Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

Databases and Information Systems (DIS) – Quiz 2

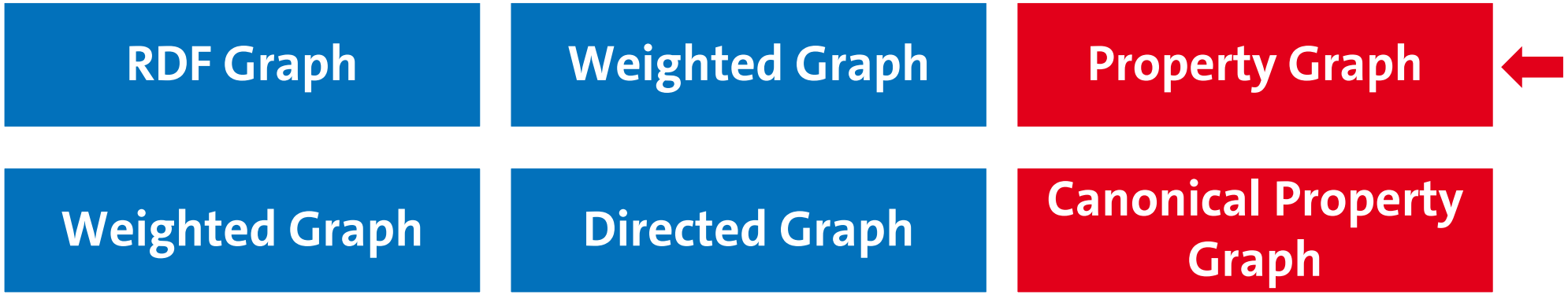
Universität Hamburg

Modern DBS: NoSQL Systems

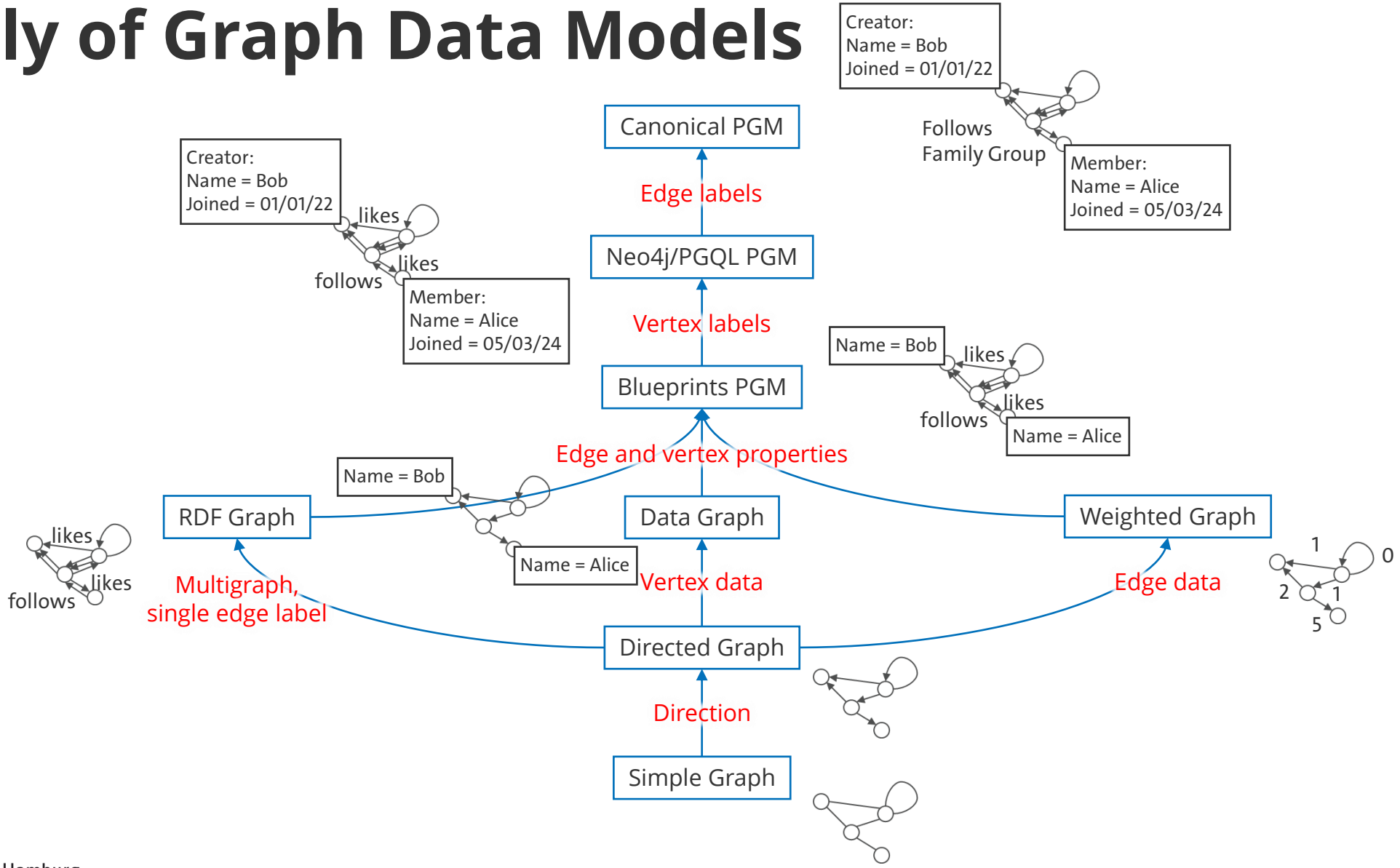
Which graph models include edge and vertex properties? Which model is used in neo4j?

- RDF Graph
- Weighted Graph
- Property Graph
- Weighted Graph
- Directed Graph
- Canonical Property Graph

Which graph models include edge and vertex properties? Which model is used in neo4j?



Family of Graph Data Models



Which of the following elements can directly be modelled using neo4j?

- Entities with properties
- Edge properties
- Multivalued properties
- N-ary relationships
- N:M relationships

Which of the following elements can directly be modelled using neo4j?

Entities with properties

Edge properties

Multivalued properties

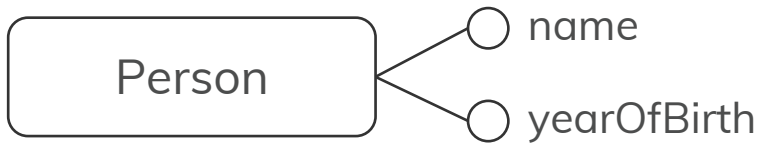
N-ary relationships

N:M relationships

Property Graph Modelling

Entity with Properties

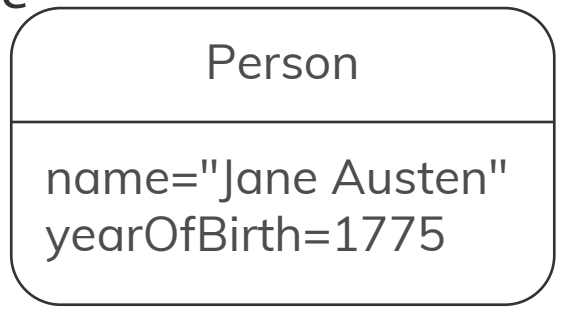
- Entity relationship model

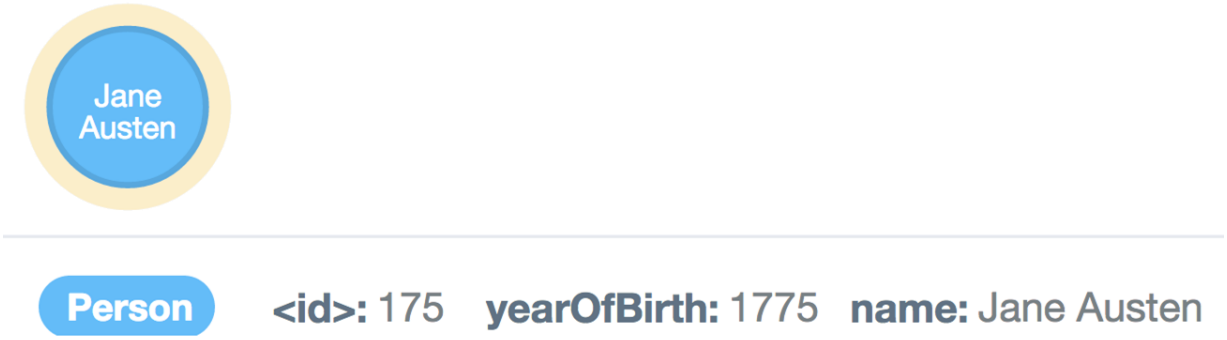


- Schema typically implicit, i.e. given with instances

```
(ja:Person { name: 'Jane Austen', yearOfBirth: 1775 })
```

- Instance

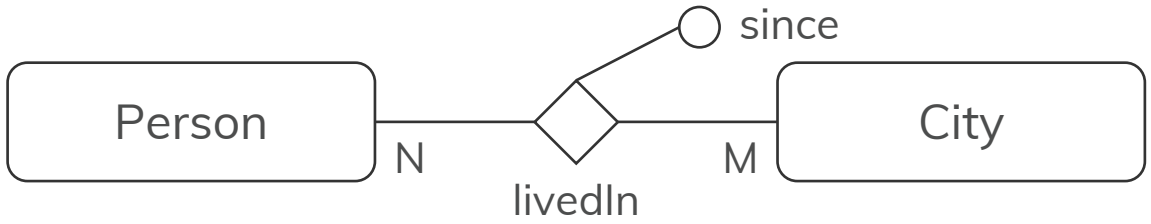




Property Graph Modelling

Relationships (N:M)

- Entity relationship model

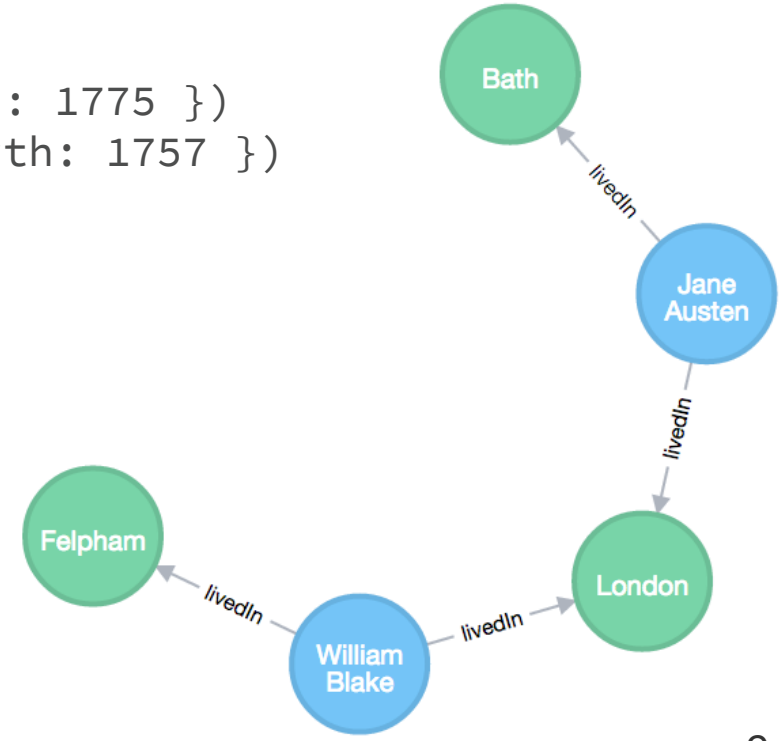


- Vertices

```
(ja:Person { name: 'Jane Austen', yearOfBirth: 1775 })
(wb:Person { name: 'William Blake', yearOfBirth: 1757 })
(lo:City {name: 'London'})
(ba:City {name: 'Bath'})
(fe:City {name: 'Felpham'})
```

- Edges

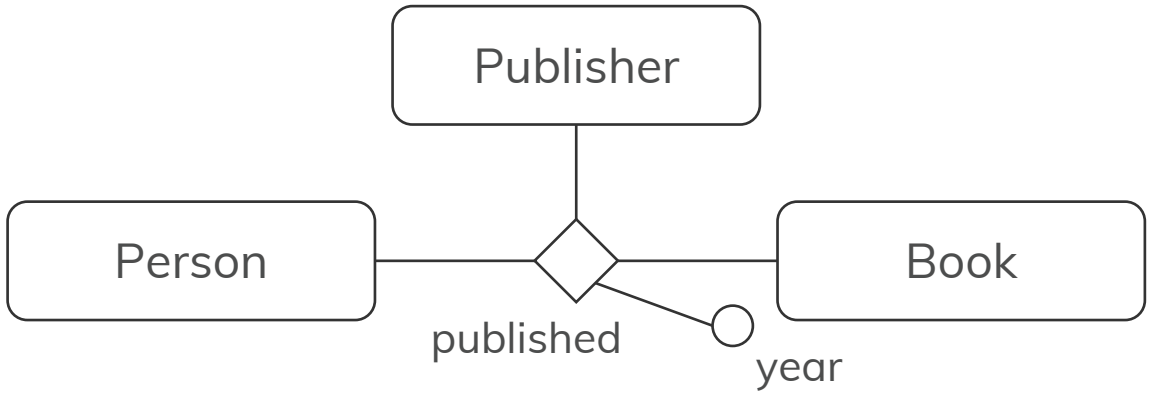
```
(ja)-[:livedIn {since: 1775}]->(lo)
(ja)-[:livedIn {since: 1800}]->(ba)
(wb)-[:livedIn {since: 1757}]->(lo)
(wb)-[:livedIn {since: 1800}]->(fe)
```



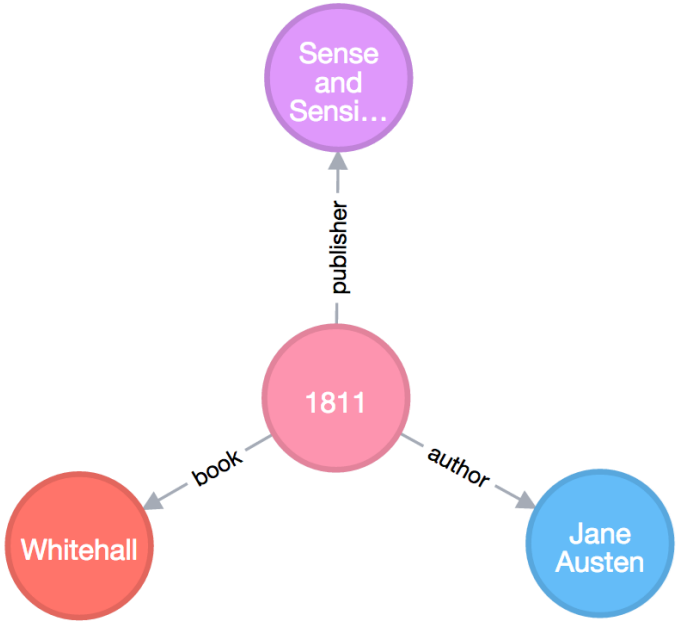
Property Graph Modelling

n-ary relationships (with n>2)

- Entity relationship model

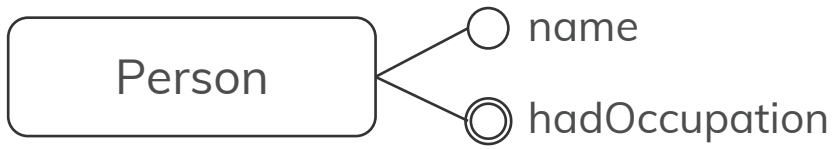



```
(ja:Person { name: 'Jane Austen', yearOfBirth: 1775 })
(wh:Publisher { name: 'Whitehall' })
(sas:Book {title: 'Sense and Sensibility' })
(pub:Publication {year: 1811 })
(pub)-[:author]->(ja)
(pub)-[:book]->(wh)
(pub)-[:publisher]->(sas)
```

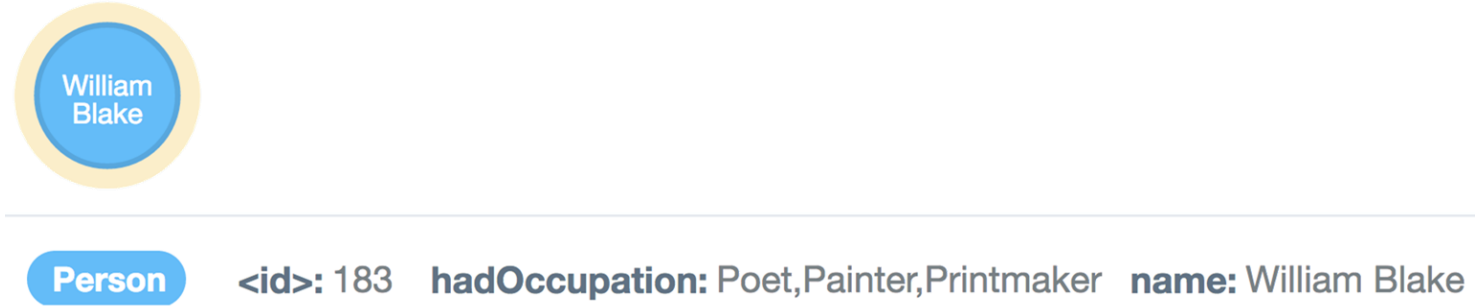


Property Graph Modelling

Multivalued properties

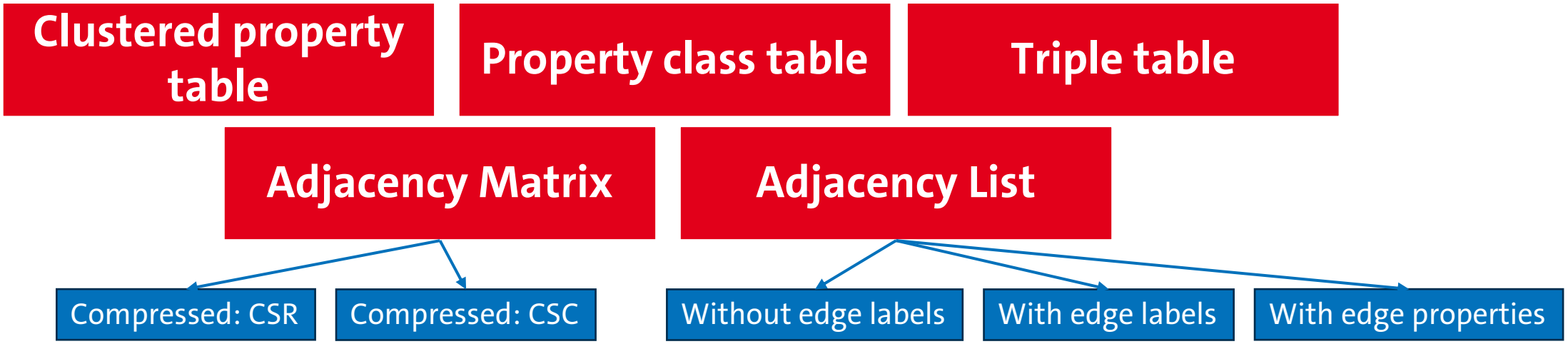



```
(wb:Person {name: 'William Blake', hadOccupation: ['Poet','Painter','Printmaker']})
```

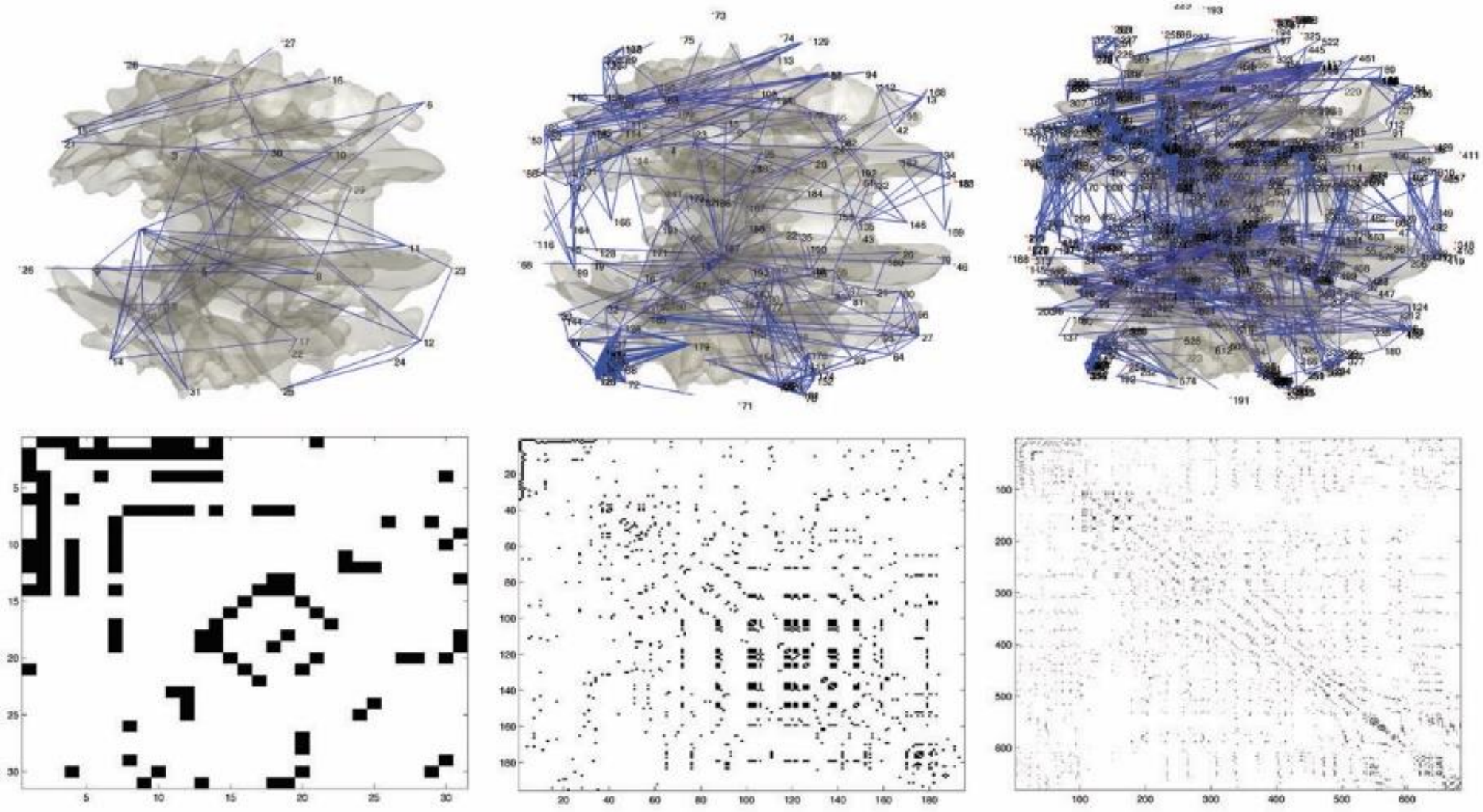


Name 3 different ways of representing graph data in a database system.

Name 3 different ways of representing graph data in a database system.



Adjacency Matrix

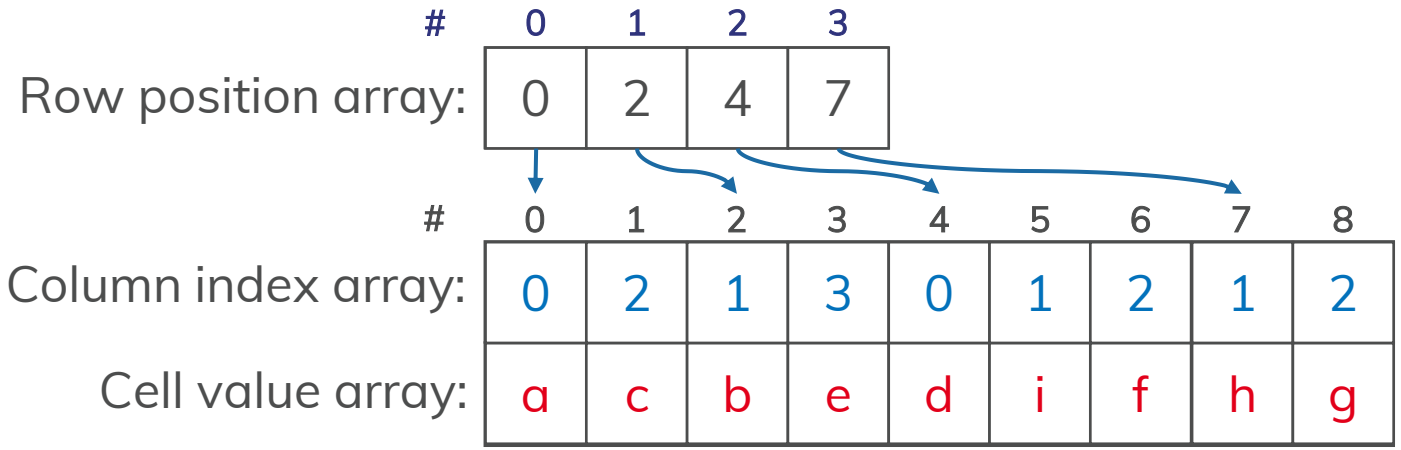


[http://brainimaging.waisman.wisc.edu/~chung/graph/admatrix.jpg]

Compress Sparse Row (CSR)

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} a & & c & \\ & b & & e \\ d & i & f & \\ & h & g & \end{pmatrix} \end{array}$$

Position for row # in other two arrays:



Representations: Adjacency List

Source vertex with outgoing edges...

$$\begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} a & & c & \\ & b & & e \\ d & i & f & \\ & h & g & \end{pmatrix} \end{array}$$

...without edge labels

0 -> (0,2)
1 -> (1,3)
2 -> (0,1,2)
3 -> (1,2)

...with edge labels

0 -> ([0,a],[2,c])
1 -> ([1,b],[3,e])
2 -> ([0,d],[1,i],[2,f])
3 -> ([1,h],[2,g])

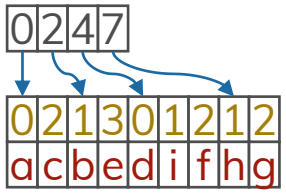
...with edge properties

0 -> ([0,a,(weight=4)],[2,c,(weight=3)])
1 -> ([1,b,(weight=3)],[3,e,(weight=2)])
2 -> ([0,d,(weight=5)],[1,i,(weight=2)],...)
3 -> ([1,h,(weight=9)],[2,g,(weight=7)])

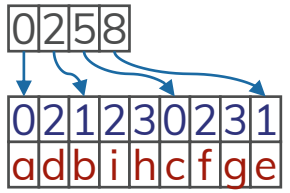
The same!

Almost the same!

Compressed Sparse Row



Compressed Sparse Column



source-oriented

target-oriented

Property Table Approaches

Triple Table

Subj.	Prop.	Obj.
ID1	type	BookType
ID1	title	"XYZ"
ID1	author	"Fox, Joe"
ID1	copyright	"2001"
ID2	type	CDType
ID2	title	"ABC"
ID2	artist	"Orr, Tim"
ID2	copyright	"1985"
ID2	language	"French"
ID3	type	BookType
ID3	title	"MNO"
ID3	language	"English"
ID4	type	DVDType
ID4	title	"DEF"
ID5	type	CDType
ID5	title	"GHI"
ID5	copyright	"1995"
ID6	type	BookType
ID6	copyright	"2004"

(Clustered) property table

Property Table

Subj.	Type	Title	copyright
ID1	BookType	"XYZ"	"2001"
ID2	CDType	"ABC"	"1985"
ID3	BookType	"MNP"	NULL
ID4	DVDType	"DEF"	NULL
ID5	CDType	"GHI"	"1995"
ID6	BookType	NULL	"2004"

Left-Over Triples

Subj.	Prop.	Obj.
ID1	author	"Fox, Joe"
ID2	artist	"Orr, Tim"
ID2	language	"French"
ID3	language	"English"

Property-Class Table

Class: BookType

Subj.	Title	Author	copyright
ID1	"XYZ"	"Fox, Joe"	"2001"
ID3	"MNP"	NULL	NULL
ID6	NULL	NULL	"2004"

Class: CDType

Subj.	Title	Artist	copyright
ID2	"ABC"	"Orr, Tim"	"1985"
ID5	"GHI"	NULL	"1995"

Left-Over Triples

Subj.	Prop.	Obj.
ID2	language	"French"
ID3	language	"English"
ID4	type	DVDType
ID4	title	"DEF"

Reduce numbers of subject-subject self joins necessary to reconstruct entities

Which of the following are distance measures in graphs?

Degree	Eccentricity
Diameter	Radius

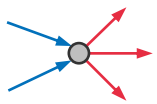
Which of the following are distance measures in graphs?

Degree	Eccentricity
Diameter	Radius

Basic Concepts: Measures

Degree (Valency)

Degree of a vertex *v*



$$\deg(v) = \deg_{out}(v) + \deg_{in}(v)$$

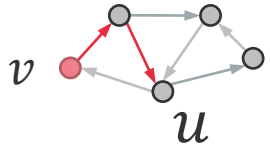
Degree of a graph *G*

$$\deg(G) = \max_{v \in V} \deg(v)$$

Degree distribution

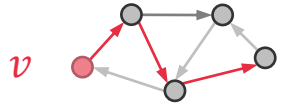
→ Probability distribution of vertex degree in a graph

Distance



→ Number of edges in a shortest path connecting two vertices

Eccentricity



→ Longest shortest path starting from v

Average Distance

→ Average shortest path
→ Average eccentricity of any vertex in the graph

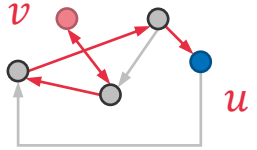
Radius

→ Minimum eccentricity of any vertex in the graph

$$r(G) = 2$$

Diameter

→ Maximum eccentricity of any vertex in the graph



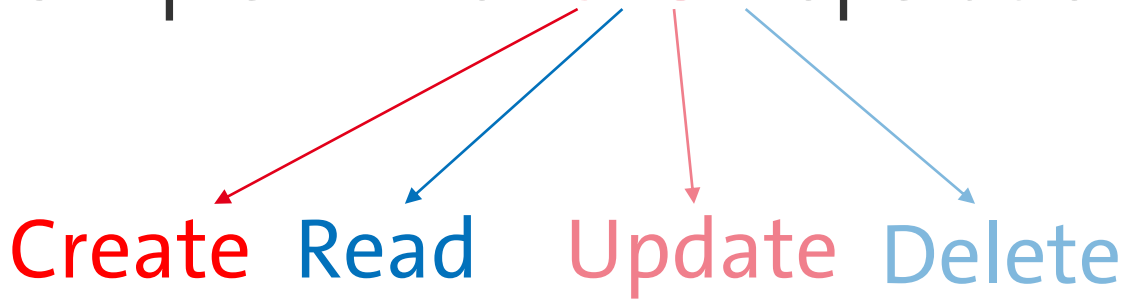
$$d(G) = 4$$

What do the letters CRUD (in relation to NoSQL DBs) stand for?

What do the letters CRUD (in relation to NoSQL DBs) stand for?

Key value store:

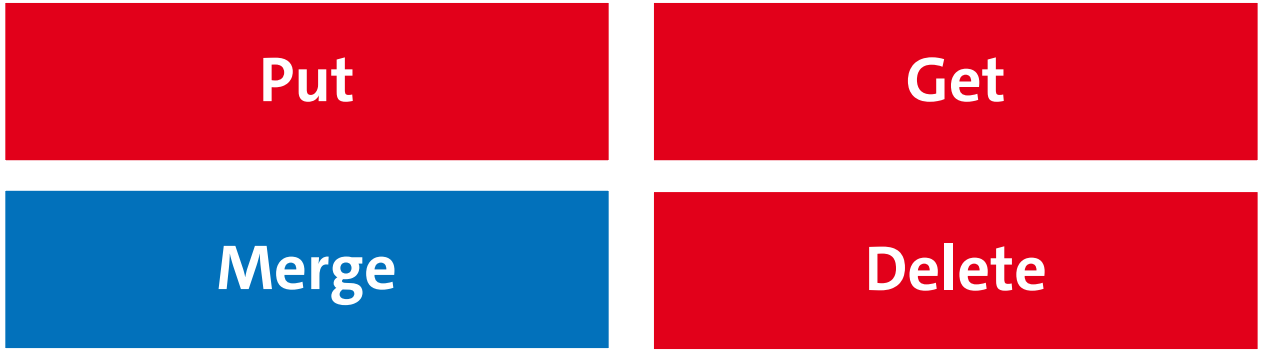
Basic key-value mapping with a simple API for **CRUD** operations



Which operations are offered by every key-value store?

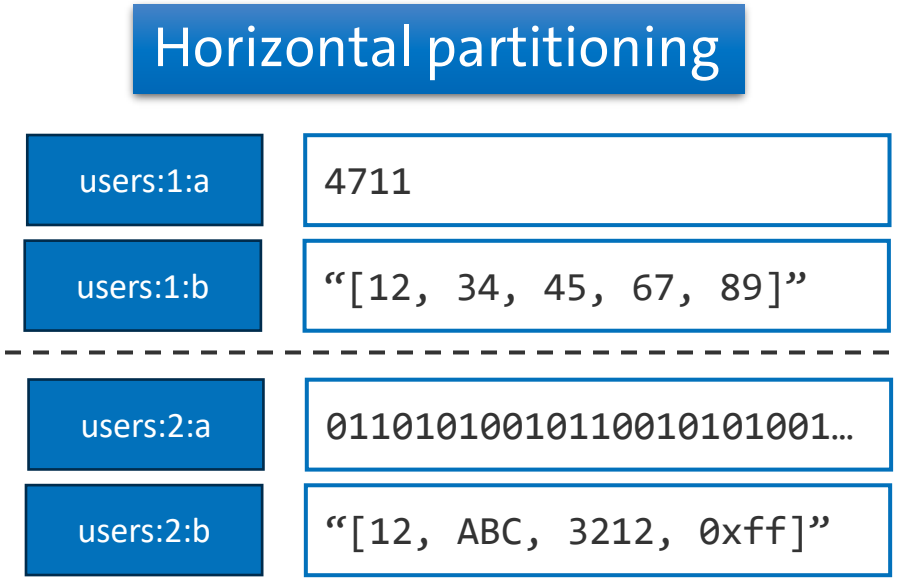
- Put
- Get
- Merge
- Delete

Which operations are offered by every key-value store?



Key-Value Stores

Basic key-value mapping with a simple API for **CRUD** operations



CRUD realized by at least 3 types of queries
Put: Add a new pair
Get: Retrieve a pair
Delete

Additional Operators implemented by some systems
Merge
MuliGet/MGet
MSet
…

Which are the variants for merging levels in an LSM tree?

Which are the variants for merging levels in an LSM tree?

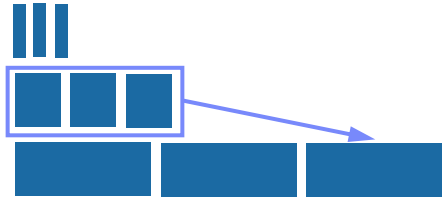
Tiering

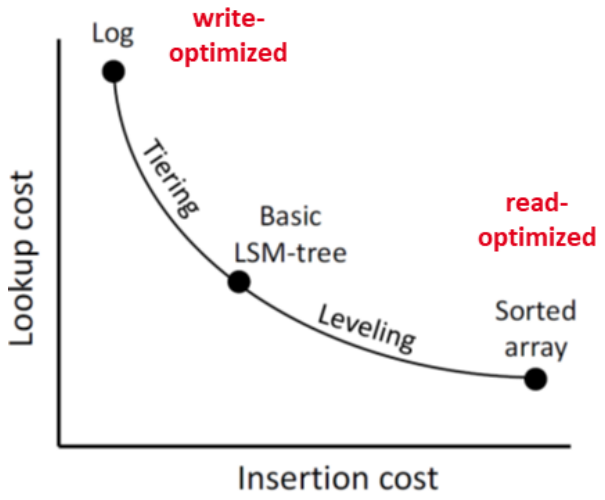
Leveling

LSM Trees: Merging

LSM Tiering

- Keep up to T-1 (sorted) runs per level L
- Merge all runs of L_i into 1 run of L_{i+1}
  - L1
  - L2
  - L3

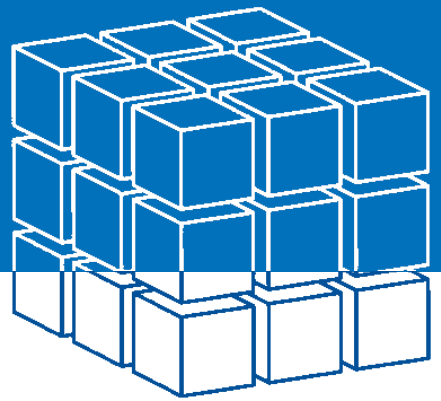




LSM Leveling

- Keep 1 (sorted) run per level L
- Sort-Merge run of L_i with L_i+1
  - L1
  - L2
  - L3

Data Warehouses and OLAP



Which of the following are characteristics of Data Warehouses?

classical

Integration of data from different source systems	No user updates
Transaction-oriented organization of data	Historic data
Analysis-oriented organization of data	Real time data

Which of the following are characteristics of Data Warehouses?

classical

Integration of data from different source systems	No user updates
Transaction-oriented organization of data	Historic data
Analysis-oriented organization of data	Real time data

Characteristics of DWHs

Analysis-oriented organization of data

Domain-oriented, models a specific application goal

Integration of data from different source systems

Integrated database, integration on structural and data level of multiple databases

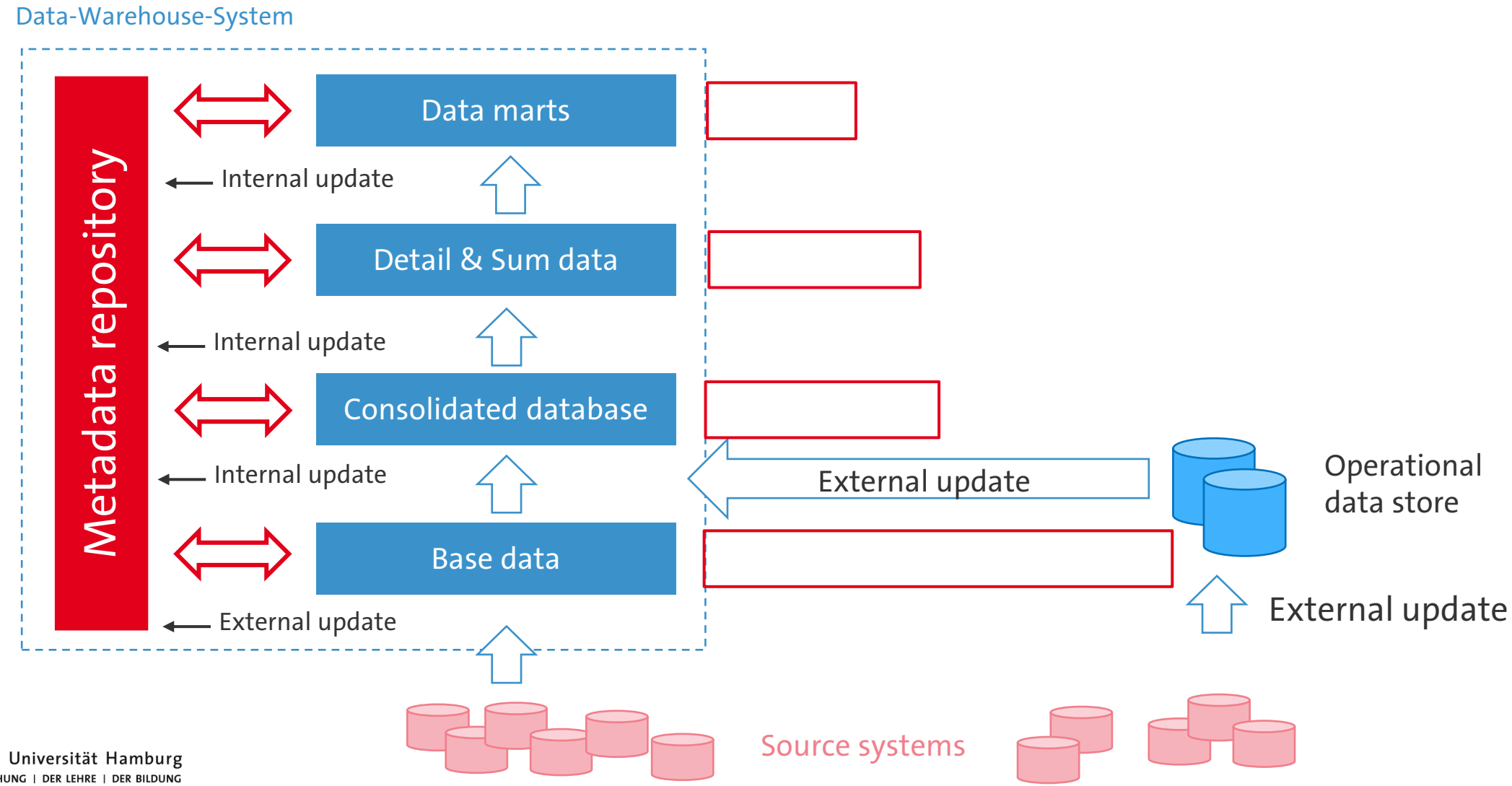
No user updates

(almost**) no updates and deletes** (technical updates / deletes only, quality assurance)

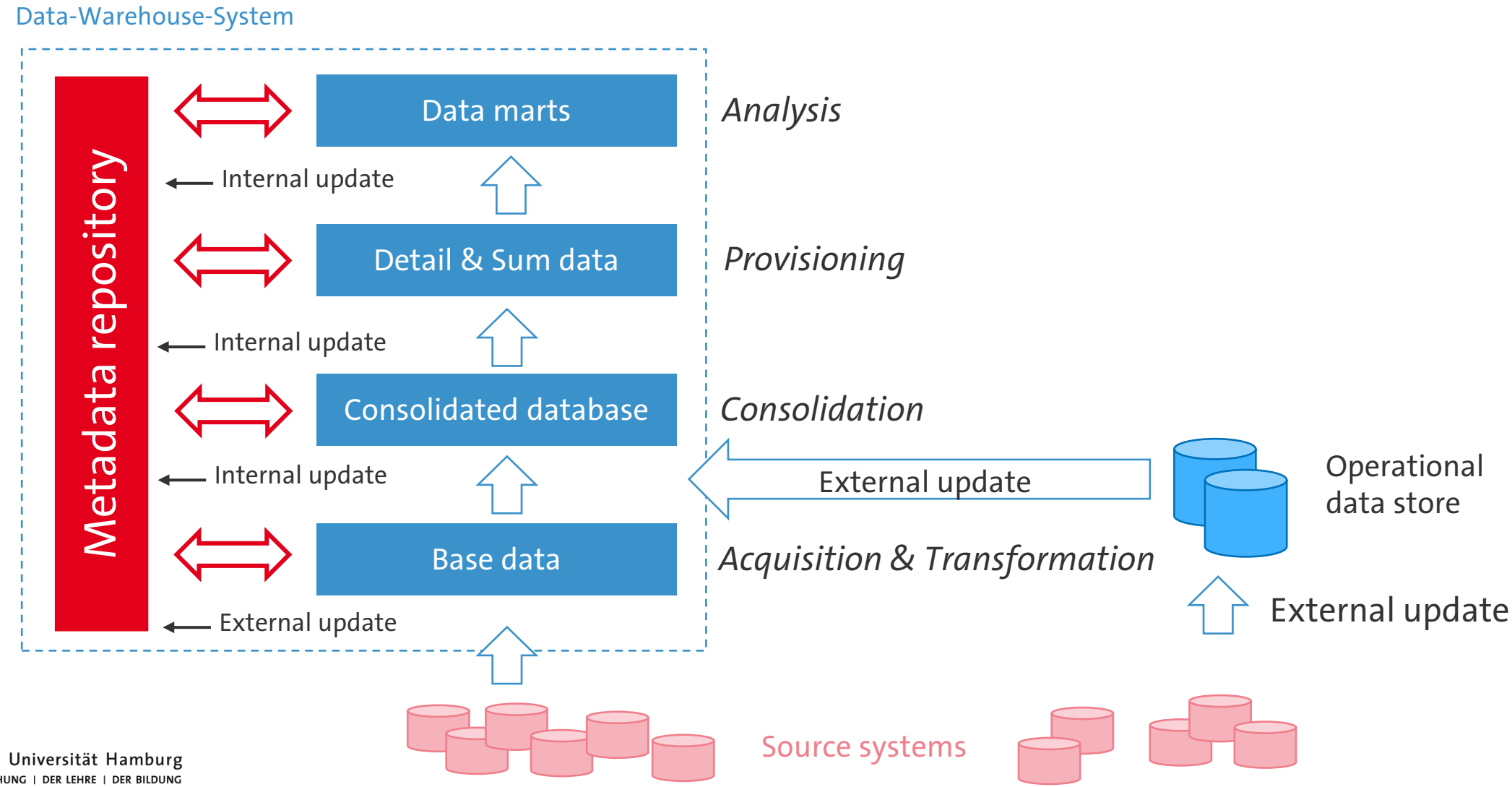
Historic data

Data is kept over a long period of time

Fill in the steps involved in creating the Data Warehouse

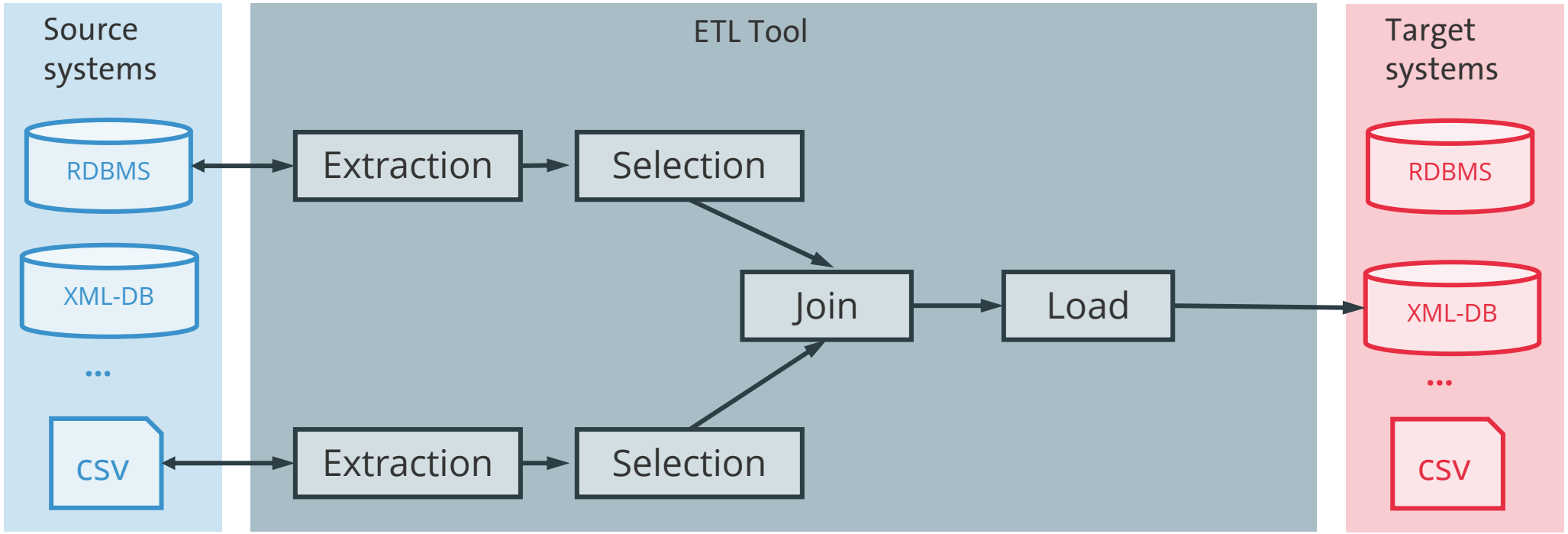


Fill in the steps involved in creating the Data Warehouse



What happens during the ETL step?

Data acquisition and transformation



Which of the following statements about ADAPT are wrong?

Different hierarchies of the same dimension can have different leaf nodes.	Attributes can be assigned to a whole cube, a dimension, and individual levels.
Self-precedence can be modelled	A cube must have exactly 3 dimensions.

Which of the following statements about ADAPT are wrong?

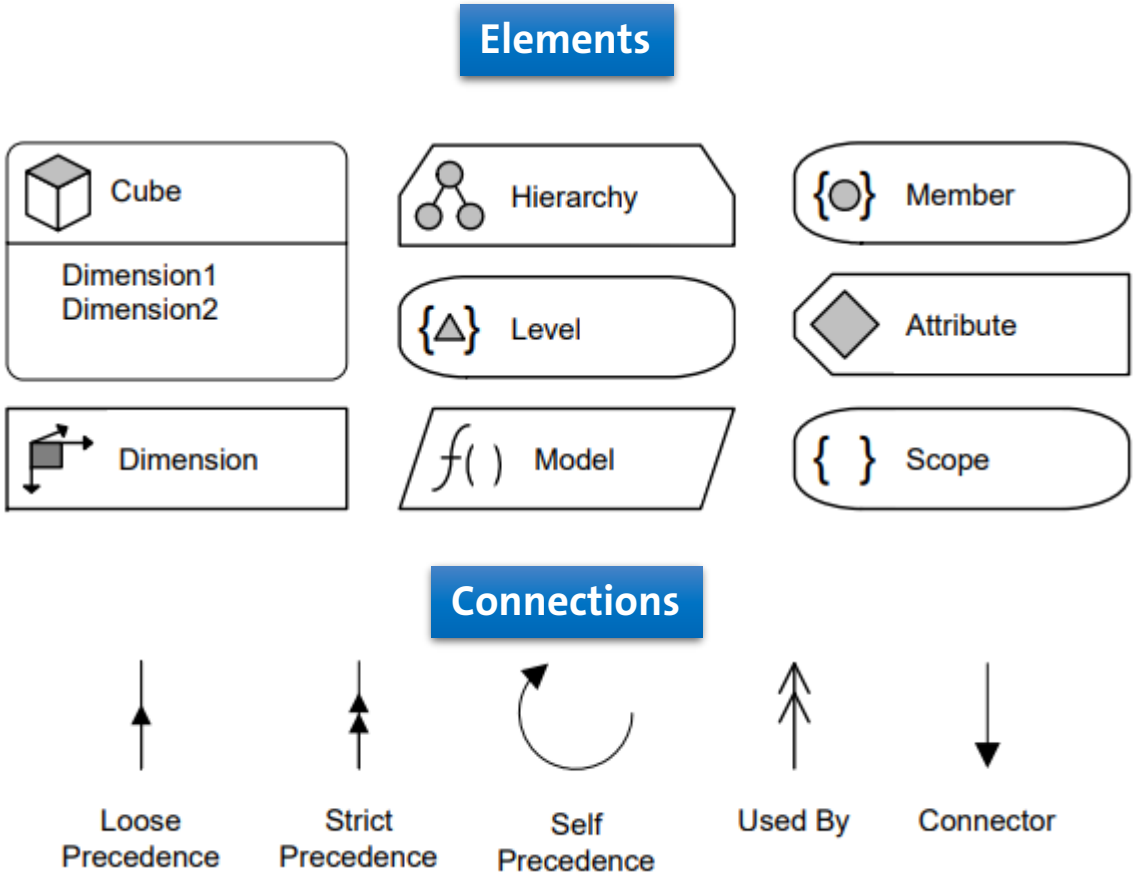
Different hierarchies of the same dimension can have different leaf nodes.

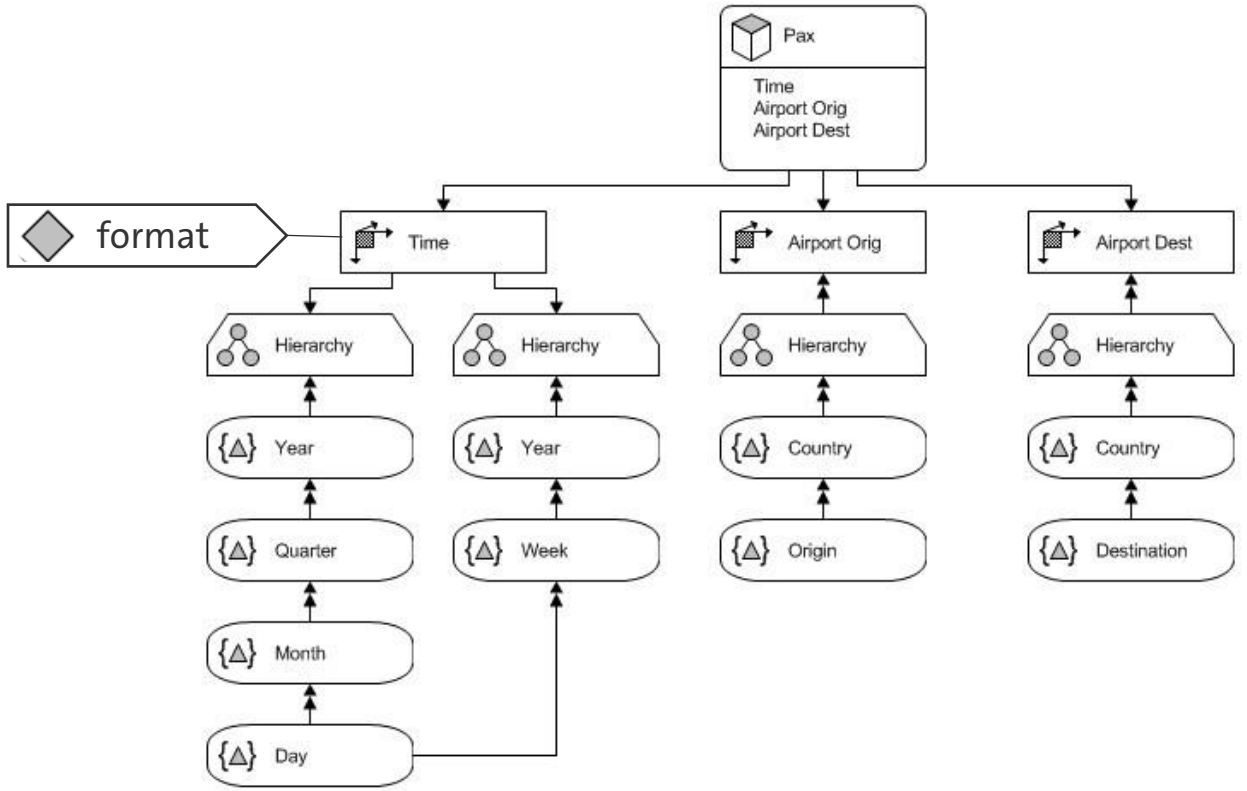
Attributes can be assigned to a whole cube, a dimension, and individual levels.

Self-precedence can be modelled

A cube must have exactly 3 dimensions.

Conceptual Models: ADAPT





Name two relational schemas typically used in OLAP and briefly explain them.

Name two relational schemas typically used in OLAP and briefly explain them.

Star schema

- One fact table and multiple dimension tables
- Dimension tables are not necessarily normalized
- Fact table is linked directly to all dimension tables via a foreign key

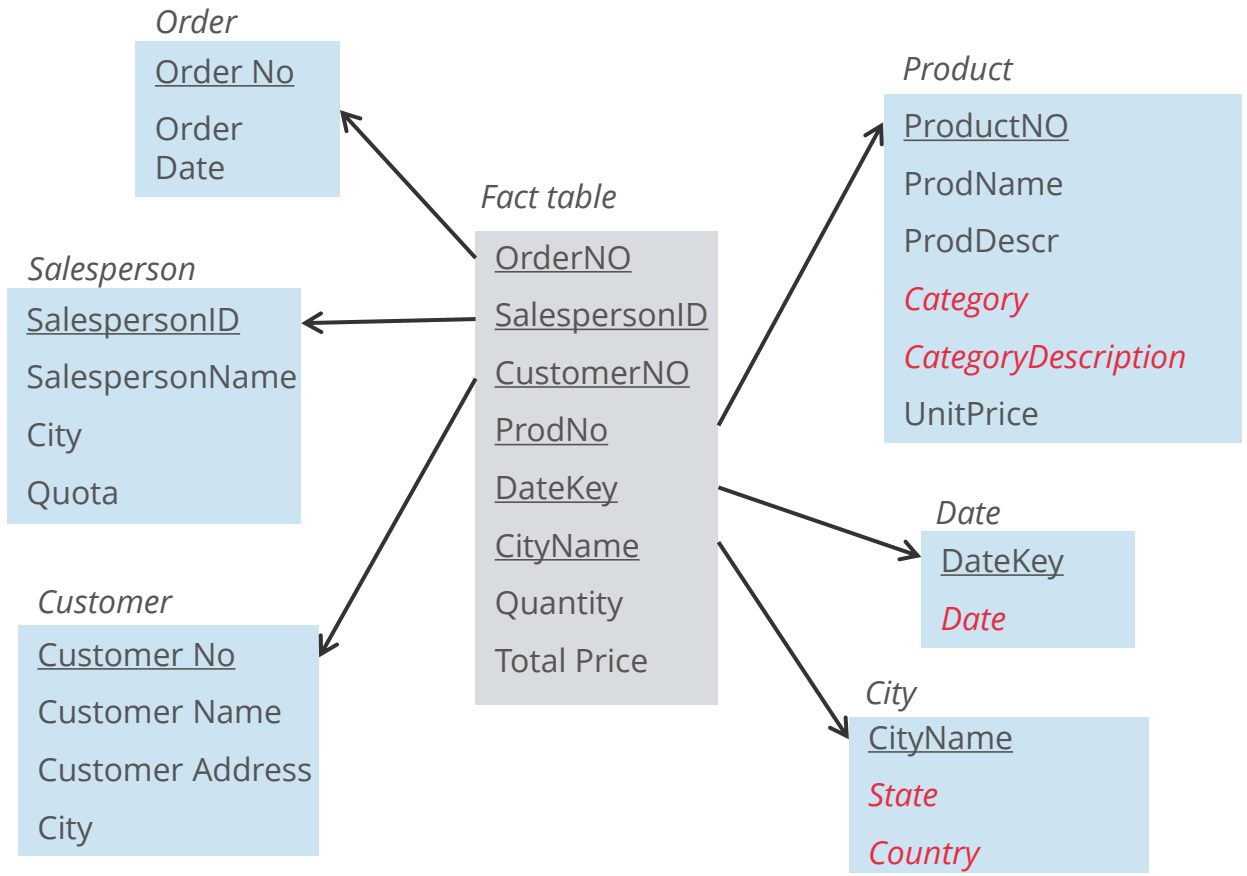
Snowflake schema

- One fact table and multiple dimension tables
  → Smaller memory footprint than star schema
- Dimension tables are normalized
  → Introduces more joins in the queries

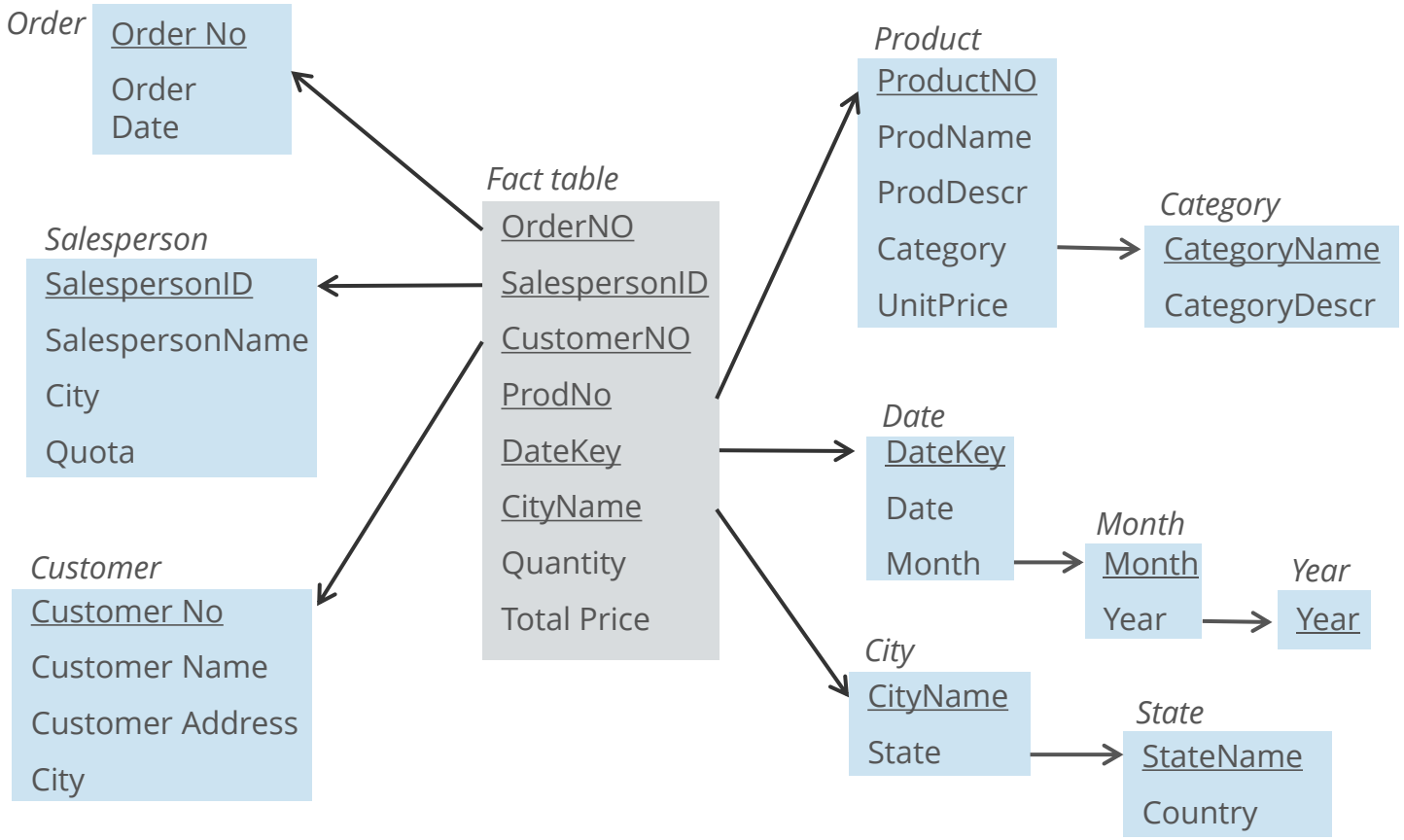
Galaxy schema

- Multiple fact tables
- Dimension tables are shared
- Joining fact tables (which are usually huge) is extremely inefficient
  → Long query execution times

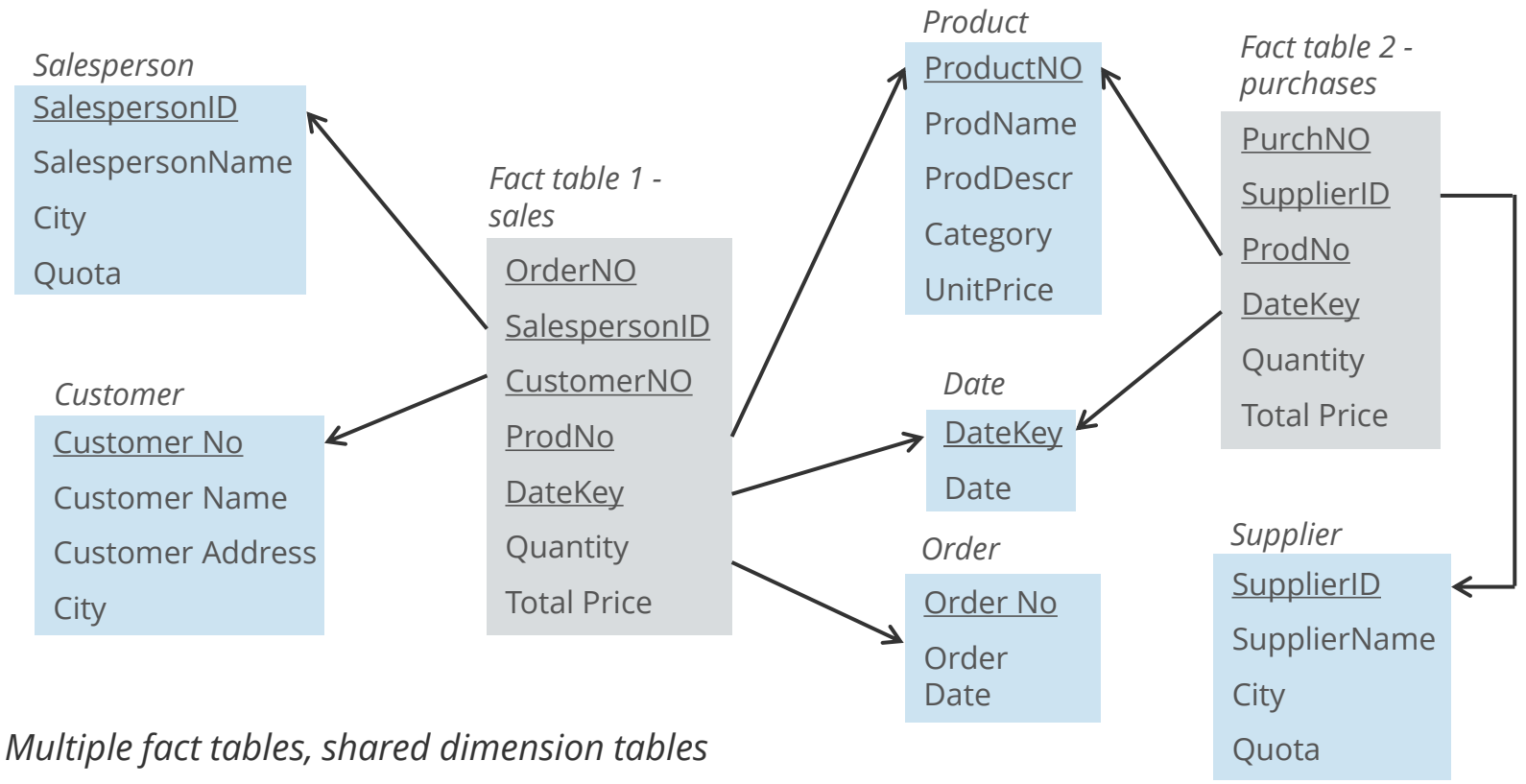
Star Schema


Order
Order No
Order Date
Fact table
OrderNO
SalespersonID
CustomerNO
ProdNo
DateKey
CityName
Quantity
Total Price
Product
ProductNO
ProdName
ProdDescr
Category
CategoryDescription
UnitPrice
Salesperson
SalespersonID
SalespersonName
City
Quota
Date
DateKey
Date
Customer
Customer No
Customer Name
Customer Address
City
City
CityName
State
Country

Snowflake Schema



Galaxy Schema



Multiple fact tables, shared dimension tables

Which of the following is/are not (a) typical method(s) of relational mapping of hierarchies?

- Horizontal Mapping
- Recursive Vertical Mapping
- Diagonal Mapping
- Iterative Horizontal Mapping
- Vertical Mapping
- Random Mapping

Which of the following is/are not (a) typical method(s) of relational mapping of hierarchies?

- Horizontal Mapping
- Recursive Vertical Mapping
- Diagonal Mapping
- Iterative Horizontal Mapping
- Vertical Mapping
- Random Mapping

Relational Mapping of Hierarchies

Vertical Mapping

Product

ID	Product	ID2
1	Flour X	1
2	Sugar Y	2
3	Water Z	3

Product group

ID	Product group	ID3
1	Flour	1
2	Sugar	1
3	Water	2

Product family

ID	Product family	ID4
1	Bakery goods	1
2	Beverages	1

Category

ID	Category
1	Food

Horizontal Mapping

Relation product

ID	Product	Product group	Product family	Category
1	Flour X	Flour	Bakery goods	Food
2	Sugar Y	Sugar	Bakery goods	Food
3	Water Z	Water	Beverages	Food

Recursive Vertical Mapping

ID	Product	Product group	Product family	Category
1	Flour X	Flour	Bakery goods	Food
2	Sugar Y	Sugar	Bakery goods	Food
3	Water Z	Water	Beverages	Food

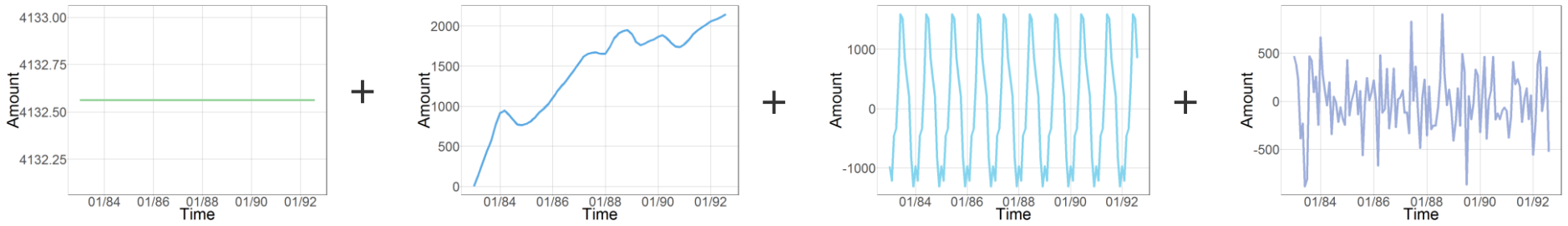
Relation product

ID	Product	ParentID
1	Food	NULL
2	Bakery goods	1
3	Beverages	1
4	Flour	2
5	Sugar	2
6	Water	3
7	Flour X	4
8	Sugar Y	5
9	Water Z	6

Data Mining

Into which components can time series typically be decomposed?

Into which components can a time series typically be decomposed?



Base + Trend + Season + Residuals

Is Piecewise aggregate approximation (PAA) a shape-based or a feature based representation for time series?

Shape-based

Feature-based

What is the difference to symbolic aggregate approximation?

Is Piecewise aggregate approximation (PAA) a shape-based or a feature based representation for time series?

Shape-based **Feature-based**

What is the difference to symbolic aggregate approximation?

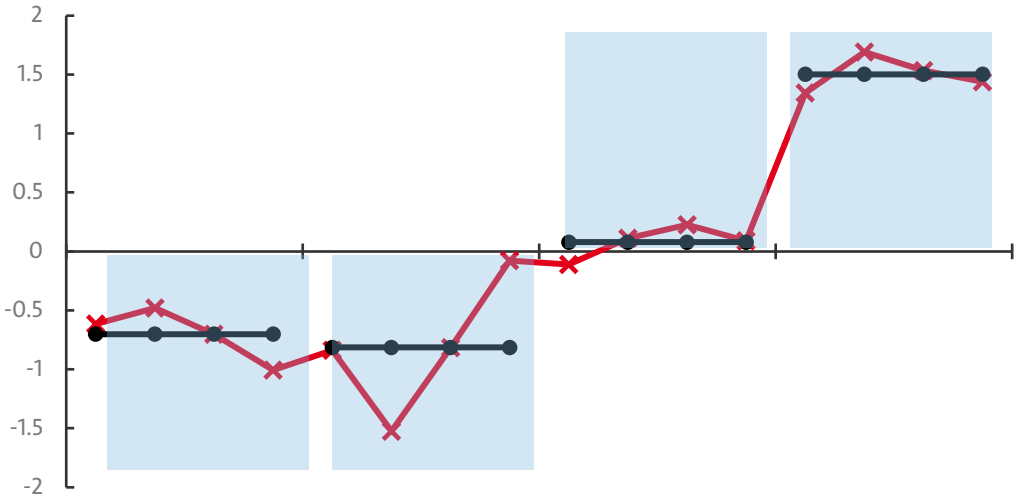
In SAX, values are discretized into an alphabet.

Shape-based Representation

- Transform a series to segments
- Time-dependent representation
- High compression

Piecewise aggregate approximation (PAA)
→ Represent each segment by its mean

Symbolic aggregate approximation (SAX)
→ Discretize the mean into an alphabet



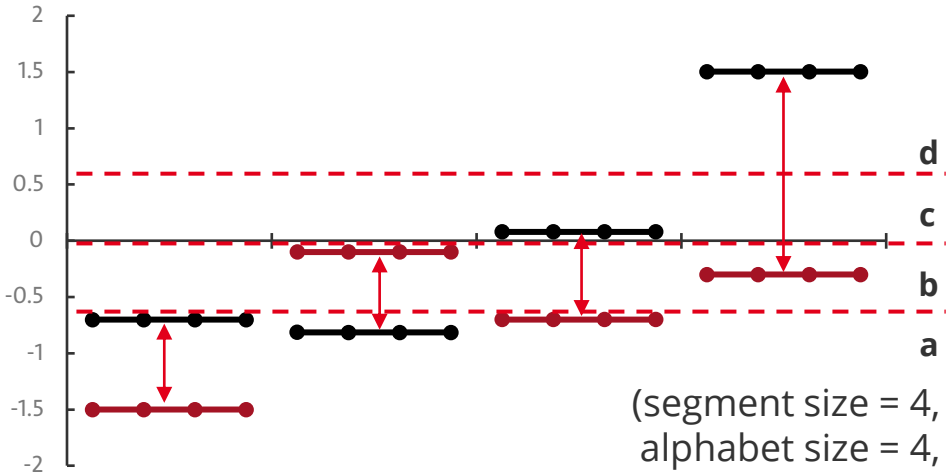
PAA: -0.70 -0.81 0.08 1.56

SAX: b b a a

(segments per series = 4, alphabet size = 2, breakpoint at 0.0)

The SAX Distance

Segments



(segment size = 4,
alphabet size = 4,
breakpoints at +/- 0.67 and 0.00)

SAX

$\hat{x}^1$	a	a	c	d
$\hat{x}^2$	a	b	a	b

$$d_{SAX}(\underline{\hat{x}}^i, \underline{\hat{x}}^j) = \sqrt{T/W} \cdot \sqrt{\sum_{w=1}^{W} cell(\hat{x}_w^i, \hat{x}_w^j)^2}$$

$cell$ returns the minimum distance of two symbols:

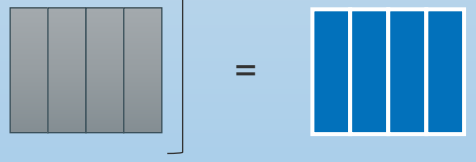
	a	b	c	d
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

PCA

Standardize all values per dimension

value-mean/standard deviation

std [ ] =

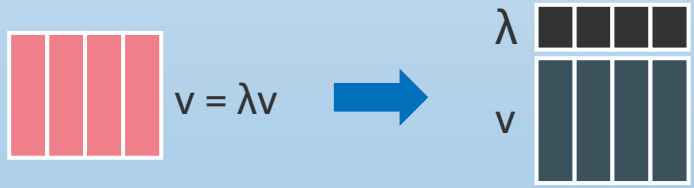


Compute the covariance matrix

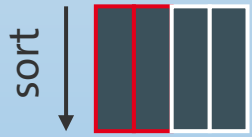
cov(x,y)=(∑(x-(mean(x))(y-mean(y))))/number of data points

cov [ ] =

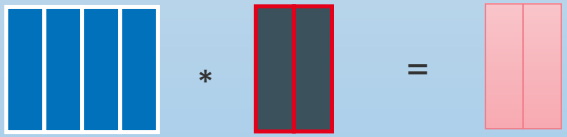
Compute Eigen values and Eigen vectors of the covariance matrix



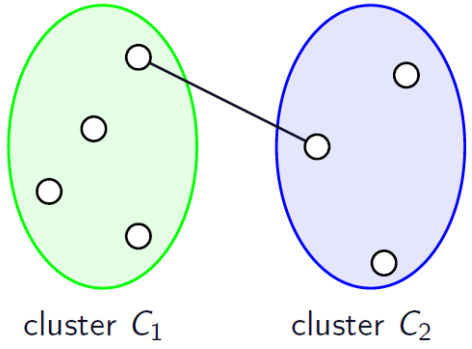
Sort elements of Eigenvectors in descending order and pick those that belong to top *x* Eigen values
→ Feature Vectors

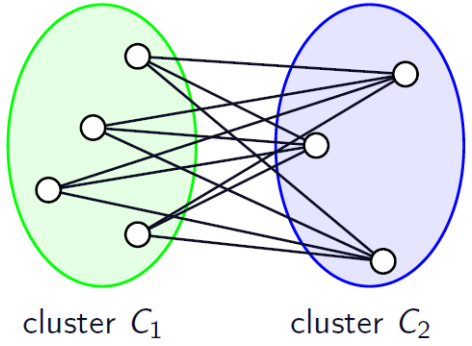


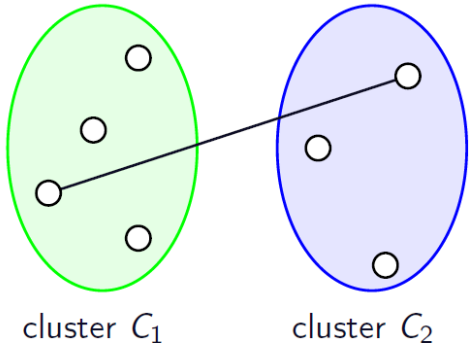
*Standardized data set * Feature vector*
= New Dataset (reduced to x dimensions)

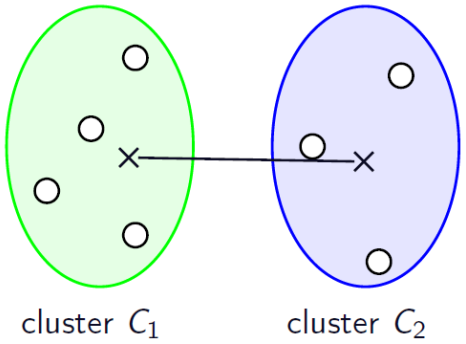


Which distance measures for clusters are shown here?

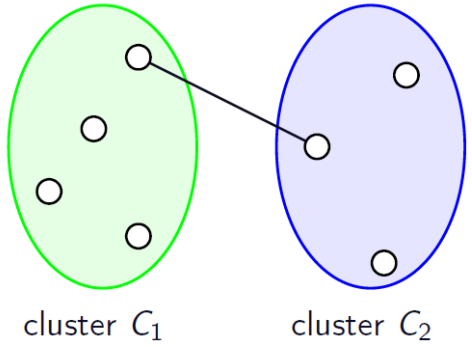




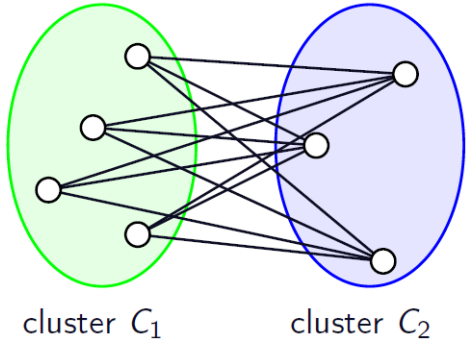




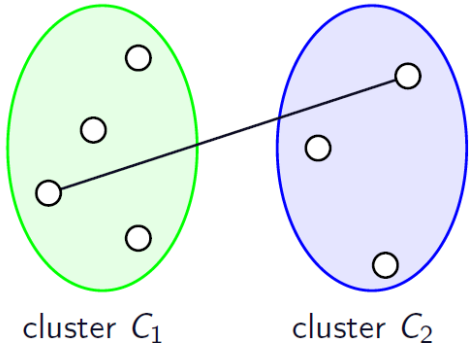
Which distance measures for clusters are shown here?



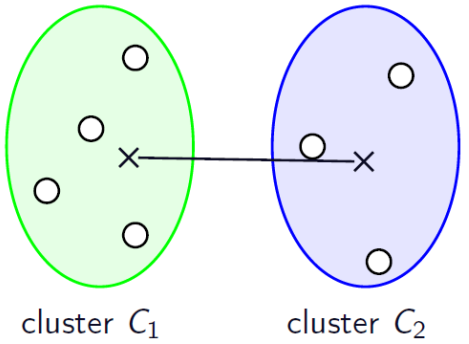
Single link



Average link



Complete link

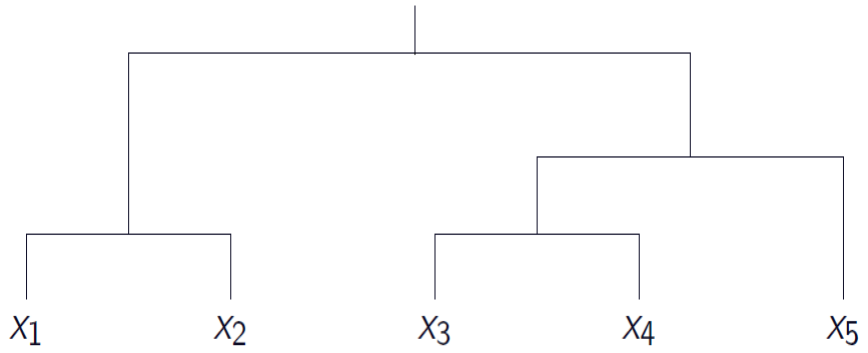


Canonical entity

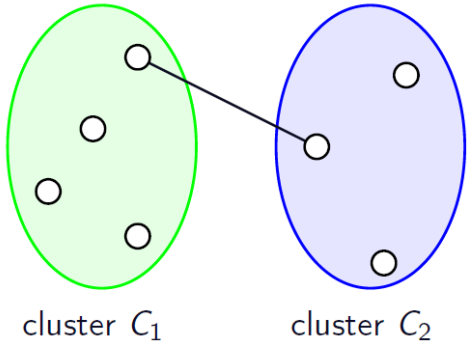
Clustering Methods

K-Means | Canopy Clustering | Hierarchical Clustering | Incremental Clustering

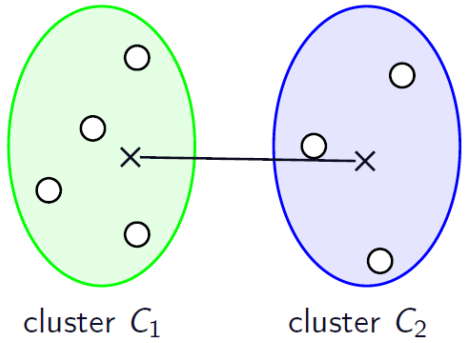
Results can be displayed as a dendrogram



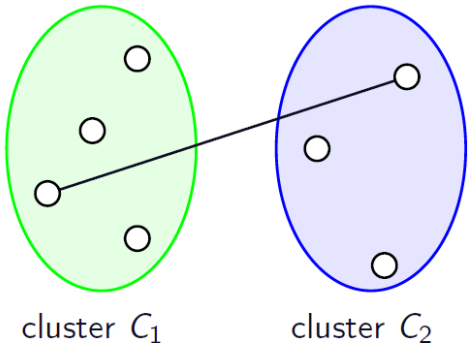
Distance measures for clusters



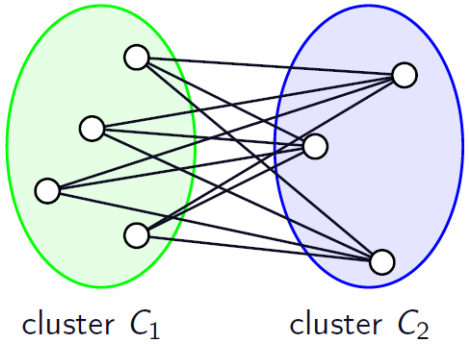
Single link
Minimal distance between two data points



Canonical entity
Distance between two cluster representatives (e.g. the centroids)

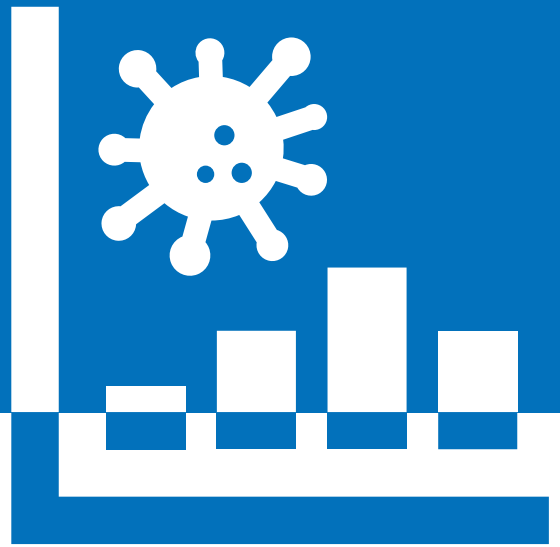


Complete link
Maximal distance between two data points

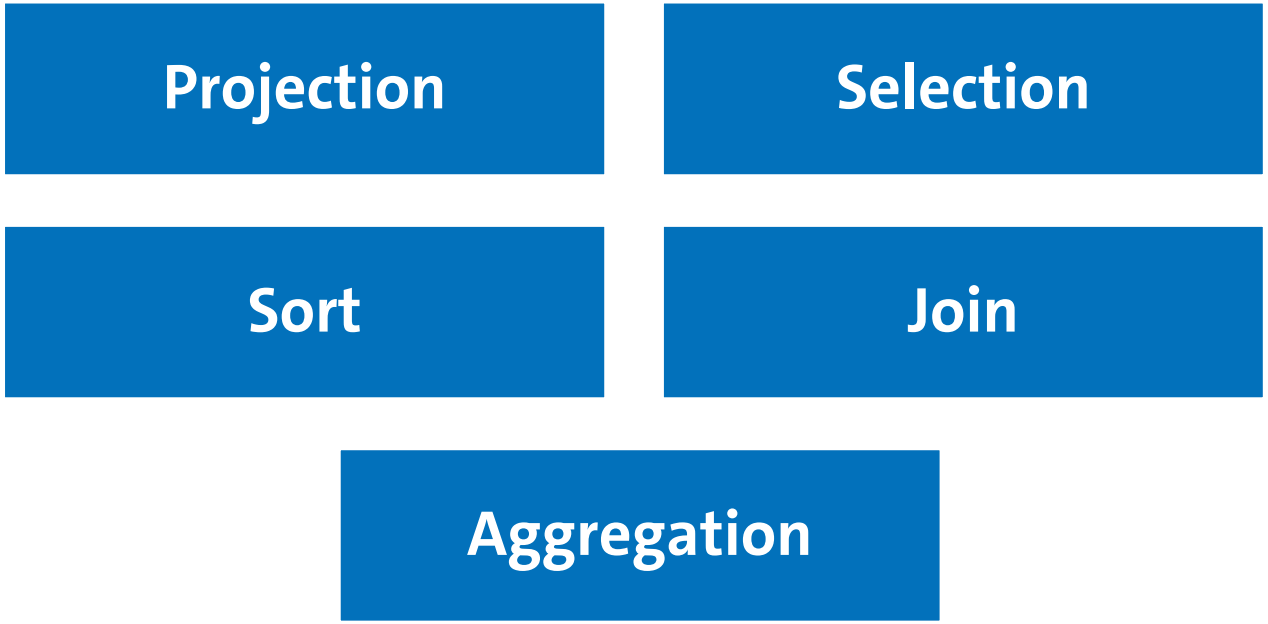


Average link
Average distance between two data points

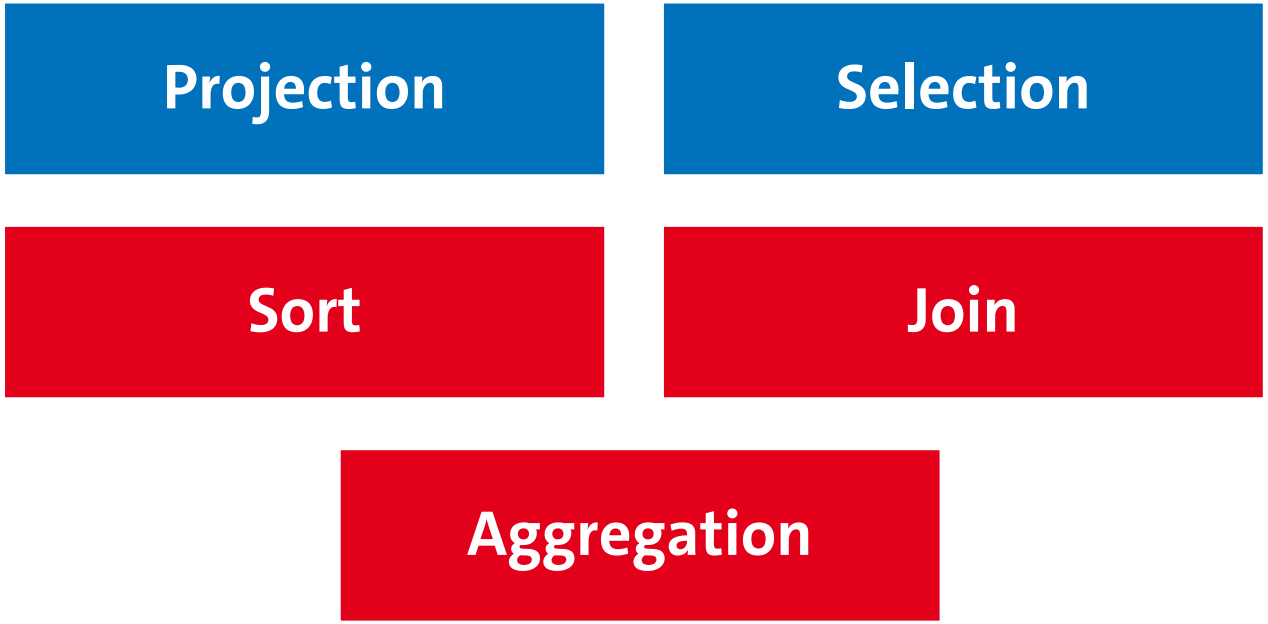
Big Data Analysis



Which operators require shuffling if they are distributed?

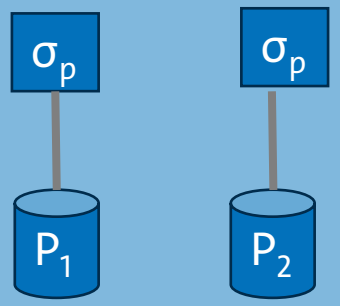


Which operators require shuffling if they are distributed?



Parallel Operators

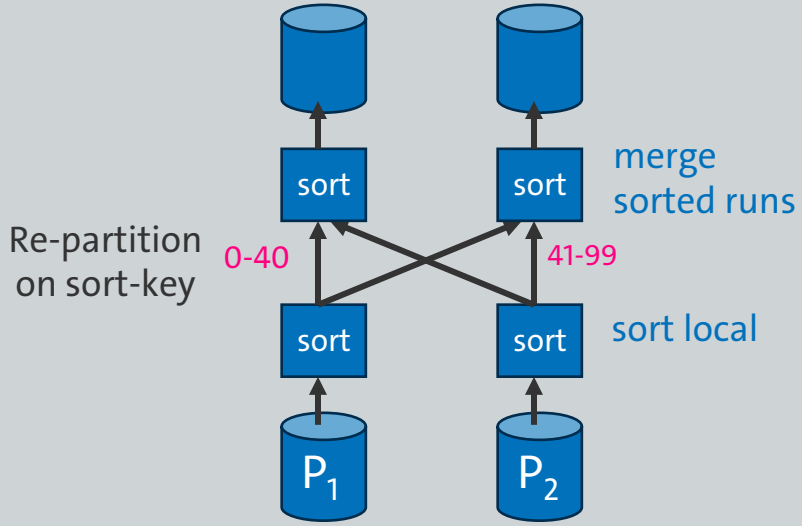
Selection



partitioned data
(on two nodes)

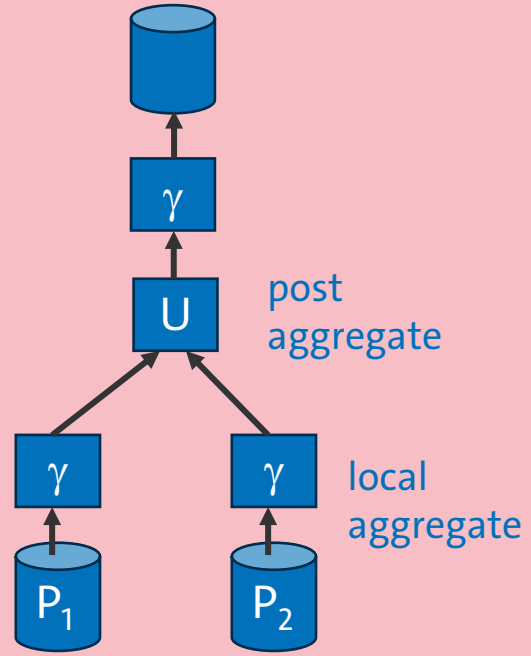
→ Projection works similarly as selection

Sort



partitioned data

Aggregation



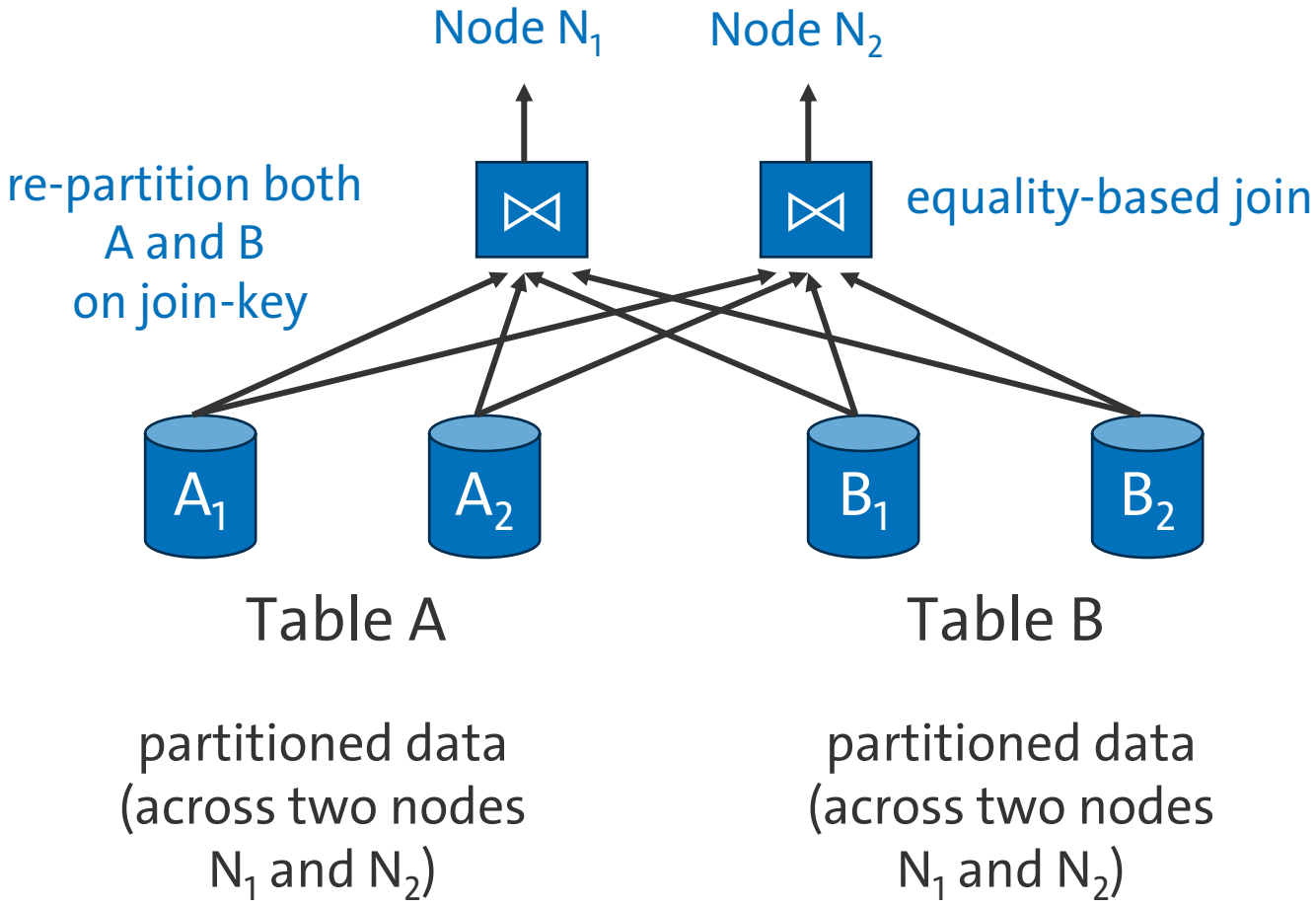
partitioned data

Join: Symmetric Repartitioning

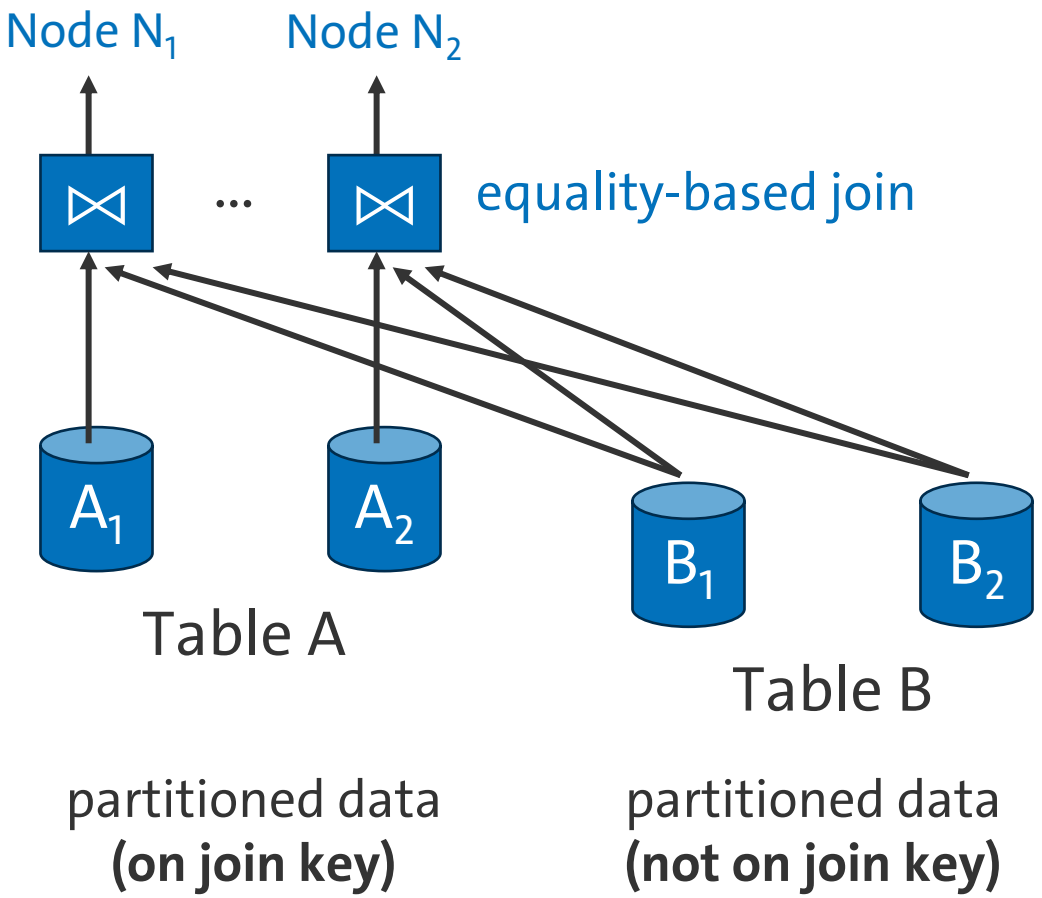
„Repartitioned join“

- Both join partners are repartitioned after the join attribute
- High communication costs

→ Avoid!



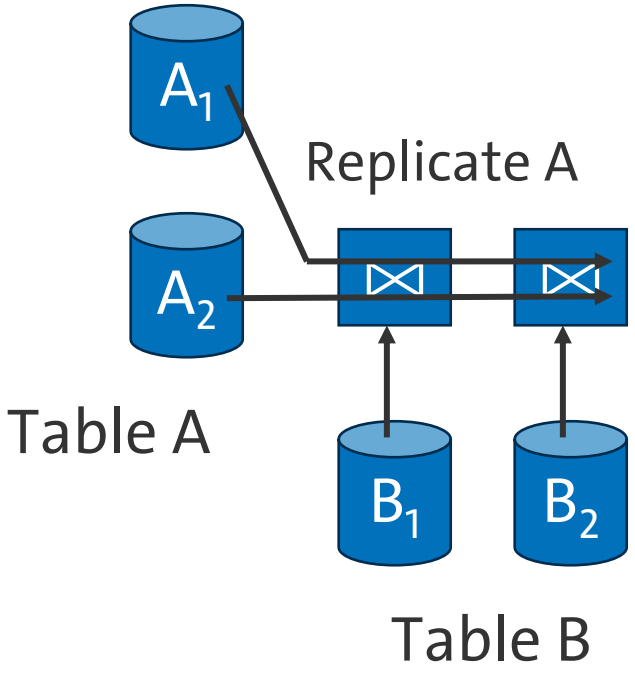
Join: Asymmetric Repartitioning



re-partition only B on join-key using same partitioning function as table A

Join: Fragment and Replicate

partitioned data
(across two nodes
A_1 and A_2)



Replicate (all fragments of) smaller table
(here: table A) to all nodes

partitioned data
(across two nodes B_1 and B_2)

Which of the following are shuffling methods for distributed operators?

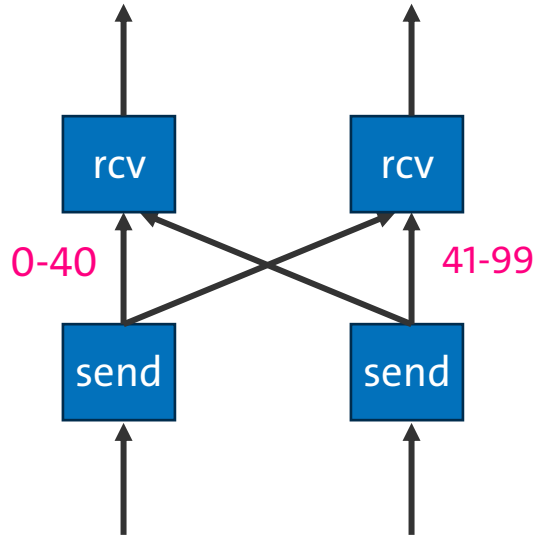
- Range-based N:M
- Hash-based N:M
- Column-based N:M
- 1:1
- N:1
- Remote N:1

Which of the following are shuffling methods for distributed operators?

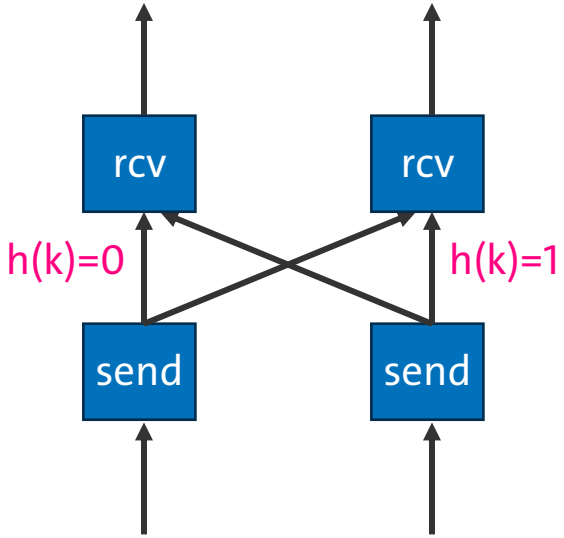
Range-based N:M	Hash-based N:M
Column-based N:M	1:1
N:1	Remote N:1

Data Shuffling: Details

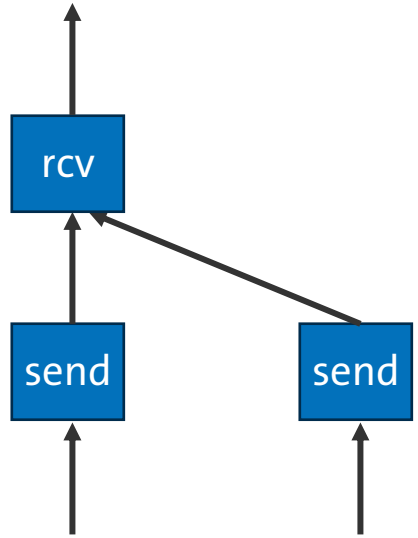
Data Shuffling can use different partitioning strategies (range vs. hash-partitioning, N:M vs. N:1) to re-partition data during query execution



Range-based N:M



Hash-based N:M



N:1 (no part. function)

Name 3 possible reasons for data skew

Name 3 possible reasons for data skew

Different partition sizes

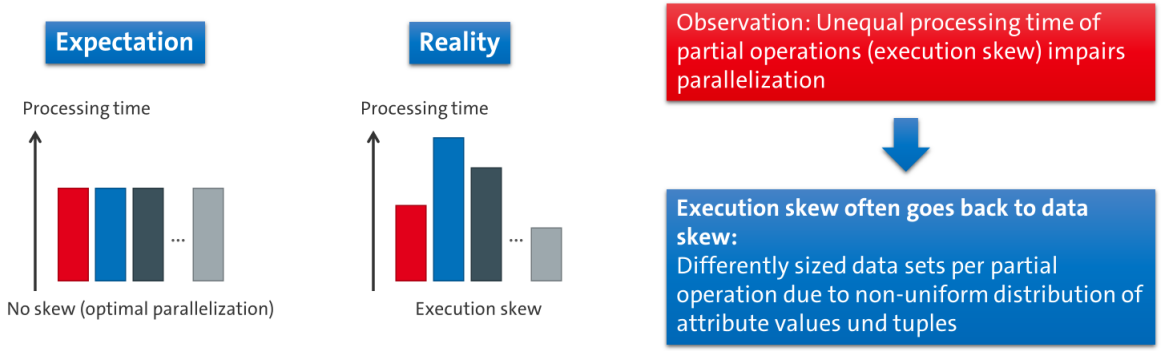
Distribution function leading to different fragment sizes

Different hit rates per partition in range queries

...

Name 3 possible reasons for data skew

The Data Skew



See handout!

What is the main difference between the Kappa architecture and the Lambda architecture?

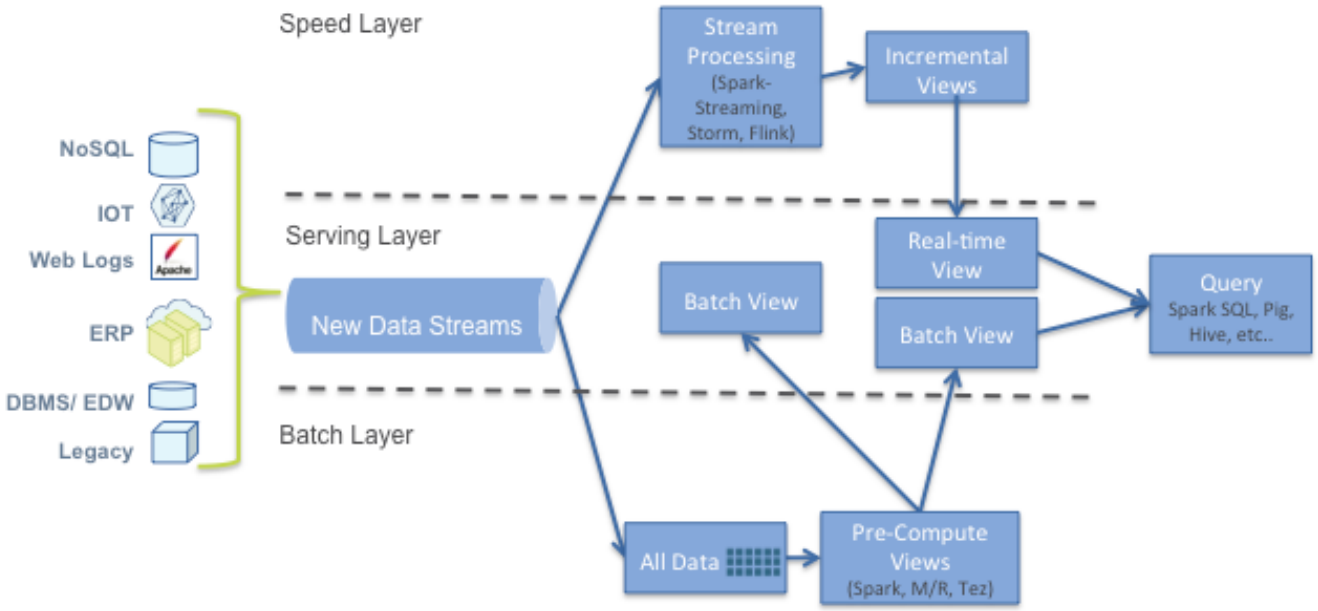
What is the main difference between the Kappa architecture and the Lambda architecture?

The Kappa architecture only has a speed layer while the Lambda architecture also has a batch layer

Lambda Architecture

→ *see lecture 16*

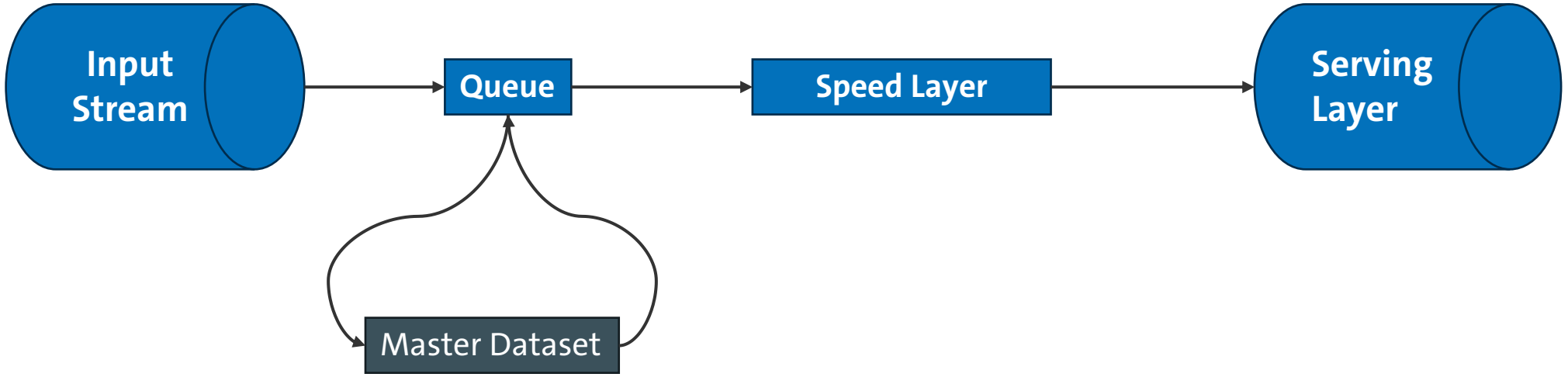
→ Batch & Stream processing



Disadvantage
2 code bases & 2 deployments, e.g. Hadoop & Storm

Kappa Architecture

→ No batch layer!



Disadvantage
Real time processing only

Which functions must be implemented by the application developer when using a map-reduce framework?

Map

Shuffle

Reduce

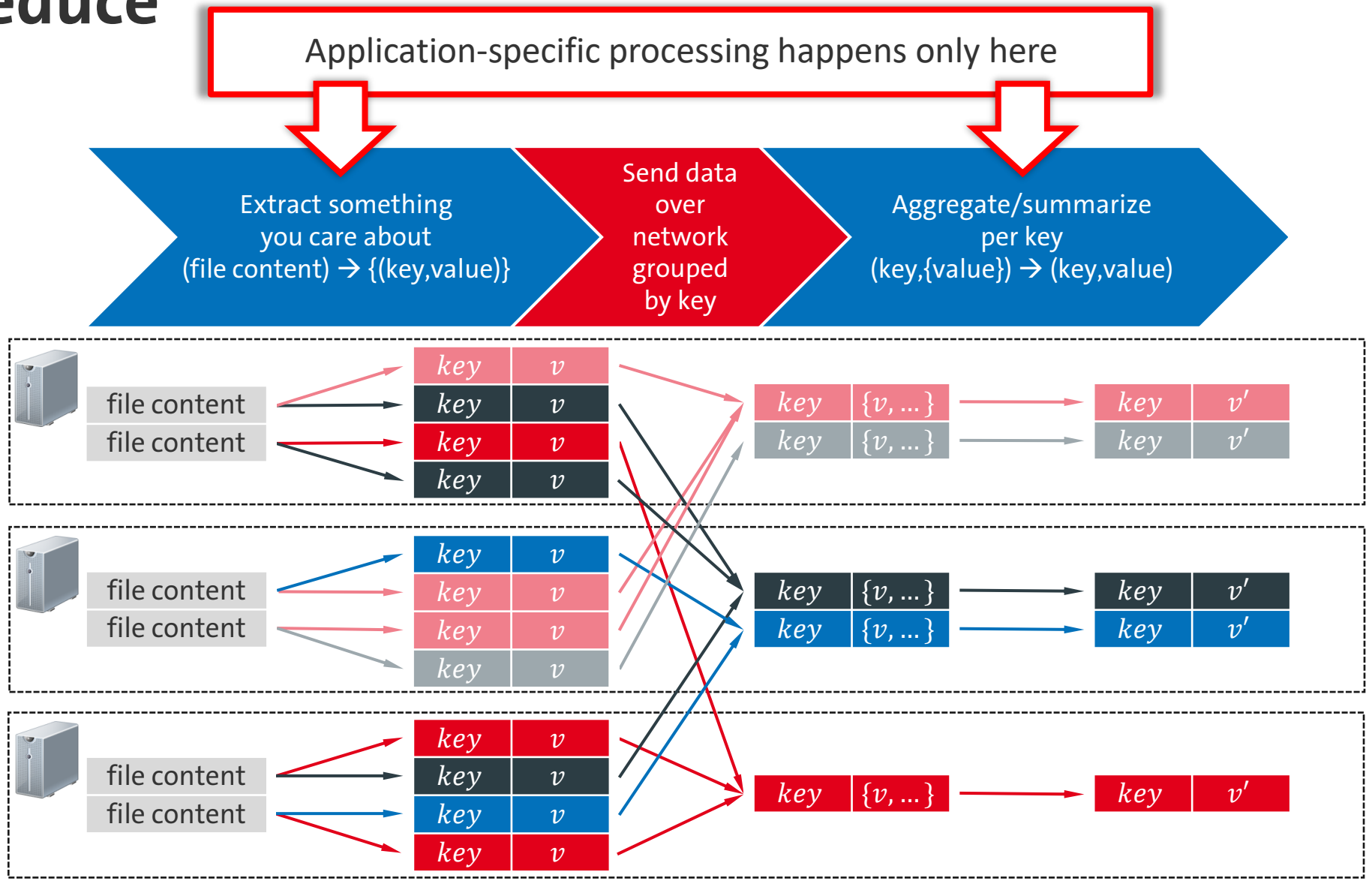
Which functions must be implemented by the application developer when using a map-reduce framework?

Map

Shuffle

Reduce

Map Reduce


Application-specific processing happens only here
Extract something you care about (file content) → {(key,value)}
Send data over network grouped by key
Aggregate/summarize per key (key,{value}) → (key,value)
file content
file content
file content
file content
file content
file content
key
v
{v, ...}
v'

Name 3 frameworks for batch processing

Framework Overview