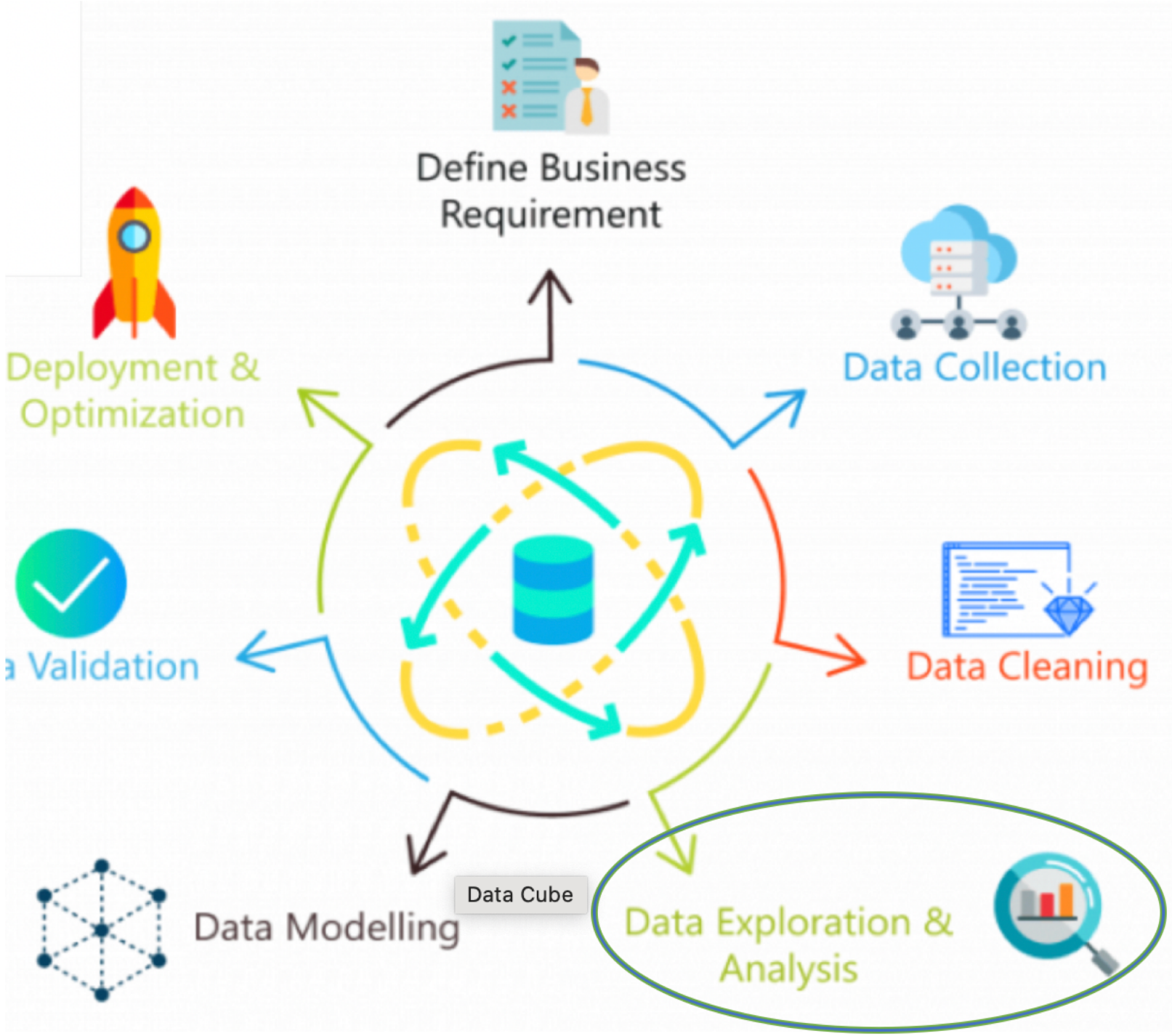
BI Data Collections

Data Structures, Aggregation and Granularity

by tdi@ek.dk


Define Business Requirement
Data Collection
Data Cleaning
Data Exploration & Analysis
Data Modelling
Data Cube
Validation
Deployment & Optimization

Agenda

Applying Descriptive Statistics in BI

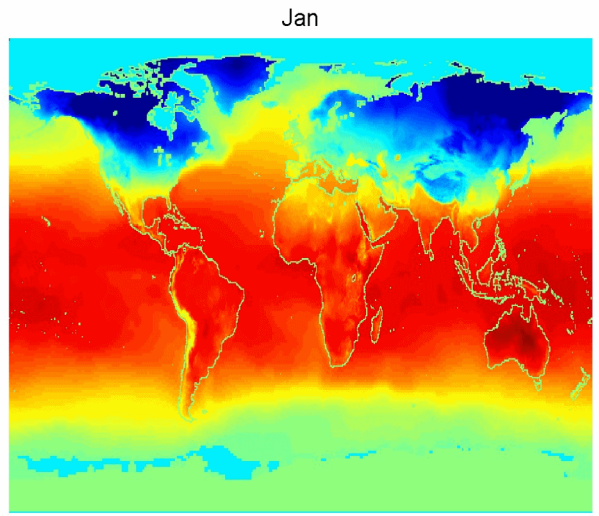
- Data Cube as BI Data Structures
  - dimensions and measures
- Operations with Data Cube
  - exploring data granularity
- Exploring Data Variables by Role in BI analysis
  - dependent and independent variables
  - linear vs non-linear dependency

Data Collections

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

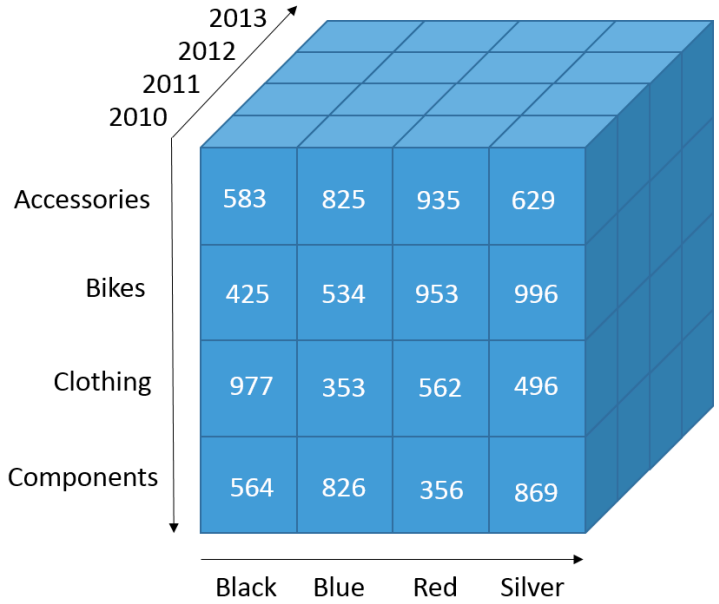
- Set of observations – one per row
- Each observation is a set of values of attributes
- Each value belongs to some of the data categories discussed earlier
- Data needs to be structured
- Some data needs to be transformed before analysis

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

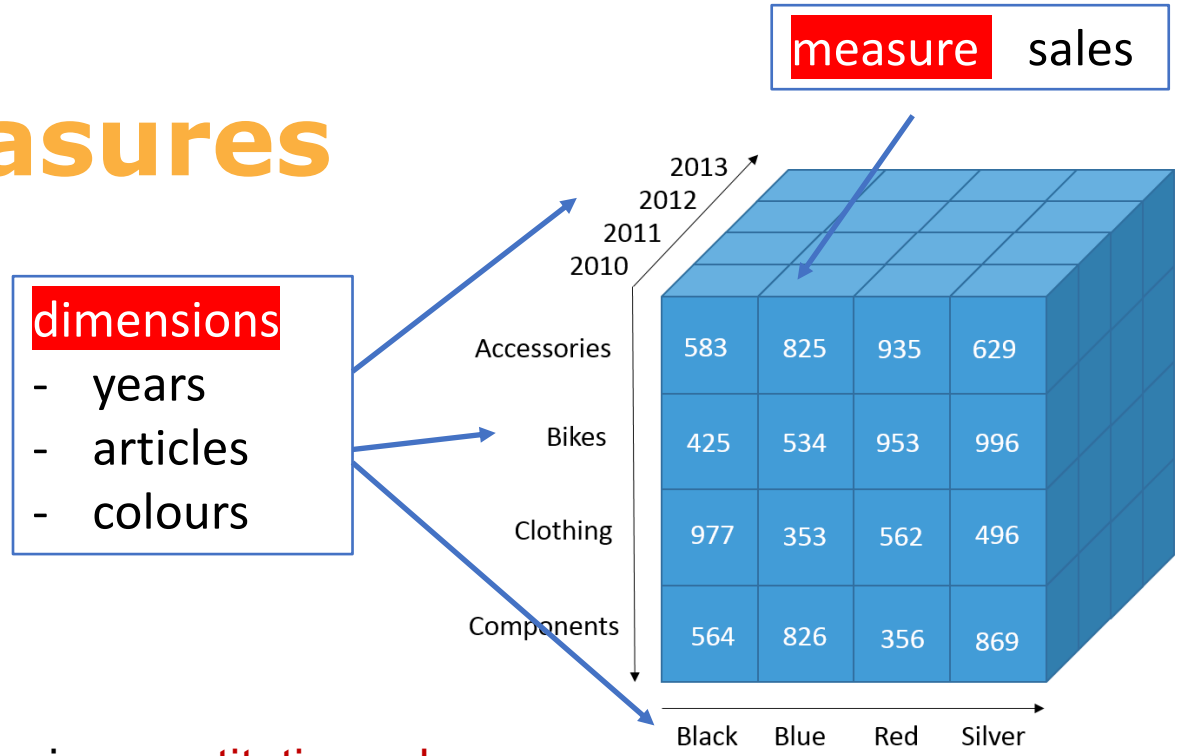


Data Cube

- In computer programming terms, Data Cube is multi-dimensional array of data
- In BI, it is a schema optimized for analytical queries
- It consist of measures (computations like a count of orders) that have one or more dimensions.
- The dimensions are commonly accepted labels, relevant to the business
- The measures are computed numbers – pre-processed from the raw data



Dimensions and Measures

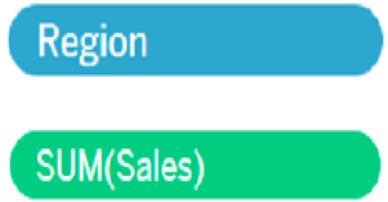


Dimensions

- contain qualitative values (names, dates, geographical data, ...)
- used to categorise, segment, and reveal the details in the data
- affect the level of detail in the view

Measures

- contain numeric, quantitative values that can be measured
- can be aggregated
- aggregation to a measure is applied by default



Operations with Cubes Level of Details

Slice

- filter and analyse data according to a single dimension
- you can take slices from any angle of your cube

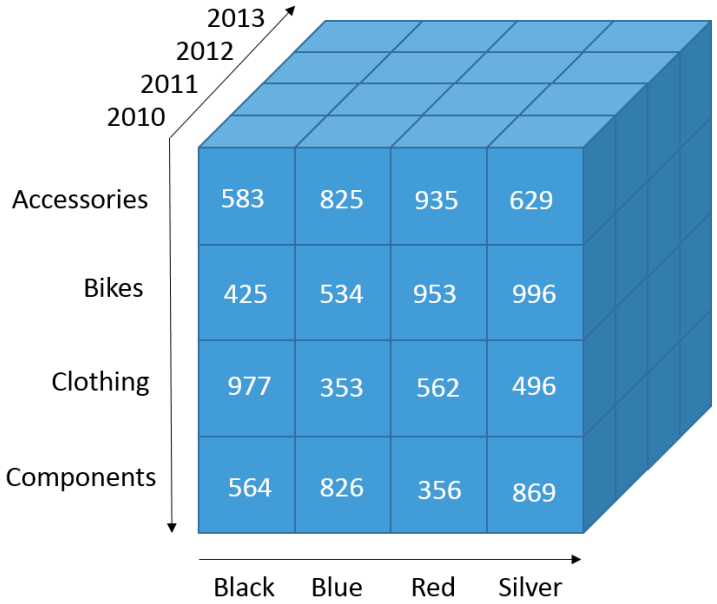
Example

- view all data in specific year

Dice

- zoom in on the values of a particular subset of our data
- cutting a smaller cube to get a closer look at more than one dimension

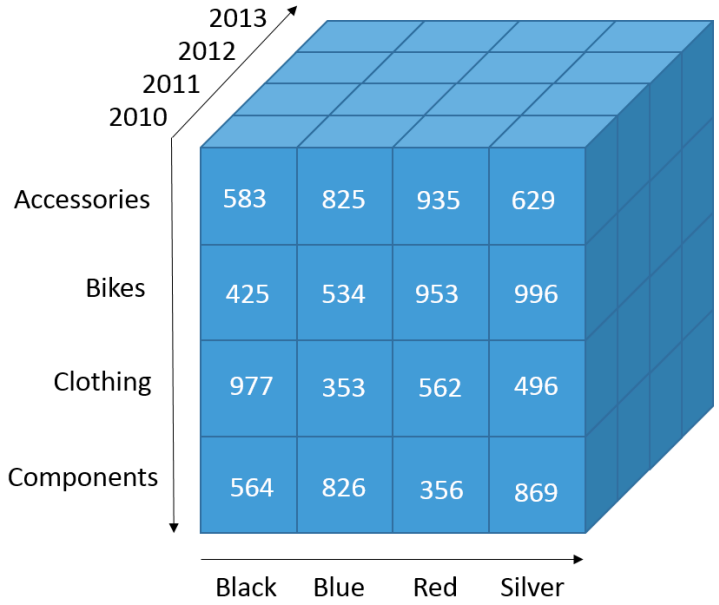
- Example
- view sell of blue bikes along the years



Operations with Cubes Level of Details

Pivoting

- viewing the data from a different angles
- swapping the row and column fields



before

Quarter	Samsung	Nokia	Apple
Q1 '12	89.2800	83.1600	33.1200
Q2 '12	90.4300	83.4200	28.9400
Q3 '12	97.9600	82.3000	24.6200
Q4 '12	106.9600	85.0500	43.4600
Q1 '13	100.6600	63.2200	38.3300
Q2 '13	107.5300	60.9500	31.9000
Q3 '13	117.0500	63.0500	30.3300
Q4' 13	119.2100	63.5800	50.2200

after

Quarter	Pivot Field Names	Pivot Field Values
Q4 '12	Apple	43.460
Q1 '13	Apple	38.330
Q2 '13	Apple	31.900
Q3 '13	Apple	30.330
Q4' 13	Apple	50.220
Q1 '10	Nokia	110.110
Q2 '10	Nokia	111.470
Q3 '10	Nokia	117.460
Q4 '10	Nokia	122.280

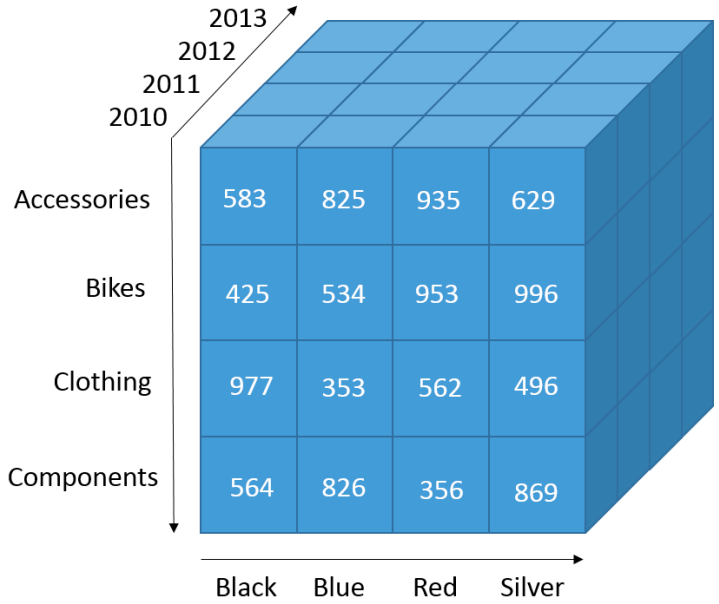
Operations with Cubes Level of Details

Drill-down

- viewing data at a level of increased detail

Examples

- first, view by year and then, months
- first, sum by country, and then by cities



Roll-up

- viewing data with decreasing detail
- opposite of drill-down
- aggregated view of data

Data Aggregation and Granularity

Aggregation refers to how data is combined

- can show the larger trends in the data
- quantitative fields are aggregated

Granularity refers to how detailed the data is

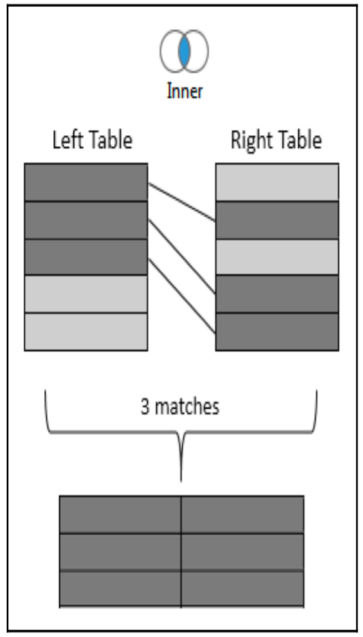
- low and high granularity

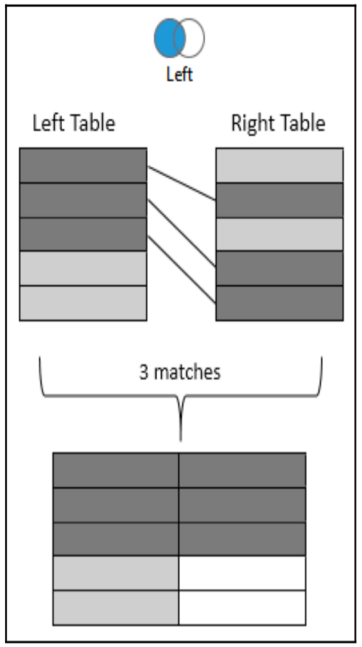
Similar to SQL DB

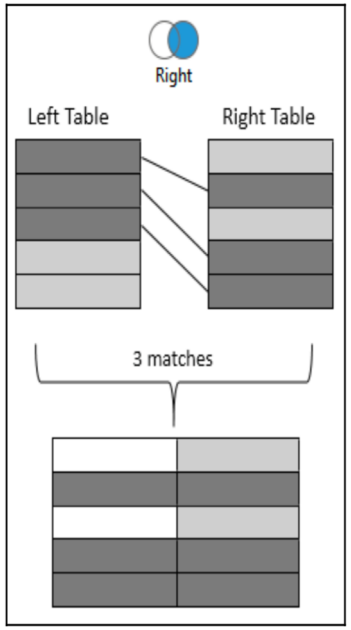
Two types

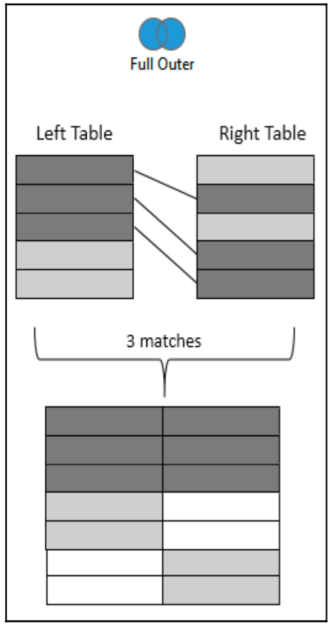
Joins - link sources on row-by-row scale - aggregate data from one source with matching data from another source

Blends - links sources on aggregate level - independent processing of data from two sources and consequent aggregation of the results









Data Aggregation and Granularity

Aggregation refers to how data is combined

- can show the larger trends in the data
- quantitative fields are aggregated

Granularity refers to how detailed the data is

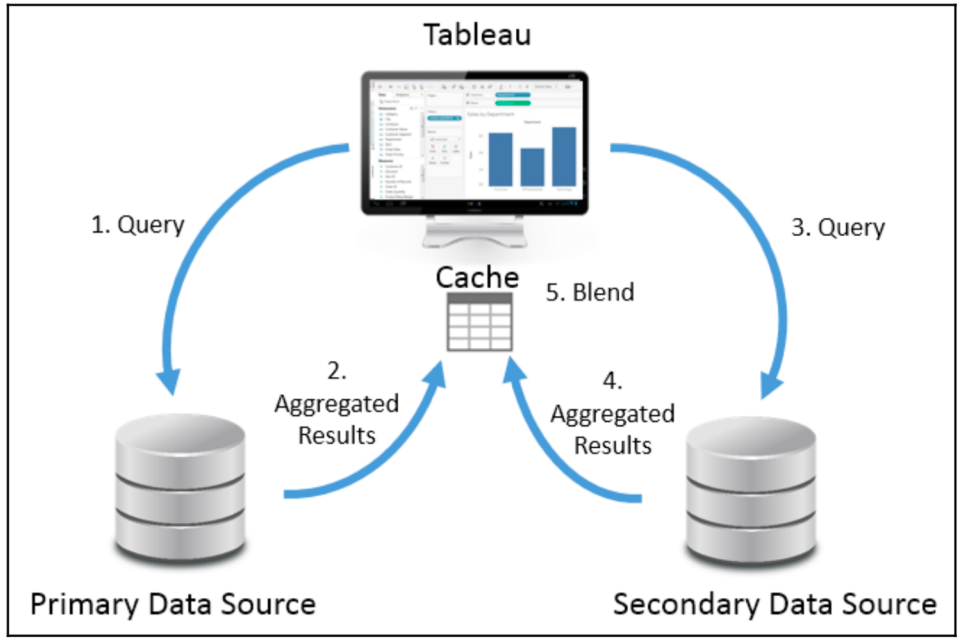
- low and high granularity

Similar to SQL DB

Two types

Joins - link sources on row-by-row scale - aggregate data from one source with matching data from another source

Blends - links sources on aggregate level - independent processing of data from two sources and consequent aggregation of the results



Data Restructuring

Employee	2/5/2020	2/6/2020	2/7/2020	2/8/2020	2/9/2020
Christine	10	10	10	10	10
Tristan	10				
Lily	10				10
Jamal	10		10		

Employee	Date	Parking Fee
Christine	2/5/2020	10
Christine	2/6/2020	10
Christine	2/7/2020	10
Christine	2/8/2020	10
Christine	2/9/2020	10
Tristan	2/5/2020	10
Lily	2/5/2020	10
Lily	2/9/2020	10
Jamal	2/5/2020	10
Jamal	2/7/2020	10

Airline
American Airlines: AA
Delta Airlines: DL
JetBlue Airways: B6
United Airlines: UA

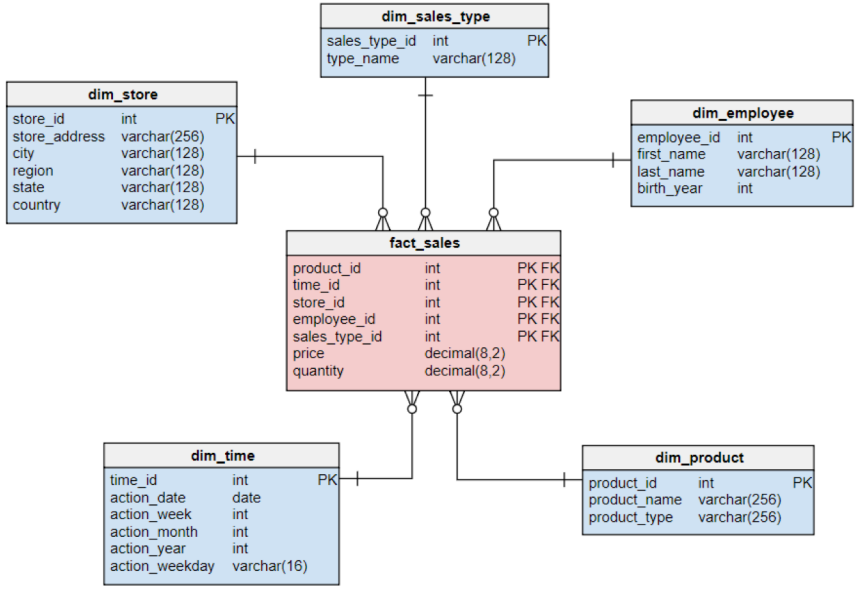
Pivoting – converting columns into rows

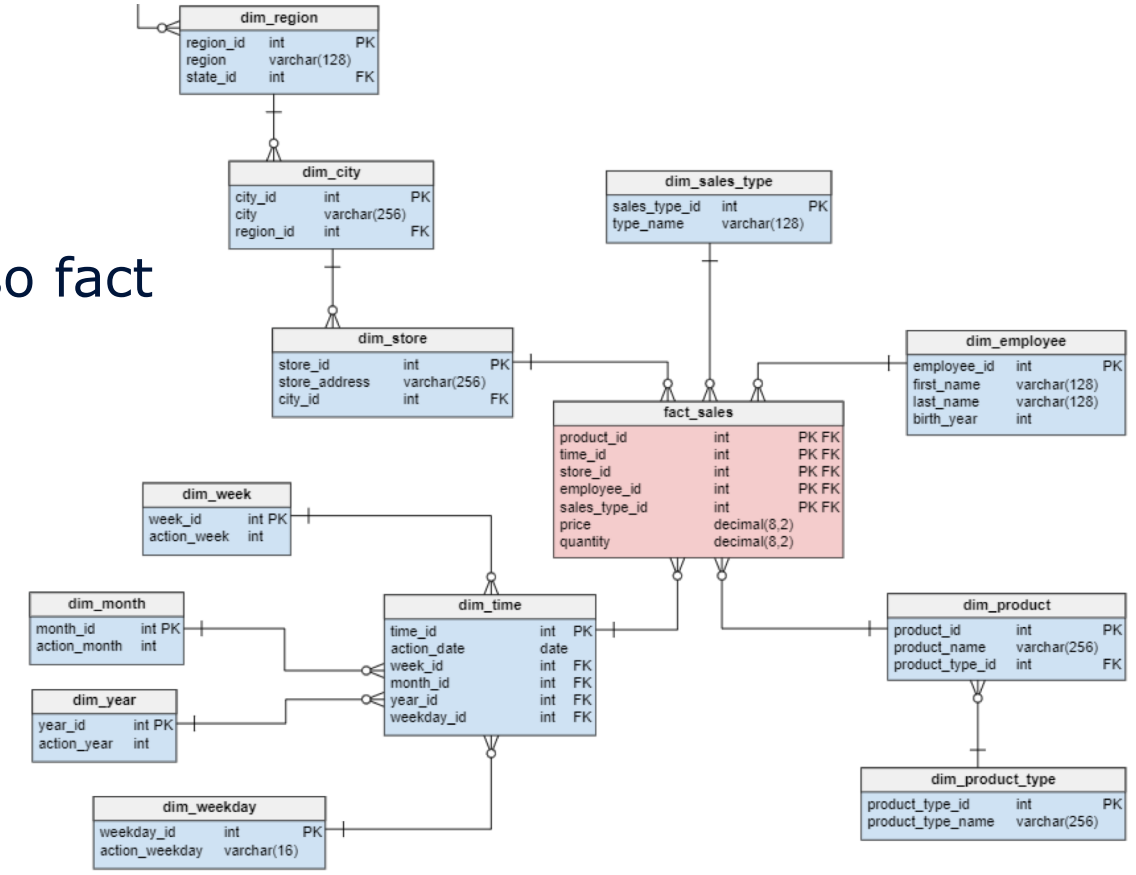
Splitting - separating a column that contains multiple pieces of information into multiple columns, one for each piece of information

Airline Name	Airline ID
American Airlines	AA
Delta Airlines	DL
JetBlue Airways	B6
United Airlines	UA

Cube-Based Data Structures

- Star schema
  - Fact table – stores measures
  - Dimension tables
- Snowflake schema - dimensions are also fact tables



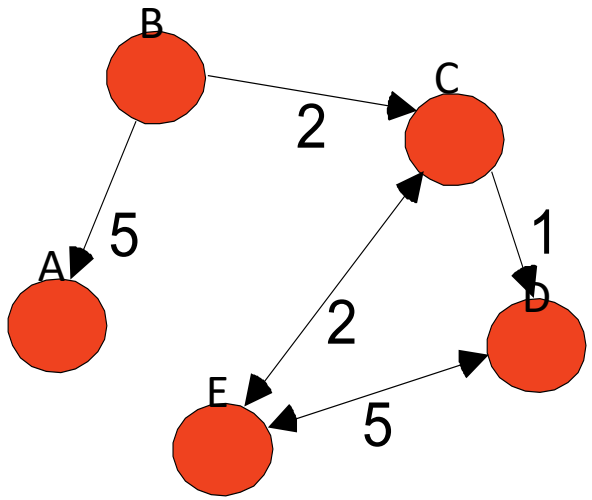


https://datawarehouseinfo.com/data-warehouse-star-schema-vs-snowflake-schema/

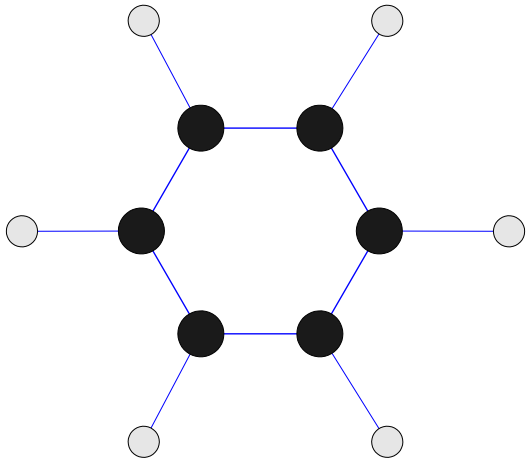
Graph-Based Data Structures

Graph and Sparse Matrix

has the property that only nonzero elements are stored in the underlying data structure



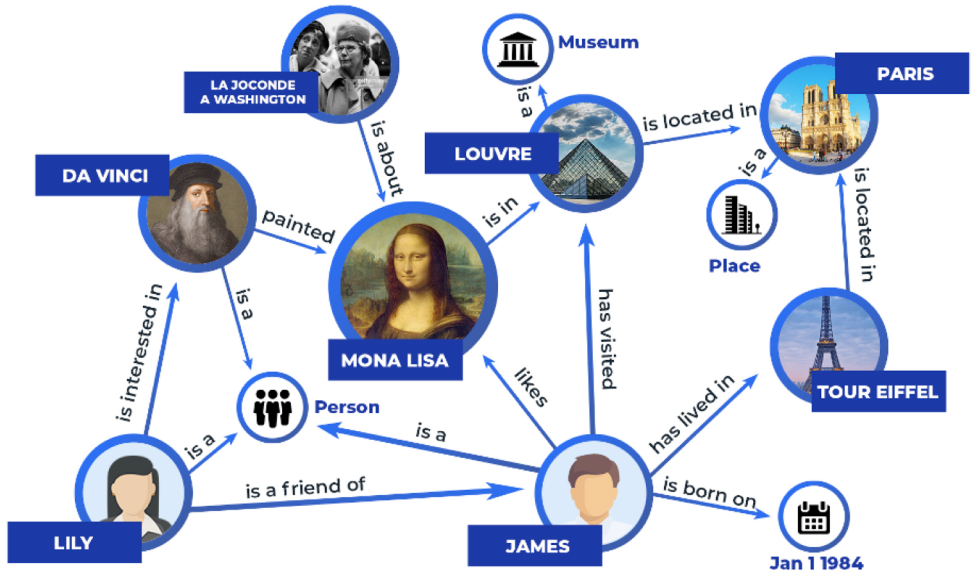
	A	B	C	D	E
A	0	0	0	0	0
B	1	0	1	0	0
C	0	0	0	1	1
D	0	0	0	0	1
E	0	0	1	1	0



Graph-Based Data Structures

Knowledge Graph

- the structure of modern AI applications
- collects all data in a specific domain area
- consists of nodes and relations (edges)
- Both the nodes and relations have type and properties
- Each node and each relation have values of the type properties (in key:value pairs)



Data Variables by Role

Dependent and Independent Variables

Variables by Role

- Independent Variables
  - Explanatory variables, or predicting variables, hold the facts, input
- Dependent Variables
  - Results, change in response to a change in the other variables, outcome
- Other Contributing Variables
  - Other variables that are important in explaining association between dependent and independent.

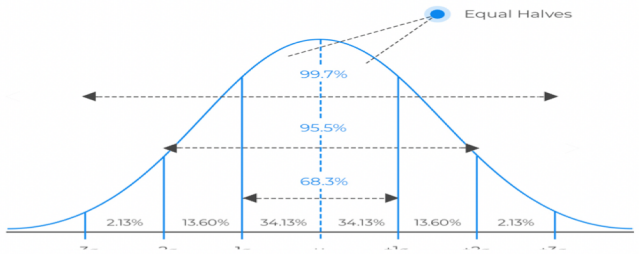
Dependency Measurement

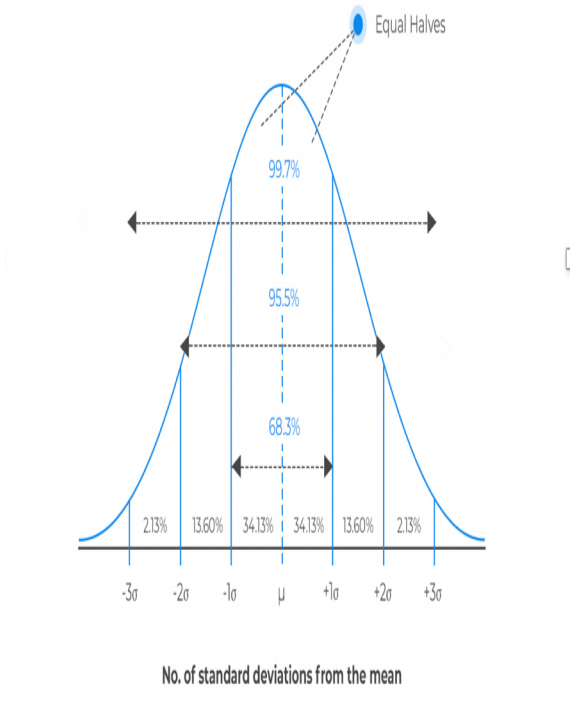
a, b, c, x1, x2, and Y are variables

a	b	c	x1	x2	y
0.96	0.38	0.45	0.49	0.36	0.15
0.65	0.09	0.83	0.22	0.25	0.81
0.79	0.84	0.57	0.81	0.48	0.21
0.71	0.41	0.79	0.27	0.53	0.53
0.16	0.43	0.01	0.63	0.96	0.90
0.46	0.17	0.56	0.65	0.42	0.75
0.15	0.54	0.02	0.26	0.31	0.52
0.62	0.69	0.88	0.17	0.49	0.16
0.12	0.21	0.86	0.35	0.86	0.37
0.16	0.40	0.79	0.38	0.75	0.34

$$Y = a*x_1 + b*x_2 + c$$

Covariance





- A metric for the extend of the dependency
  - positive covariance indicates a positive linear relationship between the variables
  - negative covariance indicates the negative

x, y – two normally distributed datasets

- Covariance defined in terms of means
  - for a sample
  - for population

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

Correlation Coefficient

- Statistical parameter, used to measure the strength and direction of the linear relationship between two numerical variables from the same observation data set
- Dimensionless - just a number in [-1, +1]
- Used only if both variables are numerical
- Estimated by a correlation coefficient r

Correlation Coefficient

Can be presented as

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

- Sx and Sy are the sample standar deviations
- Sxy is the sample covariance

or

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

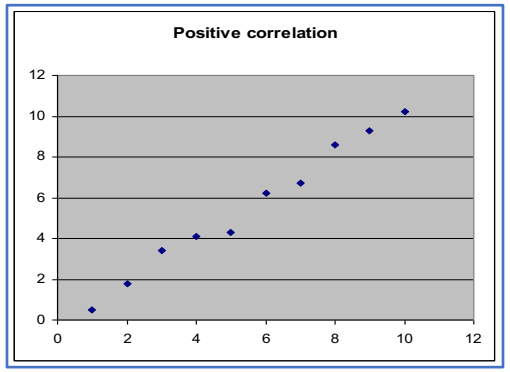
where:

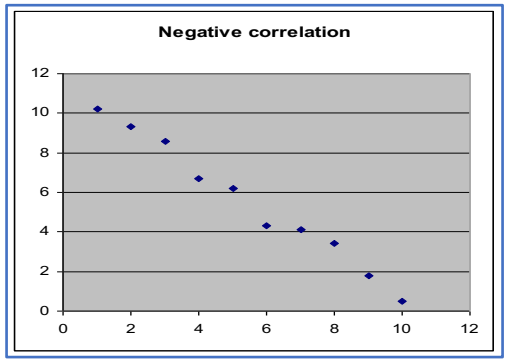
- n, x_i, y_i are defined as above
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

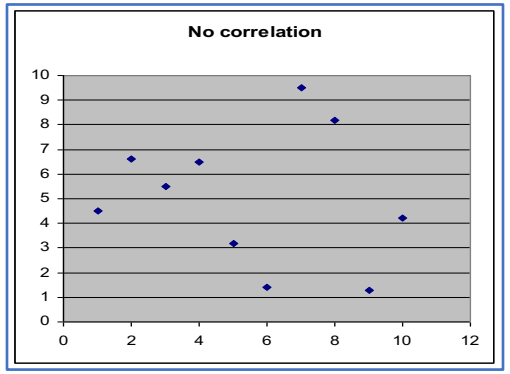
Correlation Coefficient

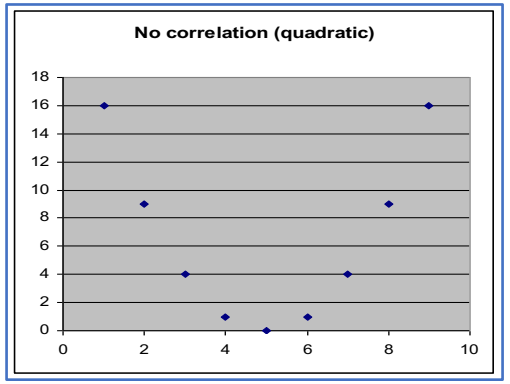
- The value of r belongs to the interval [-1, +1]
  - close to 1 indicates that the variables are positively linearly related and the scatter plot of the two samples falls almost along a straight line with positive slope
  - close to -1 indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope
  - correlation coefficient is close to 0 indicates a weak linear relationship

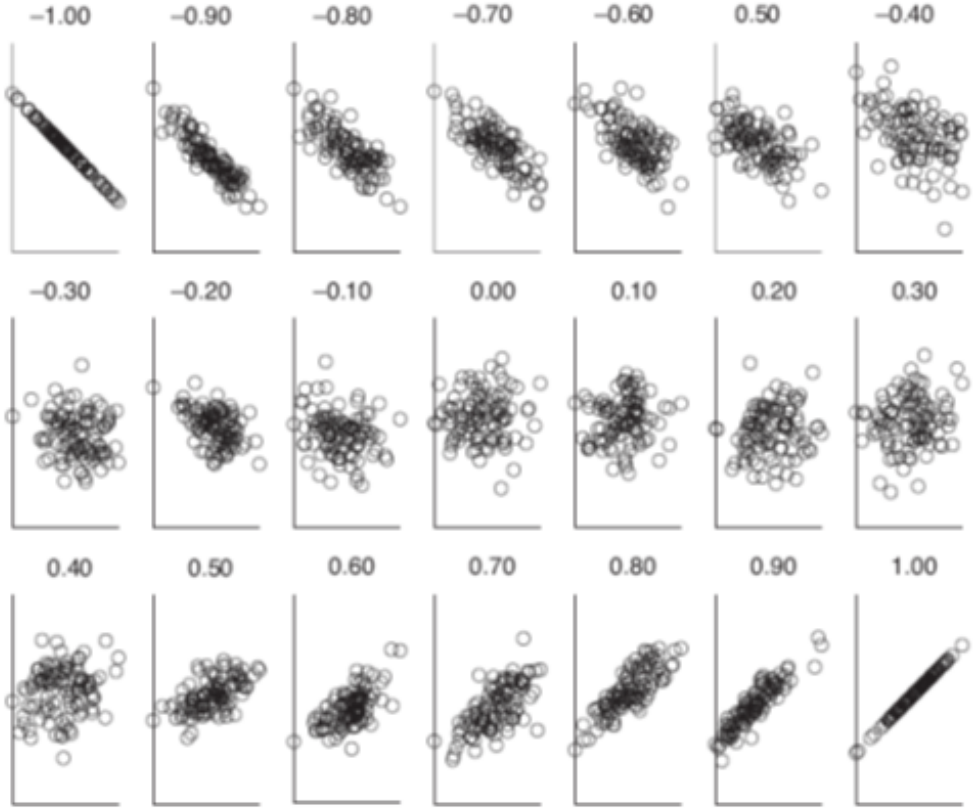
Correlation Coefficient Visually











Hands on Data Cubes