15
Philippe	White	Philippe White	heisenberg	gr2	english	2021	t2	8
Camille	Red	Camille Red	saul	gr3	software	2021	t3	17
Camille	Red	Camille Red	saul	gr3	software	2021	t2	18
Camille	Red	Camille Red	saul	gr3	english	2021	t1	2
Camille	Red	Camille Red	saul	gr3	english	2021	t2	4
Philippe	Black	Philippe Black	gus	gr3	software	2021	t3	18
Philippe	Black	Philippe Black	gus	gr3	english	2021	t1	6

78% almost crossed

83% almost crossed

1.5 %
almost rooted

coupled

derived

unique

Ratio ratecpl

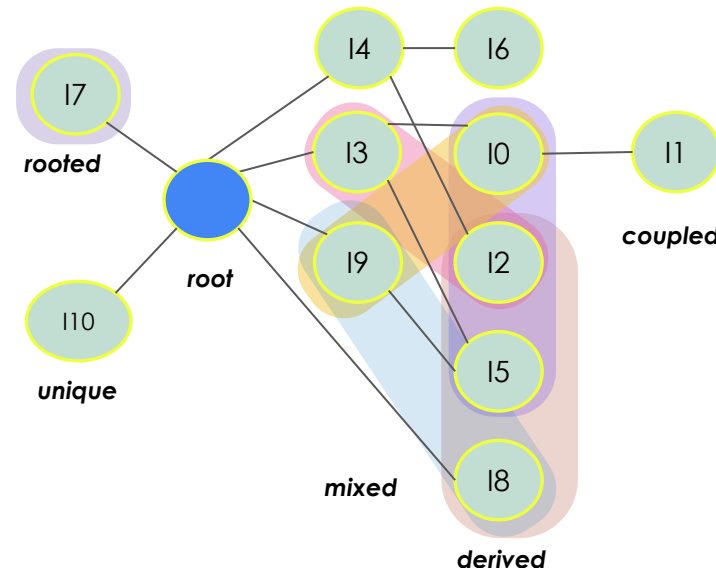
- Full name – Examen : 78 %
- Score – Root : 1,5 %
- Course – Examen : 83 %

Partition example

product	plants	plts	quantity	price	price level	group	id	supplier	location	valid
apple	fruit	fr	1 kg	1	low	fruit1	1001	sup1	fr	ok
apple	fruit	fr	10 kg	10	low	fruit10	1002	sup1	gb	ok
orange	fruit	fr	1 kg	2	high	fruit1	1003	sup1	es	ok
orange	fruit	fr	10 kg	20	high	veget	1004	sup2	ch	ok
peppers	vegetable	ve	1 kg	1.5	low	veget	1005	sup2	gb	ok
peppers	vegetable	ve	10 kg	15	low	veget	1006	sup2	fr	ok
carrot	vegetable	ve	1 kg	1.5	high	veget	1007	sup2	es	ok
carrot	vegetable	ve	10 kg	20	high	veget	1008	sup1	ch	ok

Derived tree :

- 1: root-derived (8)
 - 3 : product (4 - 4)
 - 0 : plants (2 - 2)
 - 1 : plts (0 - 2)
 - 5 : price level (2 - 2)
- 4 : price (2 - 6)
 - 2 : quantity (4 - 2)
 - 6 : group (3 - 3)
- 7 : id (0 - 8)
- 8 : supplier (6 - 2)
- 9 : location (4 - 4)
- 10 : valid (7 - 1)



Partitions :

['plants', 'price level', 'quantity'],
 ['price level', 'quantity', 'supplier'],
 ['location', 'plants'],
 ['location', 'supplier'],
 ['product', 'quantity'],
 ['id']