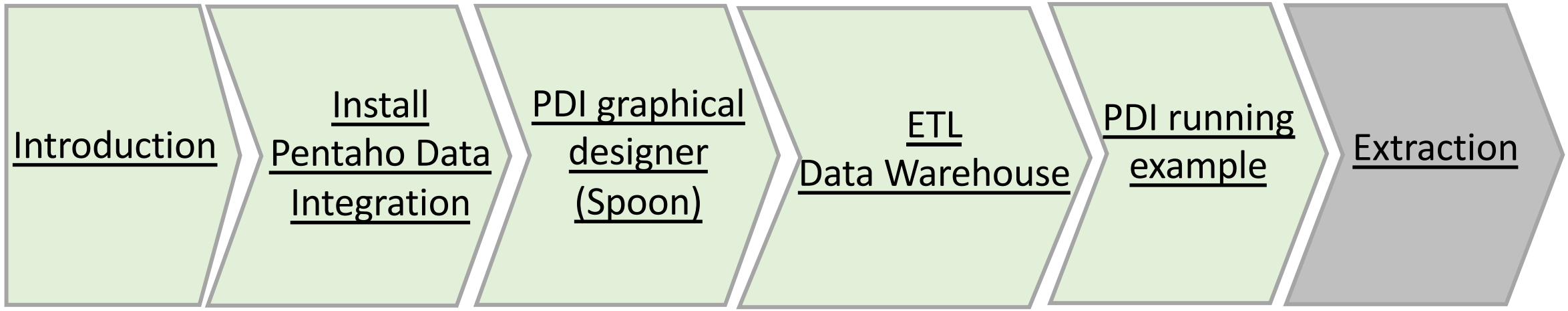
DATA ANALYTICS
(Data Warehouse)
Pentaho Data Integration

Luca Cinelli, PhD

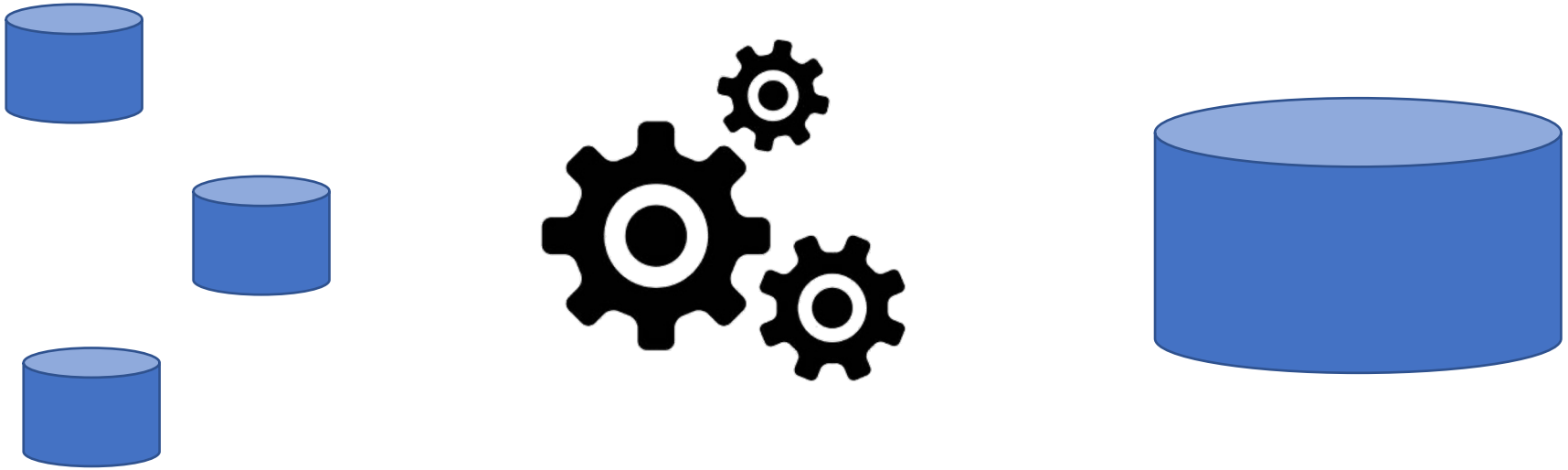
luca.cinelli@unical.it

Outline


Introduction
Install Pentaho Data Integration
PDI graphical designer (Spoon)
ETL Data Warehouse
PDI running example
Extraction

Introduction

pentaho

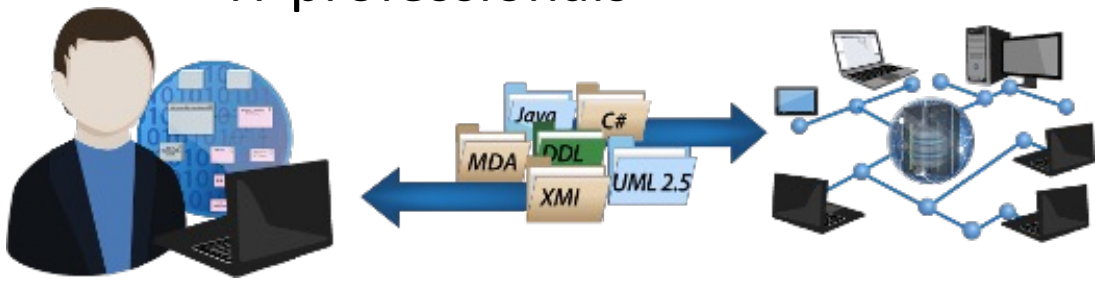


Why this course?

- For those who are curious…

University students

IT professionals



Project management

Business analysts

Pentaho BI Suite

- The Pentaho Business Intelligence Suite is **a collection of software** applications intended to create and deliver solutions for **decision making**
- The main functional areas covered by the suite are:
  - **Data Integration**. It is used to integrate information from different data sources (applications, databases, files). ***Pentaho Data Integration (PDI)*** is the tool that provides this functionality. PDI interacts with the rest of the tools (reading OLAP cubes, generating Pentaho reports, doing data mining with R Executor Scripts and the Cpython Script Executor)
  - **Analysis.** The analysis engine serves multidimensional analysis. It's provided by the ***Mondrian OLAP*** server.
  - **Data Mining**. It is used for running data through algorithms in order to understand the business and do predictive analysis. Data mining is possible thanks to ***Weka project***.
  - **Reporting**. The reporting engine allows designing, creating, and distributing reports in various known formats (HTML, PDF, and so on), from different kinds of sources. *In the Enterprise Edition of Pentaho, you can also generate interactive Reports*.
  - **Dashboards**. They are used to monitor and analyze Key Performance Indicators (KPIs). ***CTools*** is a set of tools and components created to help the user to *build custom dashboards on top of Pentaho*. There are specific CTools for different purposes, including ***a Community Dashboard Editor (CDE)***, a very powerful charting library (CCC), and a plugin for accessing data with great flexibility (CDA), among others. While the Ctools allow to develop advanced and custom dashboards, there is a *Dashboard Designer, available only in Pentaho Enterprise Edition, that allows to build dashboards in an easy way.*

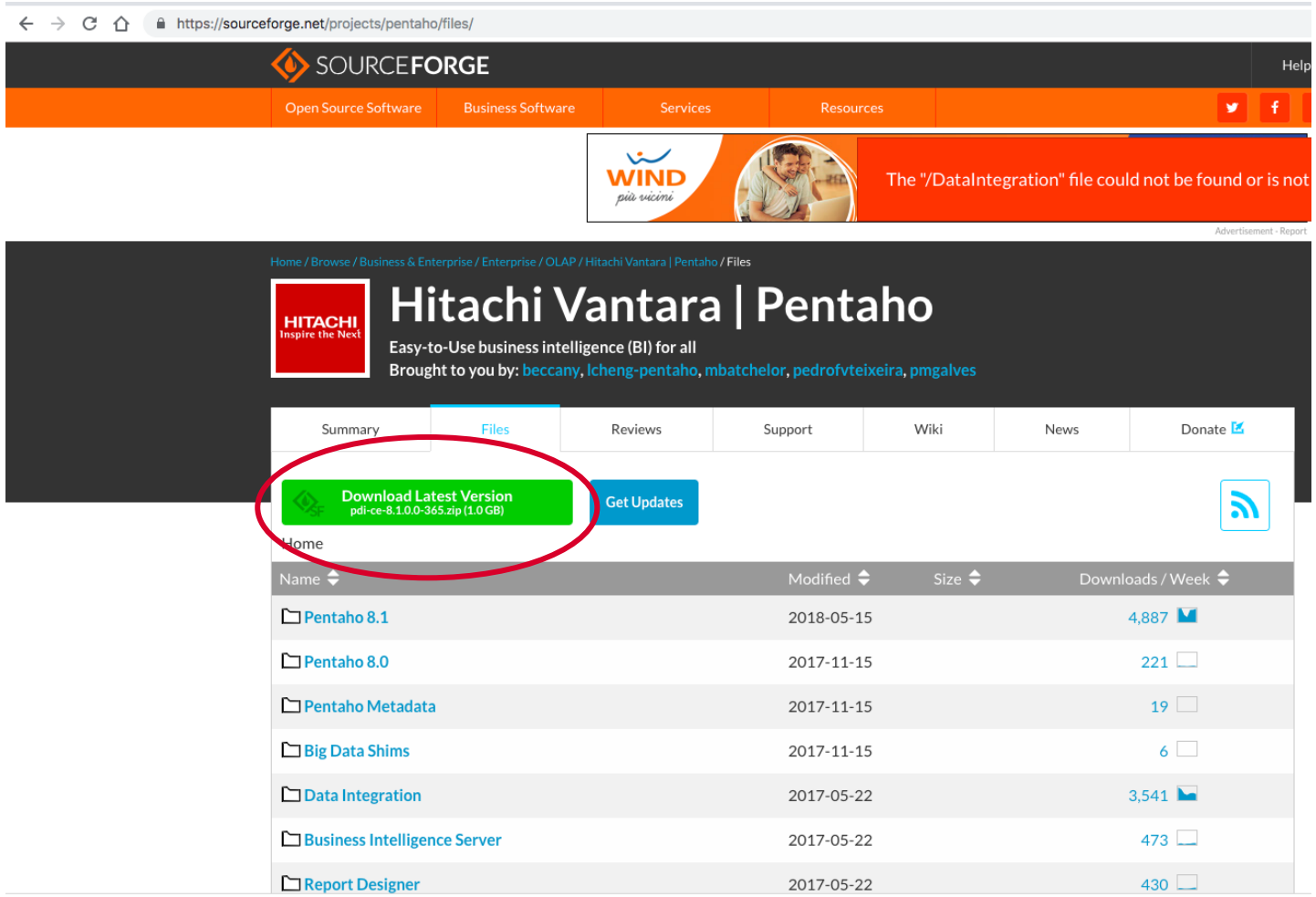
can be used standalone but also integrated

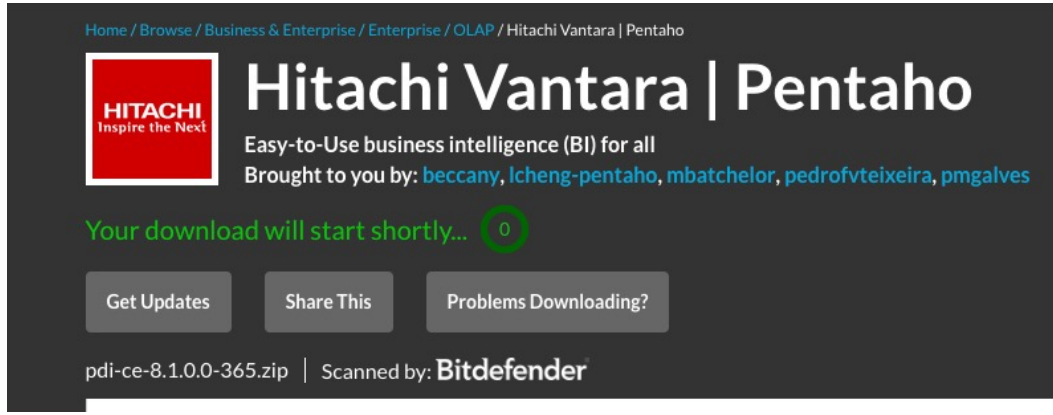
Install Pentaho Data Integration and other useful related software

PDI installation

- Download Link
  - **https://sourceforge.net/projects/pentaho/**
  - **Download** the zip file (Current latest version is 9)
  - Unzip it in a folder of you choice
  - Check java version
    - cmd→java -version (1.8.0 ok for PDI 9)
- This is the **community edition**.
- There exist an enterprise edition, it has some additional features
  - https://www.hitachivantara.com/en-us/video/pentaho-community-edition-vs-enterprise-edition.html

PDI installation





Auxiliary software installation

- A good text editor for viewing outputs of transformations and logs. E.g.:
  - Sublime Text
  - For Windows: Notepad++
- A spreadsheets editor. E.g.:
  - OpenOffice Calc.
  - Microsoft Excel
  - Numbers
- Databases: E.g. open source database engines:
  - MySQL
  - PostgreSQL
- Visual software to administer and query the database:
  - For PostgreSQL: PgAgmin
  - Work with generic database engine, included PostgreSQL: Squirrel SQL Client

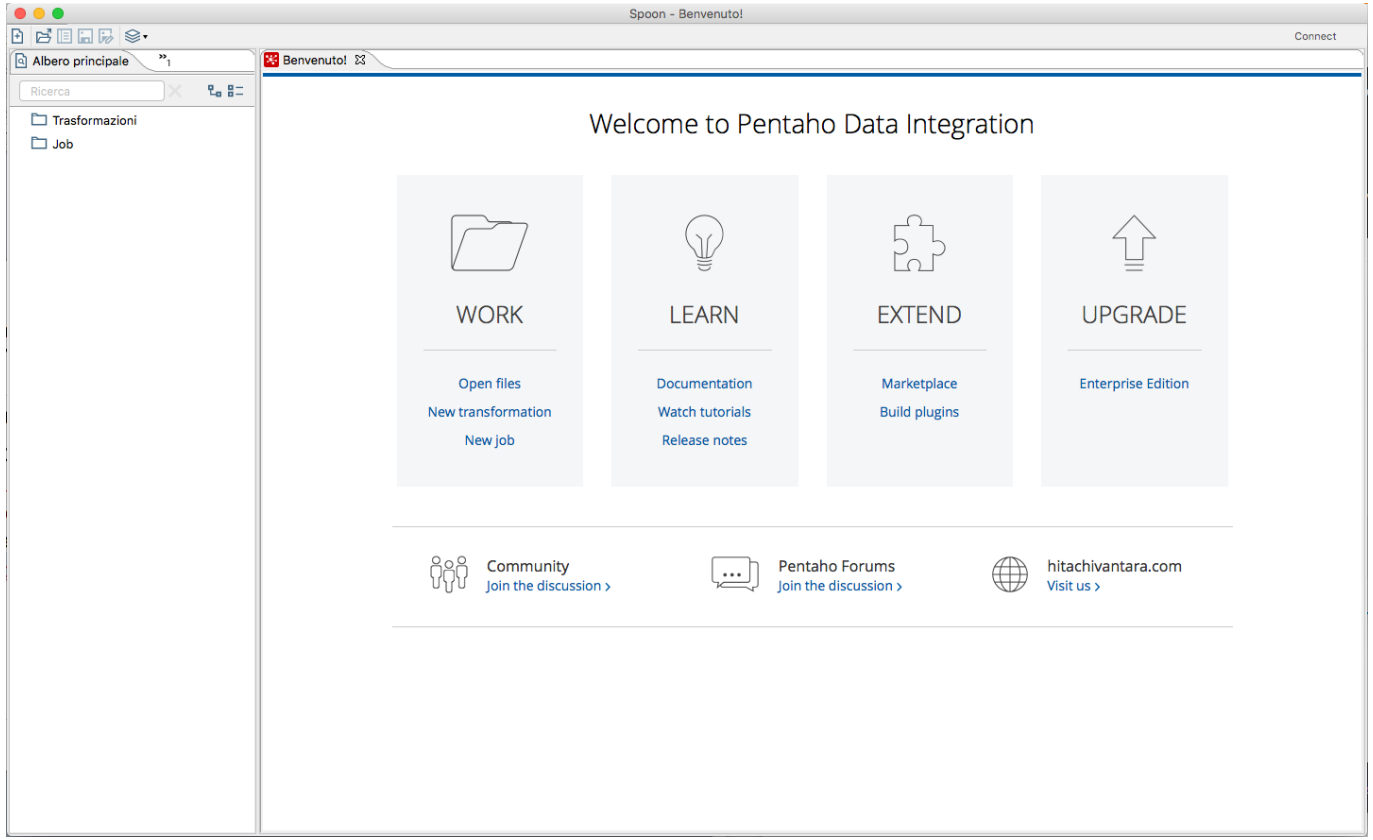
PDI graphical designer (Spoon)

Spoon

- Working with the data, it is possible by using the graphical environment
- Spoon is PDI’s desktop designer tool

Starting Spoon

- Windows: run Spoon.bat from within the PDI install directory
- Other platforms (such as Unix, Linux, Mac OS) open a Terminal and type spoon.sh
  - Suppose the folder with PDI is “data-integration”
  - >> cd data-integration
  - >> sh spoon.sh



- If Spoon doesn’t start as expected:
  - Windows: run SpoonDebug.bat
  - Other platforms: spoonDebug.sh

Main menu

Main Toolbar

View the structure of Transformations and Jobs

Data Integration File Edit View Action Tools Help

Spoon - Welcome!

Connect

View Design

Welcome! Transformation 1 Job 1

Search

- Transformations
  - Transformation 1
    - Run configurations
    - Database connections
    - Steps
    - Hops
    - Partition schemas
    - Slave server
    - Kettle cluster schemas
    - Data Services
    - Hadoop clusters
- Jobs
  - Job 1
    - Run configurations
    - Database connections
    - Job entries
    - Hadoop clusters
    - Slave server

Welcome to Pentaho Data Integration

WORK

Open files
New transformation
New job

LEARN

Documentation
Watch tutorials
Release notes

EXTEND

Marketplace
Build plugins

UPGRADE

Enterprise Edition

Community
Join the discussion ›

Pentaho Forums
Join the discussion ›

hitachivantara.com
Visit us ›

Transformati on Steps Tree

Data Integration File Edit View Action Tools Help

Spoon - Transformation 1

Connect

View Design

Welcome! Transformation 1 Job 1

100%

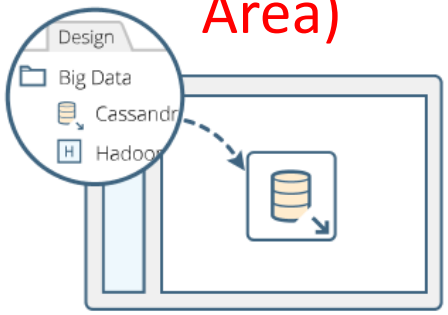
Transformation Toolbar

Has some options that make sense while working with datasets (e.g. preview)

Search

- Input
- Output
- Streaming
- Transform
- Utility
- Flow
- Scripting
- Pentaho Server
- Lookup
- Joins
- Data Warehouse
- Validation
- Statistics
- Big Data
- Agile
- Cryptography
- Palo
- Open ERP
- Job
- Mapping
- Bulk loading
- Inline
- Experimental
- Deprecated
- History

Canvas (Work Area)



Drag & Drop a Step

Also try shift + double-click

Step metrics tab (Execution results window) to see metrics for each step

Job Entries Tree

Job Toolbar

Has no options that only make sense while working with datasets (e.g. preview makes no sense to preview data in jobs)

Canvas (Work Area)

Job metrics tab (Execution results window) to see metrics for each job entry

PDI running example

Example

- What is a transformation and what is a job?
  - Transformations for ETL
  - Jobs for supporting activities like file management and emailing
- Aim is to show the capability of PDI

Example: scenario

- **Analytics Manager at a furniture sales company**
- John wants to collect **sales data** to make a sales **dashboard**. The file John receives is incomplete. Also, the job is repetitive, i.e. the file is received every week. Therefore, John wants to automate the process of sales data collection.

Example: Automation Steps

1. Check whether there is a **sales file** in **sales folder**
2. If file is available, import data from file
3. Identify the rows with **missing data** and create a separate file for them
4. Upload the complete file as an **Excel file in dashboard folder**
5. **Send** the incomplete file to the sales manager for rectification

Extract

Transform

Load

Steps 2,3 and 4 will be done in a PDI Transformation.

Step 1+ the transformation obtained from 2,3,and 4 + step 5 will be run as part of PDI Job

Extract, Transform, and Load

Data Warehouse

ETL

ETL stands for Extract, Transform and Load

1. **Extract** is reading data from data sources
2. **Transform** is processing the data
3. **Load** is writing the data to the destination

Data Warehouse

- A **data warehouse** collates data from a wide range of sources within an organization.
- **NOT** operational database.
- **NOT** updated frequently
- For the purpose of **analytics**

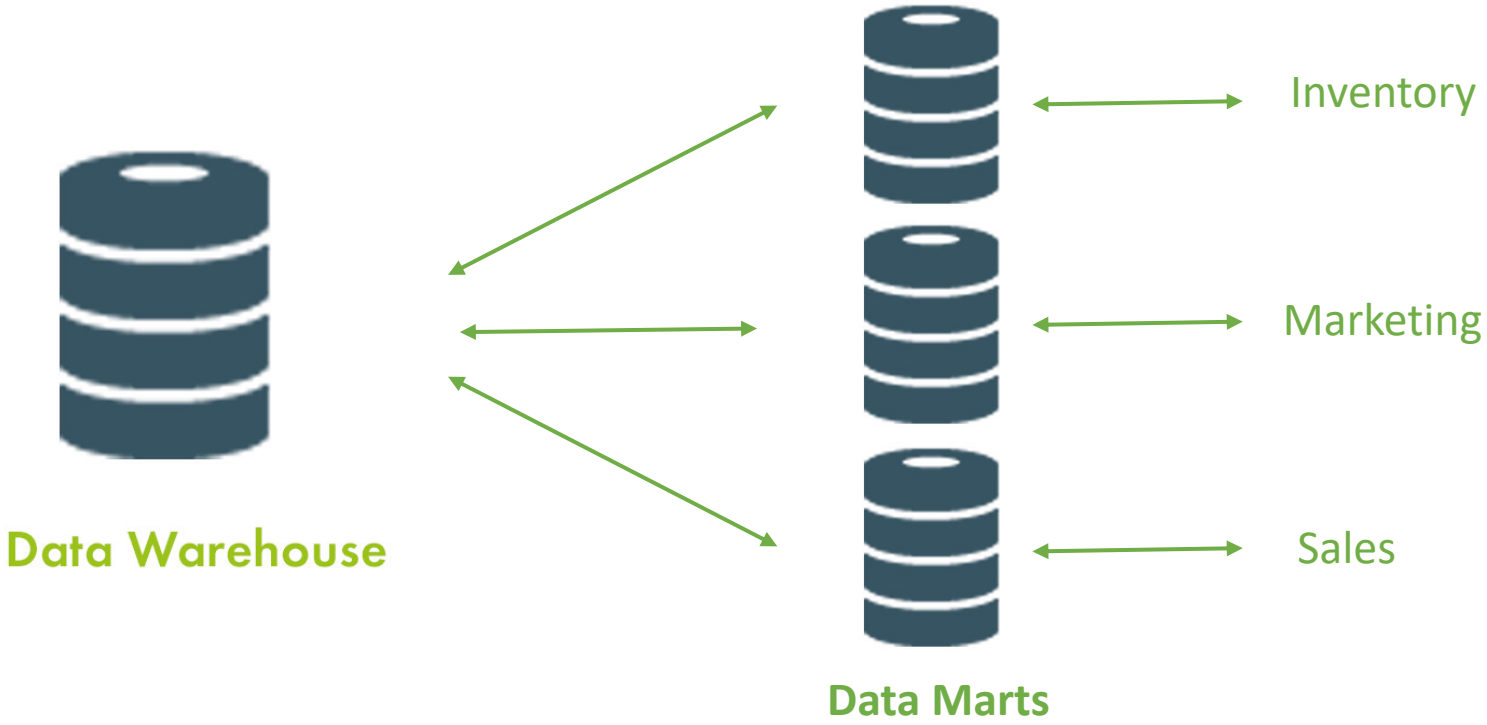
Data Warehouse: differences with OLTP

- OLTP stands for OnLineTransaction Processing.
- Used for adding/ updating one/ few rows of data at a time

CHARACTERISTIC	OLTP	DATA WAREHOUSE
System scope/view	Single business process	Multiple business subjects
Data sources	One	Many
Data model	Static	Dynamic
Dominant query type	Insert/update	Read
Data volume per transaction	Small	Big
Data volume	Small/medium	Large
Data currency	Current timestamp	Seconds to days old
Bulk load/insert/update	No	Yes
Full history available	No	Yes
Response times	< 1 second	< 10 seconds
System availability	24/7	8/5
Typical user	Front office	Staff
Number of users	Large	Small/medium

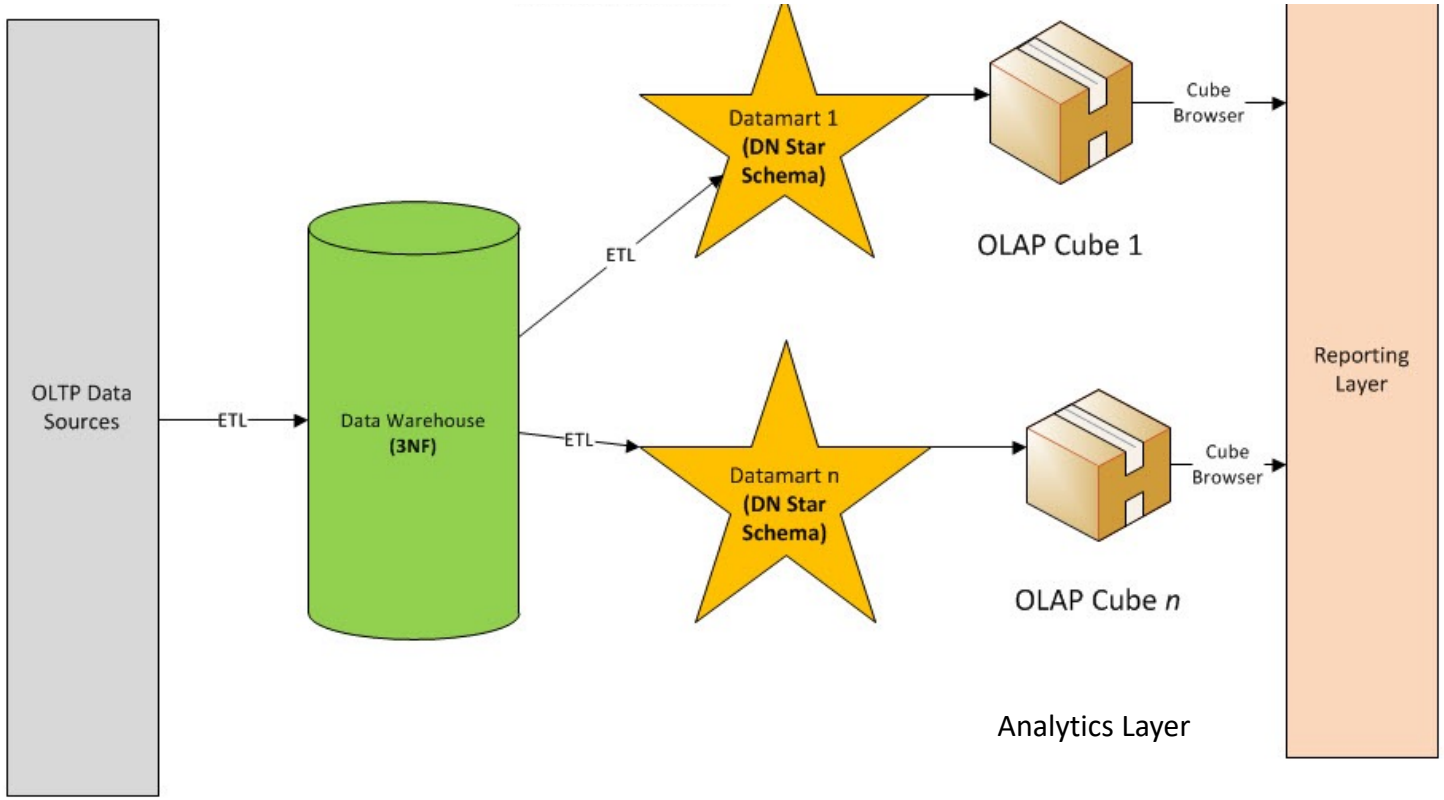
Data Warehouse: differences with Data Mart

- **Data mart** contain repositories of summarized data collected for analysis on a **specific unit** within an organization, for example, the sales, finance, operations, marketing department.



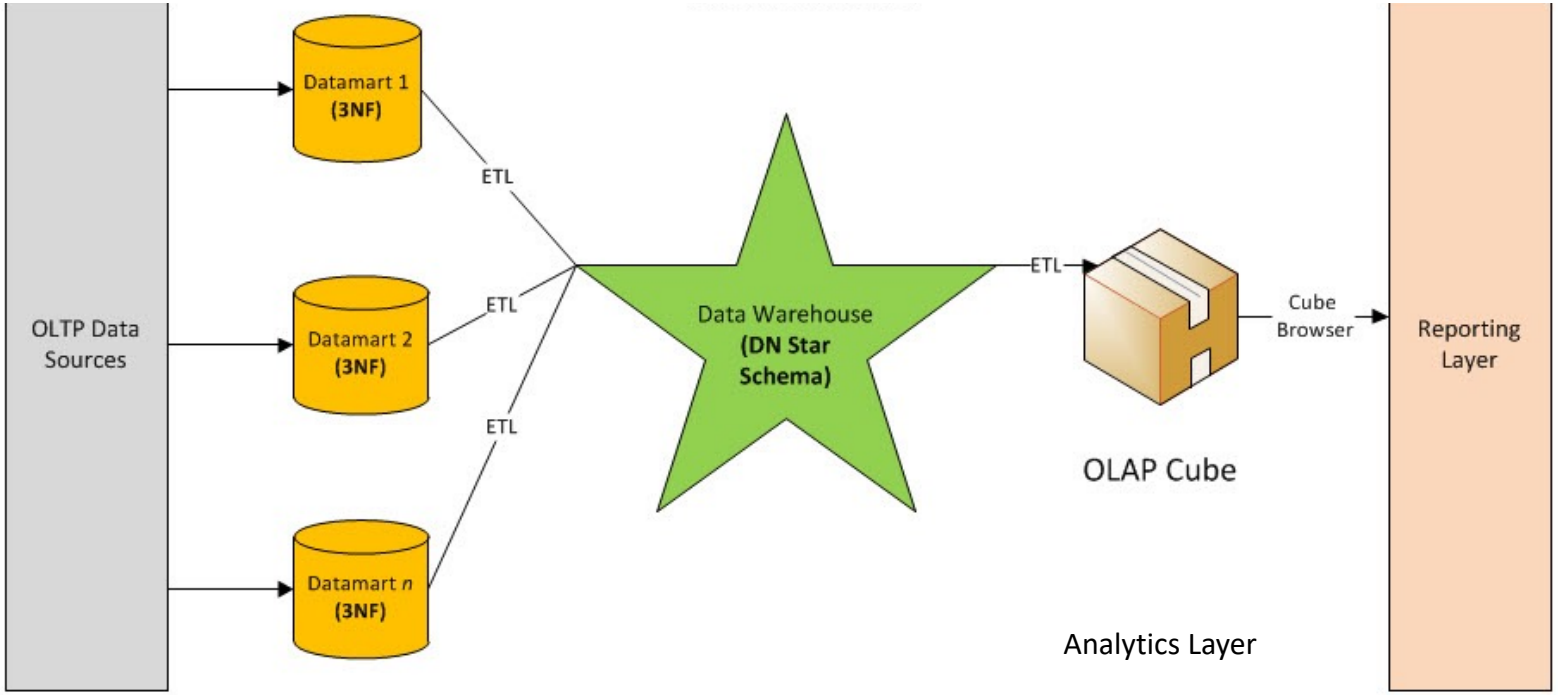
Data Warehouse: Common architectures

- **Inmon model:** Organizations create DW, Data marts are created from DW, Analytics is applied on Data marts



Data Warehouse: Common architectures

- **Kimball model:** Organizations create Data marts, DW are created from data marts, Analytics is applied on DW



ETL vs ELT

- **ETL – Extract, Transform, Load**

- **ELT – Extract, Load, Transform**

ETL vs ELT

ETL –Extract, Transform, Load	ELT –Extract, Load, Transform
Source and target databases are different (e.g, Oracle source, SAP target databases)	Source and target databases are same (e.g., Oracle source and target database)
Data volume is small or moderate	Data volume is large
Data transformation are compute-intensive	Data transformations are less complex
Data is structured	Data is unstructured