INTRODUZIONE AI BIG DATA

Prof. Flavio Venturini
Lez.02 Principi di Data Governance e DFM

PRINCIPI DI DATA GOVERNANCE E DFM

Outline

→ What is Data Governance?

→ Data Governance principles

→ Information systems modelling concepts

→ Dimensional Fact Modelling

WHAT IS DATA GOVERNANCE?

It is the planning and execution of policies, practices, and projects that acquire, control, protect, deliver, and enhance the value of data and information assets

DG is for all data

Structured and unstructured data have different life-cycles for analytical purposes

Structured Data life cycle

→ Acquisition

→ Cleansing / Discovery

→ Normalization

→ Aggregation (by relevant dimensions)

→ User access optimisation

UNIVERSITÀ TELEMATICA
INTERNAZIONALE UNINETTUNO

INTRODUZIONE AI BIG DATA

Prof. Flavio Venturini
Lez.02 Principi di Data Governance e DFM

Unstructured Data life cycle

→ Acquisition

→ Audio split / Video Frame identification / Text correction

→ Image recognition / Audio deciphering / Text Tagging

→ Text Indexing

→ Text Entity extraction

→ Text Semantic analysis

→ Text Sentiment analysis

→ Video / Audio metadata inference

Review Questions

→ Is Data Governance important for both structured and unstructured data ?

→ Explain the differences in the structured and unstructured data life cycles

---

DATA GOVERNANCE PRINCIPLES

Root principle

The objective of governing information is to provide value to the organization by processing data as quickly as is practical while keeping the quality as high as is practical and as secure as is practical

---

Retention

Information should be retained whenever physically possible within the constraints of government legislation, corporate ethics, and privacy commitments

---

UNIVERSITÀ TELEMATICA
INTERNAZIONALE UNINETTUNO

Prof. Flavio Venturini
Lez.02 Principi di Data Governance e DFM

Quality

→ Data quality is to be measured in absolute terms

→ Decision makers not only need access to data, but they also need to understand its timing, reconciliation, completeness and accuracy

Access

Any staff member can access information unless there is a specific commercial, legal, or ethical reason why the information should not be made available to this individual

Custody

→ Every item of data requires unique and ultimate ownership by a single role and person

→ A matrix of responsibilities should be managed that ensures that issues or conflicts always have an ultimate point of escalation

Compliance

Fulfill the obligations of external regulations from international, national, regional, and local governments

Mapping

→ The capture and recording of data at rest and data in motion throughout the organization allows to monitor and improve business effectiveness

→ What is the source ?

Meaning

Managing it actively reduces ambiguity, redundancy, and inconsistency, which relates directly to the quality of information
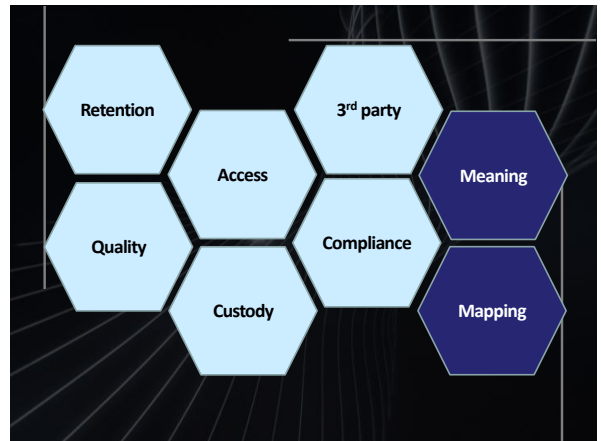
UNIVERSITÀ TELEMATICA
INTERNAZIONALE UNINETTUNO

INTRODUZIONE AI BIG DATA

Prof. Flavio Venturini
Lez.02 Principi di Data Governance e DFM

3rd party control

→ Third-party data plays an expanding role in big data
  → Providers of data
  → Subscribers of data
  → Outsourced delivery

Review Questions

What is the key data governance principle for ensuring the whole organization shares the same view on data?

INFORMATION SYSTEMS MODELLING BASICS

Data modelling techniques

→ Context diagrams
→ Entity relationship diagrams
→ Data flow diagrams
→ Dimensional Fact Model

Context diagrams - CD

→ A context diagram clarifies the interfaces and boundaries of the system

→ It also shows the interactions with other systems and users

UNIVERSITÀ TELEMATICA
INTERNAZIONALE UNINETTUNO

INTRODUZIONE AI BIG DATA

Prof. Flavio Venturini
Lez.02 Principi di Data Governance e DFM

Entity relationship - ERD

It is a graphical representation that shows the relationship between people, objects, places, concepts or events within an information system

Data flow diagram - DFD

It is a graphical representation of the "flow" of data through an information system, modelling its process aspects

Review Questions

What do these three modelling techniques have in common?

DIMENSIONAL FACT MODELLING

Dimensional Fact Model-DFM

It is a representation of facts, measures, dimensions and hierarchies used in a system to allow users and other systems access them

Fact

A fact is a focus of interest for the data analysis process typically, it models an event occurring in the enterprise world
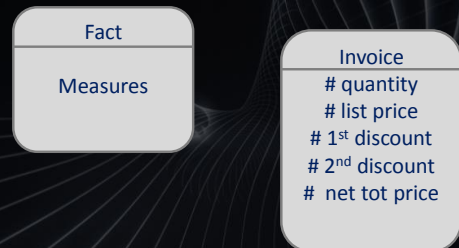(e.g. sales, shipments, hires)

INTRODUZIONE AI BIG DATA

Prof. Flavio Venturini
Lez.02 Principi di Data Governance e DFM

Measure

→ Measures are continuously valued (numerical) attributes which describe the fact from different points of view

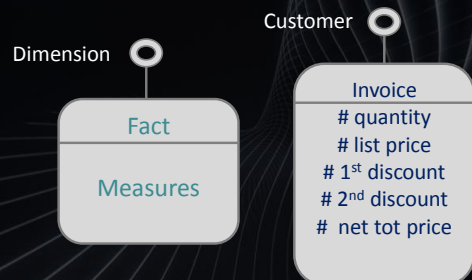→ E.g. each sale is measured by its revenue

Fact and measure syntax

Fact

Measures

Invoice
quantity
list price
1st discount
2nd discount
net tot price

Dimension

→ Dimensions are discrete attributes which determine the minimum granularity adopted to represent facts

→ typical dimensions for the sale fact are product, store and date

Dimension syntax

Customer

Dimension

Fact

Measures

Invoice
quantity
list price
1st discount
2nd discount
net tot price

Hierarchy

Hierarchies are made up of discrete dimension attributes linked by -to-one relationships, and determine how facts may be aggregated and selected significantly for the analysis process
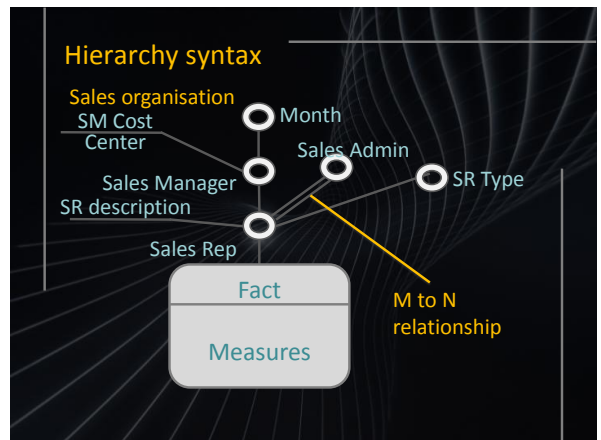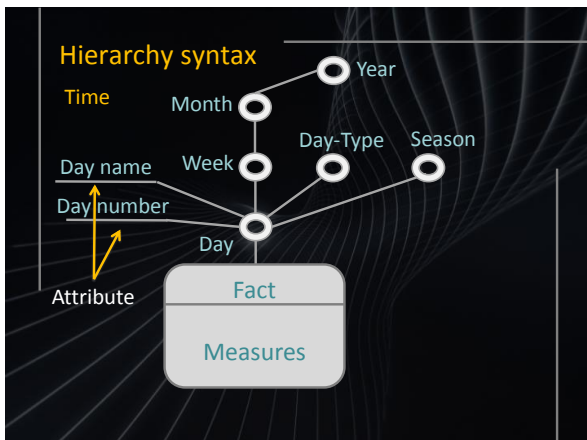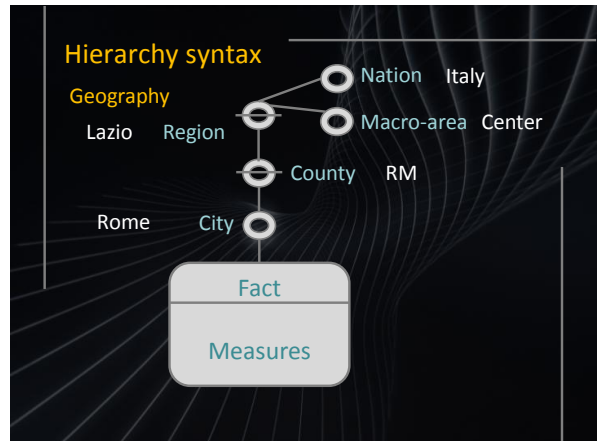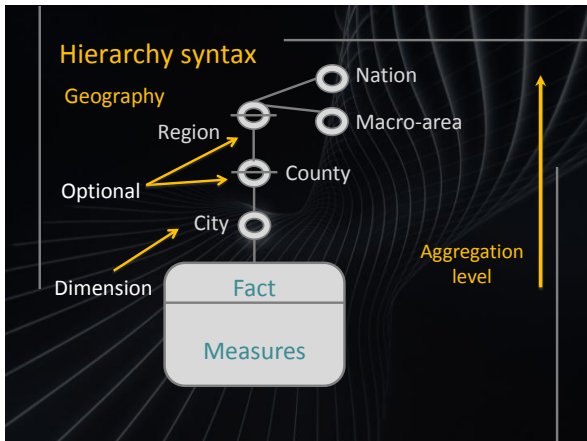
Conformed Hierarchies

→ Conformed Hierarchies are hierarchies common to more than a fact schema

→ E.g. Time hierarchy, geography hierarchy

UNIVERSITÀ TELEMATICA
INTERNAZIONALE UNINETTUNO

INTRODUZIONE AI BIG DATA

Prof. Flavio Venturini
Lez.02 Principi di Data Governance e DFM

Hierarchy syntax

Geography

Nation
Macro-area
Region
Optional
County
City
Dimension
Fact
Measures
Aggregation level

Hierarchy syntax

Geography

Nation   Italy
Macro-area   Center
Lazio   Region
County   RM
Rome   City
Fact
Measures

Hierarchy syntax

Time

Year
Month
Day-Type   Season
Week
Day name
Day number
Day
Attribute
Fact
Measures

Hierarchy syntax

Sales organisation
SM Cost Center
Month
Sales Admin
Sales Manager
SR description
SR Type
Sales Rep
Fact
Measures
M to N relationship

DFM Importance

→ It allows to understand which data is available for external consumption and how it can be accessed

→ Enables sharing and understanding of information

DFM and BIG Data Governance

Meaning

→ Shared data glossary from a user view point

Mapping

→ Single view of published data

UNIVERSITÀ TELEMATICA
INTERNAZIONALE UNINETTUNO

Copyright © Università Telematica Internazionale UNINETTUNO

7

INTRODUZIONE AI BIG DATA

Prof. Flavio Venturini
Lez.02 Principi di Data Governance e DFM

Review Questions

→ What is the key difference between ER and DFM?

→ Why the usage of DFM is key for setting up a proper Data Governance?

SUMMARY QUESTIONS

→ Is Data Governance important for both structured and unstructured data?

→ Explain the differences in the structured and unstructured data life cycles

→ What is the key data governance principle for ensuring the whole organization shares the same view on data?

→ What do these three modelling techniques have in common?

→ What is the key difference between ER and DFM?

→ Why the usage of DFM is key for setting up a proper Data Governance?