Data Warehouse Modelling

Lecture for INFO8002 - Large-scale Data Systems, ULiège

Olivier Bui Quôc – 27/11/2018

HERMES
A HUMAN WAY OF MANAGING SERVICES

I’m Olivier Bui Quoc

Liège Université / HEC Liège

Data Architect @ KBL

olivier.buiquoc@hermes-ecs.com

Agenda

- Introduction
- Data Warehouse
- Relational vs Dimensional
- Approaches
- Data Modelling
- Size Estimation
- Example
- BI and Big Data

Introduction

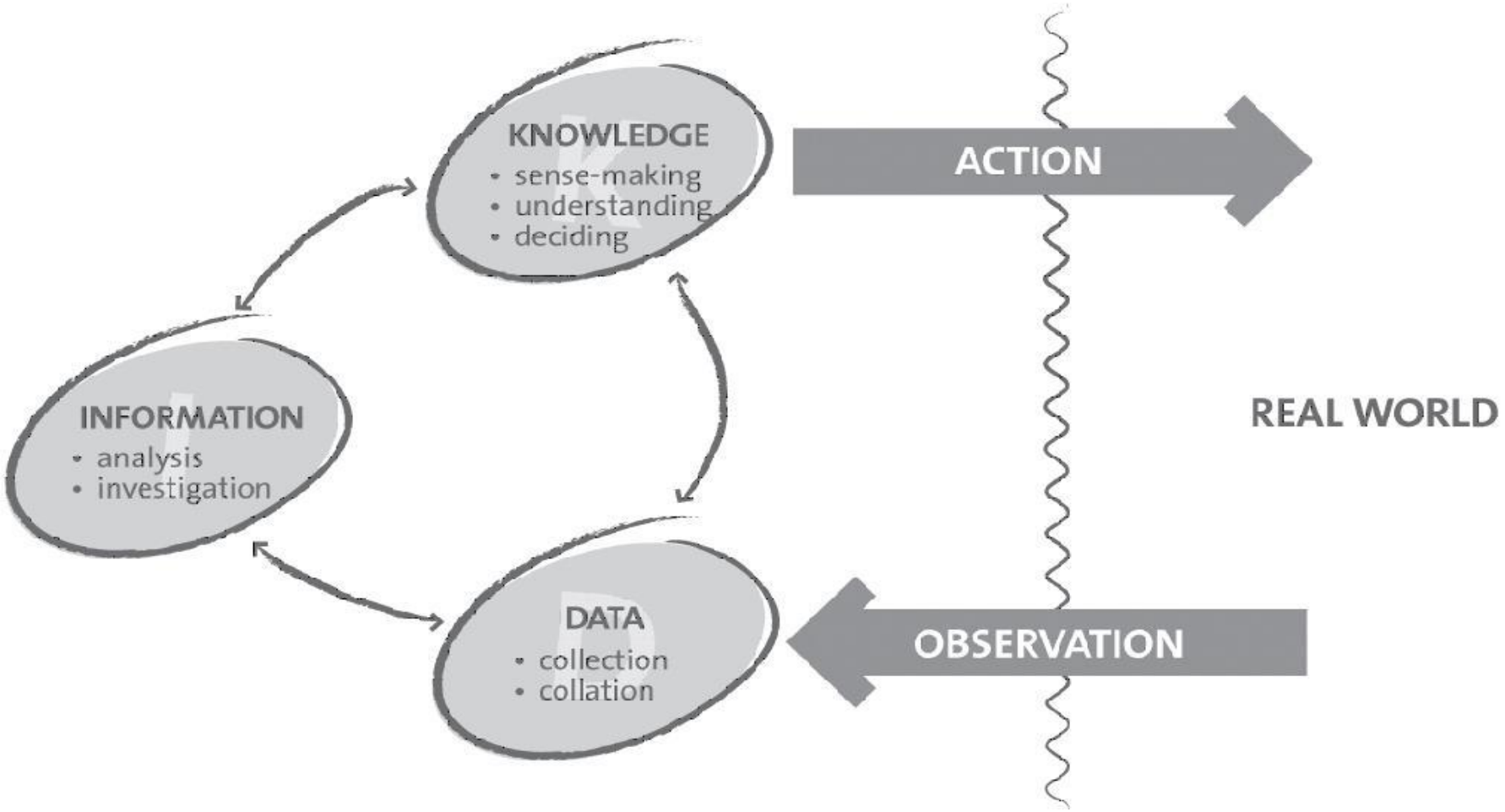
Business Intelligence ?

Business Intelligence

Business intelligence (BI)
is a broad category of applications and technologies
for gathering, storing, analyzing, and providing access to data
to help enterprise users **make better business decisions**.

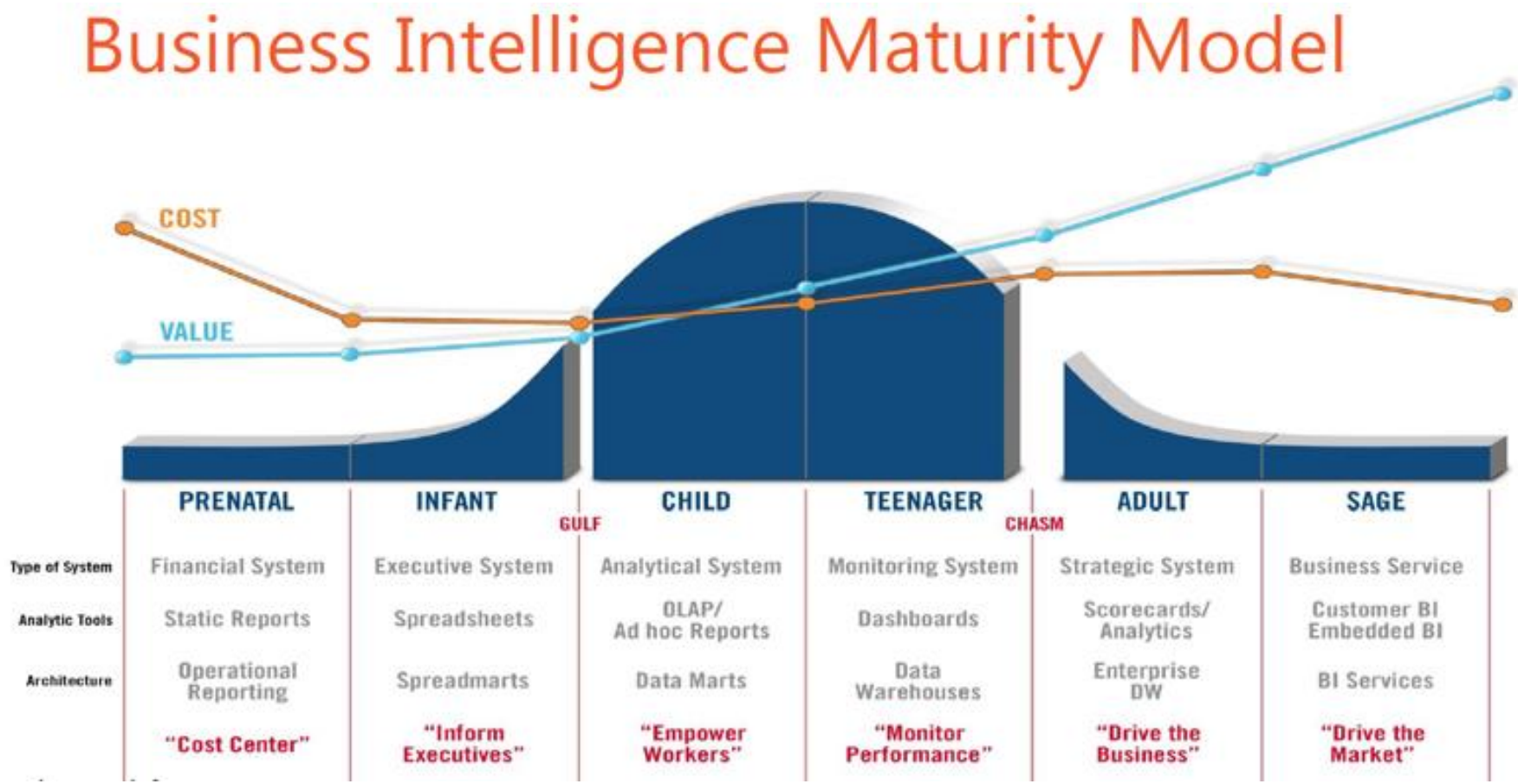
Source: SearchCrm.com

The BI Process



The BI Process

BI Maturity Model



Context

- Business users have built their own analysis and reporting solutions
- Data processing is complex and « manual »
- Data consistency is not always ensured
- Lack of compliance with IT standards (process and tools)
- Lack of automation
- IT is reduced to the role of Data Provider with few business insight
- Data is not available in a easy and ergonomic way for analysis

Data Warehouse

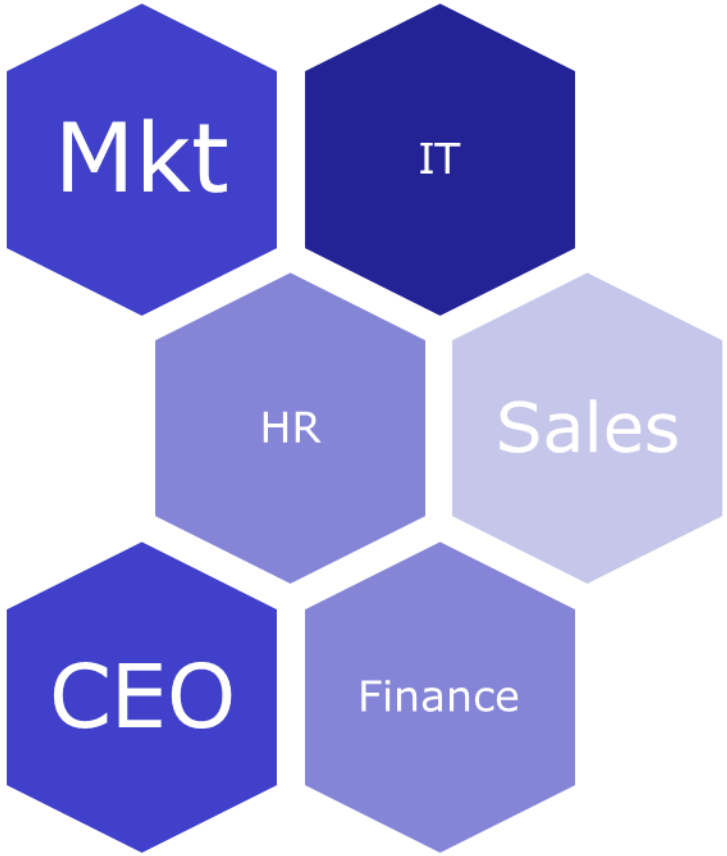
What is a Data Warehouse?

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process.

- Bill Inmon

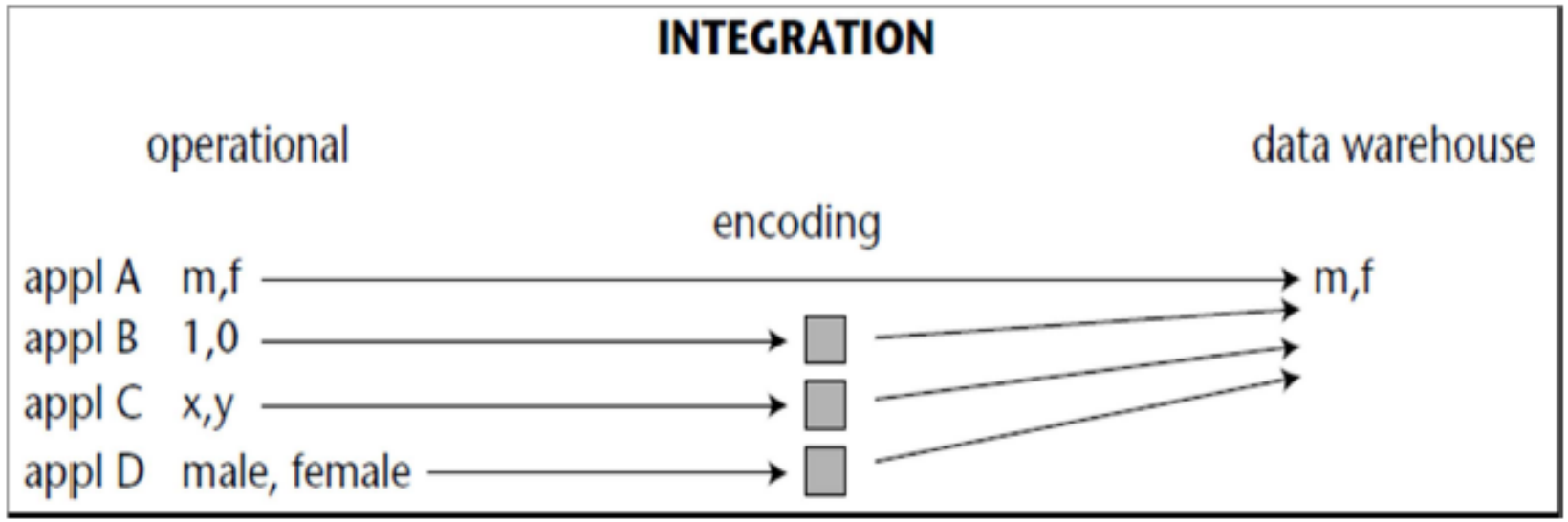
Data Warehouse

- **Subject-Oriented:** A data warehouse can be used to analyze a particular subject area.



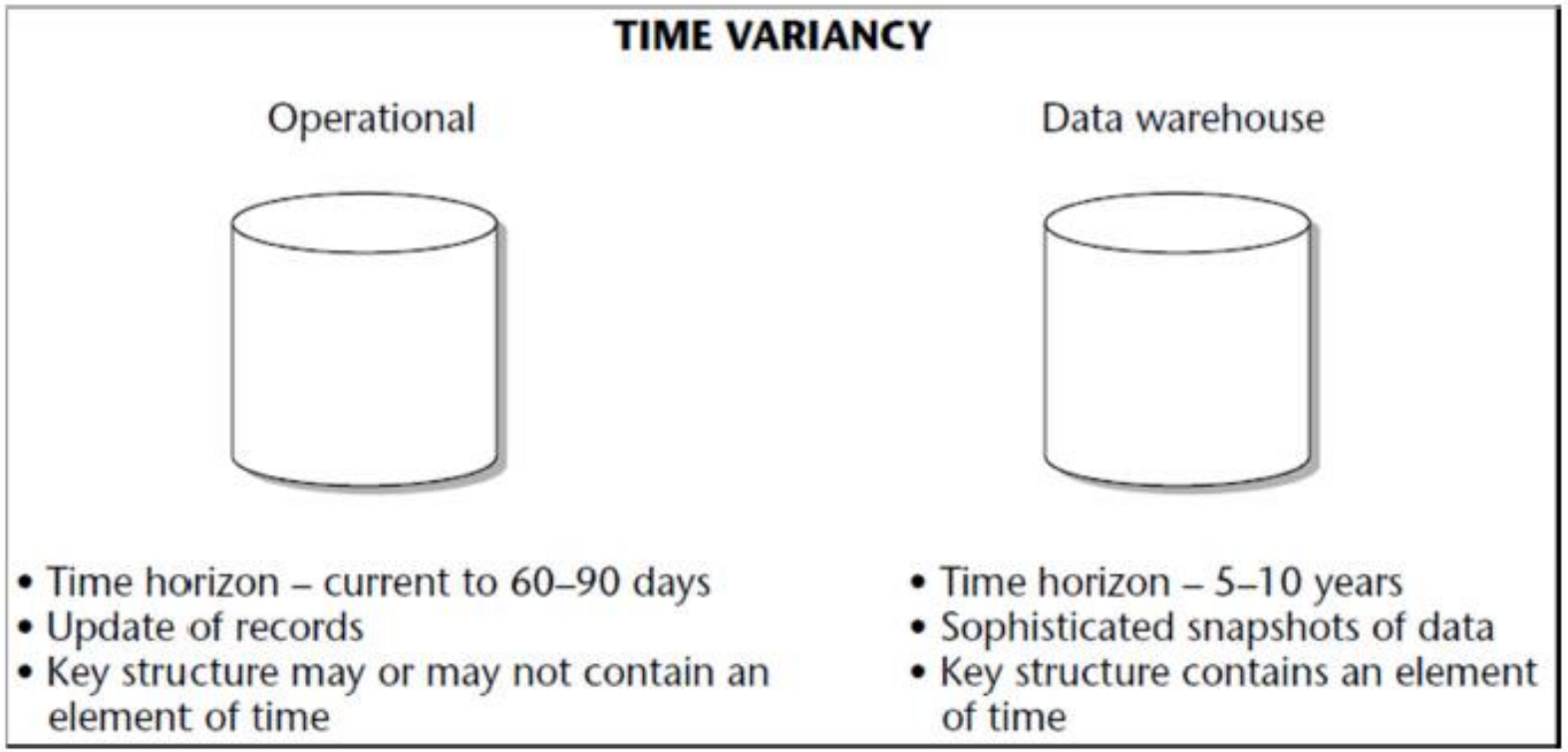
Data Warehouse

- **Integrated:** A data warehouse integrates data from multiple data sources. There will be only a single way of identifying a same concept shared by sources.



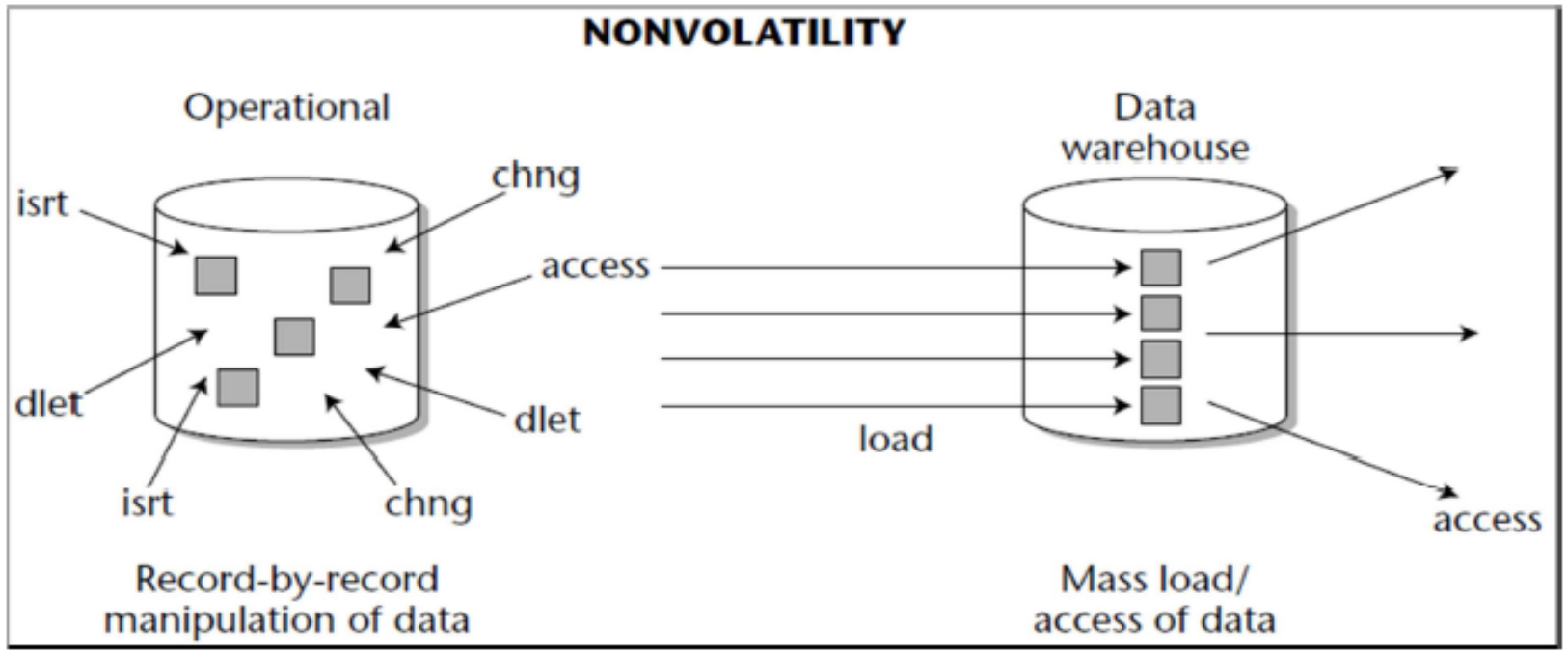
Data Warehouse

- **Time-Variant:** Historical data is kept in a data warehouse. This contrasts with a transactions system, where often only the most recent data is kept.



Data Warehouse

- **Non-volatile:** Once data is in the data warehouse, it will not change. So, historical data in a data warehouse should never be altered.



What is a Data Mart?

A data mart is a subject-oriented subset of data stored within the overall data warehouse, for the needs of a specific team, section or department within the business organization.

Data Mart

- A data mart makes specific subsets of data available to a set group of users so they have quicker access to the relevant data.
- Data marts make it much easier for individual departments to access key data insights more quickly as they are tailored for their needs
- They help prevent departments from interfering with each other’s data within the organization.

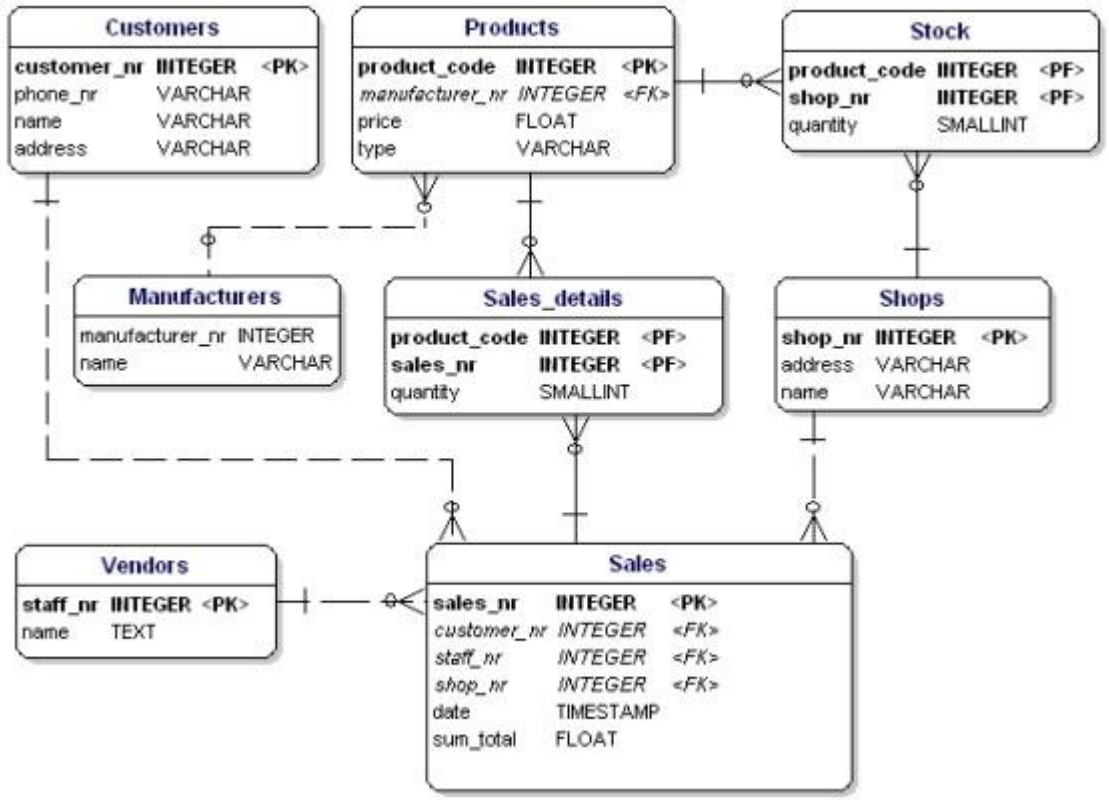
Relational vs Dimensional

OLTP vs OLAP

	OLTP	OLAP
Uses	RDBMS	Data Warehouse
Operations	Update	Analysis
Access types	Read/Write	Read
Level of analysis	Elementary	Global
Amount of data processed	Small	Large
Orientation	Row	Multi-dimensional
Database size	Small (GB)	Large (TB)
Data freshness	Most recent	Historical

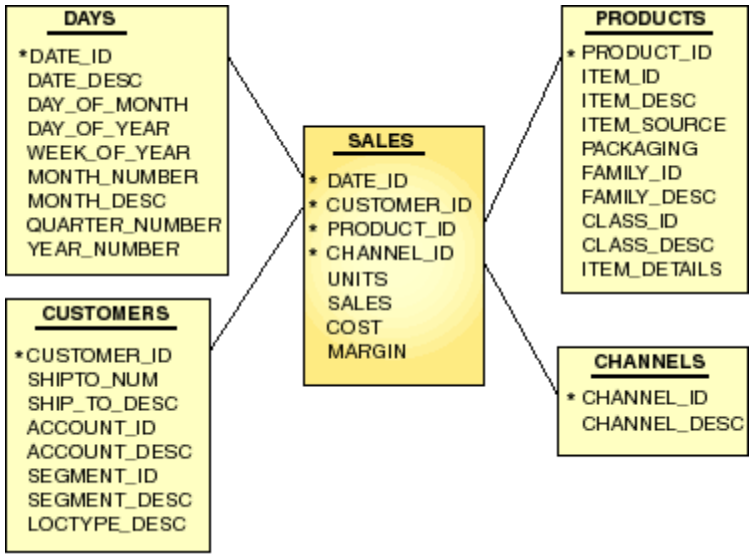
From operational to dimensional

Operational Data Model



Normalization

Dimensional Data Model

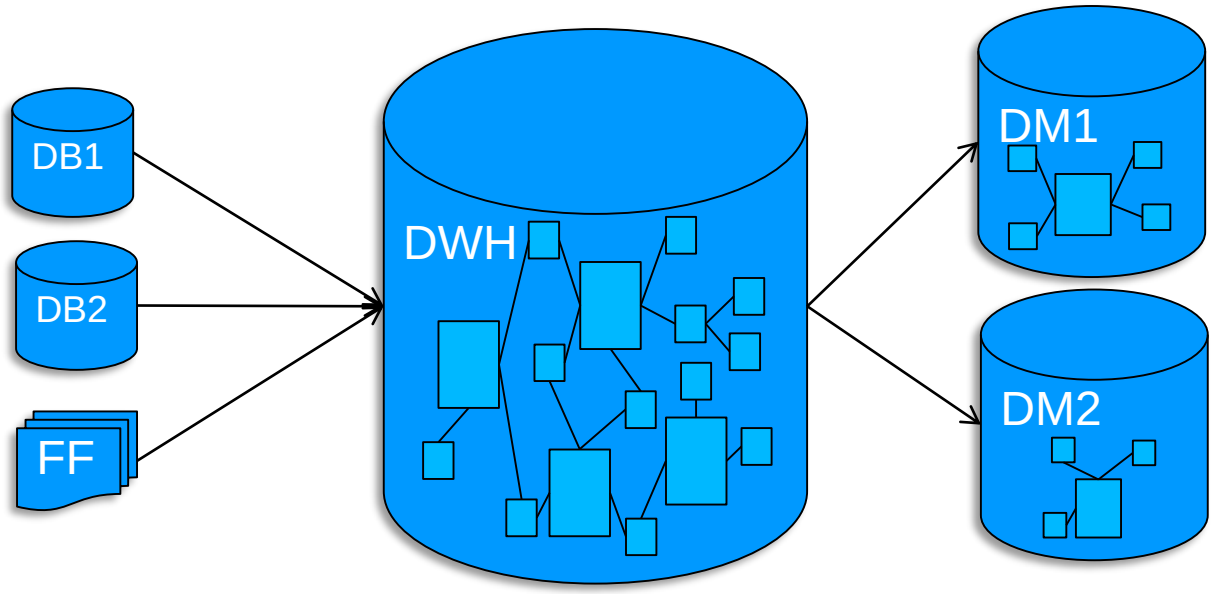


Allows redundancy for performance aspects

Approaches

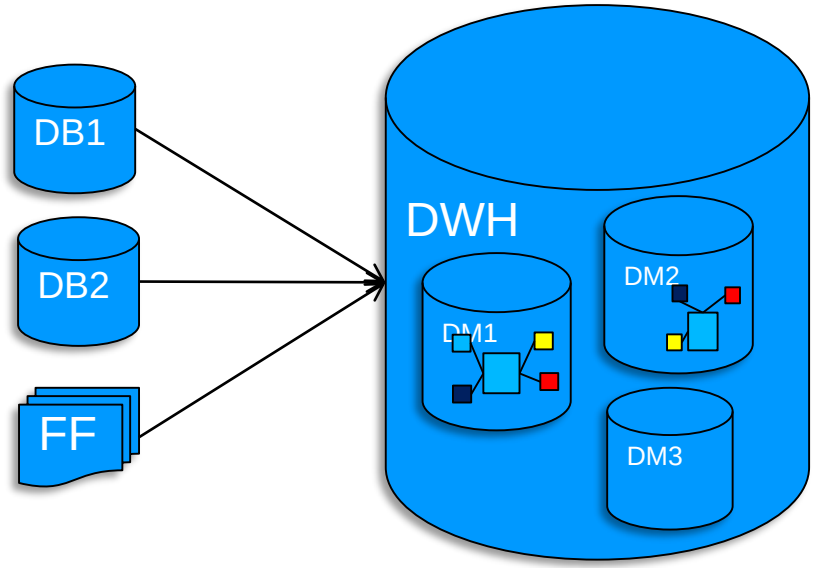
Bill Inmon

- Top-down
- A data warehouse is a subject oriented, nonvolatile, integrated, time variant collection of data in support of management's decisions



Ralph Kimball

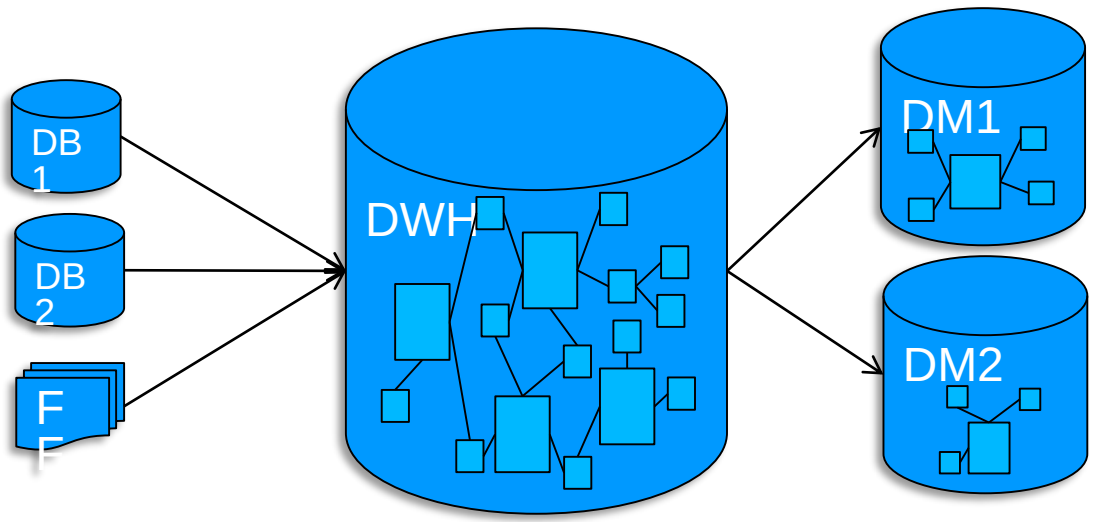
- Bottom-up
- A copy of transaction data specifically structured for query and analysis

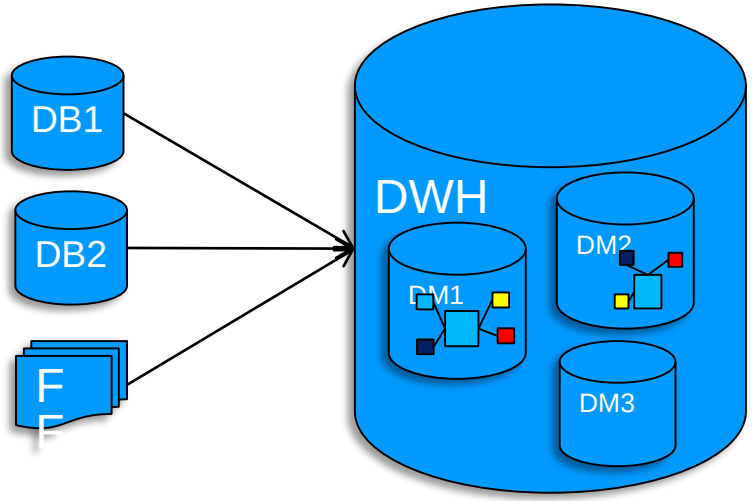


Bus architecture,
conformed dimensions

Approaches comparison

Pro's and Con's ?

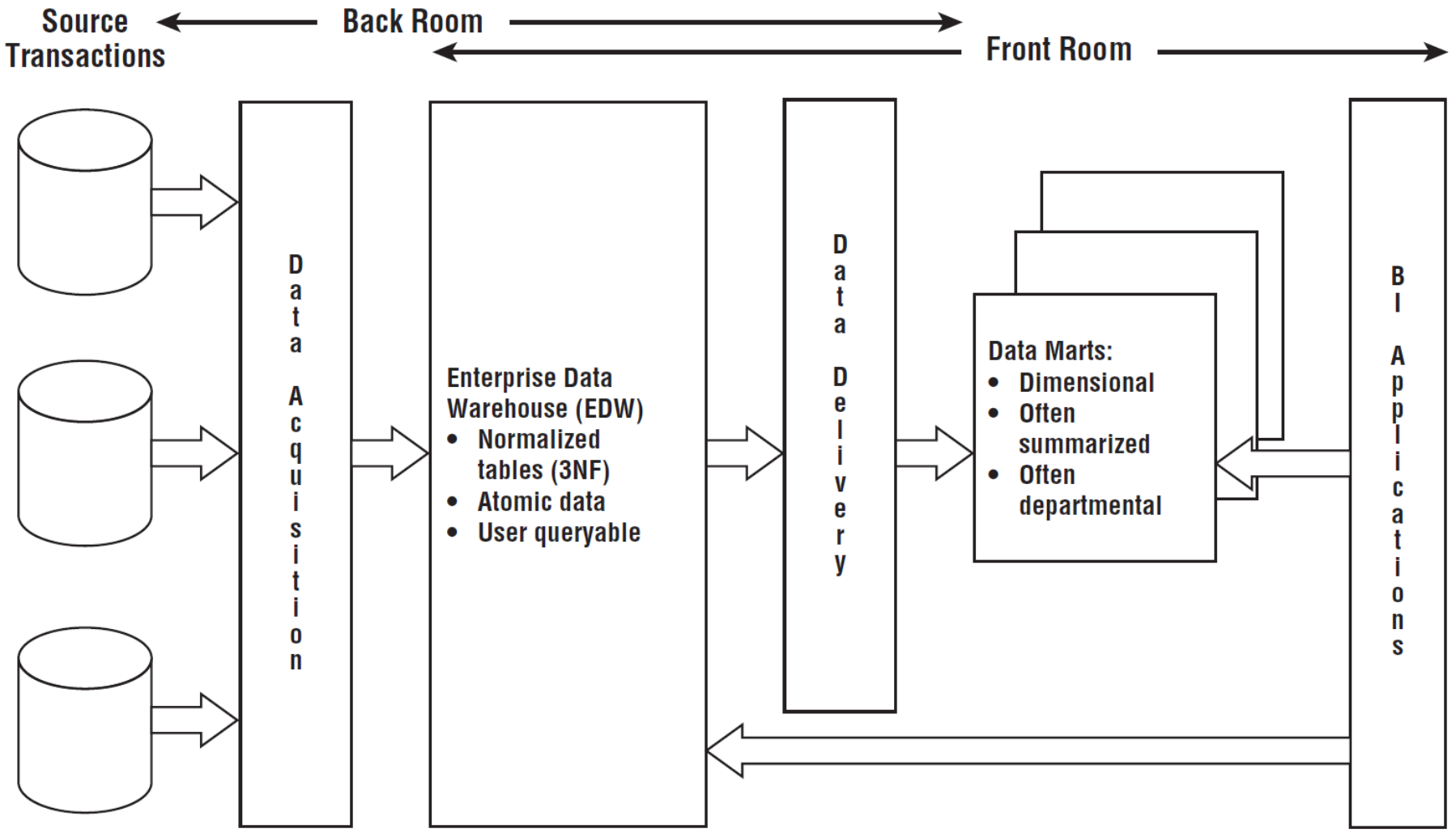




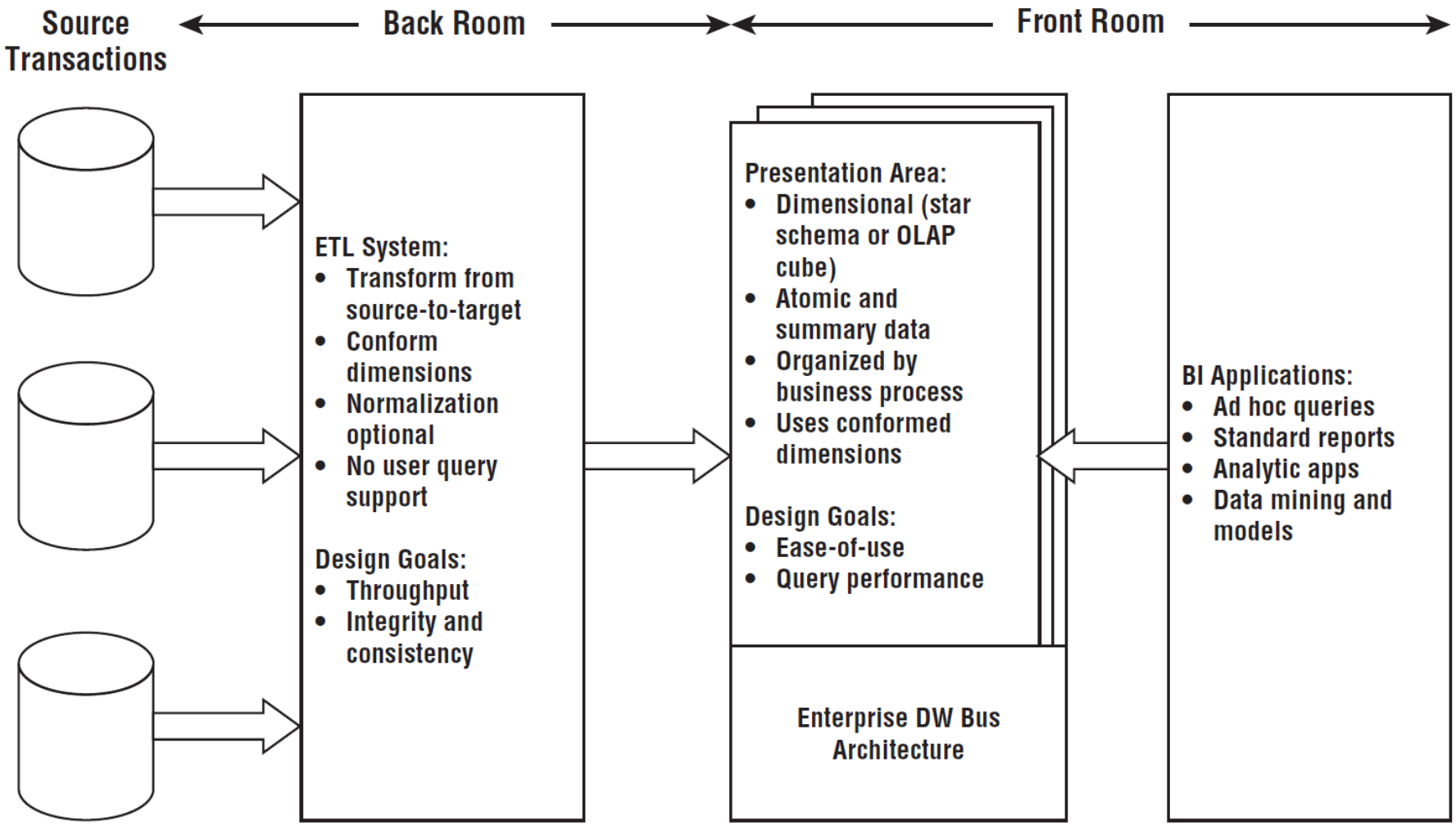
Approaches comparison

	Pro's	Con's
Bill Inmon (Top-down)	Enterprise-wide integration	Quite complex Longer start-up time Low user accessibility
Ralph Kimball (Bottom-up)	Fairly simple Time to delivery High user accessibility	Scalability with growing number of datamarts

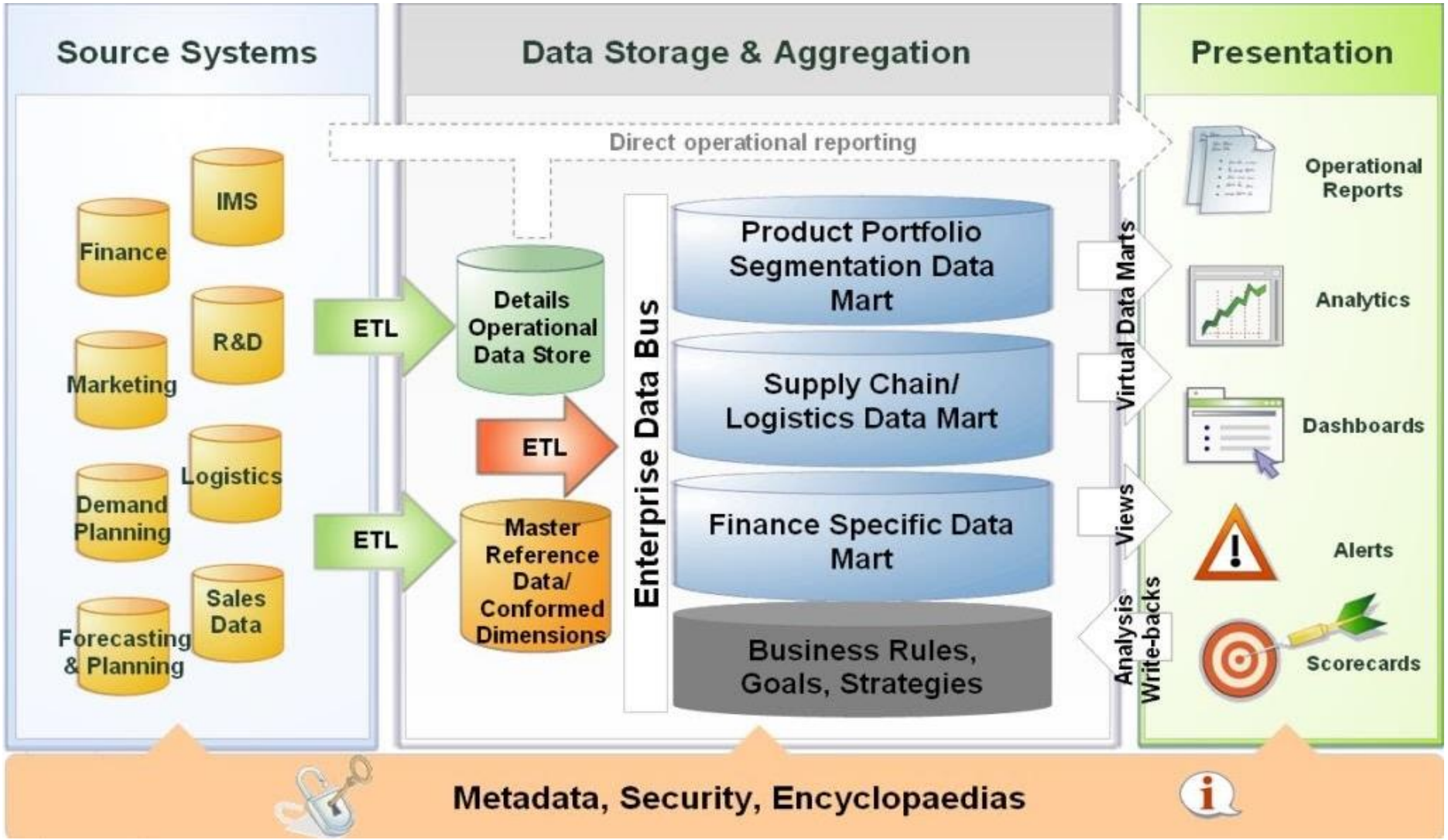
Architecture : Inmon



Architecture : Kimball



Architectures

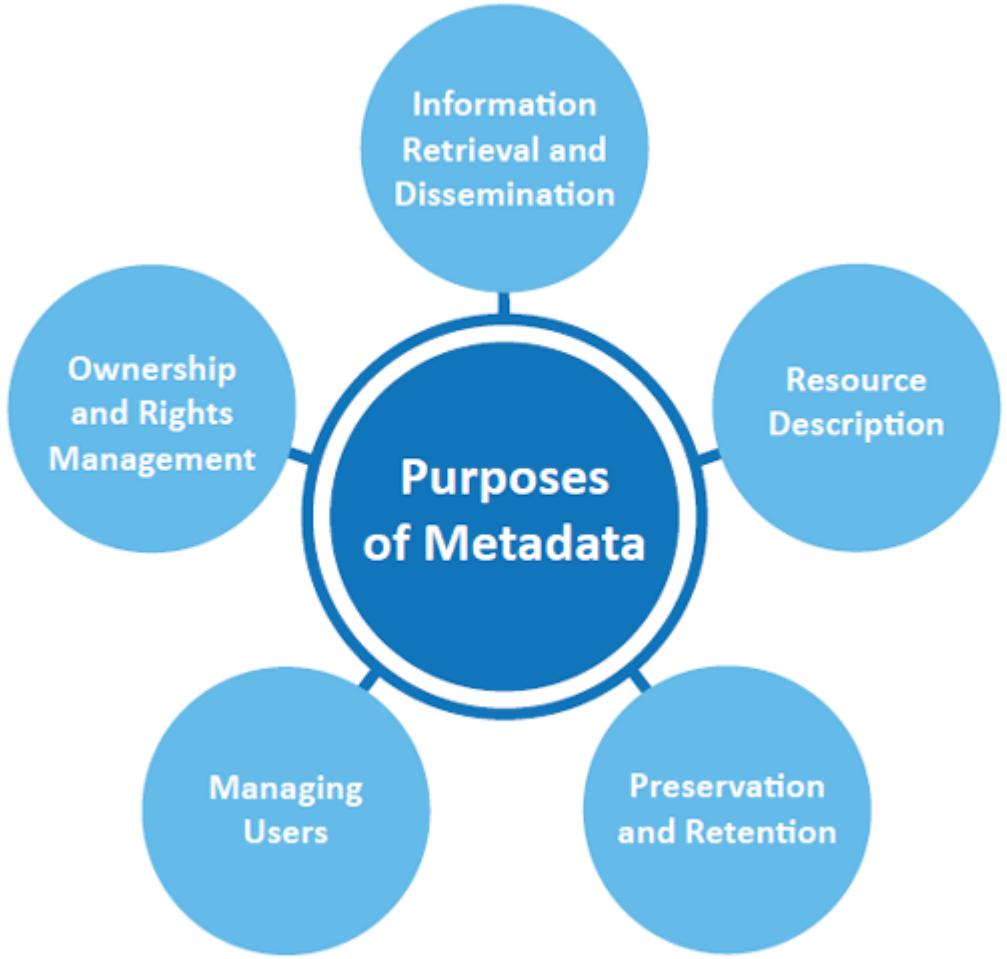


Metadata

Metadata
/ˈmɛtədeɪtə/

Noun
noun: **meta-data**

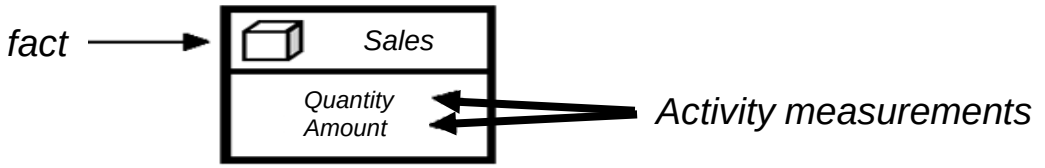
a set of data that describes and gives information about other data.

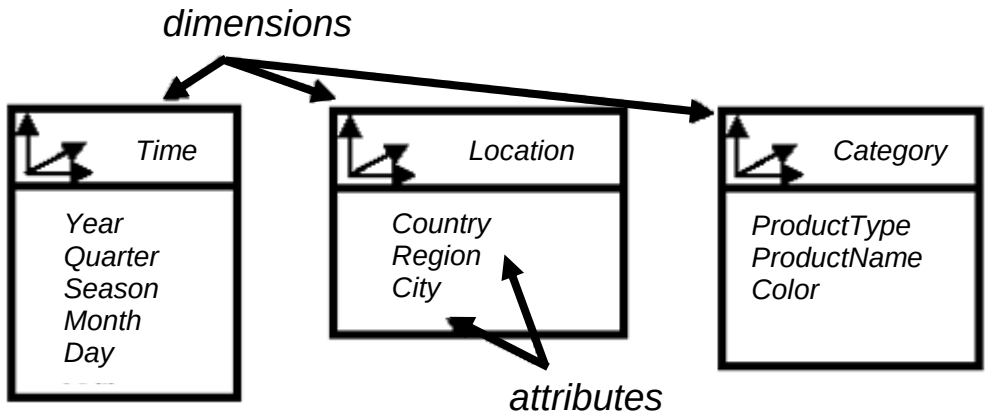


Data Modelling

Facts and Dimensions

- Fact tables
  - Contain the data related to a particular business process.
  - Each row represents a **single event** associated with a process and contains the **measurement data** associated with that event.
- Dimension tables
  - Contain the data describing the events of the fact tables.
  - They **categorize the measures** in order to enables users to answer business questions.





Facts and Dimensions

- Dimensions
  - Axis of analysis
  - Qualitative
  - Smaller size

  - Columns
    - Primary Key (Surrogate)
    - Natural Key
    - Attributes
    - Some technical fields

Product Dimension
Product Key (PK)
SKU Number (Natural Key)
Product Description
Brand Name
Category Name
Department Name
Package Type
Package Size
Abrasive Indicator
Weight
Weight Unit of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
...

Facts and Dimensions

- Facts
  - Measures
  - Quantitative
  - Large size

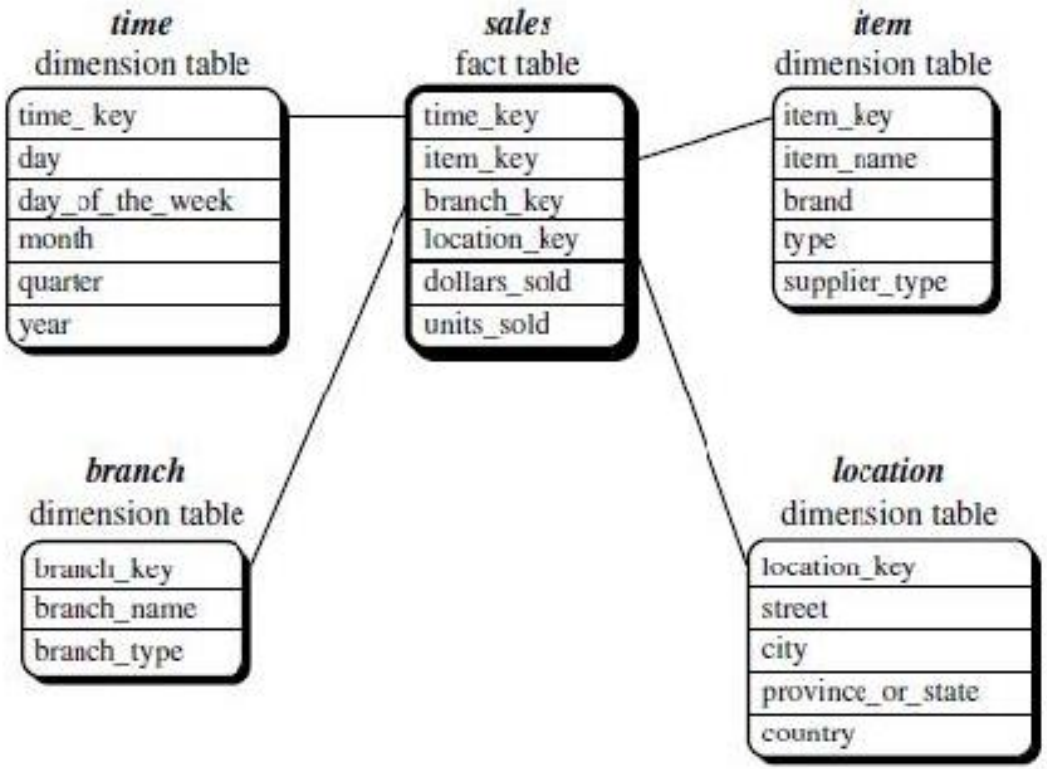
  - Columns
    - Foreign Keys to dimensions
    - Measures

Retail Sales Fact
Date Key (FK)
Product Key (FK)
Store Key (FK)
Promotion Key (FK)
Customer Key (FK)
Clerk Key (FK)
Transaction #
Sales Dollars
Sales Units

Models

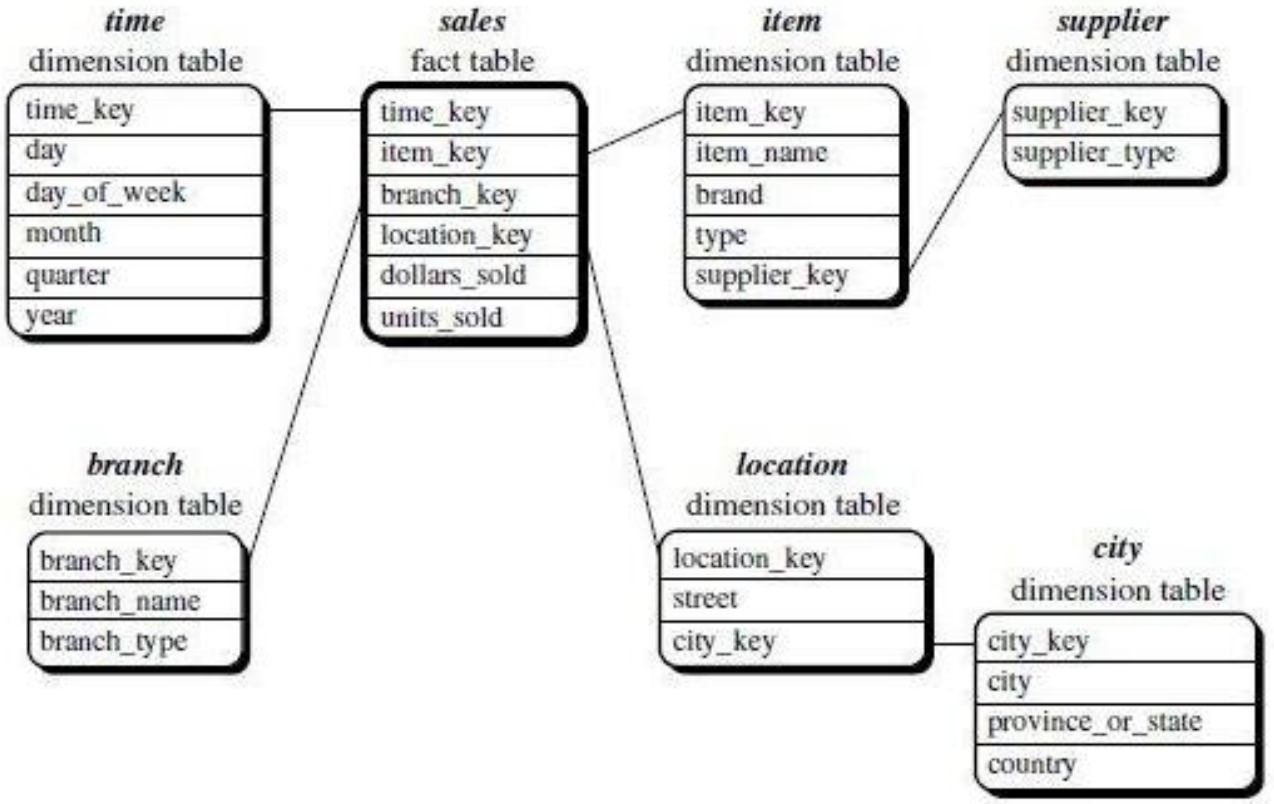
Data Models

- Star schema
  - Fact and dimensions



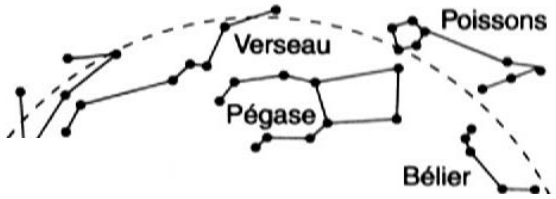
Data Models

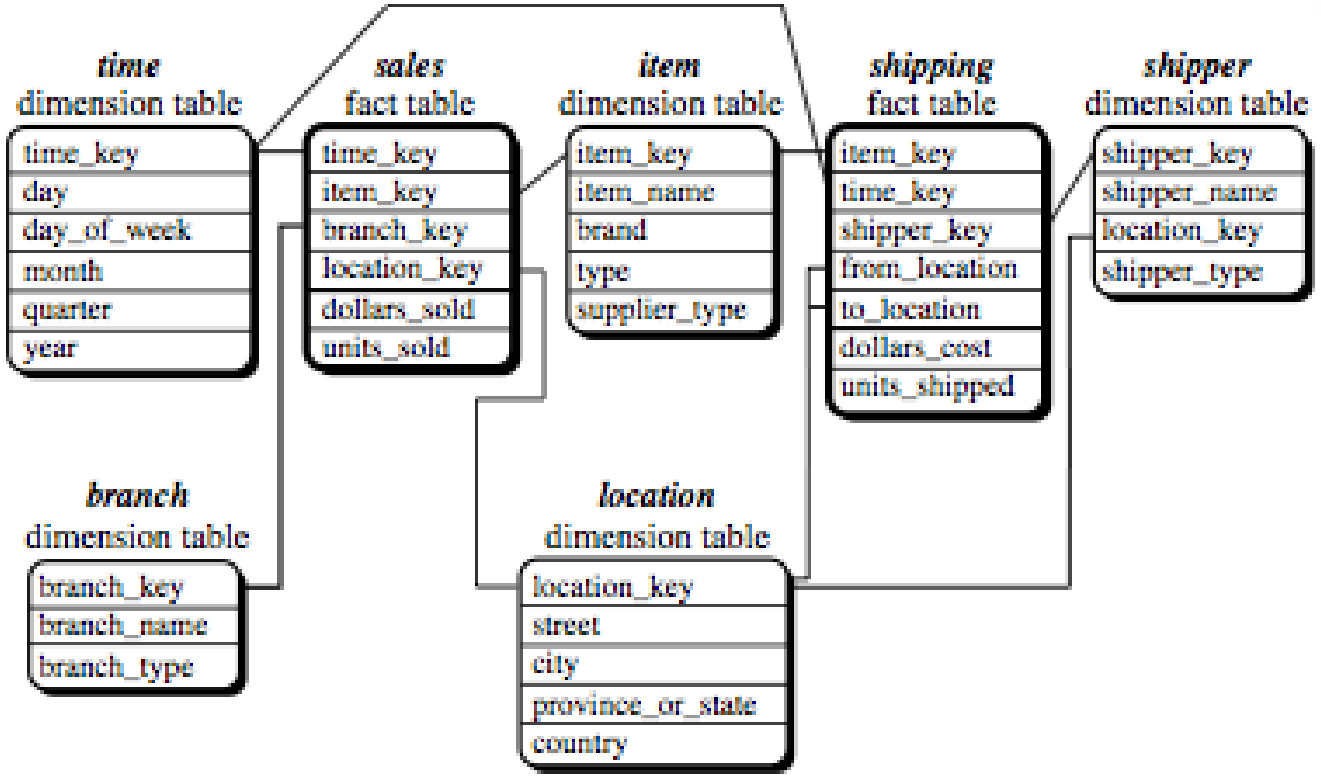
- Snowflake
  - Hierarchies



Data Models

- Constellation
  - Conformed dimensions

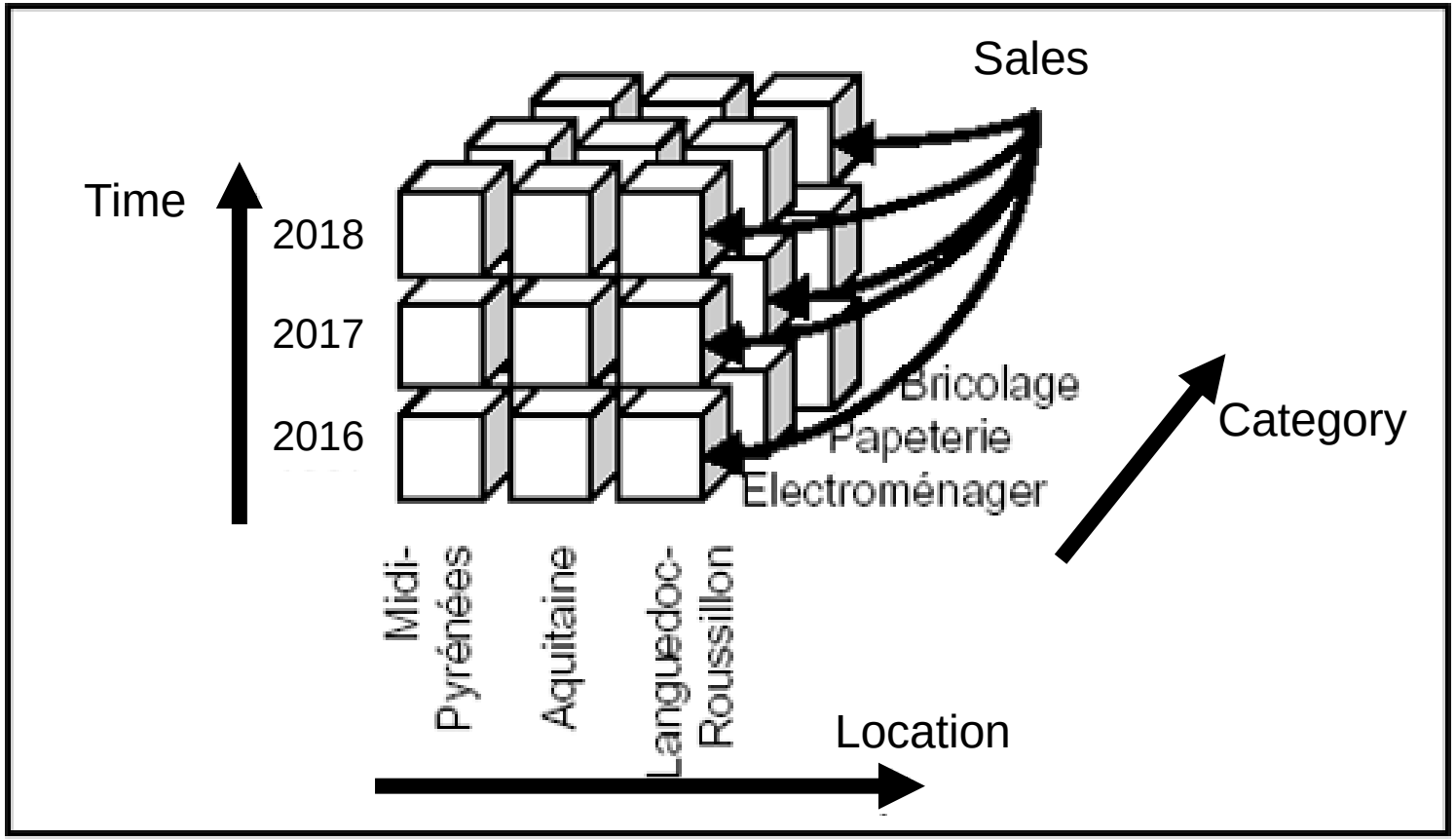




Cube operations

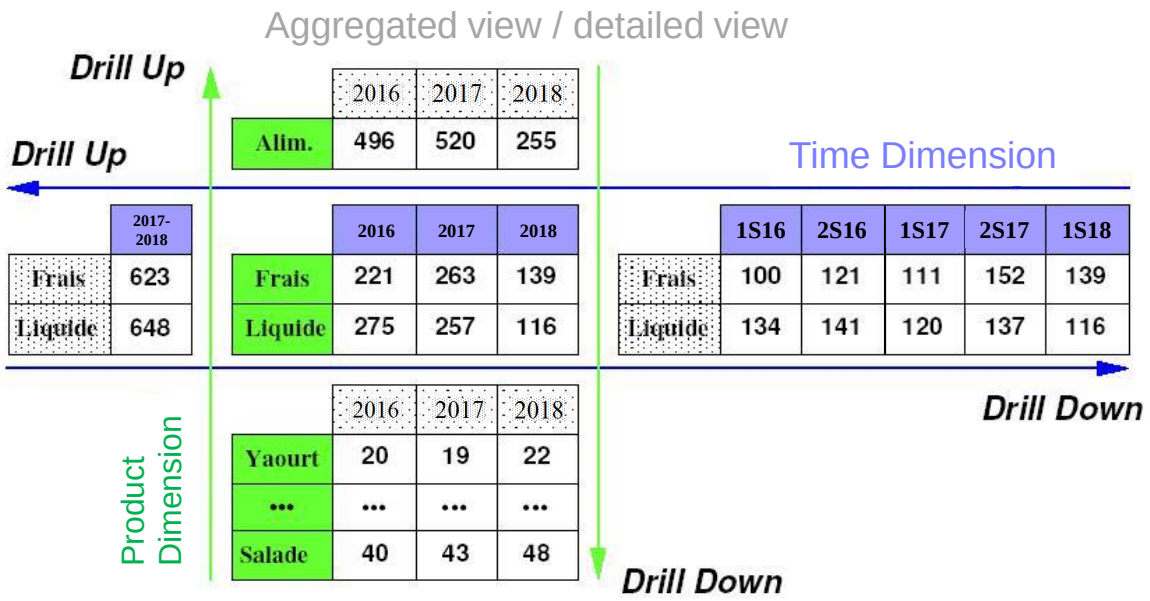
Cube operations

- Cube representation



Cube operations

- Drill
  - Drill-down
  - Roll-up
  - Drill-across

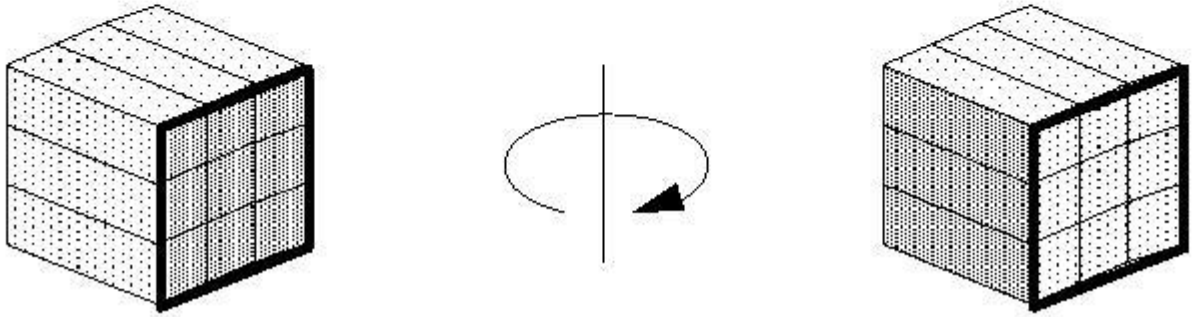


Cube operations

- Rotate

	2016	2017	2018
Frais	221	263	139
Liquide	275	257	116

	2016	2017	2018
NordPdC	101	120	52
IdF	395	400	203

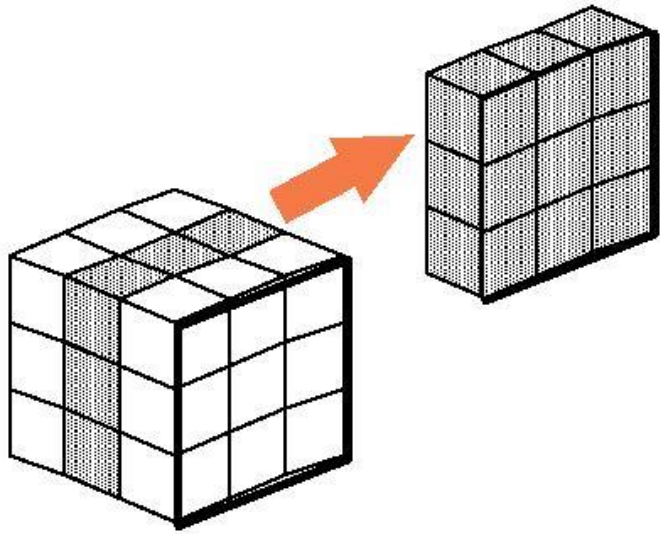


Cube operations

- Slice

		2016	2017	2018
Frais	IdF	220	265	284
	Province	225	245	240
Liquide	IdF	163	152	145
	Province	187	174	184

		2016
Frais	IdF	265
	Province	245
Liquide	IdF	152
	Province	174

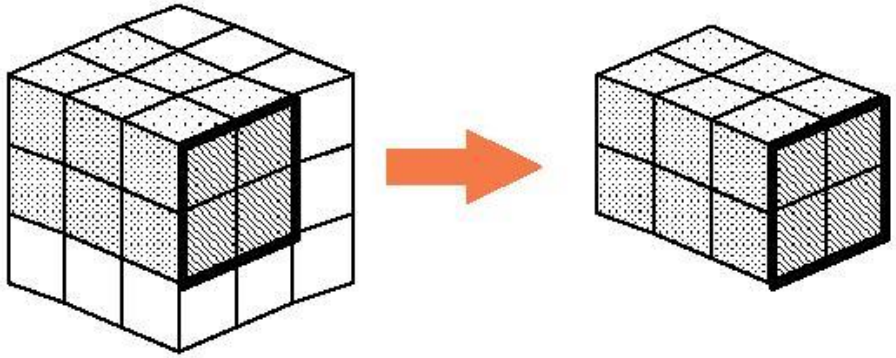


Cube operations

- Dice/Scope

		2016	2017	2018
Frais	IdF	220	265	284
	Province	225	245	240
Liquide	IdF	163	152	145
	Province	187	174	184

		2016	2017
Frais	IdF	220	265
	Province	225	245



How to keep track of history ?

Facts

- Transaction
- Periodic snapshot
- Accumulating snapshot
  - Process, pipeline

SCD : Slowly Changing Dimension

- 3 main types
  - SCD1 : update
  - SCD 2 : add new row (versioning)
  - SCD 3 : original/previous and current versions
- 5 other types have been defined to address issues such as:
  - SCD2 is growing too fast
  - Provide historical and current versions of data

SCD1 : Update

Original row in Product dimension:

Product Key	SKU (NK)	Product Description	Department Name
12345	ABC922-Z	IntelliKidz	Education

Updated row in Product dimension:

Product Key	SKU (NK)	Product Description	Department Name
12345	ABC922-Z	IntelliKidz	Strategy

SCD2 : Add new row (Versioning)

Original row in Product dimension:

Product Key	SKU (NK)	Product Description	Department Name	...	Row Effective Date	Row Expiration Date	Current Row Indicator
12345	ABC922-Z	IntelliKidz	Education	...	2012-01-01	9999-12-31	Current

Rows in Product dimension following department reassignment:

Product Key	SKU (NK)	Product Description	Department Name	...	Row Effective Date	Row Expiration Date	Current Row Indicator
12345	ABC922-Z	IntelliKidz	Education	...	2012-01-01	2013-01-31	Expired
25984	ABC922-Z	IntelliKidz	Strategy	...	2013-02-01	9999-12-31	Current

SCD3 : Original/Previous and current versions

Original row in Product dimension:

Product Key	SKU (NK)	Product Description	Department Name
12345	ABC922-Z	IntelliKidz	Education

Updated row in Product dimension:

Product Key	SKU (NK)	Product Description	Department Name	Prior Department Name
12345	ABC922-Z	IntelliKidz	Strategy	Education

SCD : Summary

SCD Type	Dimension Table Action	Impact on Fact Analysis
Type 0	No change to attribute value	Facts associated with attribute's original value
Type 1	Overwrite attribute value	Facts associated with attribute's current value
Type 2	Add new dimension row for profile with new attribute value	Facts associated with attribute value in effect when fact occurred
Type 3	Add new column to preserve attribute's current and prior values	Facts associated with both current and prior attribute alternative values
Type 4	Add mini-dimension table containing rapidly changing attributes	Facts associated with rapidly changing attributes in effect when fact occurred
Type 5	Add type 4 mini-dimension, along with overwritten type 1 mini-dimension key in base dimension	Facts associated with rapidly changing attributes in effect when fact occurred, plus current rapidly changing attribute values
Type 6	Add type 1 overwritten attributes to type 2 dimension row, and overwrite all prior dimension rows	Facts associated with attribute value in effect when fact occurred, plus current values
Type 7	Add type 2 dimension row with new attribute value, plus view limited to current rows and/or attribute values	Facts associated with attribute value in effect when fact occurred, plus current values

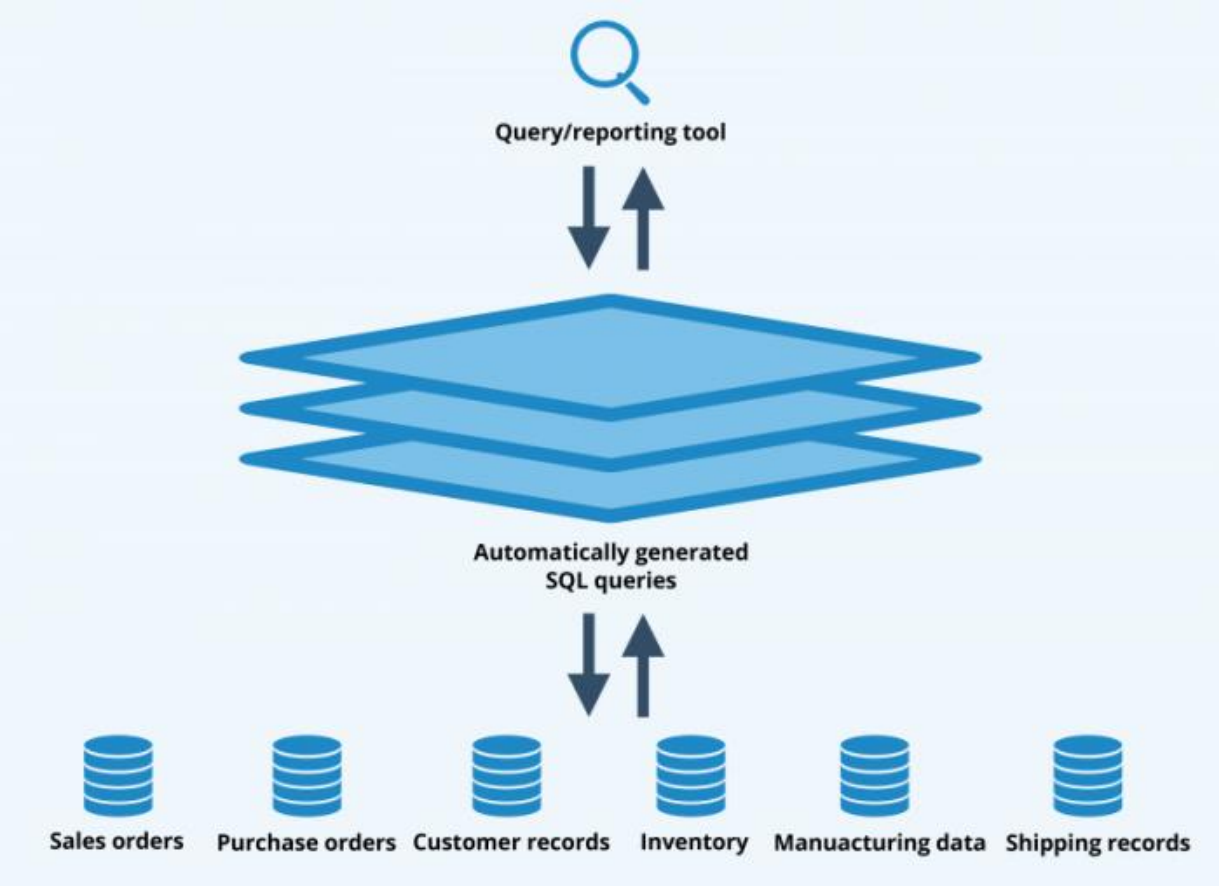
Source: www.kimballgroup.com

What about the Business Users ?

Semantic Layer

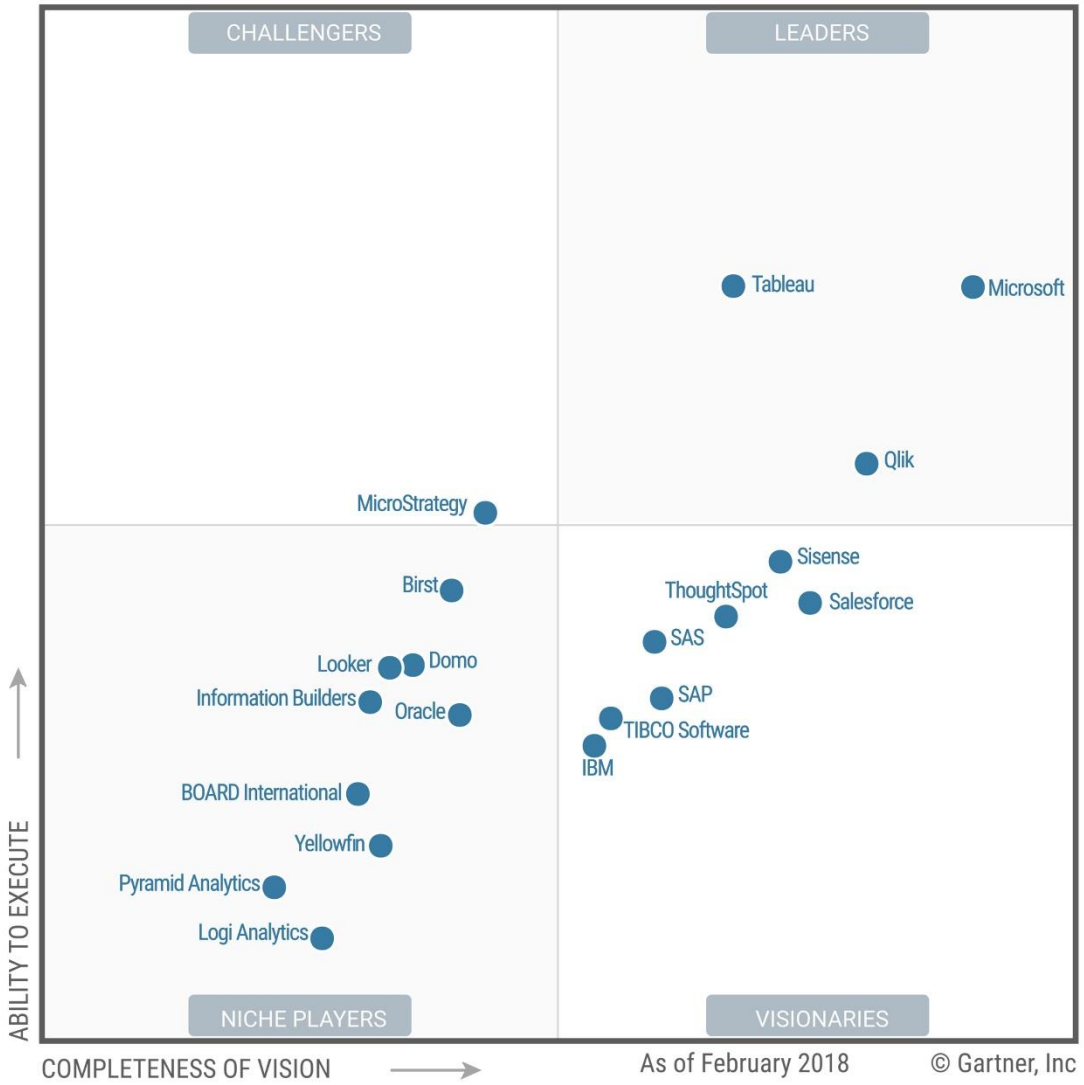
- Business representation of corporate data that helps end users access data autonomously using common business terms.
- Maps complex data into familiar business terms such as product, customer, or revenue to offer a unified, consolidated view of data across the organization

- Wikipedia



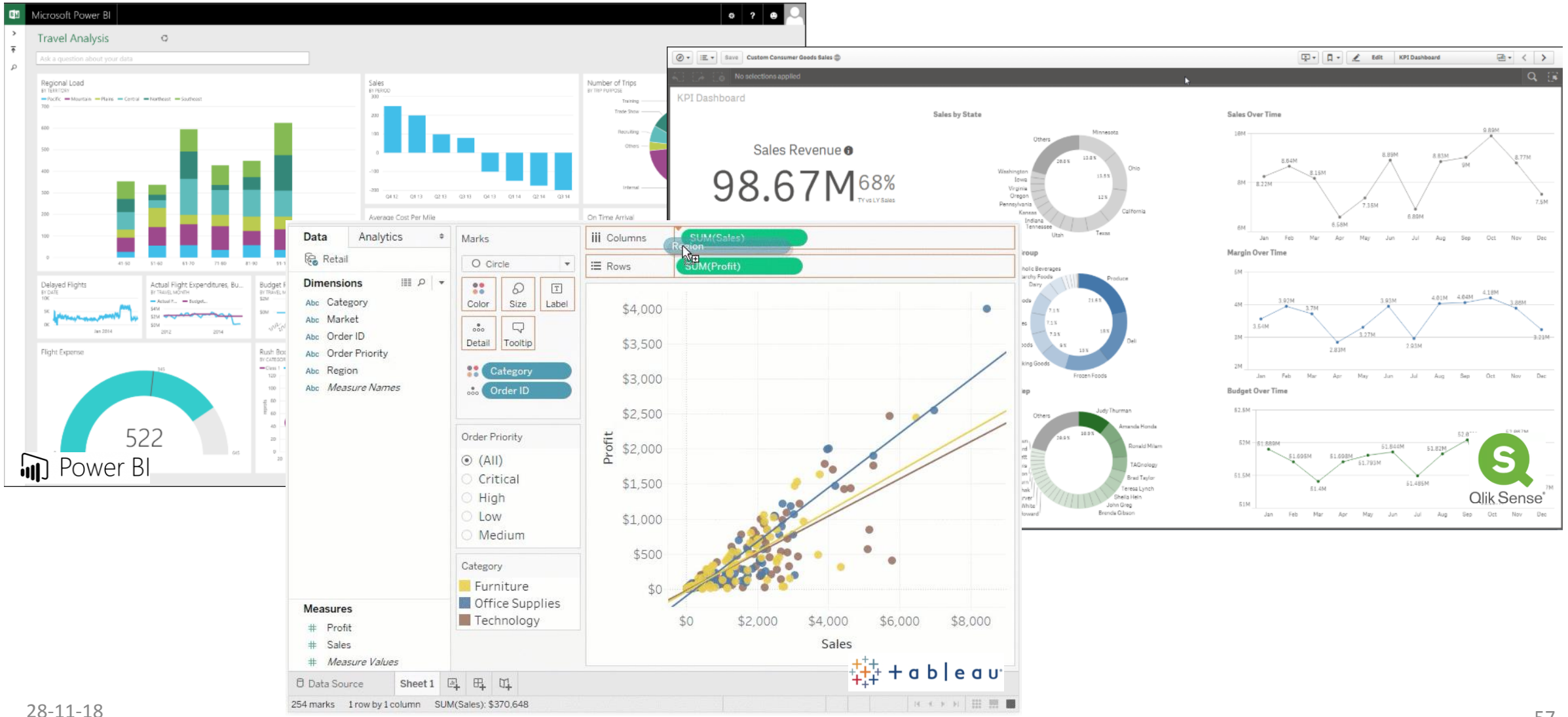
The market in 2018

Figure 1. Magic Quadrant for Analytics and Business Intelligence Platforms



Source: Gartner (February 2018)

The market in 2018: the leaders



Process & Size Estimation

Dimensional Design Process

1. Select the business process
2. Declare the grain
3. Identify the Dimensions
4. Identify the Facts

Size Estimation

Steps	Example : supermarket
Identify granularity	Retention time : 4 years Grain : daily records # Stores : 300 # Products : 200.000 references (10% sold each day)
Count expected events	4 x 365 x 300 x 20.000 = 8,76 billion rows
Count number of fields	3 for FK, 5 measures (4 bytes each)
Compute size estimation	8,76 x 10^9 x 8 x 4 bytes
Total	280 GB

Example

Retail Store

- HQ of a large grocery chain
- 100 stores across 5 states
- Each store has several departments: grocery, frozen foods, bakery, etc.
- 60.000 individual products
- Point of sale ticket example

Allstar Grocery
123 Loon Street
Green Prairie, MN 55555
(952) 555-1212

Store: 0022
Cashier: 00245409/Alan

0030503347 Baked Well Multigrain Muffins	2.50
2120201195 Diet Cola 12-pack	4.99
Saved $.50 off $5.49	
0070806048 Sparkly Toothpaste	1.99
Coupon $.30 off $2.29	
2840201912 SoySoy Milk Quart	3.19
TOTAL	12.67
AMOUNT TENDERED	
CASH	12.67
ITEM COUNT: 4	

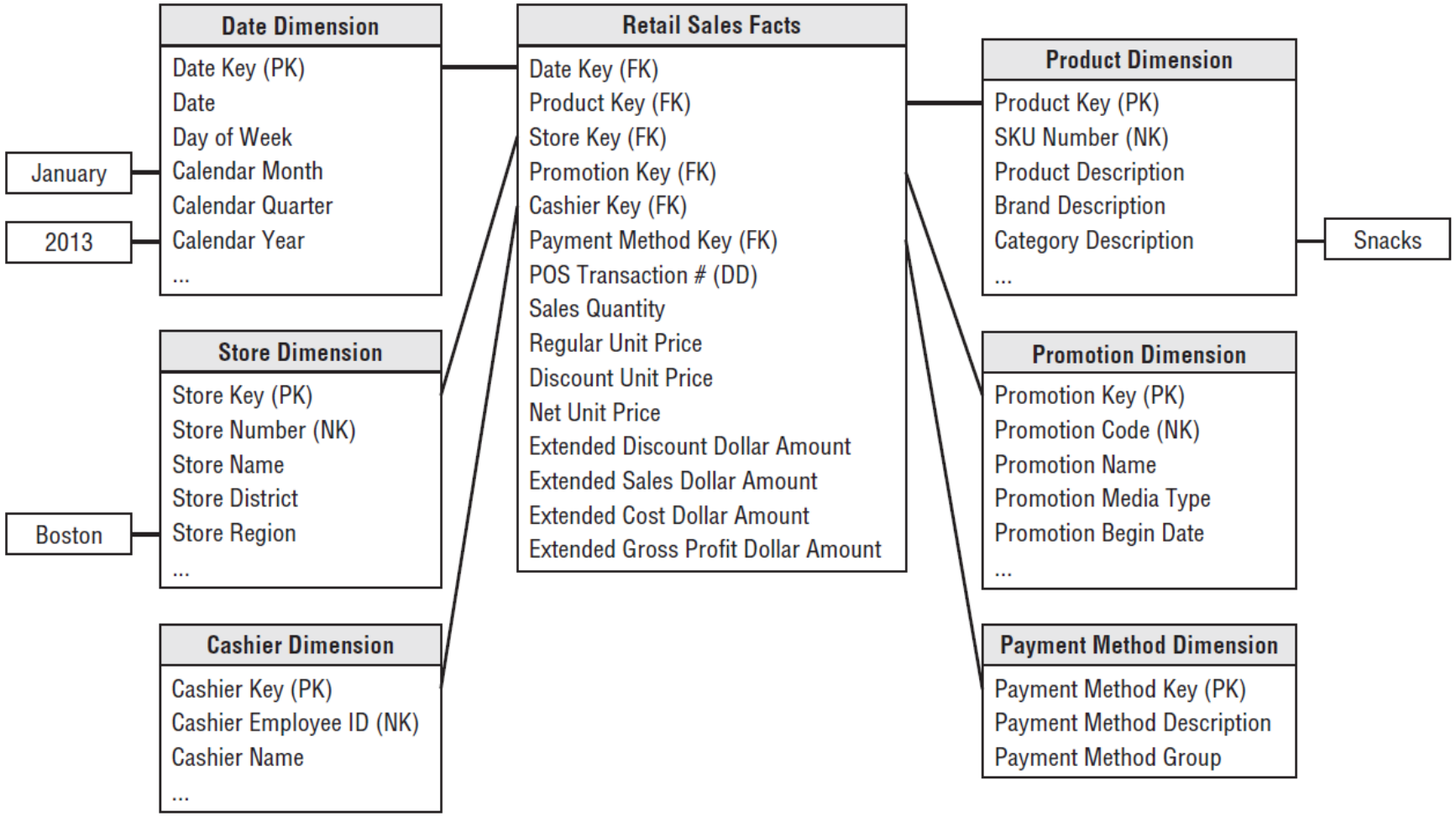
Transaction: 649 4/15/2013 10:56 AM

Thank you for shopping at Allstar

00649002204152013002454O9

Source: www.kimballgroup.com

A Data Model



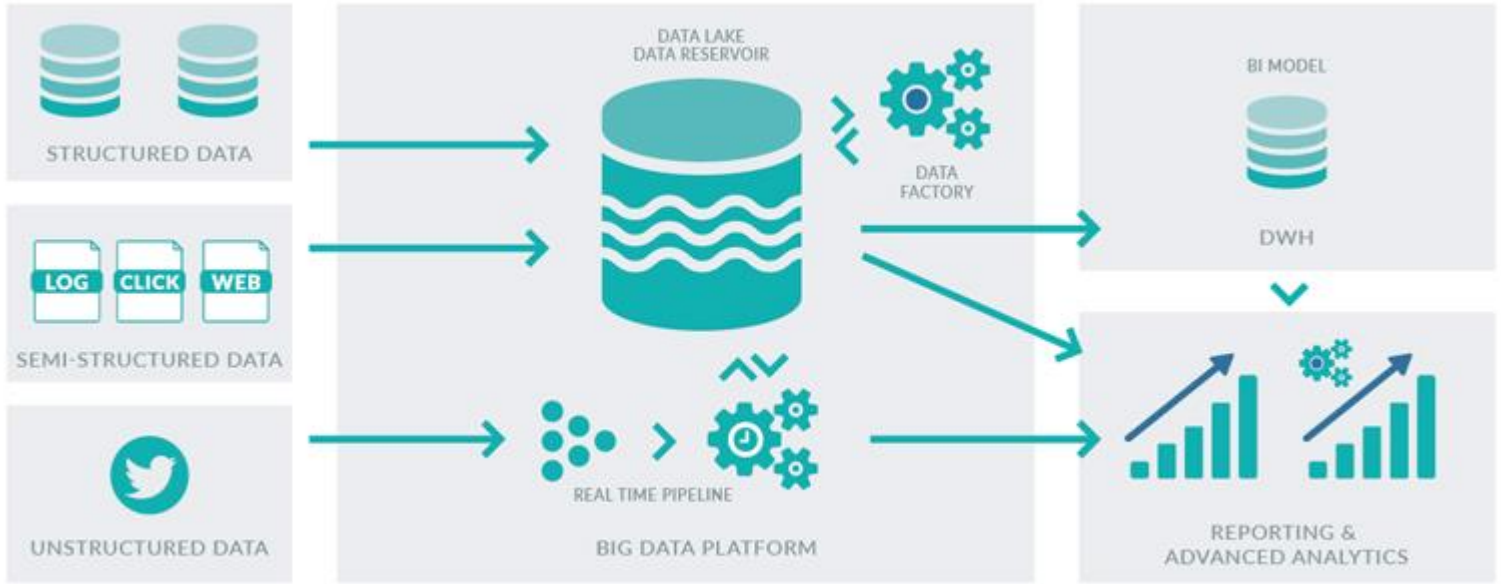
Source: www.kimballgroup.com

Business Intelligence and Big Data

Architecture overview

- From

- To



Source : www.clearpeaks.com

Business Intelligence and Big Data

	Big Data	BI
Types of analysis	Ad hoc (on read)	Prepared (on write)
Types of data	Structured, non- and semi-structured	Structured
Types of data sources	Operational, internal and external, all types of data → processing challenge	Operational, mostly internal, structured for analysis
Amounts of data	Peta/Exabytes	Terabytes
Speed of information processing	Fast on any kind of data	Fast on structured data only
Same goals	Data collecting, analyzing and reporting for better decisions Why? Where? What? How? Achieve business goals and predict the future	

What drives an architecture choice?

Conclusion

In data we trust

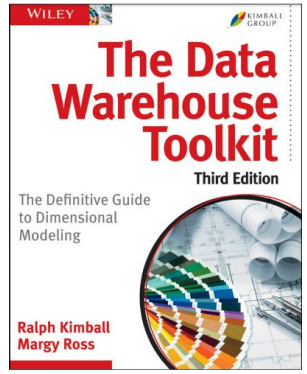
« Without data, you are just another person with an opinion »

- W. Edwards Deming

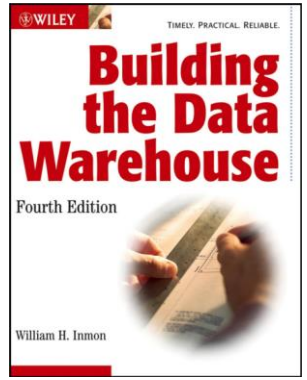
Thank you for you attention

References

- Ralph Kimball, "The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, 3rd Edition



- W. H. Inmon, «Building the Data Warehouse», ed. Wiley, 1ère édition : 1996, 4ème édition: Oct 2005, voir http://www.billinmon.com/



HERMES
A HUMAN WAY OF MANAGING SERVICES

Annexes

ETL vs ELT

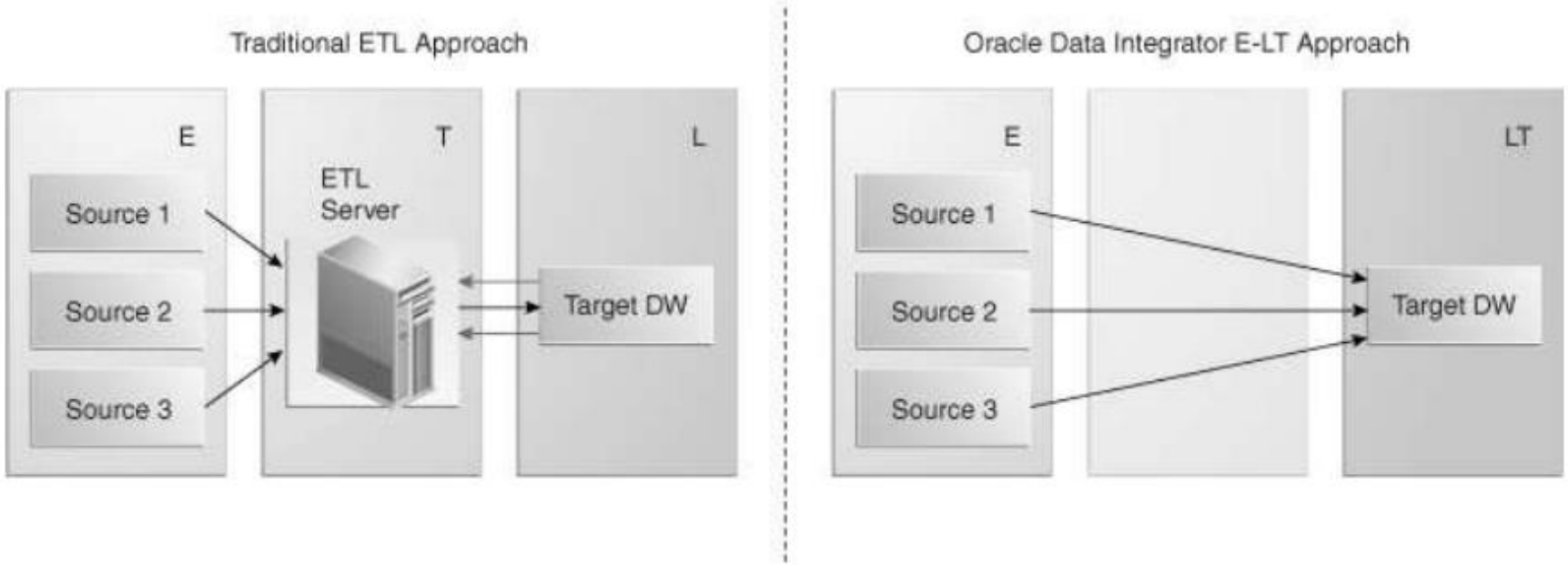


Figure 1: Traditional ETL approach compared to E-LT approach

Source: Oracle

Data mart vs Data Lake

- Data Marts
  - A data mart offers analytical capability for a restricted area of data
  - Part of a larger data warehouse system
  - Uses “schema on write” system to optimize data for analytical processing

- Data Lakes
  - Offers massive storage and agile analytics to a wide variety of users
  - Can be used with or without a data warehouse system
  - Uses “schema on read” approach to answer a business intelligence question

Data Vault

« A detail oriented, historical tracking and uniquely linked set of normalized tables that support one of more functional areas of business.

It is a hybrid approach encompassing the best of breed between 3NF and Star Schemas.

The design is flexible, scalable, consistent and adaptable to the needs of the enterprise »

- Dan Linstedt

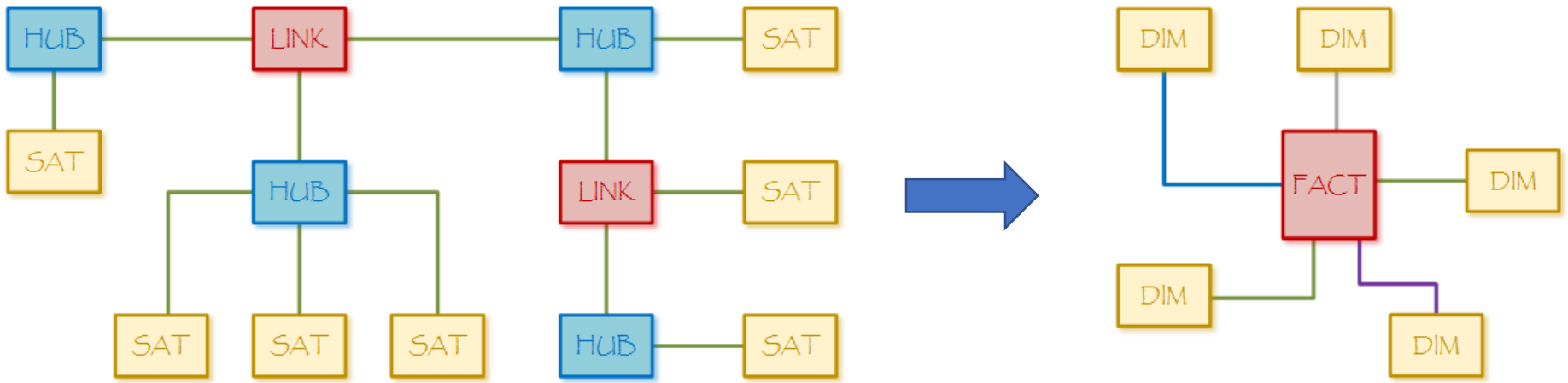
Data Vault

- Hub
  - Business entity

- Link
  - Relationship

- Satellite
  - Description

Data Vault


HUB
LINK
HUB
SAT
SAT
HUB
LINK
SAT
SAT
SAT
SAT
HUB
SAT
DIM
DIM
FACT
DIM
DIM
DIM

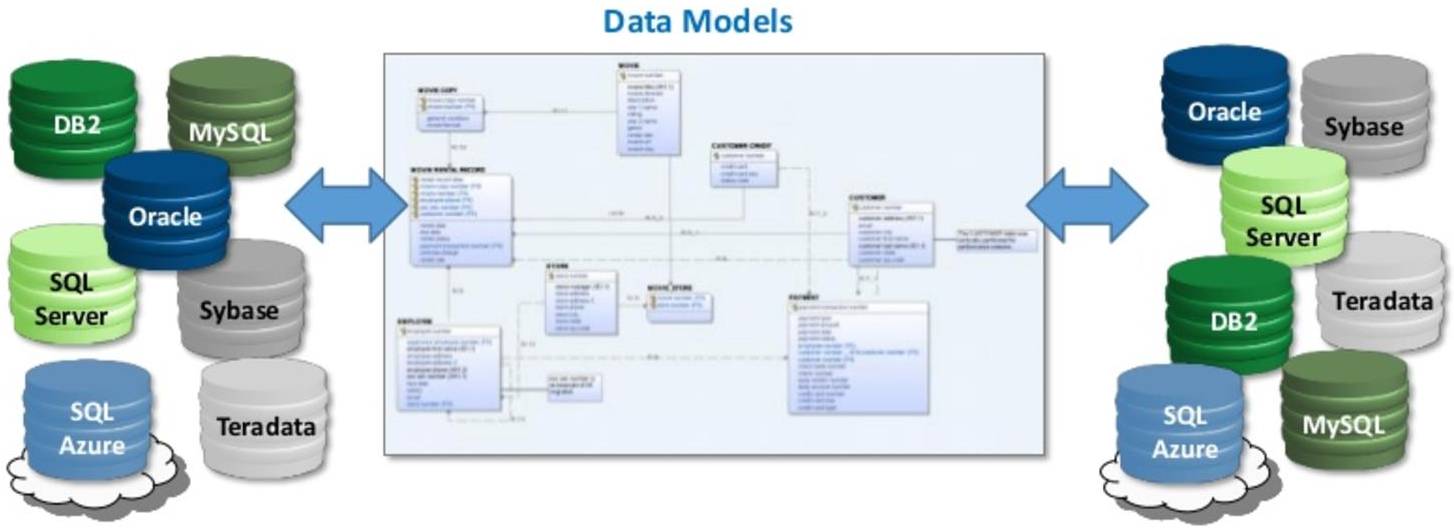
Data Vault

- Some benefits
  - Rapidly reduce business cycle time for implementing changes
  - Ease of integration: consolidate disparate data stores
  - Scale to hundreds of Terabytes or Petabytes (Big Data)
  - Near-Real-Time loads
  - Auditability: trace all data back to the source systems
  - Rapid ROI and Delivery of information to new Star Schemas

Traditional Model – “Schema on Write”

Clipper la diapositive

- With the traditional relational database paradigm, forward & reverse engineering can both create and read database structure using a graphical data model.



Big Data Model - “Schema on Read”

Clipper la diapositive

- With the Big Data and NoSQL paradigm, “Schema-on-Read” means you do not need to know how you will use your data when you are storing it.
- You do need to know how you will use your data when you are using it and model accordingly.
  - i.e. it’s not magic.
  - For example, you may first place the data on HDFS in files, then apply a table structure in Hive.
- Apache Hive provides a mechanism to project structure onto the data in Hadoop and to query that data using a SQL-like language called HiveQL (HQL).

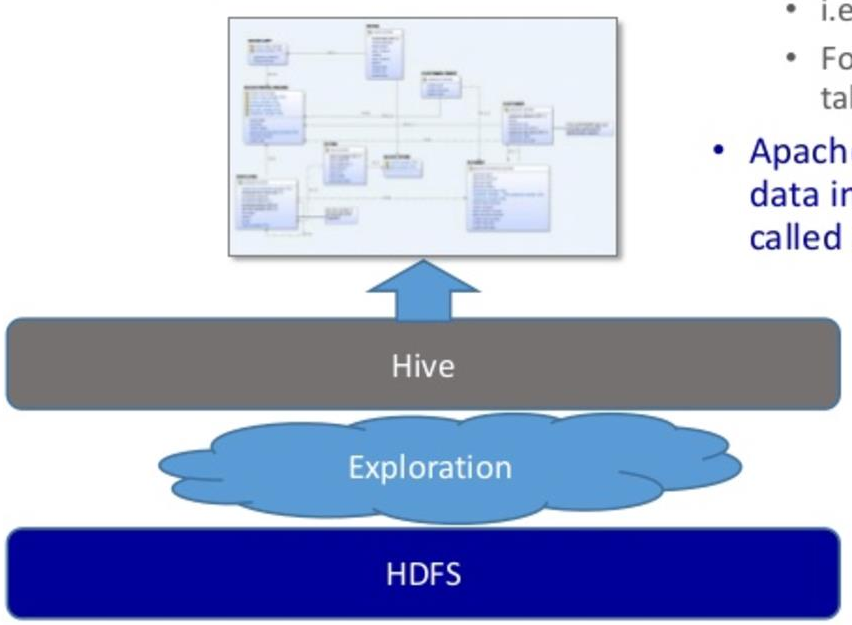


Table Structures
Create table …

Analysis
Analyze & understand the data. Build a data structure to suite your needs.

File system
hdfs dfs -put /local/path/userdump /hdfs/path/data/users