
CERN

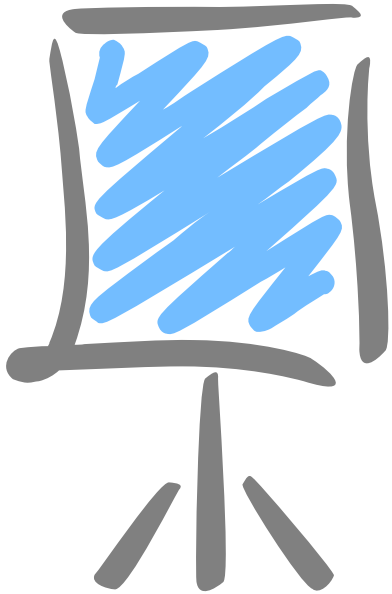
Data Warehousing and Business Intelligence

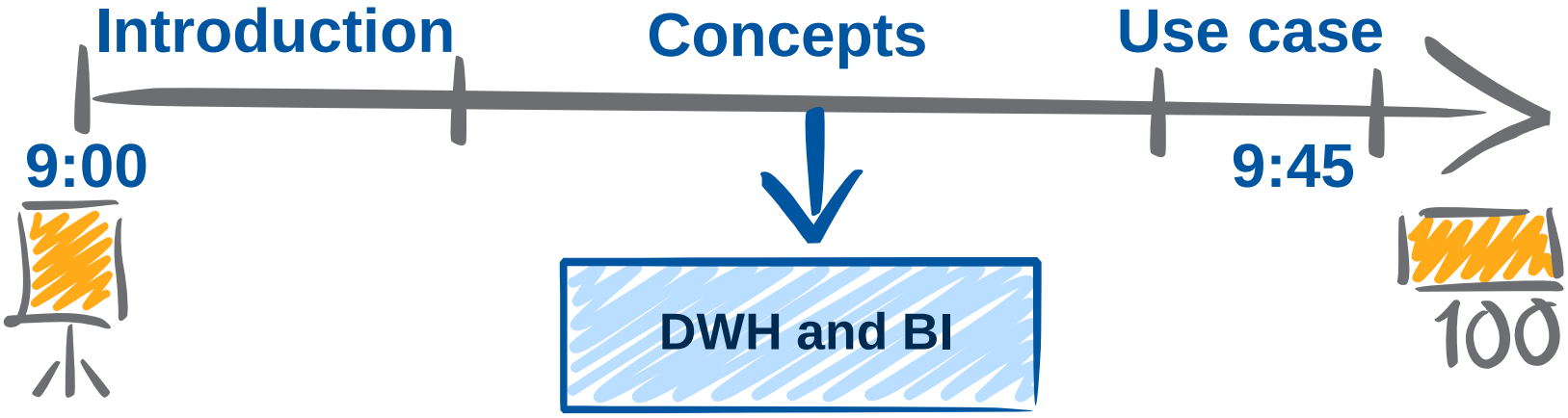
Improve strategic decision making

David Diaz Diaz CERN GS-AIS

Agenda

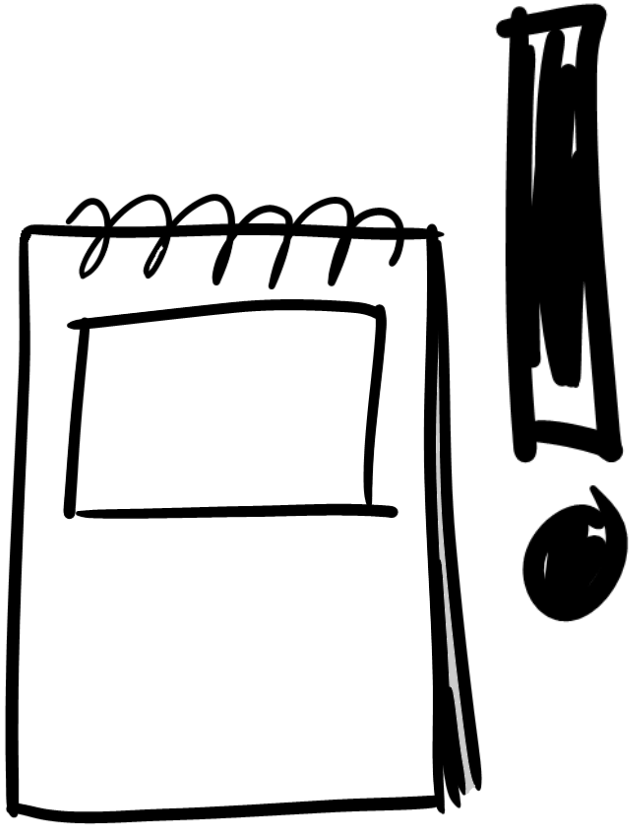
1. Introduction
2. Data Warehouse
3. Business Intelligence
4. Use case





TODO

1. Knowledge path
2. Data Warehouse
   - Dimensional model
     - Dimension tables
     - Fact table/s
   - ETL
3. Business Intelligence
   - Metadata
     - KPI
     - OLAP Cube
   - Reports and Dashboard



Introduction

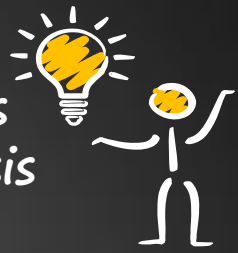
Data, Information, Knowledge, wisdom

Success

Competitive advantage

How?

The power of analysis and predictive analysis



What?

Data

Present

Past

Analyze

Forecast

Inform

Present

Future

Compare

KPI

- 90% of deliveries received on time
- 99% of payments on time

Tools

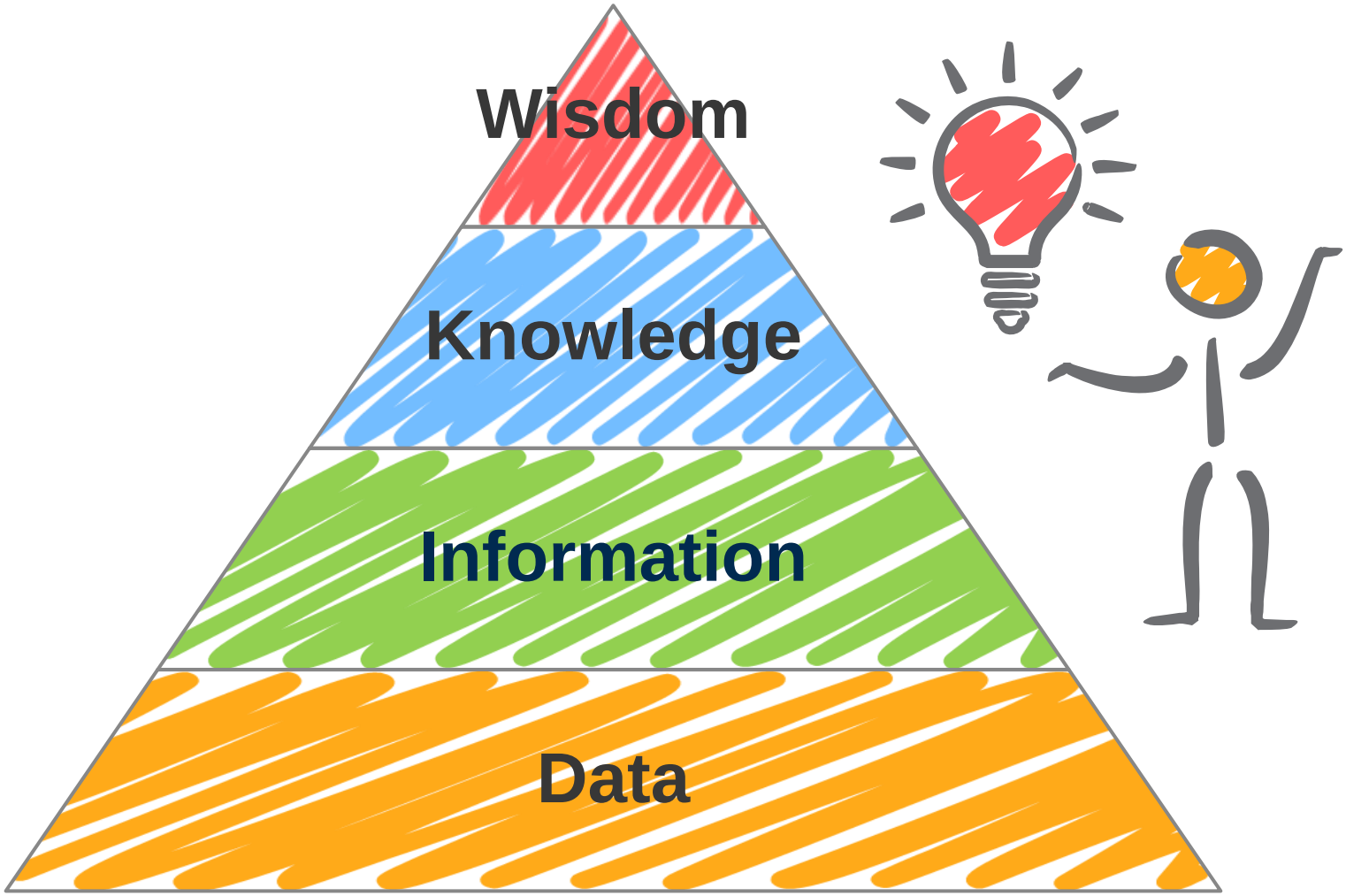
Data Warehouse

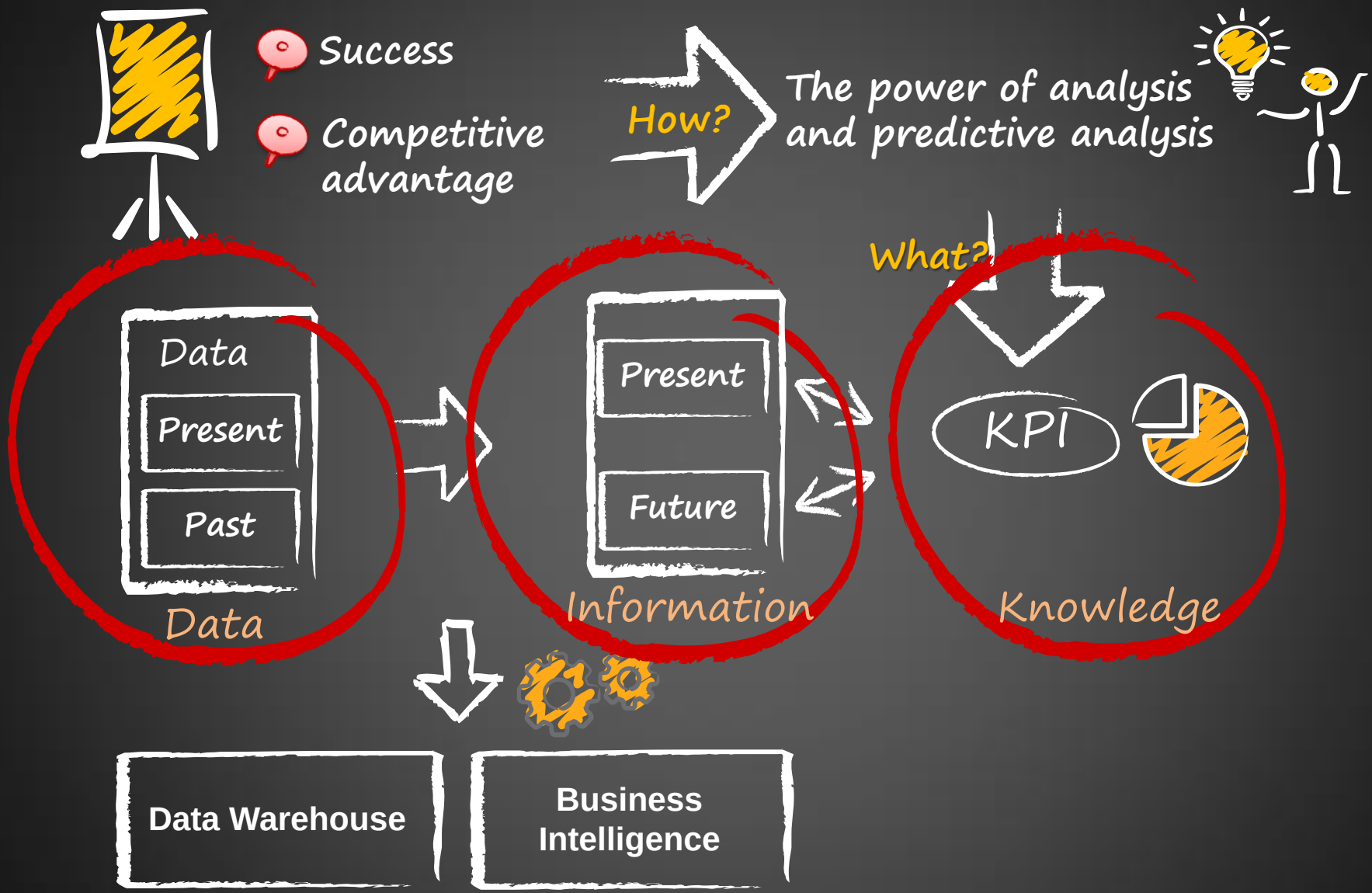
Business Intelligence

Data, Information, Knowledge

- Data
  - Items that are the most elementary descriptions of things, events, activities and transactions
  - May be internal or external
- Information
  - Organized data that has meaning and value
- Knowledge
  - Processed data or information that conveys understanding or learning applicable to a problem or activity
- Wisdom

DIKW Pyramid




Success
Competitive advantage
How?
The power of analysis and predictive analysis
What?
Data
Present
Past
Data
Present
Future
Information
KPI
Knowledge
Data Warehouse
Business Intelligence

Data, Information, Knowledge

I want to buy a new laptop to all my employees. Can I afford it? How much money is left in my budget?

Budget	Group
75000	GS-AIS
74000	GA-ASE
71000	BE-CO

Transaction	Res	Debit
21212121	Nicol	100
22222222	Regis	120
23232323	Ivica	234

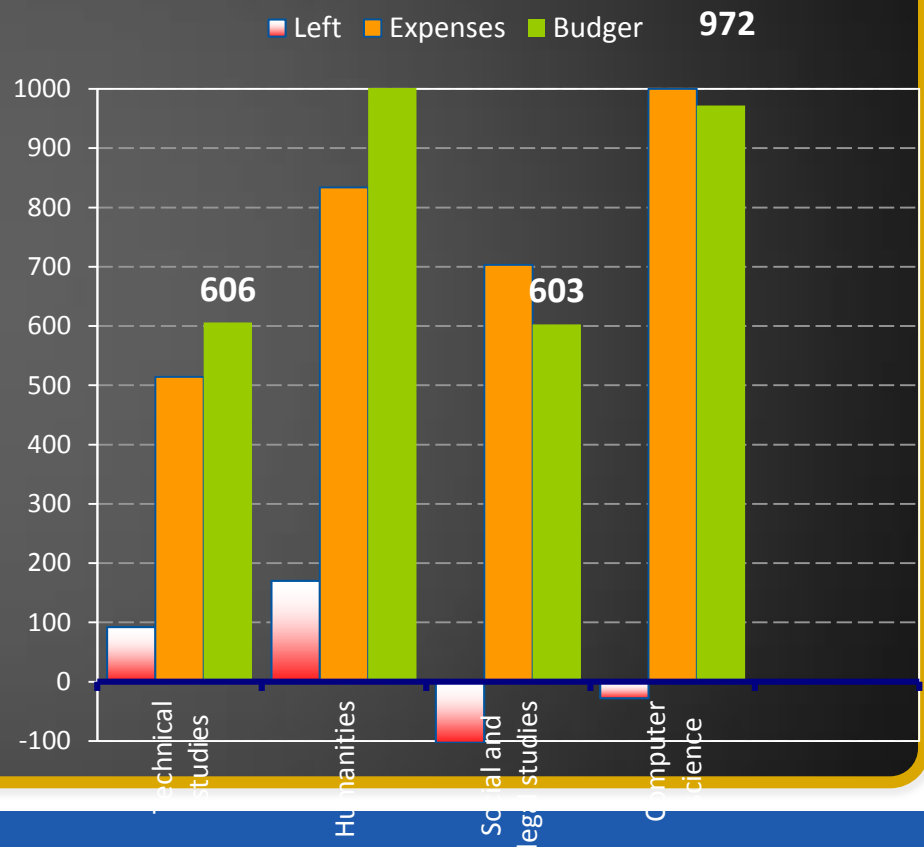
Order	Date
CA2323	23/04/13
CL2324	20/02/13
MAG23	12/04/13

Movements	Credit	Debit
21212121.1	50	100
21212121.2	60	120
21212121.3	94	234
22222222.1	35	45
22222222.2	100	50
23232323.1	32	344
23232323.2	20	78

Data, Information, Knowledge

I want to buy a new laptop to all my employees. Can I afford it? How much money is left in my budget?

Budget	Budget 2014	Expenses	Left
GS-AIS-FP	606k	514k	92k
GS-AIS-HR	1004k	834k	170k
GS-AIS-EB	603k	703k	-103k
GS-AIS-GDI	972k	1000k	-28k



Data, Information, Knowledge

- Data → Source systems
- Information → Data Warehousing
  - Consolidation and aggregation of masses of data from multiple sources into a reconciled format for reporting
- Knowledge → Business Intelligence
  - A category of applications, processes, best practices and technologies for gathering, storing, analyzing, and providing access to data to help enterprise users make better business decisions

BI and Data Warehouses

- Some companies claim you can do BI without a Data Warehouse
  - You can also use a shoe as a hammer, it works but is not very effective!

BI Needs cleaned, consistent large masses of data to work well

Solution

- Dimensional database model
  - Data Warehouse
  - Data proccess (ETL)
- Business Intelligence platform
  - Provide the most relevant indicators
- Interactive Dashboards
  - Present and visualize information using charts, pivot tables and reports
- Open data - Web services
  - Share the information between systems and everywhere

Data Warehouse

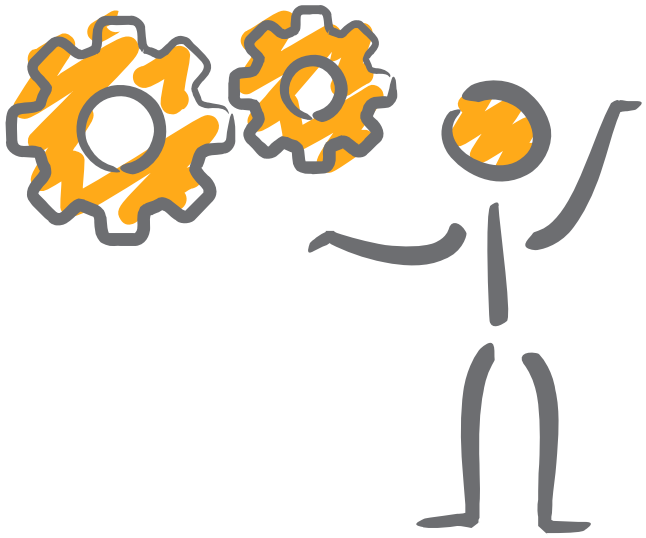
BI

KPI, Dashboards

Open data

Data Warehouse

Concepts



What is a Data Warehouse?

- Many definitions
  - DW collect relevant data from multiple different data sources, rationalize, summarize it and catalog it in a large consistent, stable, accurate, long term data stores which allows for all types of questions to be answered which otherwise would be difficult or expensive to do
  - It is the collection of key information that can be used by the business users to become more profitable
  - …
- But are these definitions sufficient ?
  - We need much more precise definition of what a data Warehouse is

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.
>
> by Bill Inmon

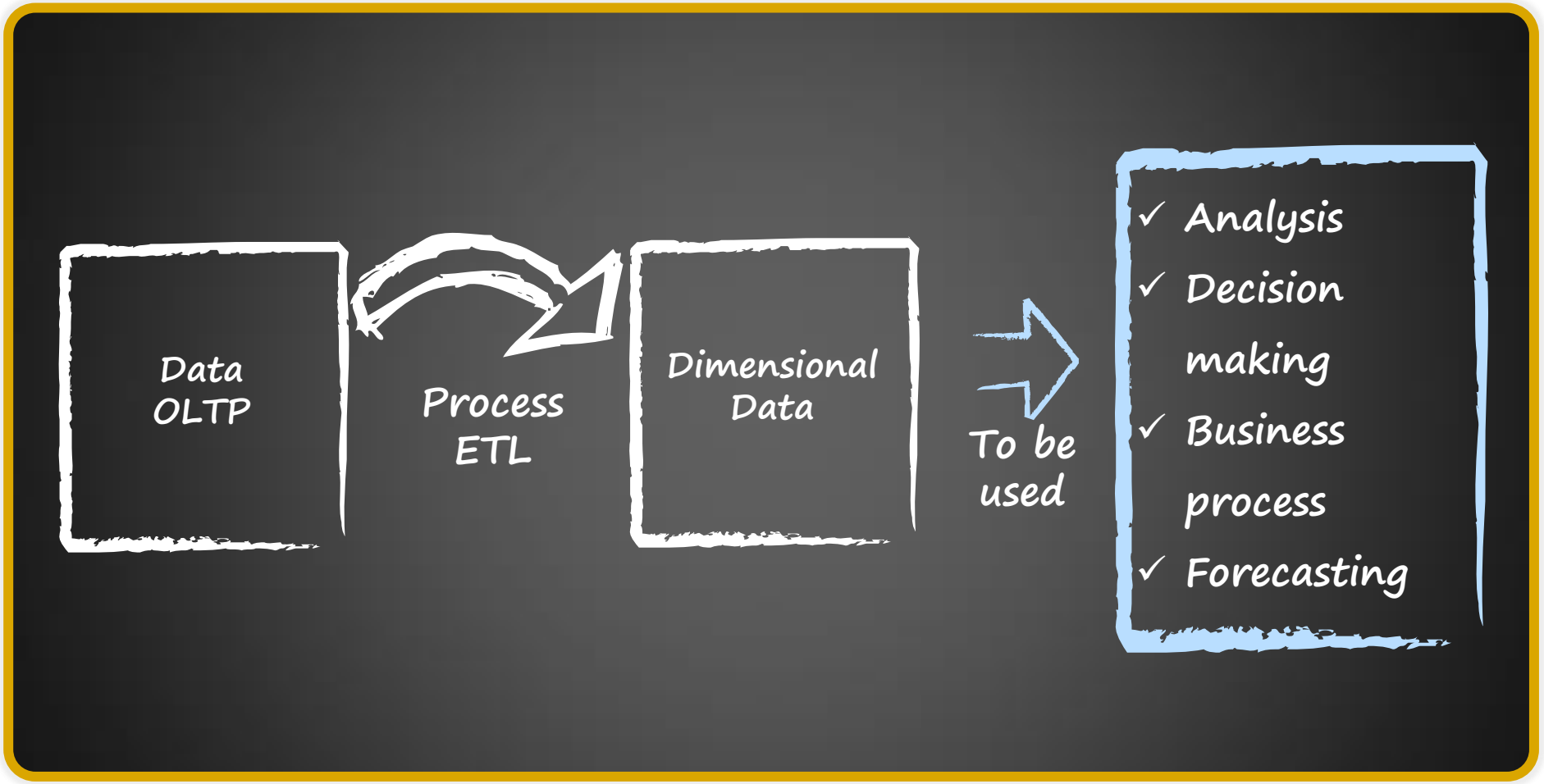
Properties of a DWH

- Subject-oriented
  - Analyze a particular subject area
    - E.g.: Finance, Human resources
- Integrated
  - Integrates data from multiple data sources
    - E.g.: Only a single way of identifying a product
- Time-Variant
  - Historical data is kept in the DWH.
    - E.g.: Retrieve data from 3 months, 6 months, 12 months, or even older
    - E.g.: It can hold all addresses associated with a customer
- Non-volatile
  - Once data is in the DWH, it will not change

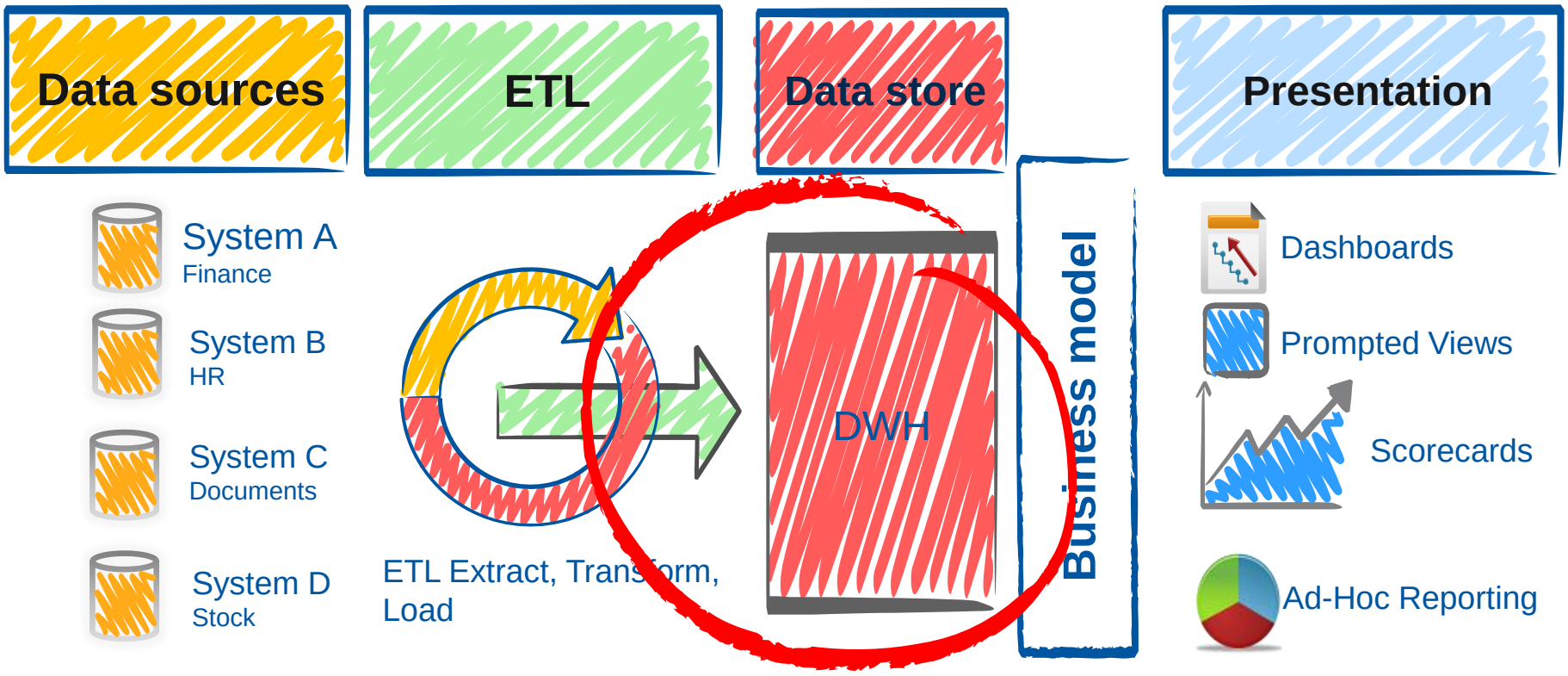
The parts of a Data Warehouse

- A Data warehouse is the
  - DATA
    - Metadata
    - Fact
    - Dimension
    - Aggregation
    - …
  - The process to make the information available
    - Load
    - Workflow
    - Query

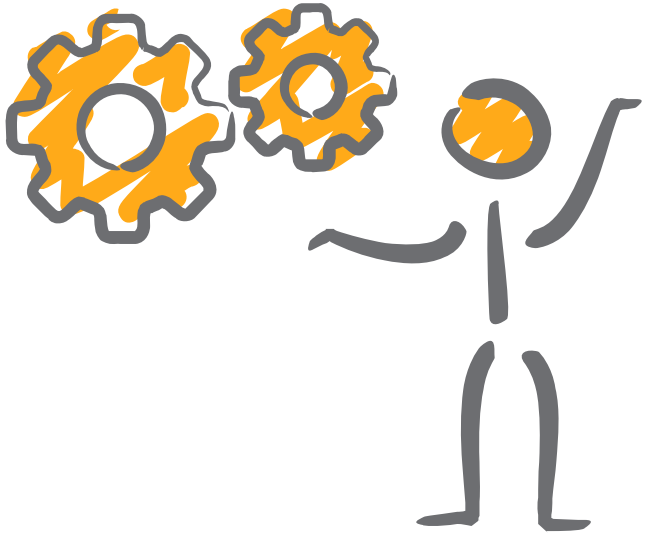
The parts of a Data Warehouse



Typical DW Architecture



Data Warehouse



The data

OLTP vs OLAP

- OLTP (On-line Transaction Processing)
  - Usually 3NF
  - Short transactions (Insert, update, delete)
- OLAP (On-line Analytical Processing)
  - Usually denormalized
  - Complex queries and aggregations
  - Historical data
  - Data Warehouse

Dimensional model

- The process and outcome of designing database schemas created to support OLAP and Data Warehousing solutions
- Schema designed to process large, complex queries
- Is composed by
  1. One or multiple fact tables
  2. A set of dimension tables

Dimensional model

	Relational Modeling	Dimensional Modeling
Data is stored in	RDBMS	RDBMS or multidimensional db
The unit of storage are	Tables	Tables or Cubes
Normalization	Data is normalized Optimized for OLTP	Data is de-normalized Optimized for OLAP
Composed by	Several tables Chains of relationships among them	Few tables and fact tables are connected to dimensional tables
Volatility/Time variant	Volatile / Variant	Non volatile / Invariant
Granularity	Detailed level	Aggregates
Query Language	SQL	MDX
Reporting	Normal Reports	User friendly, interactive, drag and drop multidimensional reports

by Divya Manduva

Dimensions (By Keyword)

Total charged in 2014 and 2013 by currency and account

- The time independent, textual and descriptive attributes by which users describe objects
- Two purposes
  - Selection of Data
  - Grouping data at a desired detail
    - Often the “by” word in a query or report

Dimensions

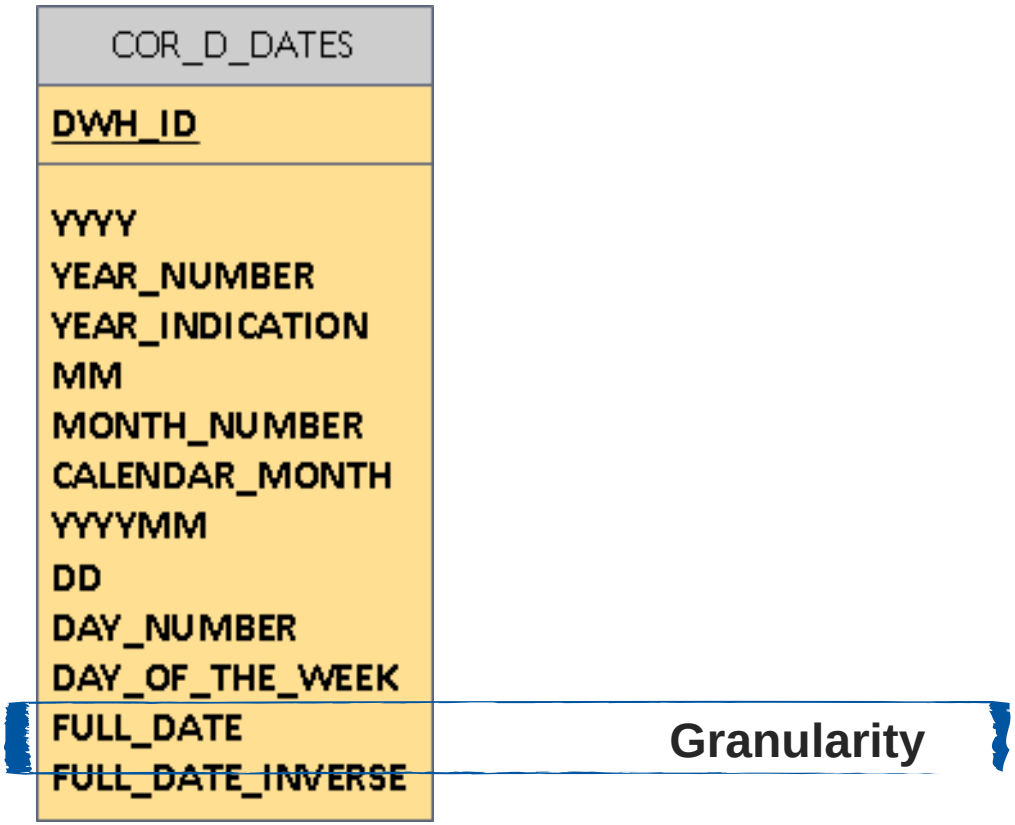
- Often employ surrogate keys
- Not uncommon to have many, many columns
- Denormalized, to reduce joins
- Granularity
  - The maximum level of detail
- E.g.:
  - Date dimension (Trimester, month, day) and sometimes time dimension (hour, min, sec)
  - Geography dimension (Country, city, postal code)
  - Organization dimension (Department, group, section)

Attributes

- Attributes are individual values that make up dimensions
  - A Time dimension may have a Month attribute, a Year attribute, and so forth
  - A Geography dimension may have a Country attribute, a Region attribute, a City attribute, and so on

Dimensions

Wide



Dimensions

DWH_ID	ORG	DEP	DES_DEPT	CGROUP	DESC_G	SECTION	DES_EN	DES_FR
4279	CERN	GS	GENERAL INFRASTRUCTURE SERVICES	GS-AIS	AD. INFORMATION S.	GS-AIS-EB	E Business	E Business
4281	CERN	GS	GENERAL INFRASTRUCTURE SERVICES	GS-AIS	AD. INFORMATION S.	GS-AIS-HR	HR Information Systems	Systemes d'Information RH
4283	CERN	GS	GENERAL INFRASTRUCTURE SERVICES	GS-AIS	AD. INFORMATION S.	GS-AIS-PM	Projects Manag. & Resource	Gestion De Projets & Planification des
4390	CERN	GS	GENERAL INFRASTRUCTURE SERVICES	GS-AIS	AD. INFORMATION S.	GS-AIS-FP	Finance & Purchasing	Finances & Achats
4391	CERN	GS	GENERAL INFRASTRUCTURE SERVICES	GS-AIS	AD. INFORMATION S.	GS-AIS-GDI	General Developmen t & Infrastr.	Developpement General & Infrastr.
-1	Na	Na	Unknown	Na	Unknown	Na	Unknown	Na
-2	Err	Err	Error	Err	Error	Err	Error	Inconnu

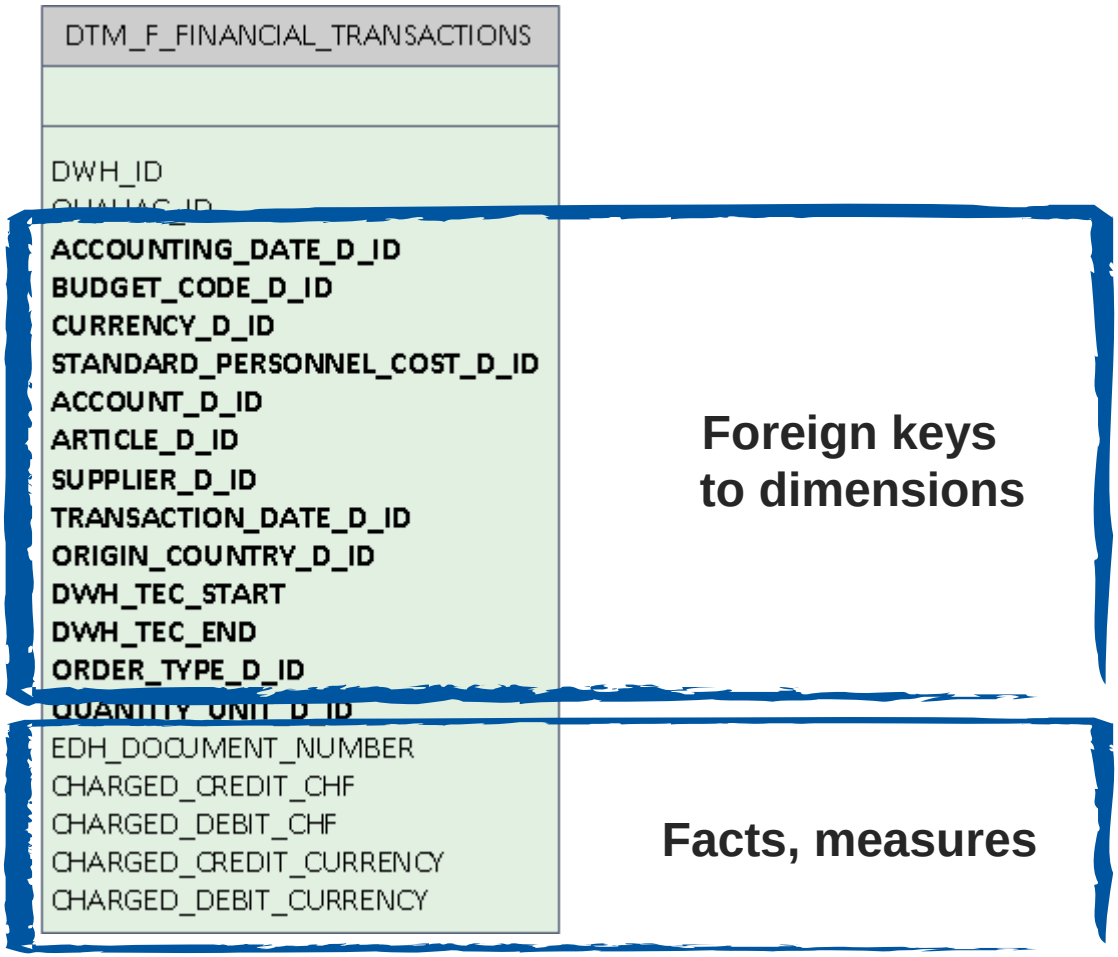
Fact (What keyword)

Total charged in 2014 and 2013 by currency and account

- Facts are the objects that represent the subject of the desired analyses
- Contains numeric, additive fields
  - Measurements of the business
- It has two types of columns
  - Contains measures
  - Foreign keys to dimension tables
- Typically narrow tables, but often very large

Example of Fact

Deep and narrow



Example Fact and dimensions

Total charged in 2014 and 2013 by currency and account

YEAR_NUMBER	CURRENCY	ACCOUNT_NAME	CHARGED
2013	CHF	6063300	21816817.61
2013	EUR	6063300	74450188.95
2013	GBP	6063300	2233807.6
2014	CHF	6063300	142347.36
2014	EUR	6063300	364598.82
2014	GBP	6063300	9021.03

- YR, ACCOUNT CODE and UNI are dimensions

Fact query example

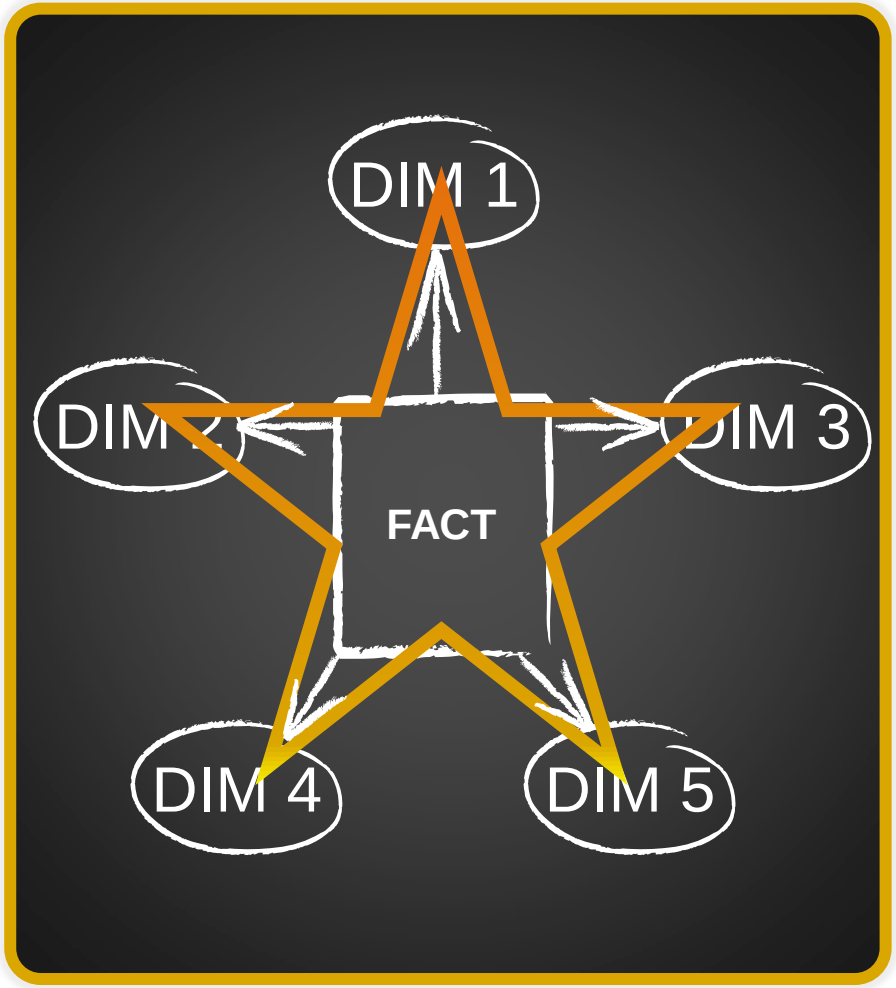
```
select dim_dates.year_number,
       dim_cur.currency_code,
       dim_accounts.account_name,
       SUM(charged_debit_currency - charged_credit_currency) charged
from dtm_f_financial_transactions fact,
       cor_d_dates dim_dates,
       cor_d_accounts dim_accounts,
       cor_d_currencies dim_cur
where fact.account_d_id = dim_accounts.dwh_id and
       fact.accounting_date_d_id = dim_dates.dwh_id and
       fact.currency_d_id = dim_cur.dwh_id
       dim_dates.year_number in (2013,2014)
group by dim_dates.year_number,
        dim_cur.currency_code,
        dim_accounts.account_name
```

Data Warehouse Schemas

- Star schema
  - Fact is in the middle of dimension tables
- Snowflake schema
  - Fact is in the middle of dimension normalized into set of smaller dimension tables.
- Fact constellations / Galaxy schema
  - Fact tables share dimension tables

Star schema

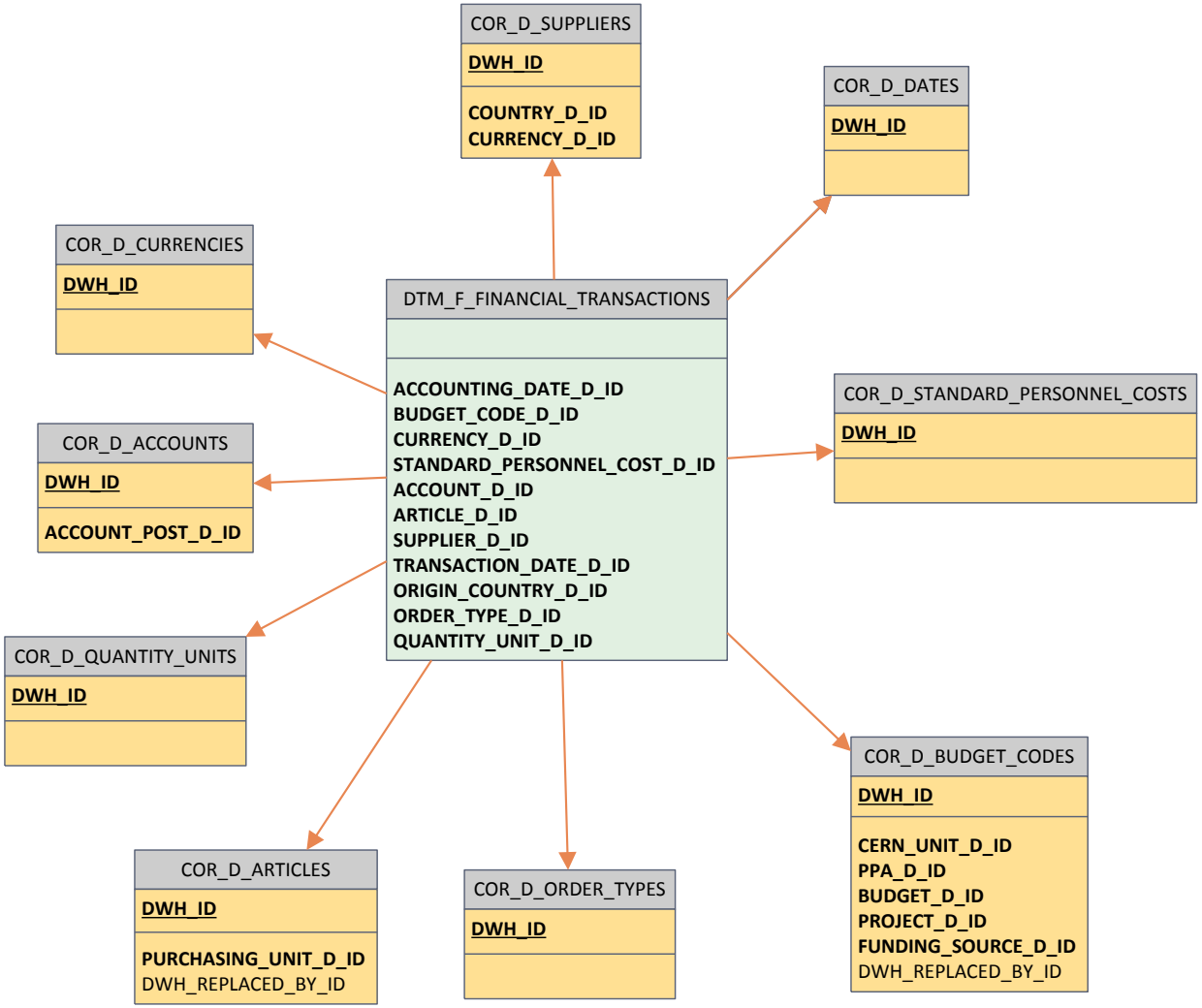
- A single, large and central table (Fact) sorrounded by multiple descriptive tables (Dimension)
  - Every fact record points to one tuple in each of the dimensions and has additional attributes



Star schema, benefits

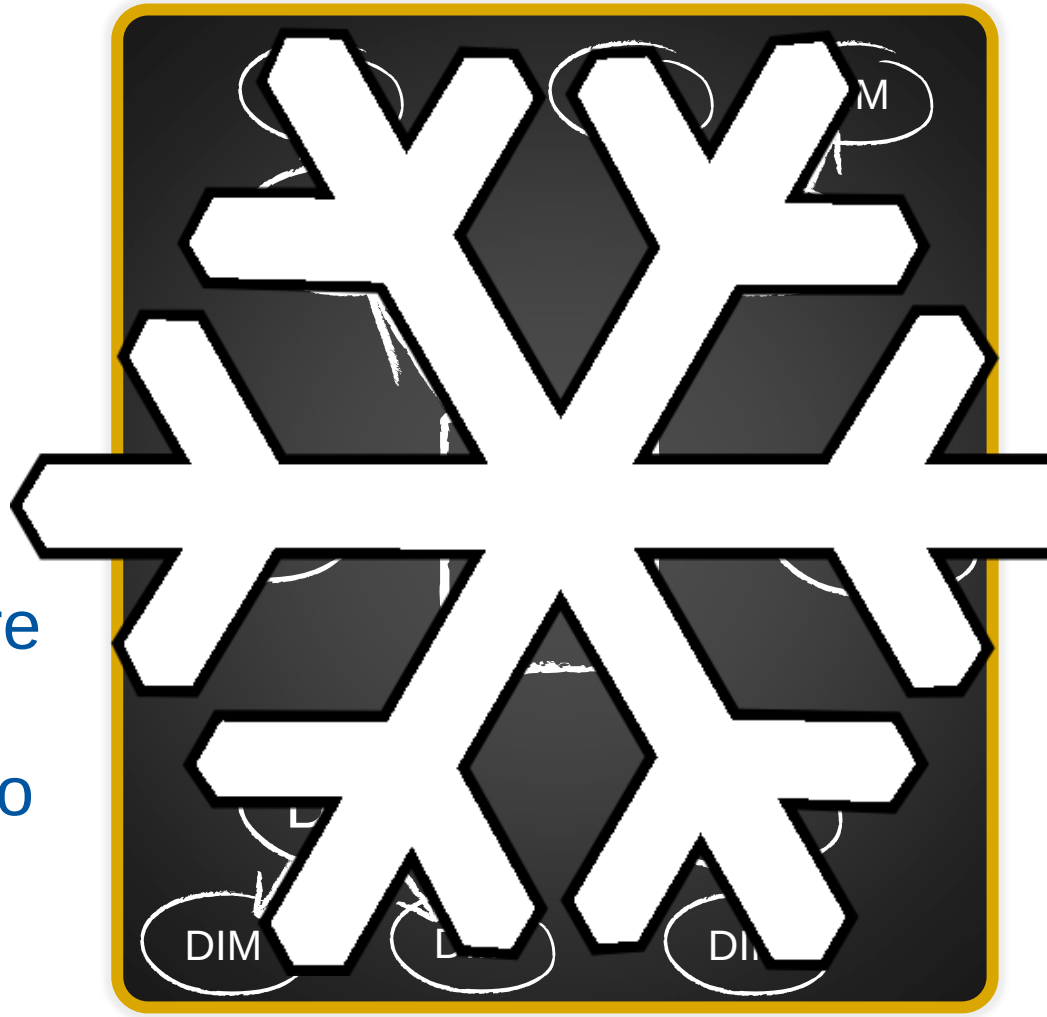
- Usability/simplicity
  - Easy to read, understand, …
  - Easy to do calculations, create hierarchies, …
- Performance
  - Reduce number of physical joins
  - Integer relantionships

Star schema, example



SnowFlake Schema

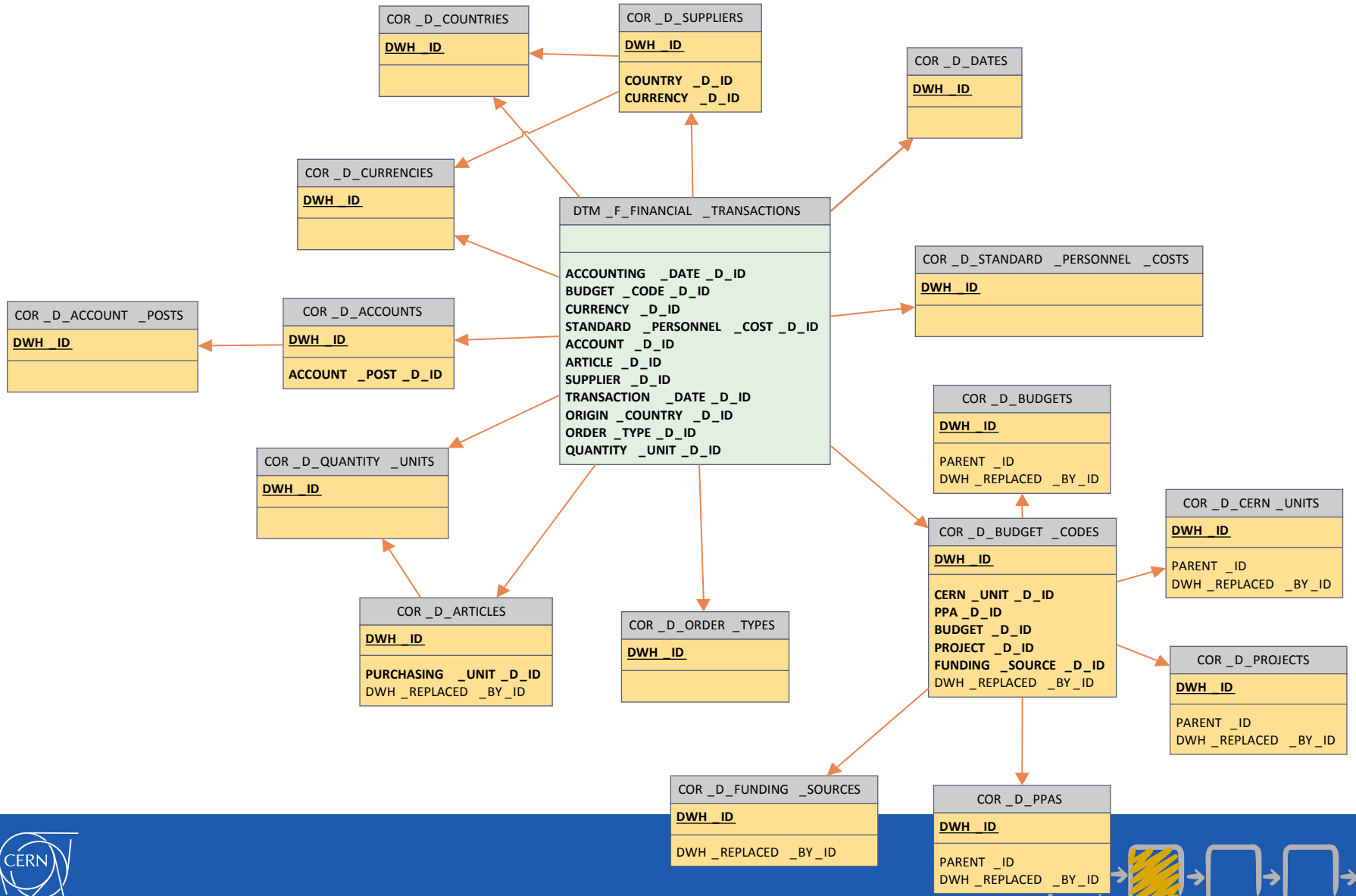
- Variant of the star schema
- A single, large and central fact table
- One or more tables for each dimension
  - Dimension tables are normalized
  - They are splitted into additional tables



SnowFlake Schema

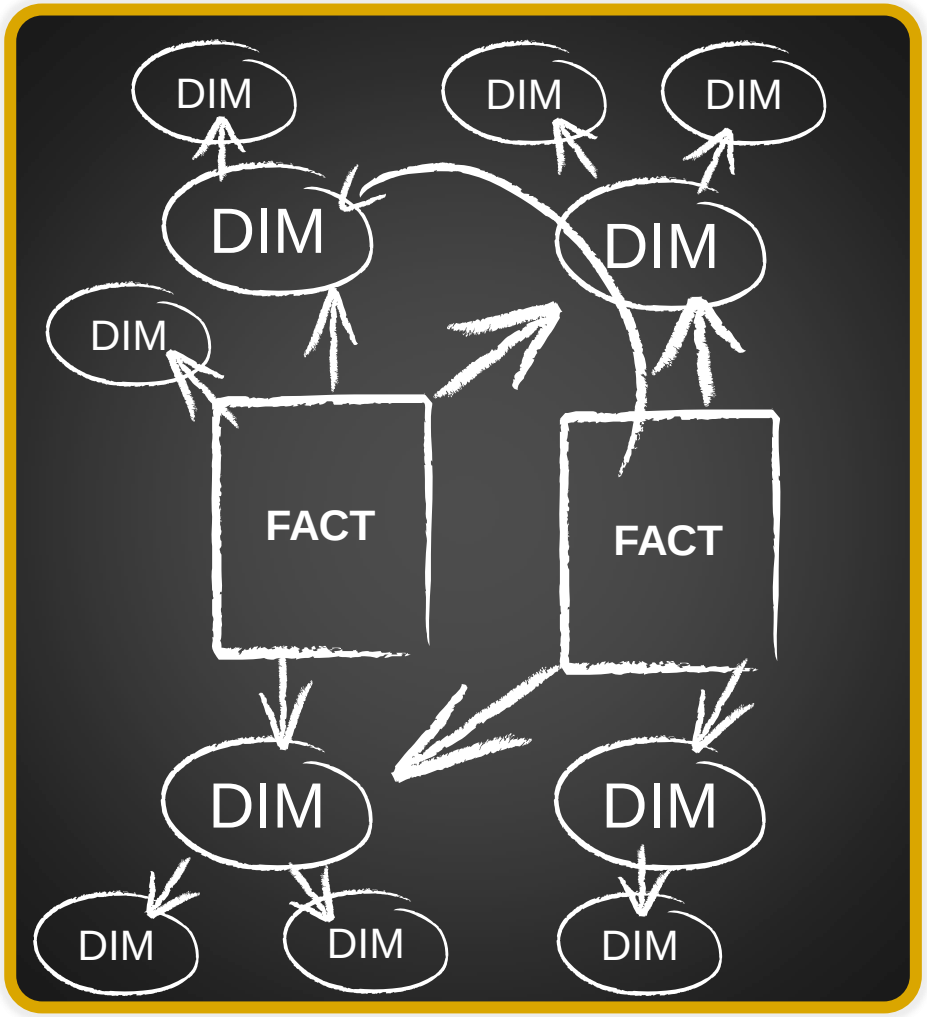
- Benefits
  - Tables are normalized
    - Avoid data redundancy
    - Save space
  - Easy to maintain
  - More flexible
- Drawbacks
  - Time consuming joins, more lookups
  - Queries are much more difficult to create

SnowFlake Schema

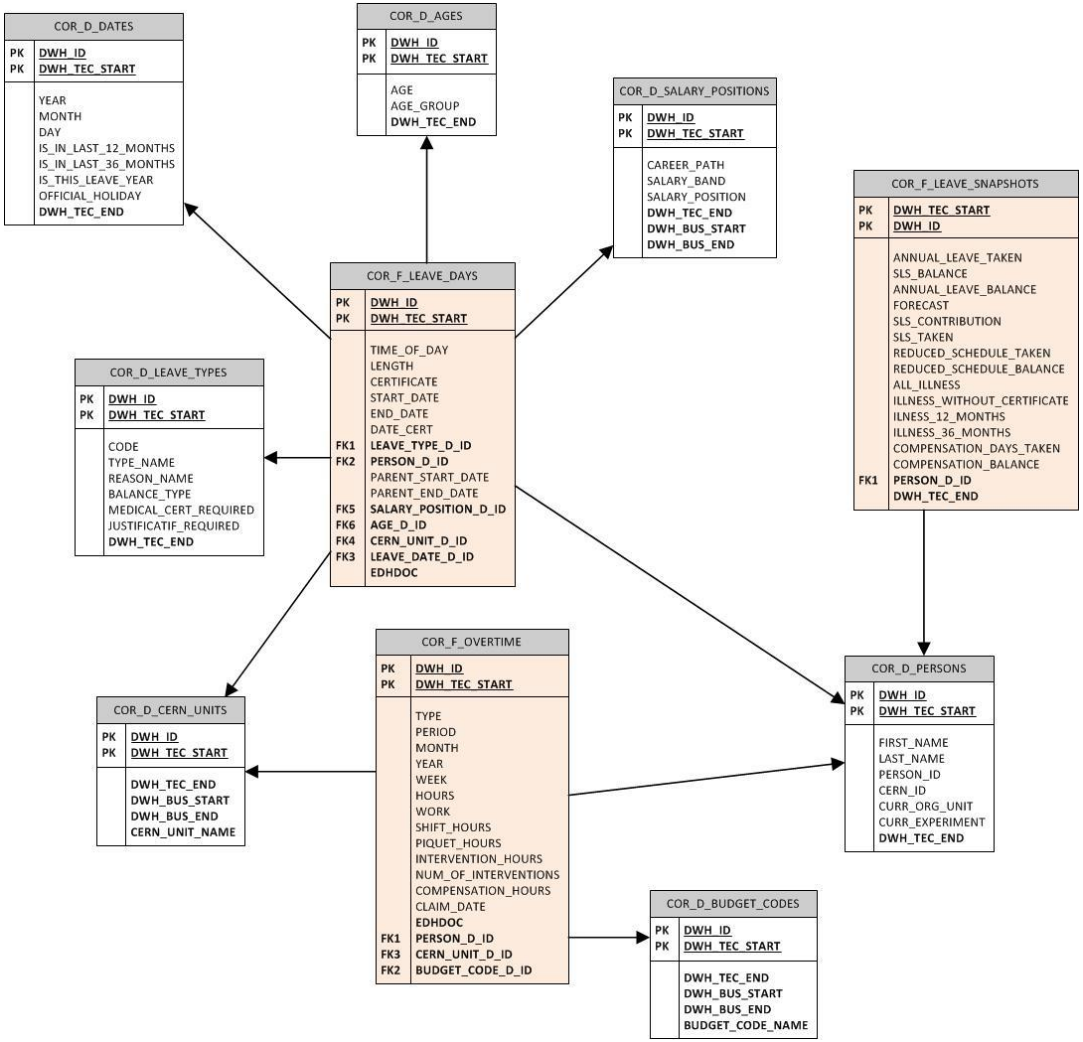


Constellation Schema

- Multiple fact tables with common dimension tables
  - Conformed dimensions
- This scheme is viewed as collection of stars
  - Used by sophisticated applications

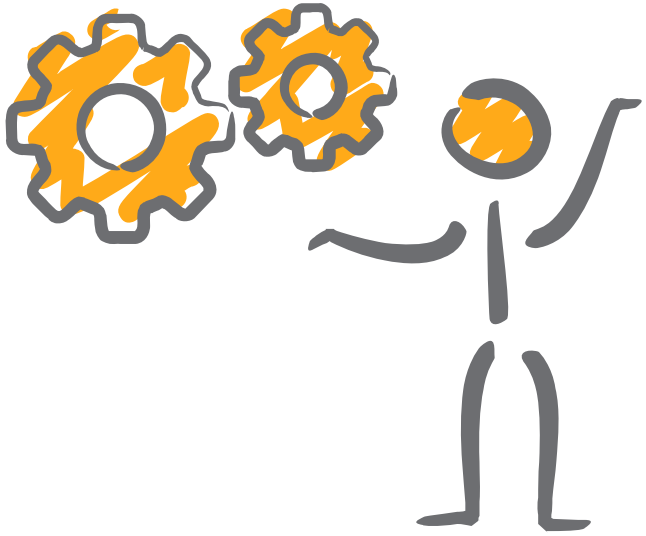


Constellation Schema



Data Warehouse

The process ETL



Extraction, Transformation, and Loading (ETL)

- The process of data consolidation is called Extraction, Transformation, and Loading

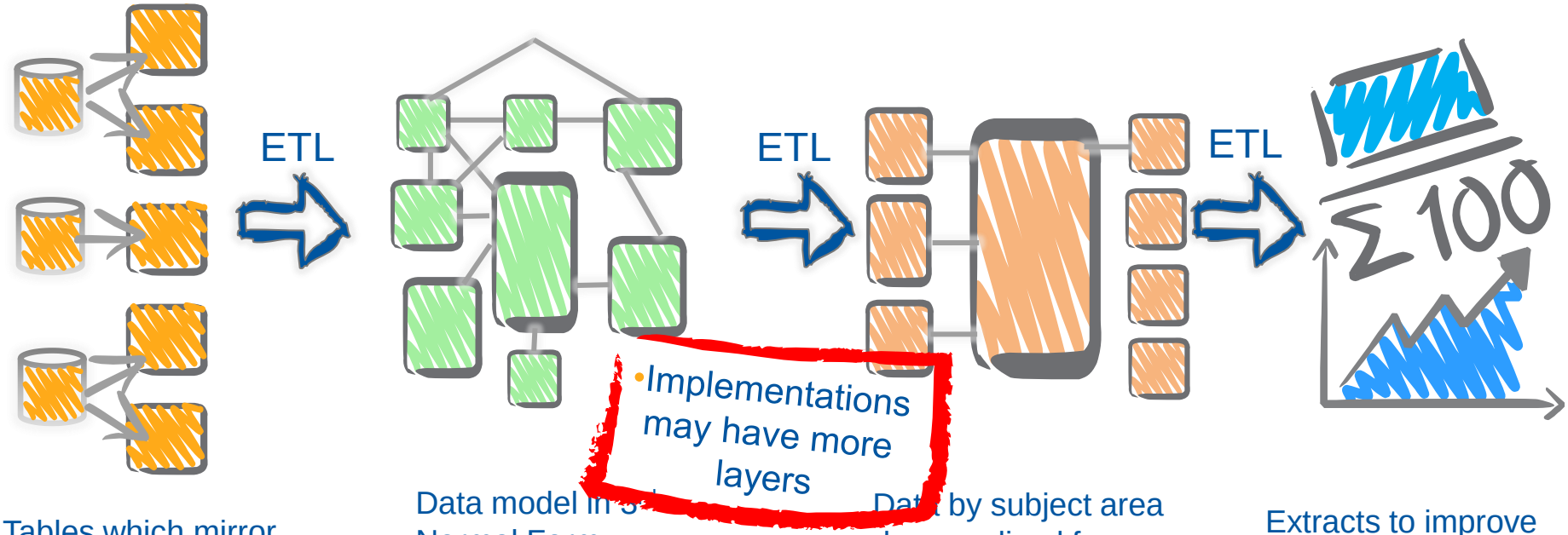
  ETL

Consolidation of Data

- Consolidate the data involves:
  - Moving it
  - Making it consistent
  - Cleaning up the data as much as possible
- Why?
  - Data is frequently stored in different formats
  - Data may be inconsistent between sources
  - Data may be dirty
    - Missing or inconsistent values

ETL Process

Staging area | **Operational Data Store (ODS)** | **Star / Snowflakes De Normalized Data** | **Aggregates for reporting**



Tables which mirror original source systems

Data model in 3rd Normal Form.

The Single Version of the Truth

Data by subject area denormalized for faster reporting and analysis => data duplication

Extracts to improve reporting performance

The Challenges

- There are several common issues inherent to any DWH project:
  - Data exists in multiple places
  - Data is not formatted to support complex analysis
  - Different people have different data needs
  - What data should be examined and in what detail
  - How will users interact with that data

Data Quality Issues

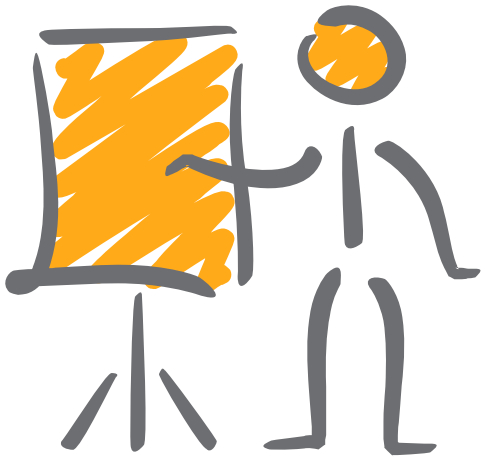
- Clean data facilitates more accurate analysis
- Many data entry systems allow free-form data entry of text values
  - E.g: the same city might be entered as Genève, Geneva, and Ginebra
- Routines to clean up data need to take into account all possible variations of bad data
  - New values in dimensions
    - Error
    - Not found

Inconsistent Data

- Data may be inconsistent
  - The money left in our budget is not the same in all systems
  - To represent True and False
    - One system may use 1 and 0
    - Another system may use T and F
  - Data stored in different countries will likely store sales in their local currency
    - These sales must be converted to a common currency

Third-Party ETL Tools

Business Intelligence



Concepts

Business Intelligence

> is a set of theories, methodologies, architectures, and technologies that…

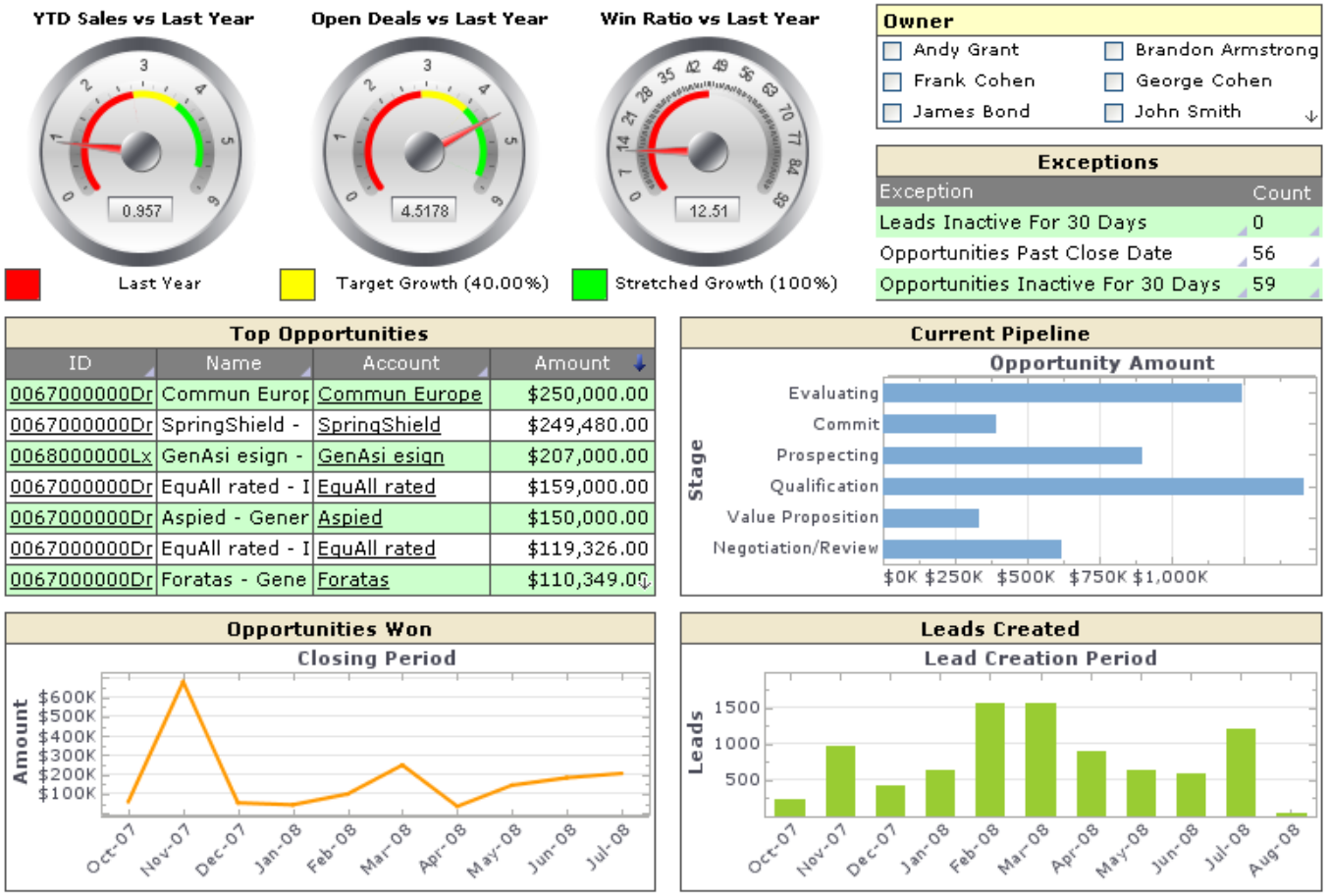
By Wikipedia

> It makes interpreting voluminous data (from a Data Warehouse) friendly, easily used for decision making

By myself

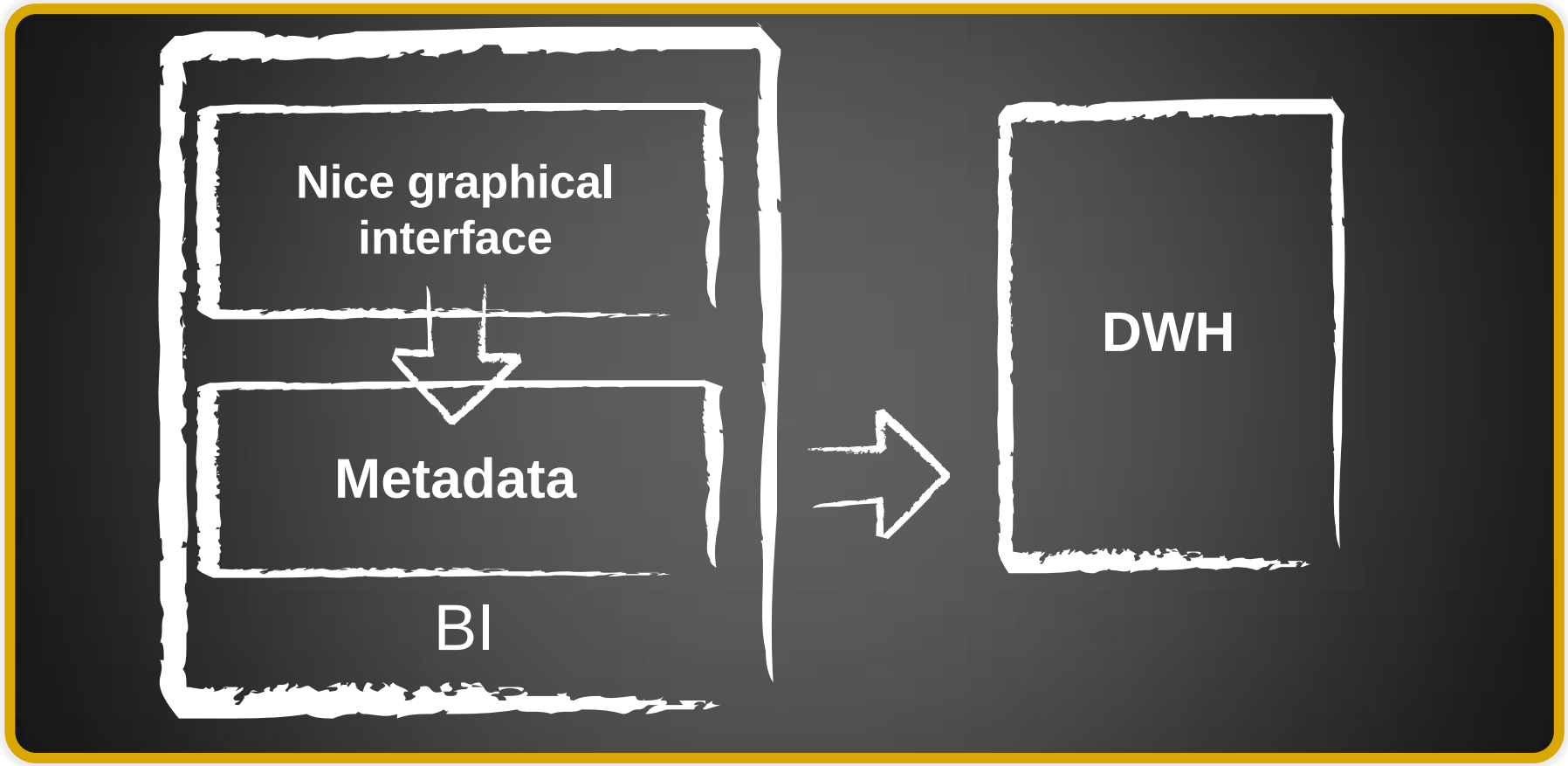
http://www.business-intelligence-secrets.com

Business Intelligence



http://www.inesoft.com

What is BI?



Why BI?

Benefits of BI

- Improve Management Processes
  - Planning
  - Control
  - Measuring
- Improve Operational Processes
  - Order processing
  - Purchasing
- Predict the Future

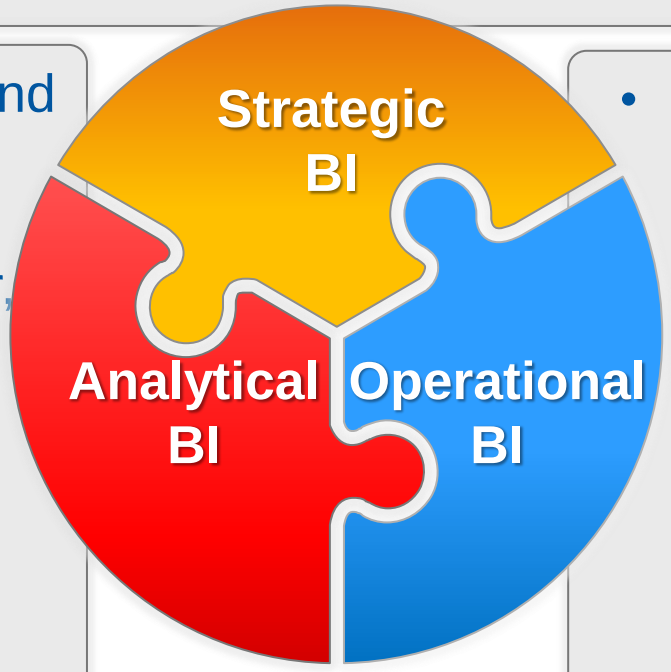
BI Golden Rules

> Get the right information to the right people at the right time

Types of BI

- Strategic (Balance, Strategic planning)
  - **Who**: Strategic leaders
  - **What**: Formulate strategy and monitor corporate performance

- Analytical (Financial and sales analysis)
  - **Who**: analysts, knowledge worker, controller
  - **What**: ad-hoc analysis



- Operational (Budgeting, Sales forecasting)
  - **Who**: Operational managers
  - **What**: Execution of strategy against objectives

Components of BI (Metadata)

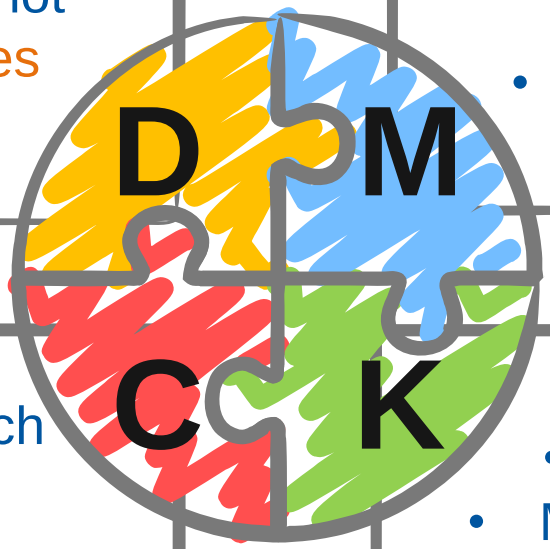
- Metadata
  - Metadata is “Information about Data”

    E.g: What is a customer Identifier, What is the definition of Revenue, How are addresses captured
- There are several items that make up a BI solution
  - Dimensions
  - Measures (=> Facts)
  - Cubes
  - Key Performance Indicators

Components of BI (Metadata)

- How you want to see the data
- Made up of attributes and may or may not include hierarchies
- Time, geography, product, account,

- What you want to see
  - Additive:
    - Expenses, sales
  - Non additive:
    - Shipment date



- Structures in which data is stored
- Access data by navigating through various dimensions

- Especial measure
- Metric tied to a target
- Results are on target or off target (10% increase in sales)

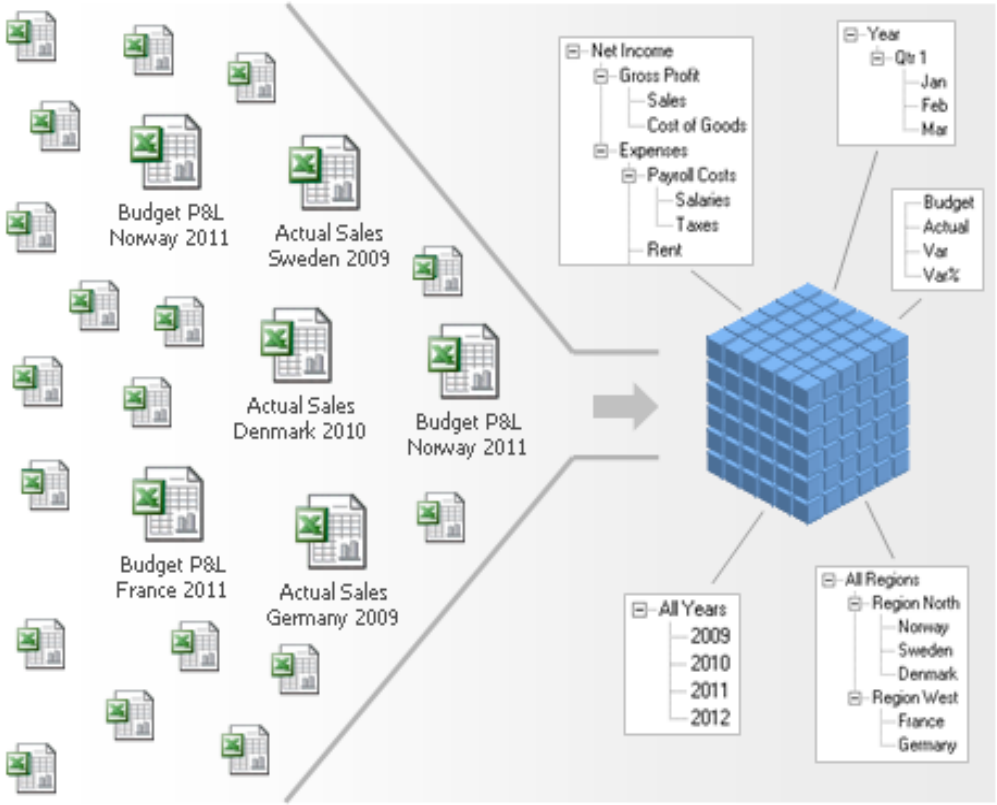
Key Performance Indicators

- Key Performance Indicators (KPIs) are typically a special type of measure
- KPIs are often what are shown on scorecards
- KPIs often contain not just the number, but also a target number
  - Used to evaluate the “health” of the value

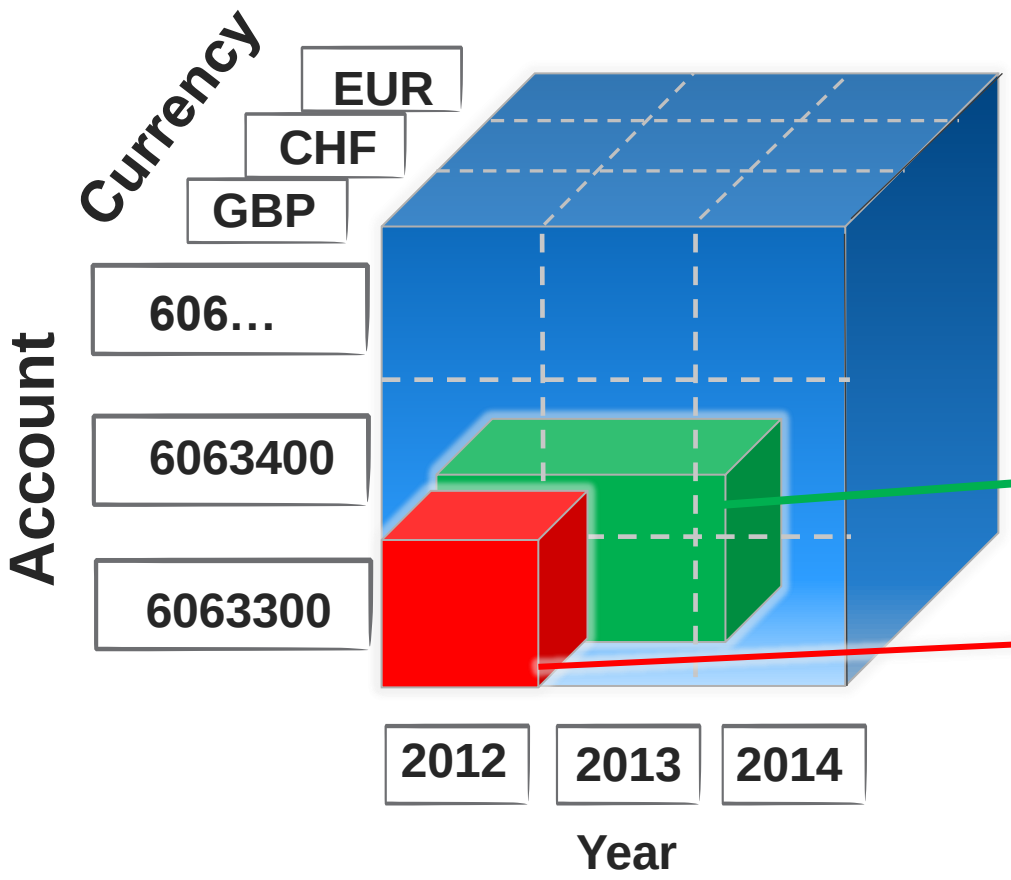
OLAP Cube

Show me the expenses by month by product for a concrete regional area

- Humans tend to think in a multidimensional way, even if they don't realize it



OLAP Cube



- Spreadsheats have only two dimensions
- Cubes have several dimensions

YR	ACCOUNT_CODE	UNI	SUM(INVOICE_AMOUNT)
2012	6063300	CHF	22769830.89
2013	6063300	CHF	3232148.76
2012	6063300	EUR	30609335.51
2013	6063300	EUR	8023806.59
2012	6063300	GBP	1437573
2013	6063300	GBP	458042.99

Approaches to Consuming BI

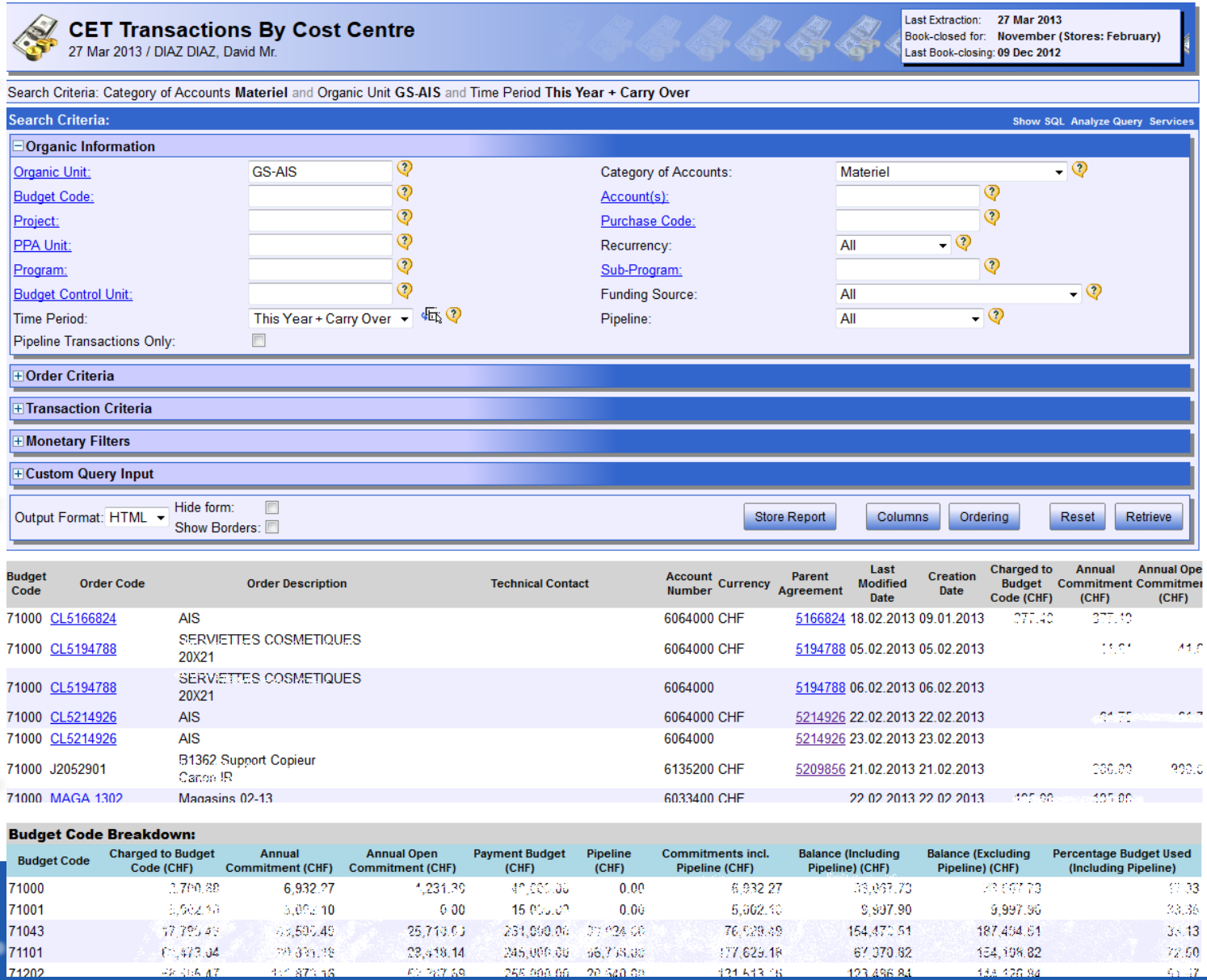
- Scorecards
  - Customized high-level views with limited analytic capabilities
- Reports
  - Standardized reports aimed at a large audience, with no or limited analytic capabilities
- Analytics Applications
  - Applications designed to allow complex data analysis
  - Dashboards contains Reports and Scorecards
- Custom Applications
  - Embed BI data within an application
  - Open Data

Report application CERN

Filters

Financial transactions

Financial budgets



CET Transactions By Cost Centre
27 Mar 2013 / DIAZ DIAZ, David Mr.

Last Extraction: 27 Mar 2013
Book-closed for: November (Stores: February)
Last Book-closing: 09 Dec 2012

Search Criteria: Category of Accounts **Materiel** and Organic Unit **GS-AIS** and Time Period **This Year + Carry Over**

Search Criteria: Show SQL Analyze Query Services

Organic Information

Organic Unit:	GS-AIS	Category of Accounts:	Materiel
Budget Code:		Account(s):	
Project:		Purchase Code:	
PPA Unit:		Recurrency:	All
Program:		Sub-Program:	
Budget Control Unit:		Funding Source:	All
Time Period:	This Year + Carry Over	Pipeline:	All
Pipeline Transactions Only:			

Order Criteria

Transaction Criteria

Monetary Filters

Custom Query Input

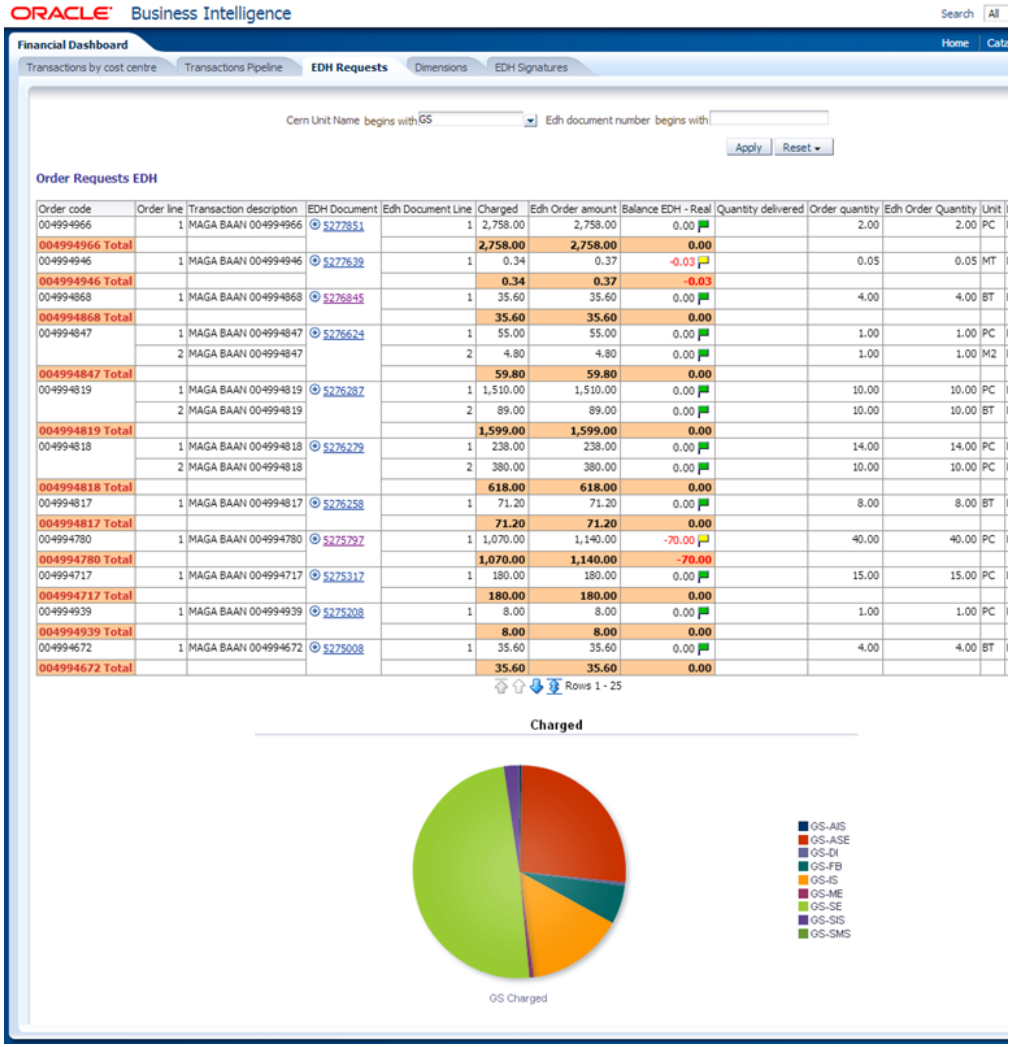
Output Format: HTML Hide form: Show Borders: Store Report Columns Ordering Reset Retrieve

Budget Code	Order Code	Order Description	Technical Contact	Account Number	Currency	Parent Agreement	Last Modified Date	Creation Date	Charged to Budget Code (CHF)	Annual Commitment (CHF)	Annual Ope Commitmer (CHF)
71000	CL5166824	AIS		6064000	CHF	5166824	18.02.2013	09.01.2013			
71000	CL5194788	SERVIETTES COSMETIQUES 20X21		6064000	CHF	5194788	05.02.2013	05.02.2013			
71000	CL5194788	SERVIETTES COSMETIQUES 20X21		6064000		5194788	06.02.2013	06.02.2013			
71000	CL5214926	AIS		6064000	CHF	5214926	22.02.2013	22.02.2013			
71000	CL5214926	AIS		6064000		5214926	23.02.2013	23.02.2013			
71000	J2052901	B1362 Support Copieur Canon IR		6135200	CHF	5209856	21.02.2013	21.02.2013			
71000	MAGA 1302	Magasins 02-13		6033400	CHF		22.02.2013	22.02.2013			

Budget Code Breakdown:

Budget Code	Charged to Budget Code (CHF)	Annual Commitment (CHF)	Annual Open Commitment (CHF)	Payment Budget (CHF)	Pipeline (CHF)	Commitments incl. Pipeline (CHF)	Balance (Including Pipeline) (CHF)	Balance (Excluding Pipeline) (CHF)	Percentage Budget Used (Including Pipeline)
71000		6,932.27			0.00	6,932.27			
71001				15	0.00	5,002.10	9,997.90	9,997.90	
71043			25,718.00			76,529.49	154,470.51	187,494.51	
71101			23,618.14	245,000.00		177,629.18	67,370.82	154,104.82	
71202									

Analityc application CERN



Why do companies need BI?

BI Solutions

GoodData

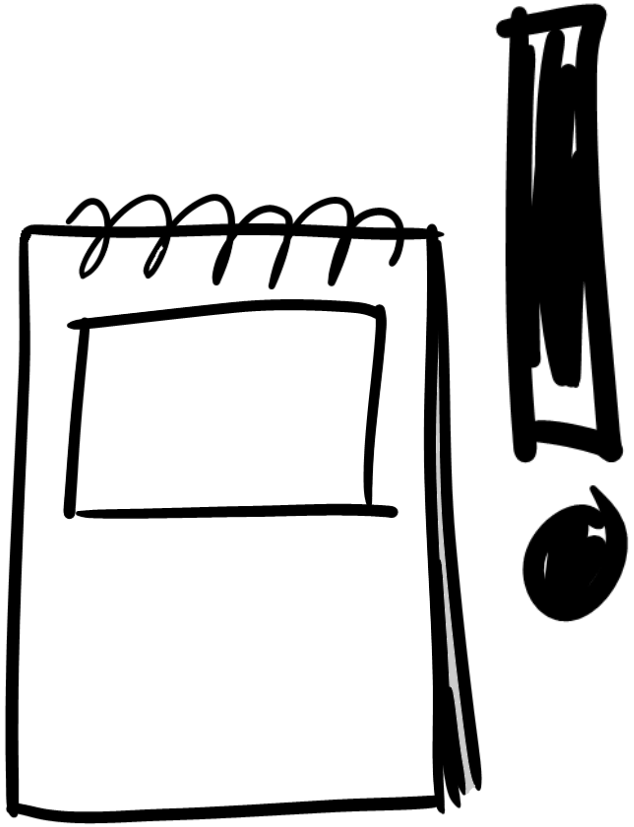
Summary

- The ETL process extracts data from source systems, transforms it and then loads it to a Data warehouse or a data mart.

- BI enhance the information stored in the DWH, making it easy to read by using interactive Reports and Dashboards

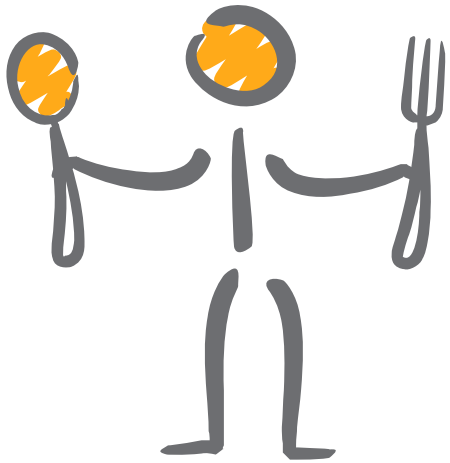
Review (Done?)

1. Knowledge path
2. Data Warehouse
   - Dimensional model
     - Dimension tables
     - Fact table/s
   - ETL
3. Business Intelligence
   - Metadata
     - KPI
     - OLAP Cube
   - Reports and Dashboard

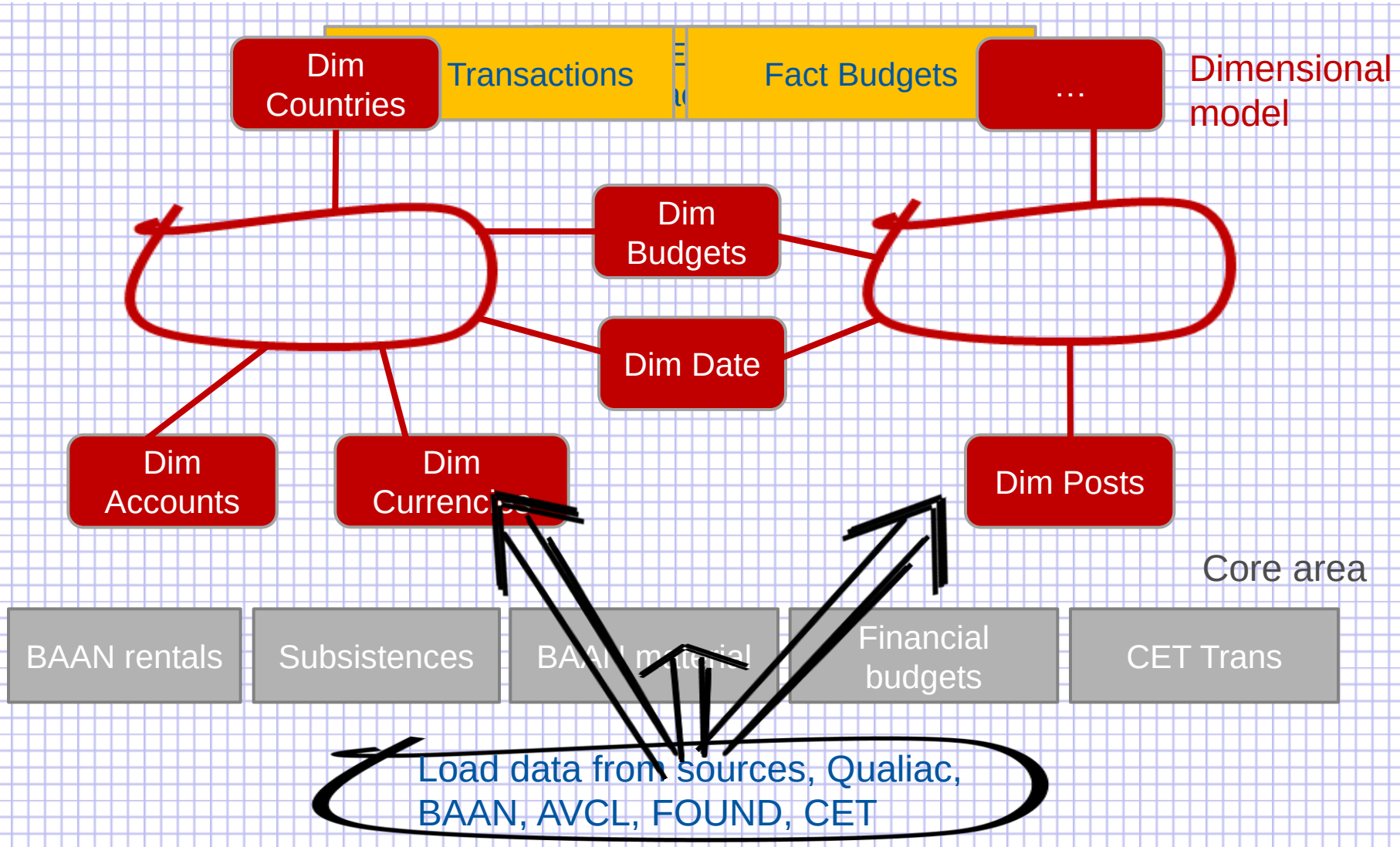


Examples

Examples

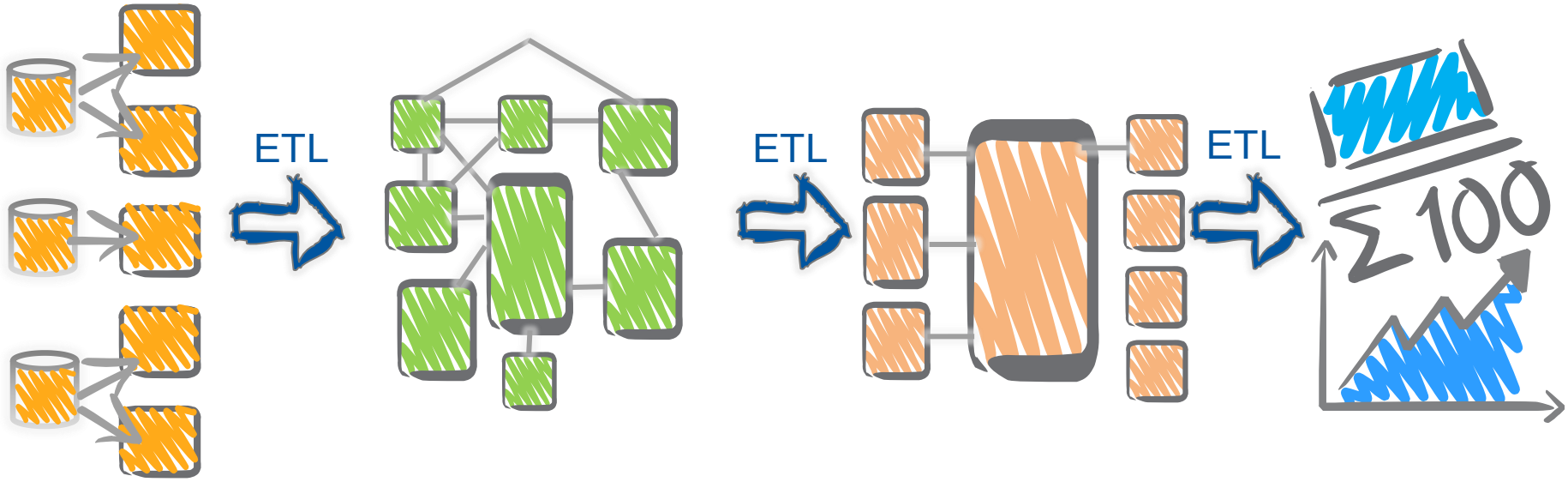


Financial BI at CERN



Inside a typical DW

Staging area | Operational Data Store (ODS) | Star / Snowflakes De Normalized Data | Aggregates for reporting



Tables which mirror original source systems

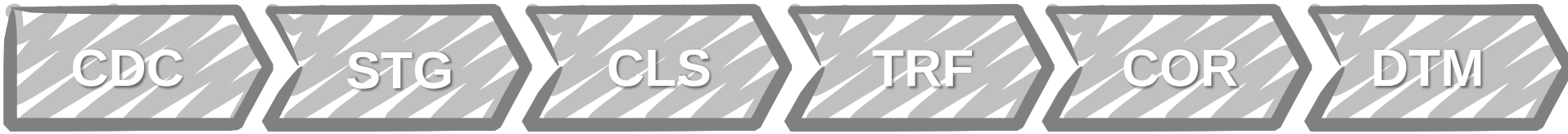
Data model in 3rd Normal Form.

The Single Version of the Truth

Data by subject area denormalized for faster reporting and analysis => data duplication

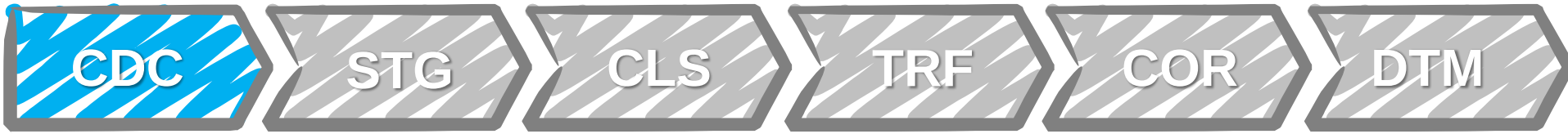
Extracts to improve reporting performance

6 areas

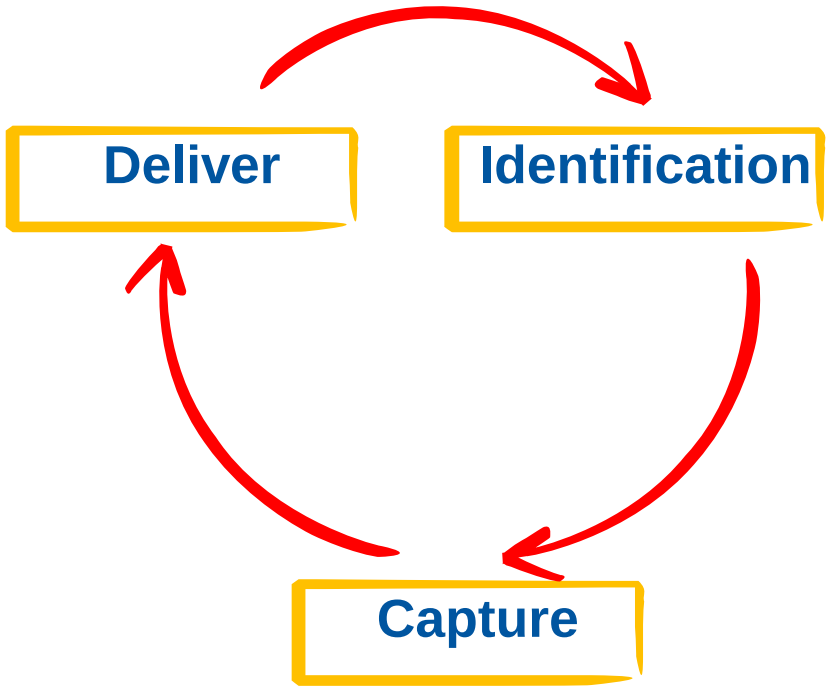


- Change data capture
- Staging
- Cleansing
- Transform
- Core
- Data marts

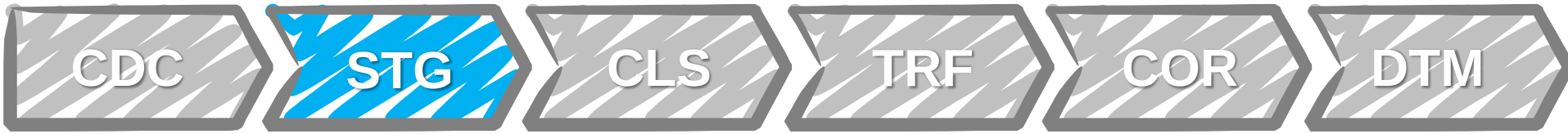
Change Data Capture



- Temporary/Transient working area
- Identify the data which has been changed
  - Reduce the volume of data moved

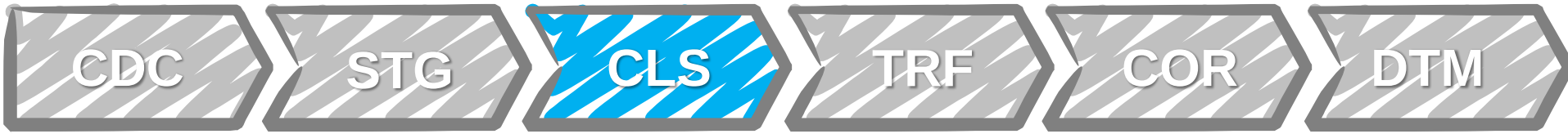


Staging



- Copy of the source data used for data processing in the ETL process
- Functions:
  - Consolidate. Data stored for further process
  - Change detection

Cleansing



- Used to clean and to transform data before transfer in core area

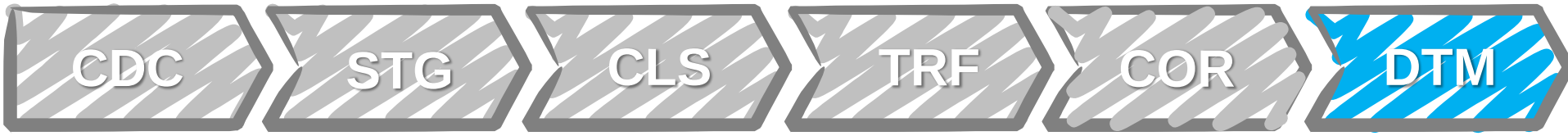
Transform

- Used to transform the data before it is transfer to the core area
  - Complex transformations
  - Merge different source tables
  - Ensure data consistency
    - Ej: Represent “dates”, “people” in the same way

Core

- Dimensional model
  - Denormalize tables
  - Lookups for dimensions
- This is a permanent area and any delete is allowed

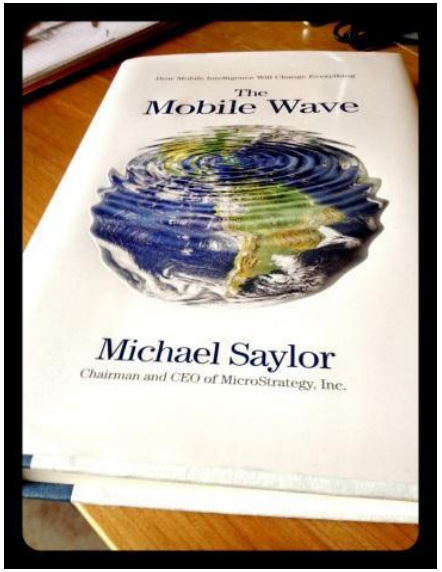
Data mart



- Data mart is a presentation layer of Data.
- Each data mart present a functional aspect of data
- This is a permanent area and any delete is allowed

Books

- Business Intelligence



- Larissa Terpeluk Moss, Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support-Applications (Addison-Wesley Information Technology)

- Data Warehouse

- Kimball, R., Ross, M., & Thornthwaite, W. (2012). *Relentlessly Practical Tools for Data Warehousing and Business Intelligence* (1.ª ed.). Wisley.

- Kimball, R., & Ross, M. (2002). The Data Warehouse Toolkit: *The Complete Guide to Dimensional Modeling* (2.ª ed.). Wiley.

Questions?