# The concepts behind “Business Intelligence”, “Analytics” & “Data Warehousing”

## The rise of Distributed Computing and Storage

Abdel Dadouche
DJZ Consulting

adadouche@hotmail.com
@adadouche

# The Origin of “Business Intelligence”

# Origin of the term “**Business Intelligence**”

- Coined by H.P. Luhn (IBM) in 1958:
  - **Business** as: “*a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera.*”
  - **Intelligence** as: “*the ability to apprehend the interrelationships of presented facts in such way as to guide action toward a desired goal*”

In Fench BI = Informatique décisionelle

Source: Luhn HP (1958) A business intelligence system. IBM Journal of Research and Development 2(4):314–319

# The beginnings of BI - 70s-90s

- Tools for editing reports, statistics, using **operational databases**
- Small decision support systems (**DSS**) based on **spreadsheets** (budget simulation, etc.)
- Expert systems based on rules, designed by extracting knowledge from one or more experts (limited interest and results)
- Specific decision support systems (DSS): based on Operational Research (OR) techniques, the simulation, optimization, ...

# The beginnings of BI – end of 90s

- In 1989, Howard Dresner (later a Gartner Group analyst) proposed a new definition:

*“Business intelligence" is as an umbrella term to describe "concepts and methods to improve business decision making by using fact-based support systems.*"

# The rise of BI – 90s – Y2K

- Computer technology allowed the development of larger warehouses of data (Data Warehouse)
- New approaches & algorithms:
  - often derived from statistics, allows to extract information from raw data
  - allowing the extraction of new or hidden information, knowledge from data
  - grouped together using data mining software
- Data from the Web ("Web Mining")

# The rise of “modern” BI : Analytics – since Y2K

*Business intelligence encompass now a broad category of applications, technologies, and processes for gathering, storing, accessing, analyzing and transforming data into accurate information and support the day to day decision-making processes*

# So, BI, but why, what for, by who, for whom?

# Let’s put some context first around Y2K

- Economy context
  - Globalization, new markets emergence
  - Increasing & disruptive competition
  - Shorter and faster decision-making processes
- Data context
  - Data decentralization (from DC to users)
  - Difficulty accessing larger set of information
  - Information has become a source of income and competitiveness and a strategic business issue
- IT context
  - Growing computing power & storage capacity
  - More efficient DBMS (MPP, in-memory, cloud etc.)
  - Open Data
- Company / Organisational context
  - More and more customer-oriented
  - Shorter and shorter design to production cycles
  - New distribution channels

DBMS: Database management systems
MPP: Massively Parallelized Processing

# BI as a platform for analysis (for whom)

- Used by decision makers to obtain in-depth knowledge of the company and to define and support their business decisions & strategies

- For example:
  - gain a competitive market advantage
  - improve the company performance and efficiency
  - faster response to changes
  - increase profitability

# BI as a platform for analysis: use cases (for what)

- Accounting : Invoice and order management
- Finance : Cash flow analysis, fraud detection
- HR : Career management, employee satisfaction
- Manufacturing : Production forecasting, cost reduction, quality controls
- Marketing : Campaign analysis
- Sales : Sales forecasting
- …

# BI as a platform for analysis (how)

- Build and keep up-to-date “warehouses” of data based on historical and multidimensional data extracted from various operational & external sources
- Extract subsets of data according to various criteria and filters
- Analyze these data sets along different axes, identify trends and correlations
- Make hypothesis and run simulations
- Share your analysis and insights with others and make decisions

# Data Warehouse Overview

# Inmon definition of a Data Warehouse (1992)

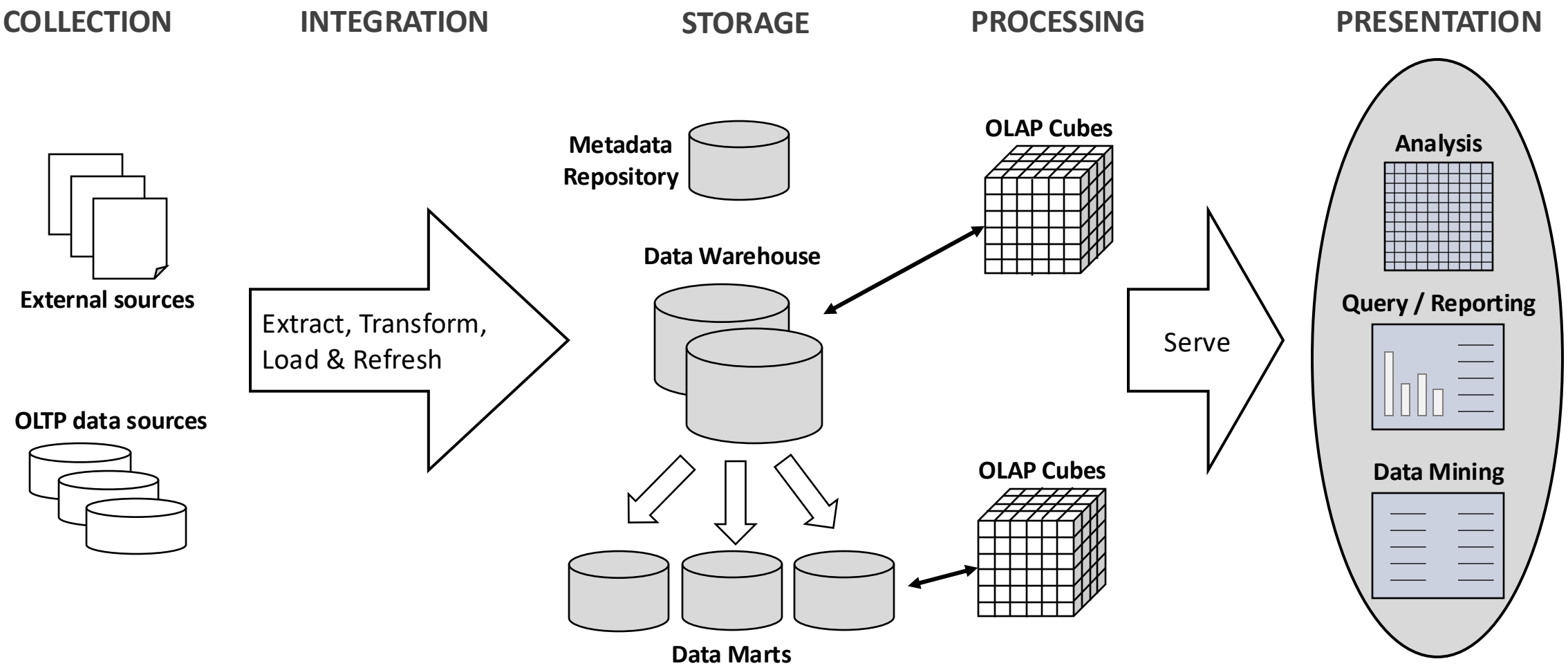
A subject oriented, nonvolatile, integrated, time variant collection of data in support of management's decisions

Source: https://en.wikipedia.org/wiki/Bill_Inmon

# Inmon definition of a Data Warehouse (1992)

- Subject-oriented:
  - data is gathered and organized based on subject or theme associated with the different functional, relevant and necessary for analysis needs
- Non-volatile:
  - data is exclusively used in read query and cannot be modified
- Integrated:
  - data results from heterogeneous data sources integration
- Time Variant:
  - data allows analysis of activity from any point in time

# A « Modern » BI architecture



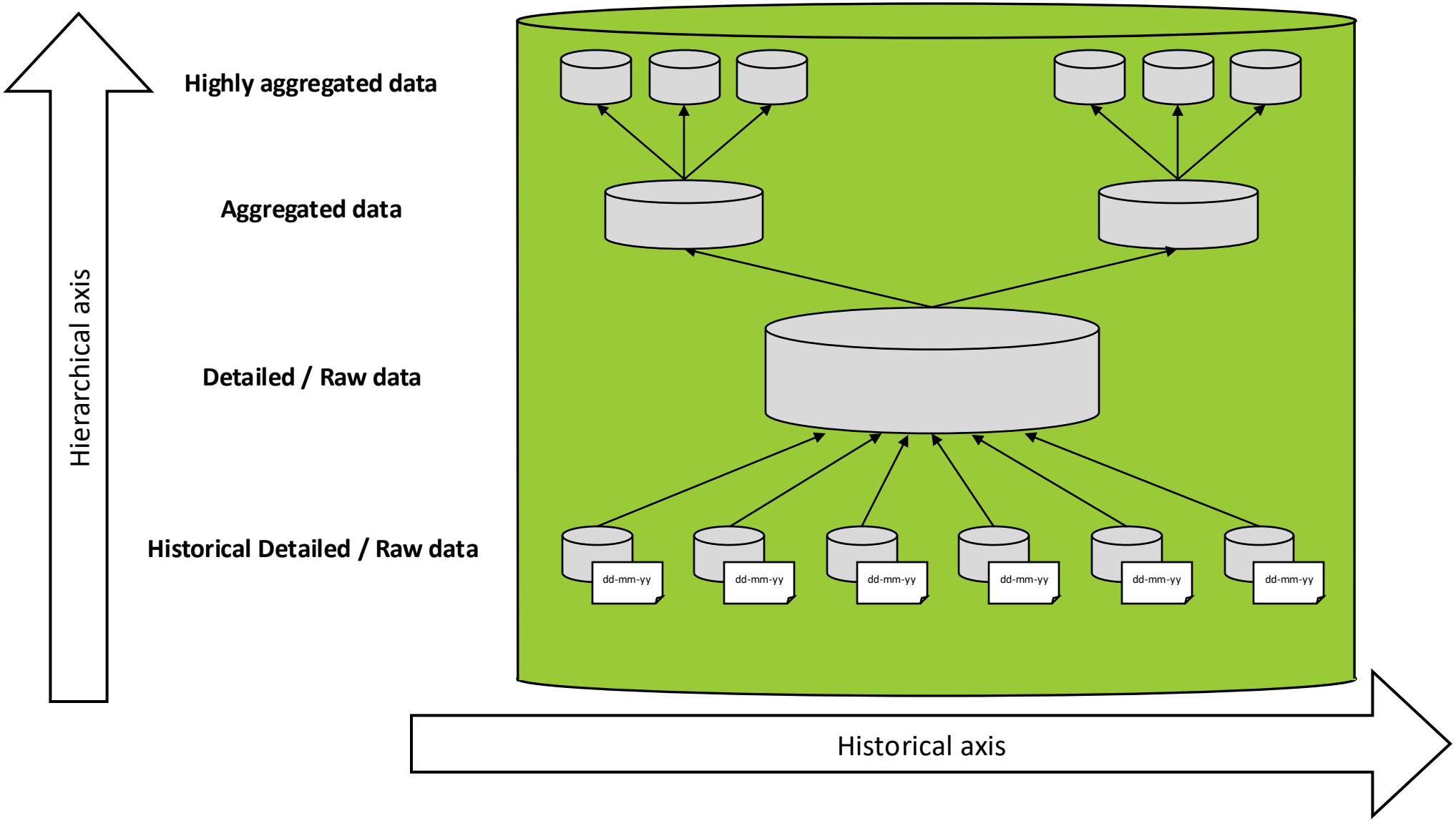

Business Intelligence Process (Wan, J., Chen, T. and Chiu, S. (2005)

# Data warehouse vs Operational databases

| Features | Operational databases | Data warehouse |
|---|---|---|
| **Data Granularity** | Detailed | Aggregated, summarized, meta data'ed (context) |
| **Data Homogeneity** | Homogeneous | Not necessarily homogeneous<br>Data integration often necessary |
| | Process oriented | Subject oriented |
| **Time variant access** | No, only the current view is accessible.<br>Data is periodically archived | Yes, data is partitioned, historized and non volatile |
| **Operations** | Many concurrent deletes, inserts and updates and simple reads | Mostly reads (complex queries) |

# Functional architecture of a data warehouse

# Hierarchical and historical axis

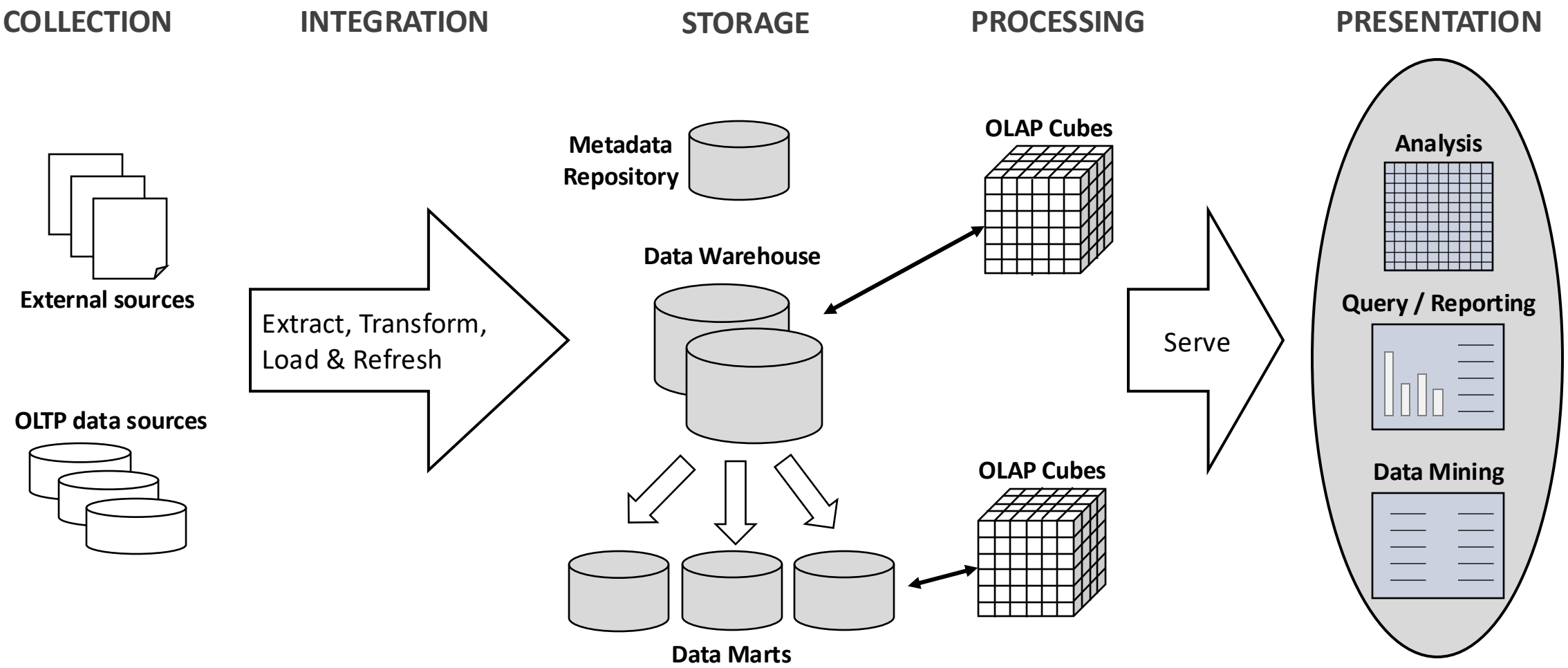


# Hierarchical and historical axis

- Hierarchical axis:
  - establishes an aggregation hierarchy :
    - highly summarized data at a higher level
    - aggregated data summarizing detailed data
    - detailed data representing the most recent events
- Historical axis:
  - includes detailed historical data snapshots representing past events

# Data Warehouse vs Data Mart

# A « Modern » BI architecture




Business Intelligence Process (Wan, J., Chen, T. and Chiu, S. (2005)

# Functional view - Data Warehouse

- Collects data from various data sources (operational database, 3rd party sources, web logs, etc.)
- Ensure durability over time and ability to reconstruct a “view” at any point in time
- Data is meant to be transformed into insight supporting decision process

# Functional view - Data Mart

- Used to support multidimensional-type analysis processes (OLAP)
- A subset of the data warehouse to support specific decision-making process :
  - for specific/functional populations of user / decision makers
  - for specific analysis needs
- Data Mart are “subject-oriented” by definition

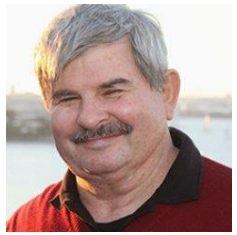
# Technical view - Data Warehouse

- Requires large investments in infrastructure and staff to manage and maintain very large data volumes containing all historical detail and aggregated data
- Is the “centralized” storage location target for all production data extracts
- A data model facilitating efficient management and historization is the key of a successful data warehouse project

# Technical view - Data Mart

- Requires smaller infrastructure and staffing investments as volume are usually reduced and is subject to shorter development cycles (months instead of years)
- Designed for decision support based on data extracted from large DW or directly from operational sources
- A data model facilitating the decision-making process is the key of a successful data mart project

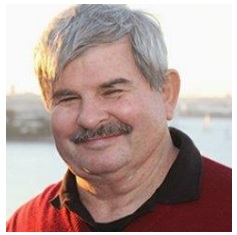
# Architectural Approach and Philosophy



# Bill Inmon vs Ralph Kimball

- Bill Inmon recommends building the data warehouse that follows the **top-down** approach.
- Start with building a big centralized enterprise data warehouse where all available data from transaction systems are consolidated into a subject-oriented, integrated, time-variant and non-volatile collection of data that supports decision making.
- Then data marts are built for analytic needs of departments.

- Ralph Kimball recommends building the data warehouse that follows the **bottom-up** approach.
- Starts with mission-critical data marts that serve analytic needs of departments.
- Then it is integrating these data marts for data consistency through a so-called information bus.
- Kimball makes uses of the dimensional model to address the needs of departments in various areas within the enterprise.

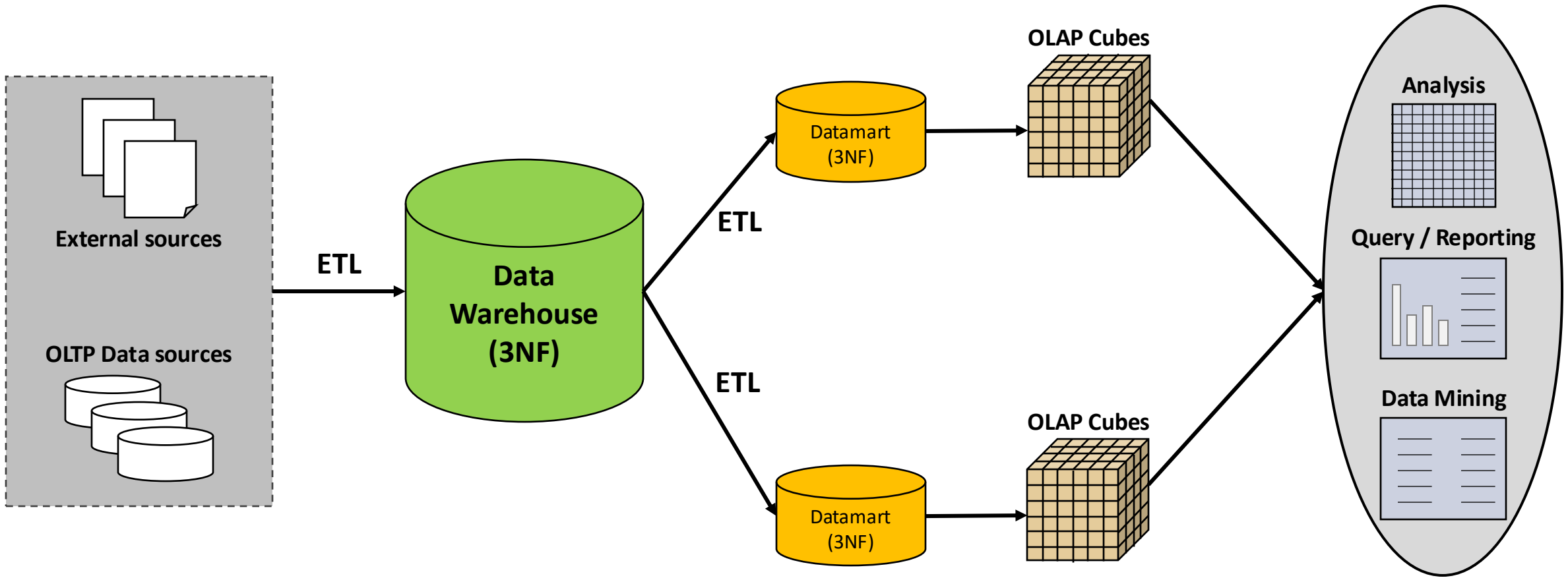


# Bill Inmon vs Ralph Kimball

| Characteristics | Inmon | Kimball |
|---|---|---|
| Business requirement support | Strategic | Tactical |
| Drivers & Users | Entreprise / Corporate | Business Areas / LOBs |
| Data structure | Data that meet multiple and varied information needs and non-metric data | KPI, business performance measures, scorecards… |
| Data sources | Changeable | Stable |
| Skill sets | Bigger team of specialists | Small team of generalists |
| Time constraint | Lower / Longer Time Scale | Immediate / Urgent |
| Cost to build | High start-up costs | Low start-up cost |
| Budget | Larger | Smaller |
| Requirements | More stable and growing | Volatile |
| Modeling approach | 3NF | Star / Snowflake |

KPI: Key Performance indicators
3NF: 3rd Normal Form

# Bill Inmon Approach




ETL: Extract, Transfom & Load
OLAP: On-line Analytical Processing

# Let’s ChatGPT “3NF” (Third Normal Form)

- **3NF** is a database normalization rule that requires a table to be in **Second Normal Form (2NF)** and ensures that all non-key columns are **only dependent on the primary key**, eliminating transitive dependencies.

- This means no column should depend on another non-key column, which helps reduce data redundancy and improve data integrity.

- Achieving 3NF simplifies database maintenance and reduces anomalies during data operations.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « 3NF » ?

# Let’s ChatGPT “2NF” (Second Normal Form)

- **2NF** is a database normalization stage that requires a table to be in **First Normal Form (1NF)** and ensures that all non-key attributes are fully functionally dependent on the entire primary key, eliminating partial dependencies.

- This means that if the primary key is composite (made of multiple columns), no non-key attribute should depend on just part of that key.

- By achieving 2NF, databases reduce redundancy and update anomalies, improving data integrity and consistency.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « 2NF » ?

# Let’s ChatGPT “1NF” (First Normal Form)

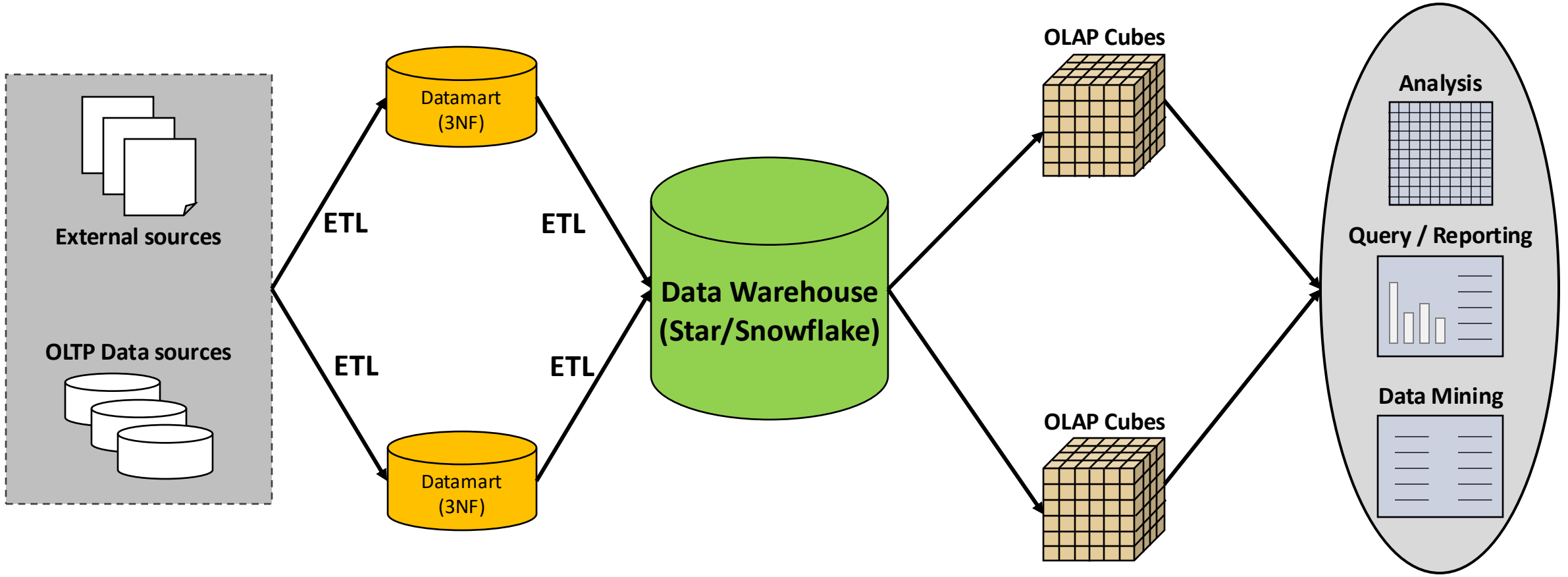
- **1NF** is the first stage of database normalization that requires a table to have **atomic (indivisible) values** in each column and ensures that each record is unique, with no repeating groups or arrays.

- This means every field contains only one value, and the table structure is organized as rows and columns without nested data.

- Meeting 1NF is essential for reducing data redundancy and enabling efficient querying and updates.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « 1NF » ?

# Ralph Kimball



# Let’s ChatGPT “Star Schema”

- A **Star Schema** is a data warehouse schema design where a central **fact table** connects directly to multiple **denormalized dimension tables**, forming a star-like shape.

- This structure simplifies queries and improves performance by minimizing the number of joins needed to retrieve data.

- Star Schemas are commonly used in business intelligence because they are easy to understand and optimize for fast data retrieval.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of « Star schema » ?

# Let’s ChatGPT “Snowflake Schema”

- A **Snowflake Schema** is a data warehouse schema where the **dimension tables are normalized** into multiple related tables, creating a more complex, branching structure that resembles a snowflake.

- This normalization reduces data redundancy and saves storage space but can make queries slower due to the increased number of joins.

- Snowflake Schemas are used when storage efficiency is prioritized over query performance.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of «  Snowflake schema » ?

# Let’s ChatGPT “Snowflake Schema”

- A **Snowflake Schema** is a data warehouse schema where the **dimension tables are normalized** into multiple related tables, creating a more complex, branching structure that resembles a snowflake.

- This normalization reduces data redundancy and saves storage space but can make queries slower due to the increased number of joins.

- Snowflake Schemas are used when storage efficiency is prioritized over query performance.

Prompt: in 3 sentences, including at most 2 sources, what is the definition of «  Snowflake schema » ?

# Let’s ChatGPT “Star” vs “Snowflake” Schema

- A **Star Schema** has a simple structure with a central fact table connected directly to denormalized dimension tables, which improves query performance by minimizing joins, while a **Snowflake Schema** normalizes dimension tables into multiple related tables, reducing data redundancy but increasing query complexity due to more joins.

- The Star Schema prioritizes fast, easy-to-understand queries, whereas the Snowflake Schema emphasizes storage efficiency and data integrity.

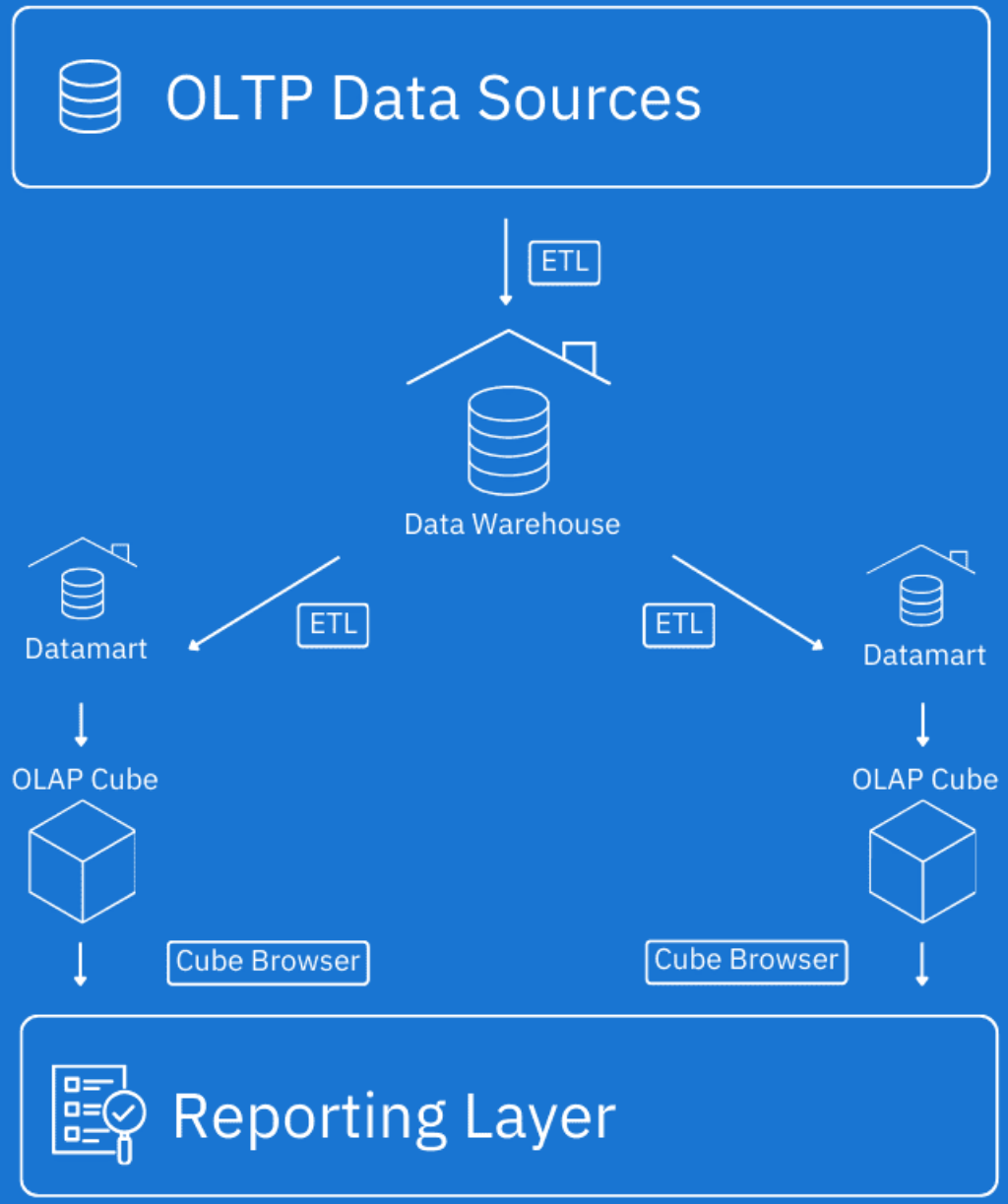
- Choosing between them depends on whether query speed or storage optimization is more important for the data warehouse.

Prompt: in 3 sentences, including at most 2 sources, what are the differences between a “Star Schema” and a «  Snowflake schema » ?

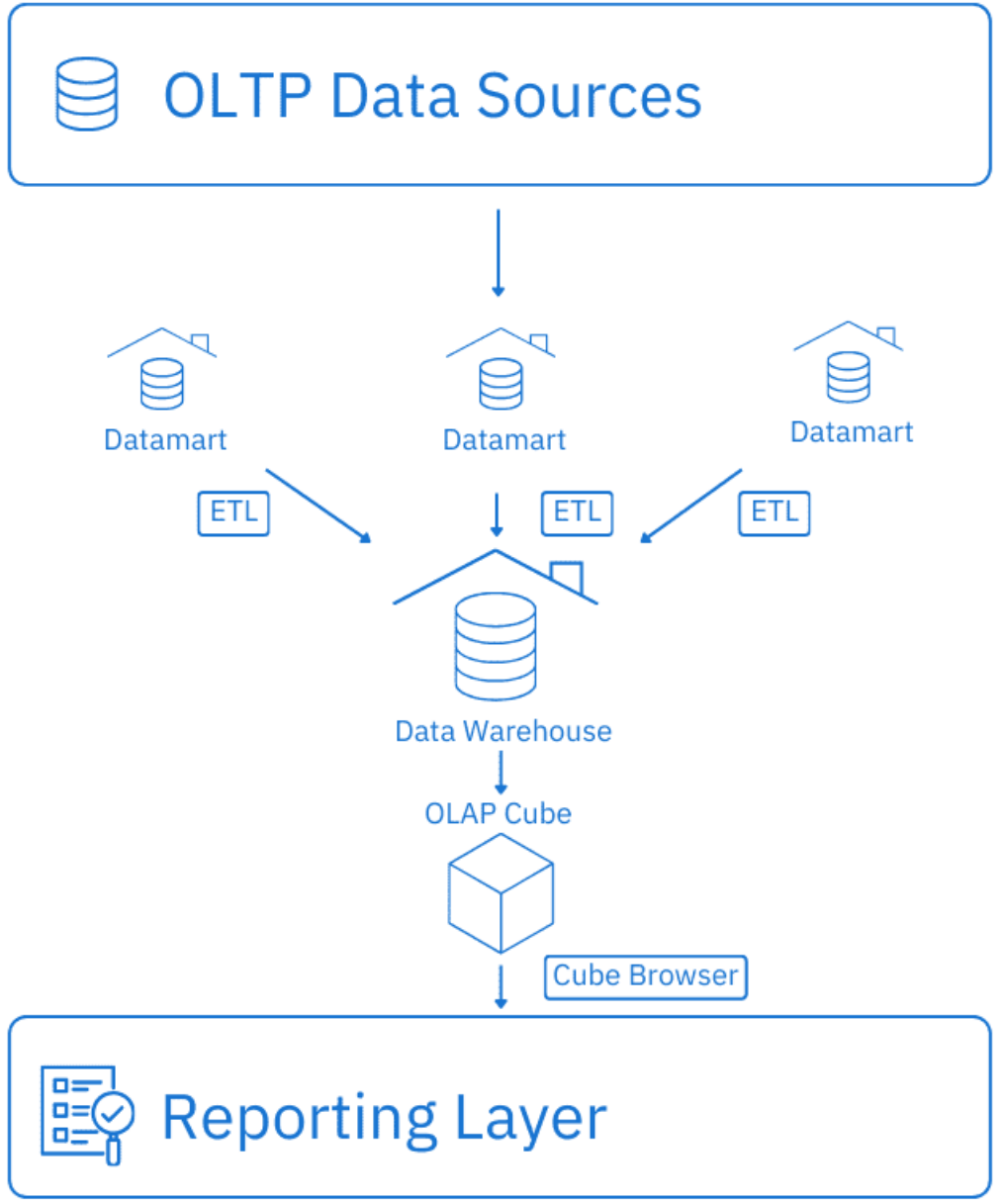
# Data Normalization Vs Denormalization

| | NORMALIZATION | DENORMALIZATION |
|---|---|---|
| **Definition** | Normalization is the process of creating a set schema to store non-redundant and consistent data | Denormalization is the process of combining the data so that it can be queried speedily |
| **Purpose** | To reduce the data redundancy and inconsistency | To achieve the faster queries execution through introducing redundancy |
| **Used in** | OLTP system, where the emphasize is on making the insert, delete and update anomalies faster and storing the quality data | OLAP system, where the emphasis is on making the search and analysis faster |
| **Integrity** | Maintained | Not a requirement |
| **Redundancy** | Eliminated | Added on purpose |
| **Table count** | Increases | Decreases |
| **Storage** | Optimized usage | Wastage |

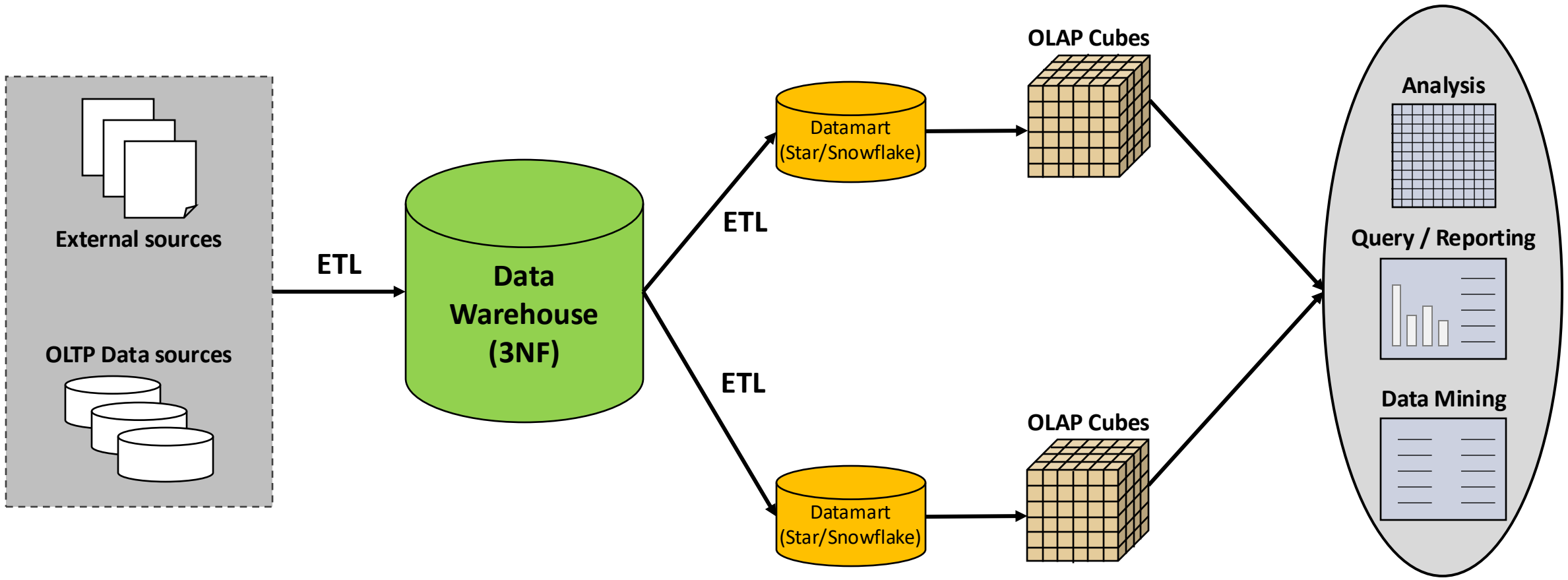
# The Inmon Model




# The Kimball Model



# At the end of the story – Hybrid approach



# Data Warehouse vs Data Mart

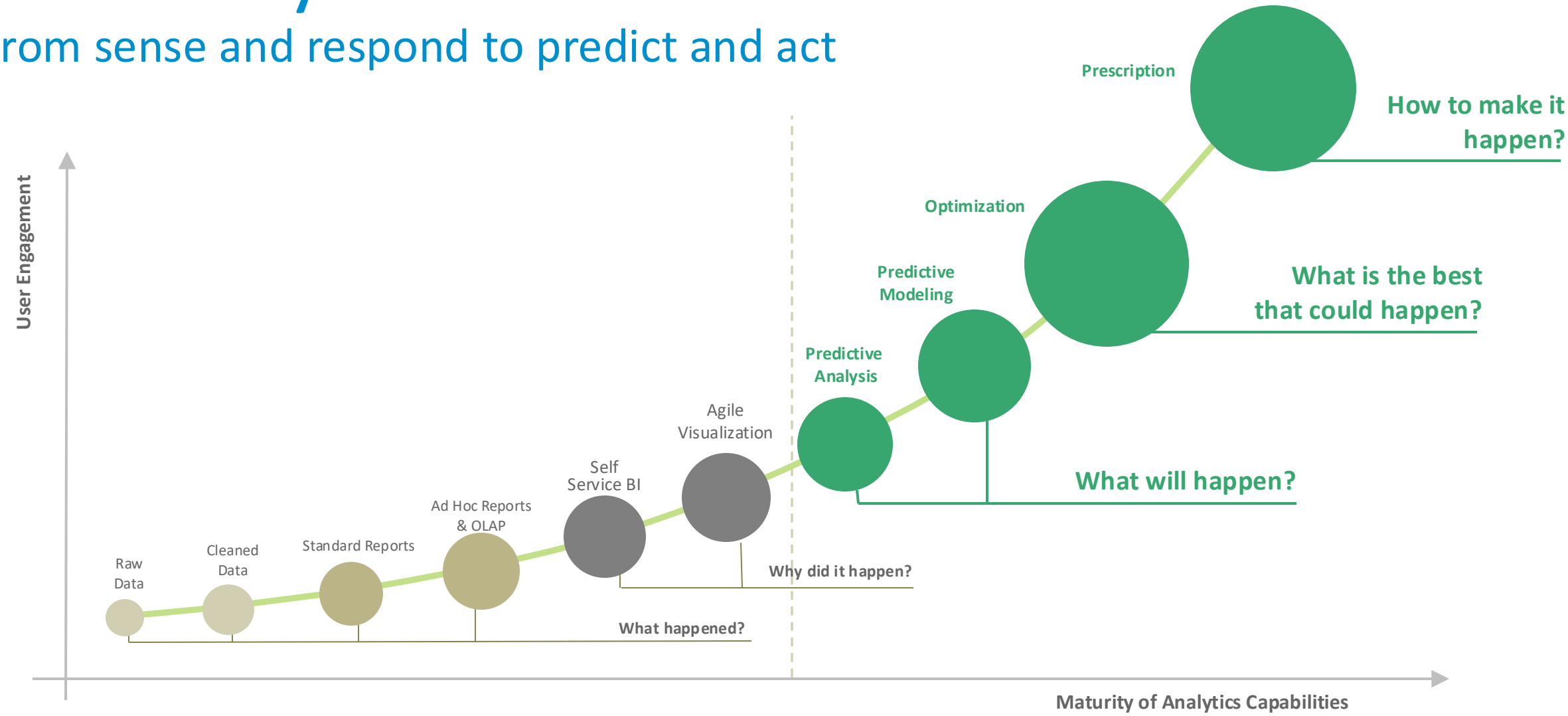
- The **Data Warehouse** is an intermediate **storage** place for different data with a view to building up the decision-making information system.
- It centralize an extremely large volume of data consolidated from the various sources.
- It includes **historical** data assembled from different departments and business units but also **external data sources**

- A **Data Mart** is a sub-element of the **Data Warehouse**.
- A **Data Mart** organize data according to business need or targeted areas and often contains information only relating to a particular LOB.
- They serve LOB users to meet their needs like in the Financial, Marketing or Sales department

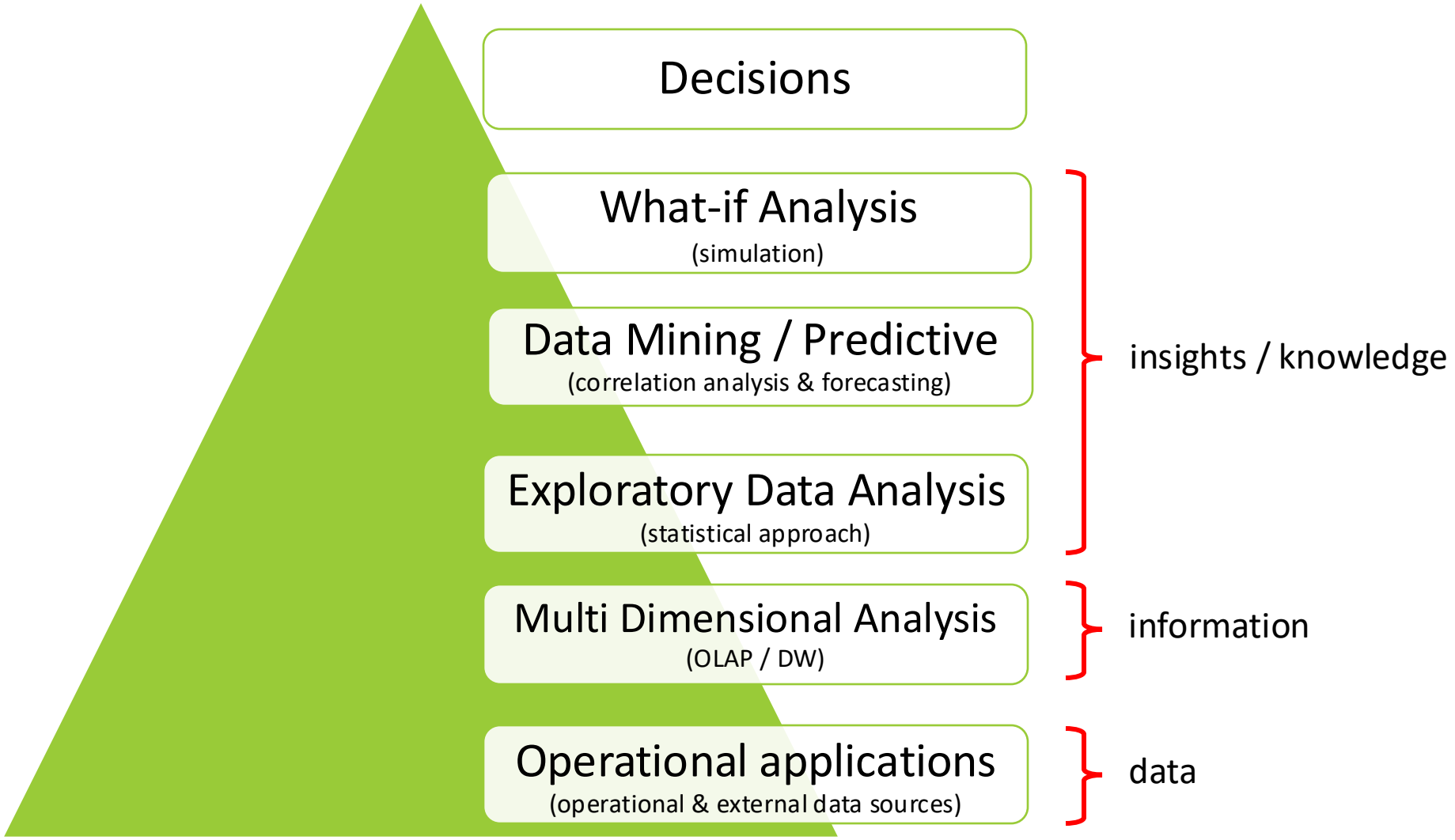
# Modern Analytics

# The Analytics Continuum

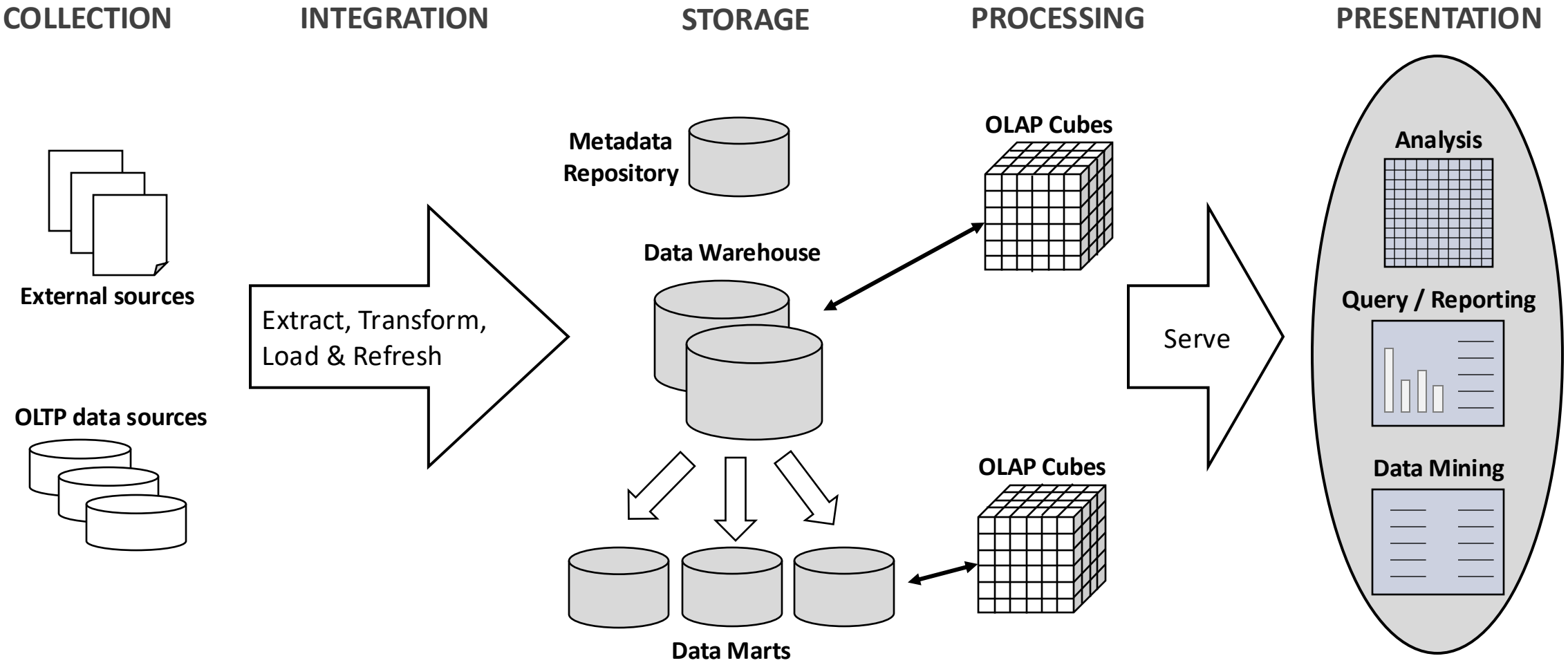
## From sense and respond to predict and act



# Modern Analytics Process



# A « Modern » Analytics architecture



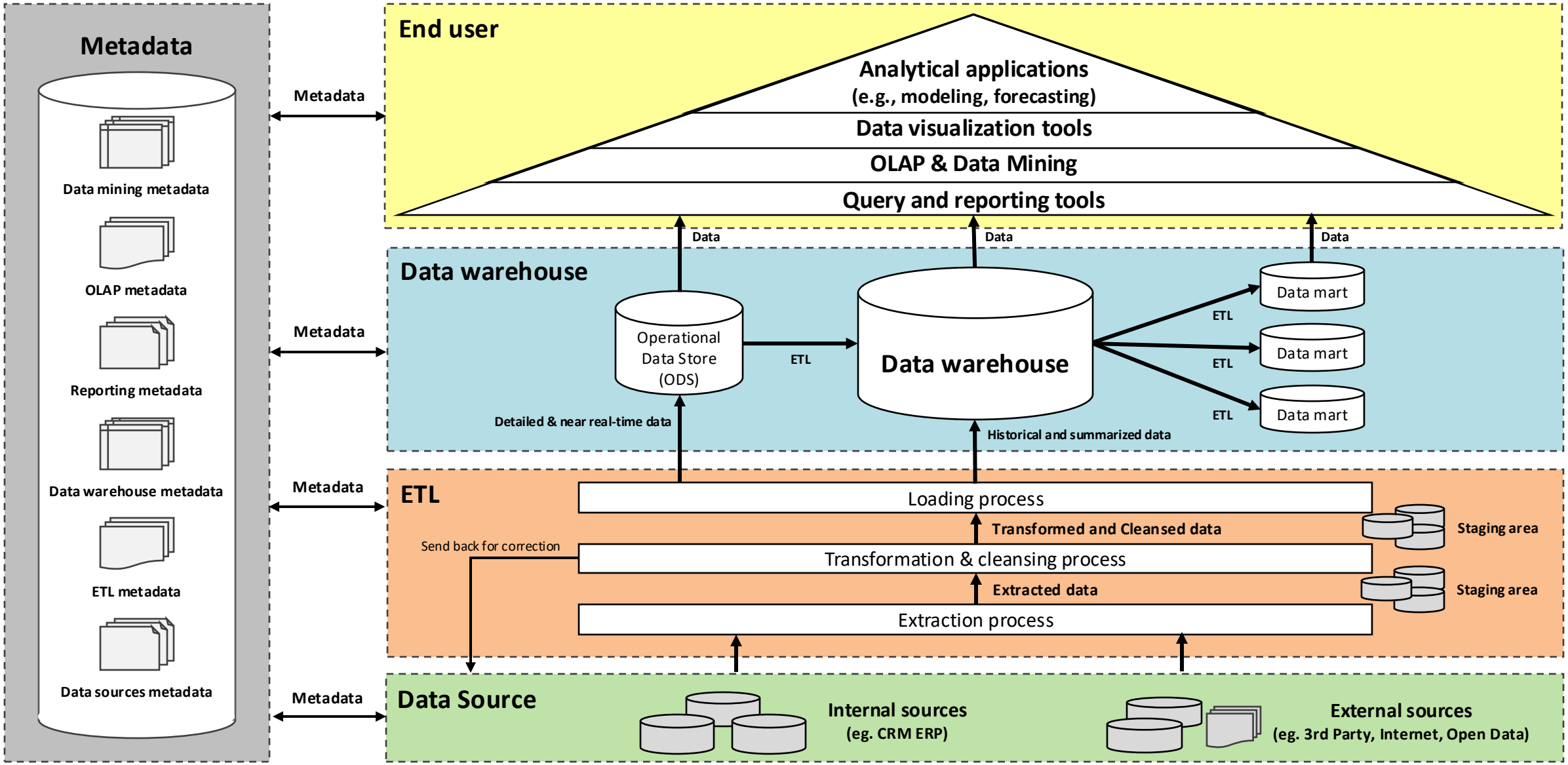

Business Intelligence Process (Wan, J., Chen, T. and Chiu, S. (2005)

# The 5 components of an Analytics architecture

- Collection
  - Identify & connect the data sources: internal, external, file storage, dbms…
- Integration
  - Extract, transform and load data (or Extract, Load then Transform)
- Storage
  - Store and organize / model the data
- Processing
  - Query and analyse data
- Presentation
  - Build visualization / dashboard to support the decision process

Source: A Five-Layered Business Intelligence Architecture (2011)

# The « modern » architecture by layers



# From Transactional to Analytical systems

Or OLTP to OLAP

# Transactional systems (OLTP)

- Support major functions of the company like production, marketing, sales, human resources, finance, accounting, research, …
- Based on traditional relational DBMS (Oracle, DB2, SQL Server etc.) to manage “operational” databases (from GB to TB)
- Enables transactional processing on line of data (OLTP - On line Transactional Processing):
  - Simple, Interactive, Concurrent, Numerous & Repetitive

# Transactional systems (OLTP)

- These OLTP processes (or transactions) usually concern data updates on a limited number of records
- OLTP processes are meant to be ACID (atomic, consistent, isolated, durable) compliant.
- This means one transaction completes before another begins.

# Transactional systems (OLTP)

- Transactional systems data models are usually complex to be easily understood by decision-makers
- The way data is organized and structured is “normalized” (3NF)
- Transactional systems operations cannot be interrupted to answer questions requiring significant calculations

# Analytical systems (OLAP)

- Access current and historical company data in real time, process it and extract the relevant information & insights without impacting / interrupting other line of businesses
- Need to consider increasing amounts of historical data (TB to PB)
- Data is “denormalized” and organized around several axes of analysis or dimensions (time, geographies, categories etc.) according to different detail levels
  - Allows interactive processing for different points of view, level of detail
- Enables analytical processing on line of data (OLAP - On line Analytical Processing):
  - Complex, Interactive, Concurrent, Reduced number & Not predictable

# OLAP vs OLTP

| Features | | OLTP | OLAP |
|---|---|---|---|
| **Concept** | **Design** | Transactional | Analytical |
| | **Modelling** | Entity-Relation | Star or Snowflake |
| **Data** | **Granularity** | Detailed | Aggregated, summarized |
| | **Nature** | Relational | Multidimensional |
| | **Updates** | Updated and up-to-date | Historical, re-calculated |
| | **Size** | GB to TB | TB to PB |
| **Processing** | **Unit** | Simple transactions | Complex queries |
| | **Access** | Read/Write | Read |
| | **Rows accessed** | Ten’s or hundred’s | Million’s or billion’s |
| | **Metric** | Transaction throughput | Response time |
| **Users** | **Type** | Operator / applications users | Analysts / business users |
| | **Number** | Thousand’s | Hundred’s |

# Key terminology (translated)

# Key terminology (translated)

| **English Terminology** | ***Terminologie française*** |
|---|---|
| Business Intelligence (BI) | *Informatique Décisionnelle (ID)* |
| Competitive Intelligence (CI) | *Intelligence Economique (IE)* |
| Customer Relationship Management (CRM) | *Gestion de la Relation Client* |
| Data Mining (DM) | *Fouille de données* |
| Data Warehouse (DW) | *Entrepôt de données (ED)* |
| Decision Support Systems (DSS) | *Systèmes d'aide à la décision (SIAD)* |
| Knowledge Discovery in databases (KDD) | *Extraction de Connaissances dans les Données* |
| On-Line Analytical Processing (OLAP) | *Traitement Analytique en ligne de données* |
| On-Line Transactionnal Processing (OLTP) | *Traitement Transactionnel en ligne de données* |

# Summary

- The concept of Business Intelligence is not new and continues to evolve
- Business Intelligence is here to help take better decisions by providing a clearer and more accurate view of the data (so, not just for creating dashboards or graphs!)
- Analytical systems are not here to replace Transactional systems (different need and purposes) – OLTP vs OLAP, Normalized vs Denormalized