

Data Warehousing

Lecture 6 Metadata and Data Marts

CITS3401
CITS5504

Zeyi Wen

Computer Science and
Software Engineering

School of Maths, Physics
and Computing

Acknowledgement: Some lecture slides are based on online sources.

- **Metadata**
 - ✓ Introduction to Metadata and Metadata Examples
 - ✓ Data Warehouse and Metadata
 - ✓ Metadata Repository and Category
 - ✓ Challenges of Metadata Management
- **Data Mart**
 - ✓ Introduction to Data Mart
 - ✓ Types of Data Mart
 - ✓ Inmon and Kimball's Data Warehouses
 - ✓ Implementing Data Mart
 - ✓ Advantage/ Disadvantage of Data Mart
 - ✓ Use Cases Example

What is metadata?

Metadata is simply **data about data**.

- leads us to detail and represent data
- helps organise, find and understand data.
- index of a book serves as a 'metadata' for the contents in the book

Relation between Metadata and Data warehouse

- Metadata is the road-map to a data warehouse
- Metadata in a data warehouse defines the warehouse objects.
- Metadata acts as a directory to locate the contents of a data warehouse

What is metadata?

Metadata is simply **data about data**.

- leads us to detail and represent data
- helps organise, find and understand data.
- index of a book serves as a 'metadata' for the contents in the book

Typical metadata elements

- Title and description
- Tags and categories
- Who created and when
- Who last modified and when
- Who can access or update

Metadata Examples: Photo

Every time you take a photo with today's cameras a bunch of metadata is gathered and saved

- date and time
- filename
- camera settings
- geolocation

A photo



Data

Filename: Tadzik.jpg
Author: Piotr Kononow
Date: August 15, 2016 6:40:10PM
5,312 × 2,988 JPEG
File: 15.9 megapixels
3,393,448 bytes
(3.2 megabytes)
Camera: Samsung SM-G920F
4.3 mm
Lens: Max aperture f/1.9
(shot wide open)
Auto exposure
Program AE
Exposure: 1/402 sec
f/1.9
ISO 40
Flash: none



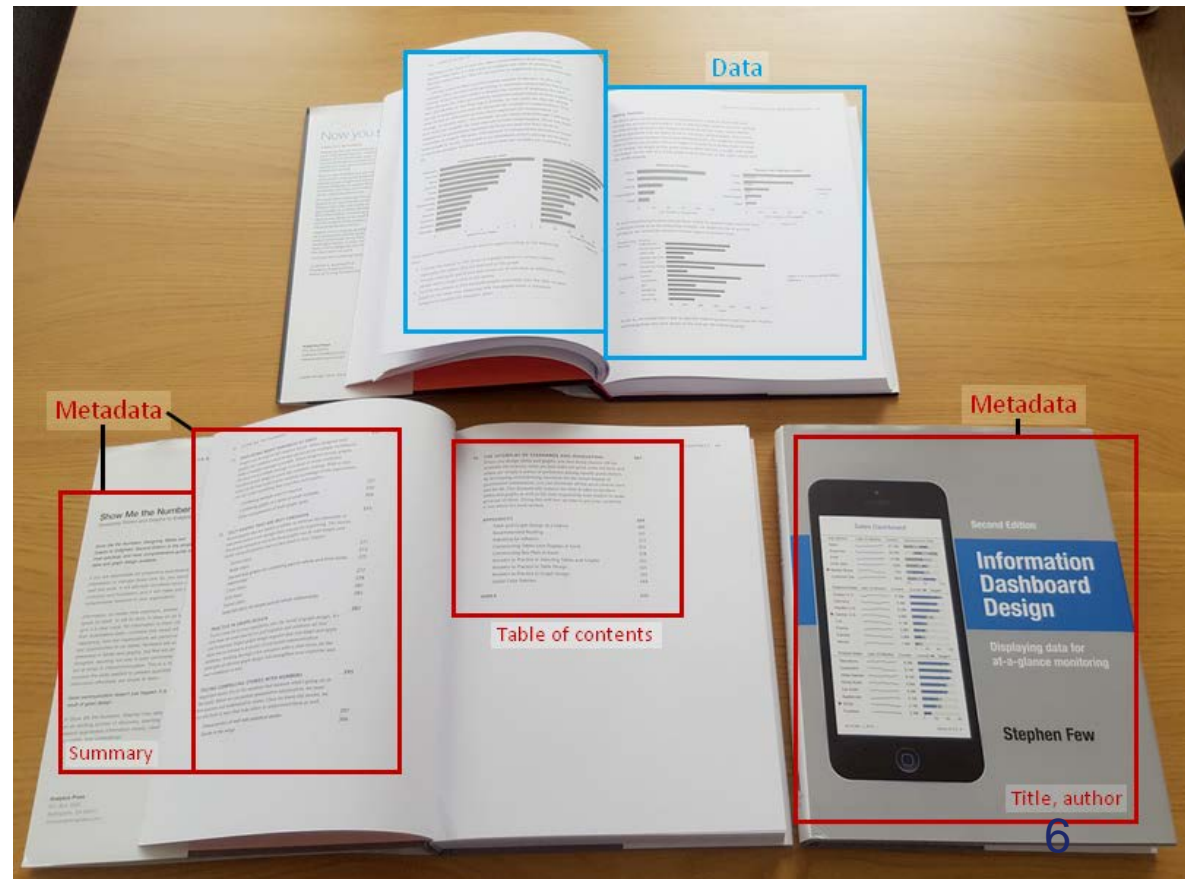
Metadata

Metadata Examples: Book

Each book has a number of standard metadata on the covers and inside. This includes:

- A title
- author name
- publisher and copyright details
- description on a back
- table of contents
- index
- page numbers

A book



Metadata Examples: Database Table

Relational database

employee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Bemdt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Data

Metadata

Column	Data Type	Description
employee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current position title, e.g. Secretary
department_id	int	Employee department. Ref: Departments
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee still

Relational databases (most common type of database) store and provide access not only data but also metadata in a structure called **data dictionary or system catalog**. It holds information about:

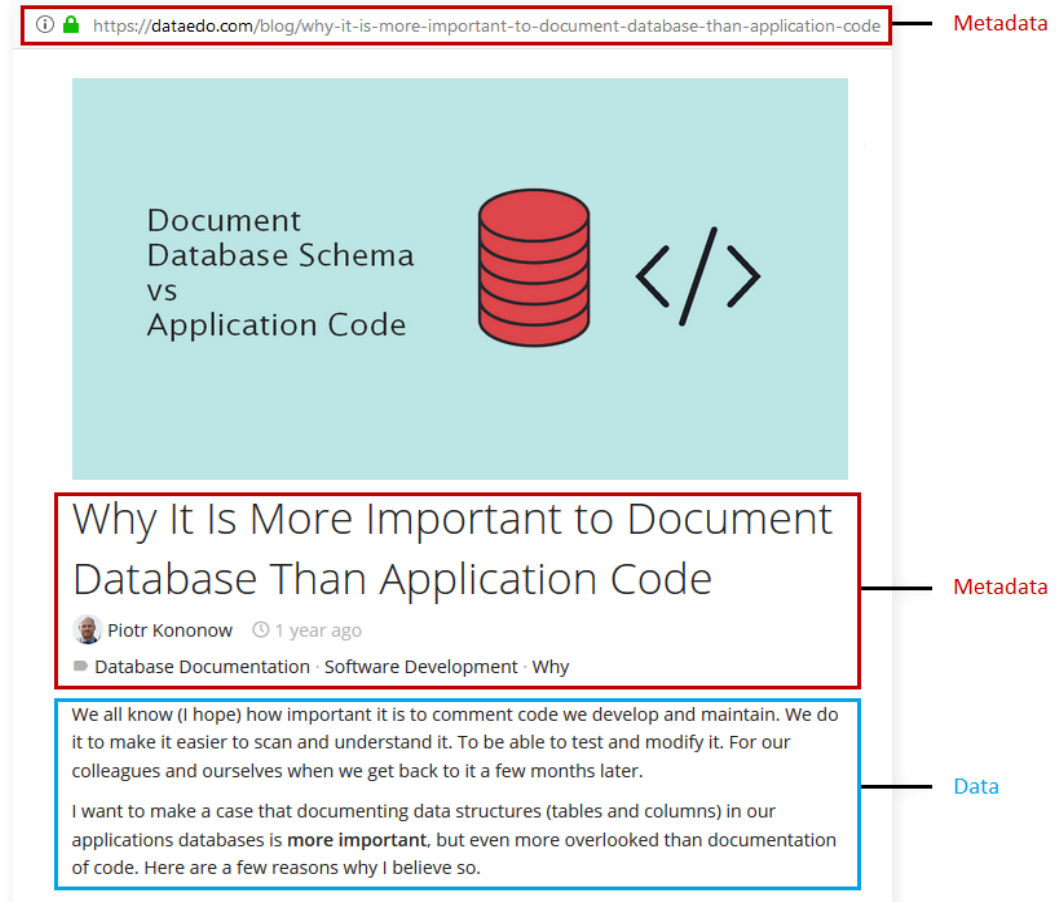
- tables
- columns,
- data types,
- constraints
- table relationships,
- and many more

Metadata Examples: Webpage

Every blog post has standard metadata fields that are usually at before first paragraph. This includes:

A blog post

- A title
- Author id/ name
- published time
- Category
- Tags



The screenshot shows a web browser address bar with the URL `https://dataedo.com/blog/why-it-is-more-important-to-document-database-than-application-code`, highlighted by a red box and labeled "Metadata". Below the address bar is a light blue header image containing the text "Document Database Schema vs Application Code", a red database cylinder icon, and a code symbol `</>`. The main content area has a title "Why It Is More Important to Document Database Than Application Code" in a red box, labeled "Metadata". Below the title is the author "Piotr Kononow" and "1 year ago". A category tag "Database Documentation · Software Development · Why" is also present. The main text of the blog post is enclosed in a blue box and labeled "Data". It begins with "We all know (I hope) how important it is to comment code we develop and maintain. We do it to make it easier to scan and understand it. To be able to test and modify it. For our colleagues and ourselves when we get back to it a few months later." and continues with "I want to make a case that documenting data structures (tables and columns) in our applications databases is **more important**, but even more overlooked than documentation of code. Here are a few reasons why I believe so."

- **Metadata**

- ✓ Introduction to Metadata and Metadata Examples
- ✓ **Data Warehouse and Metadata**
- ✓ Metadata Repository and Category
- ✓ Challenges of Metadata Management

- **Data Mart**

- ✓ Introduction to Data Mart
- ✓ Types of Data Mart
- ✓ Inmon and Kimball's Data Warehouses
- ✓ Implementing Data Mart
- ✓ Advantage/ Disadvantage of Data Mart
- ✓ Use Cases Example

A Data Warehouse has specific metadata requirements.

Metadata that describes tables typically includes:

- Physical and Logical Name
- Type: Fact, Dimension
- Role: Legacy, OLTP, Stage
- DBMS: DB2, Informix, MS SQL Server, Oracle,
- Definition/Description
- Location

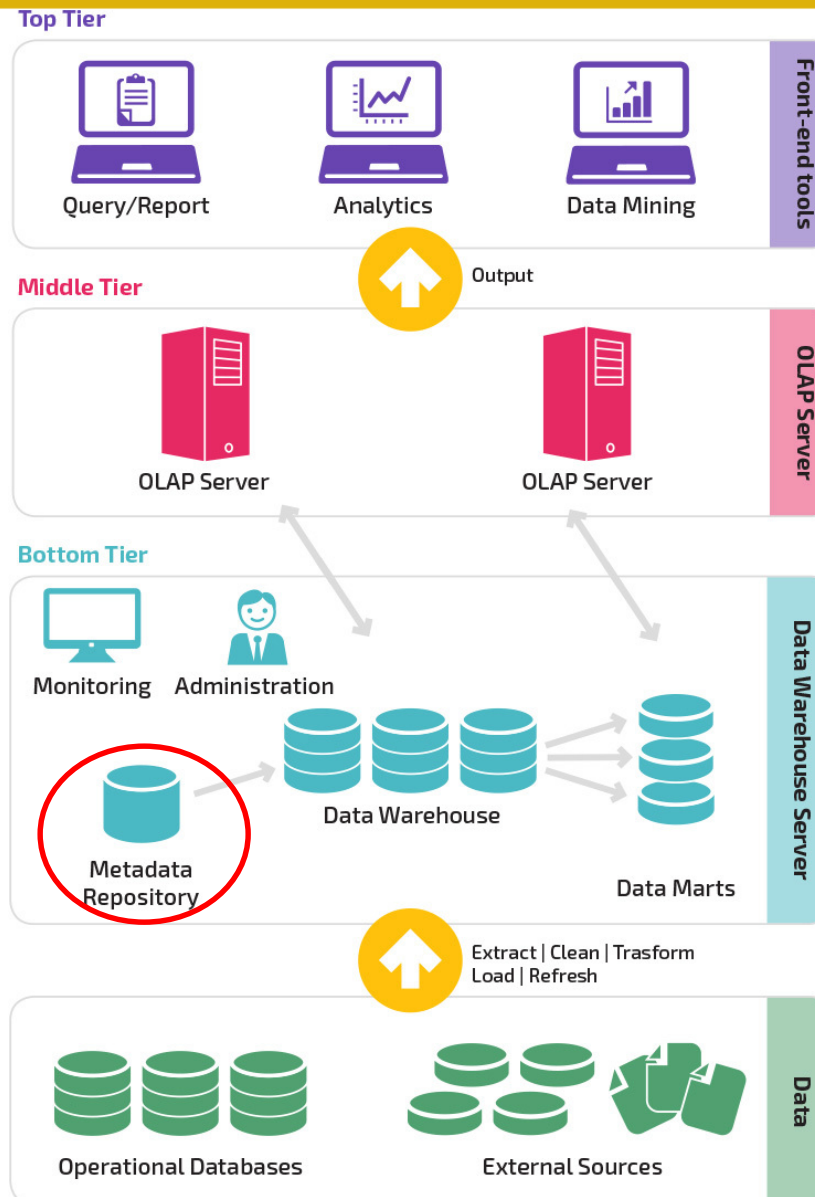
Metadata describes columns within tables:

- Physical and Logical Name
- Order in Table
- Datatype
- Length
- Default Value
- Nullable/Required
- Edit Rules

Data warehousing metadata is best managed through a combination of concerned individual, process and tools:

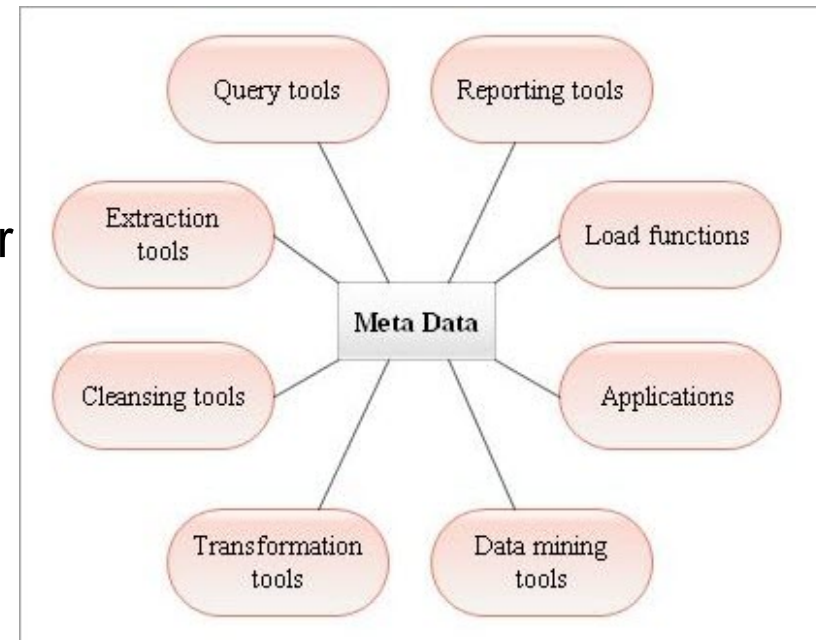
- The **individual** side requires that people be trained in the importance and use of metadata.
- The **process** side incorporates metadata management into the data warehousing and business intelligence life cycle.
- Metadata can be managed through individual **tools**:
 - ☐ Metadata manager/repository
 - ☐ Metadata extract tools
 - ☐ Data modelling and ETL tools
 - ☐ BI Reporting tools

DW Architecture and Metadata



Metadata has a very important **role** in a **data warehouse**:

- helps the decision support system to **locate** the **contents** of the data warehouse
- helps in **decision support system** for mapping of data when data is **transformed** from **operational** environment to **data warehouse** environment
- used for **query tools**
- used in **extraction**, **cleansing**, **reporting**, and **transformation** tools.
- plays an important role in **loading functions**



- **Metadata**
 - ✓ Introduction to Metadata and Metadata Examples
 - ✓ Data Warehouse and Metadata
 - ✓ **Metadata Repository and Category**
 - ✓ Challenges of Metadata Management
- **Data Mart**
 - ✓ Introduction to Data Mart
 - ✓ Types of Data Mart
 - ✓ Inmon and Kimball's Data Warehouses
 - ✓ Implementing Data Mart
 - ✓ Advantage/ Disadvantage of Data Mart
 - ✓ Use Cases Example

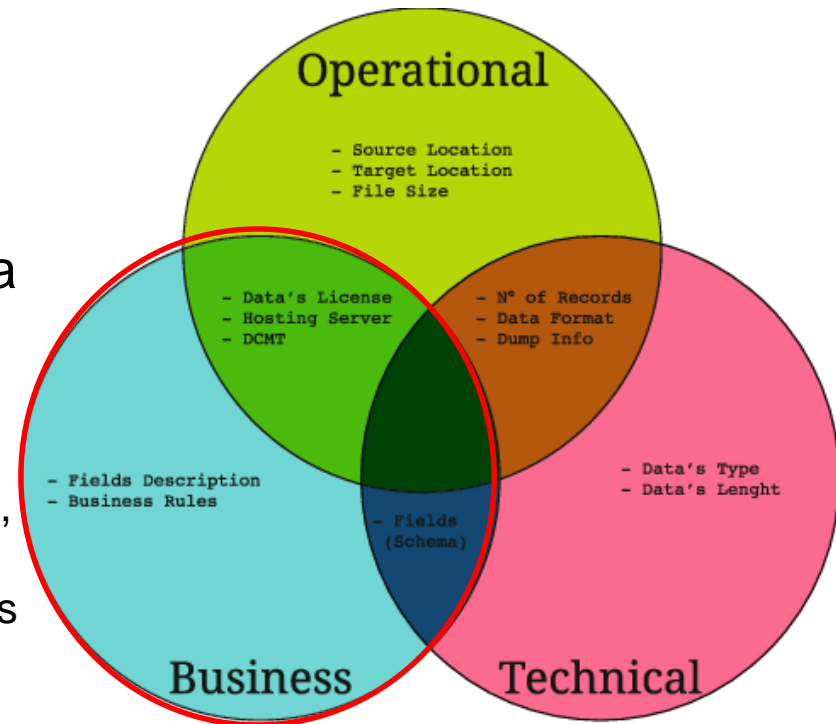
- Metadata is collectively organised in a catalogue called metadata repository.
- It stores **descriptive information** about data model used to store and share.

Each Data Warehouse has one or multiple repositories that hold the following metadata in them:

- **Definition** of the structure of the data warehouse
- **Structure** of the **tables** involved in the warehouse
- **Description** of the dataflow through the warehouse
- **Outcomes** of all querying process e.g., indexing, ETL, etc.
- Information about **who** accesses objects and when
- includes **aggregation, summarising**, etc.

Broadly categorised into three categories: Business metadata, Operational Metadata, and Technical Metadata

- **1) Business Metadata** includes data ownership information, business definition, and changing policies.
 - ☐ Description of information from the business perspective (e.g., weekly sales, or budget variance reports)
 - ☐ Contains high-level definitions of all fields present in the data warehouse, information about cubes, aggregates, Data Marts.
 - ☐ addressed to and used by the data warehouse users
 - ☐ report querying authors, cubes creators, data managers, testers, analysts



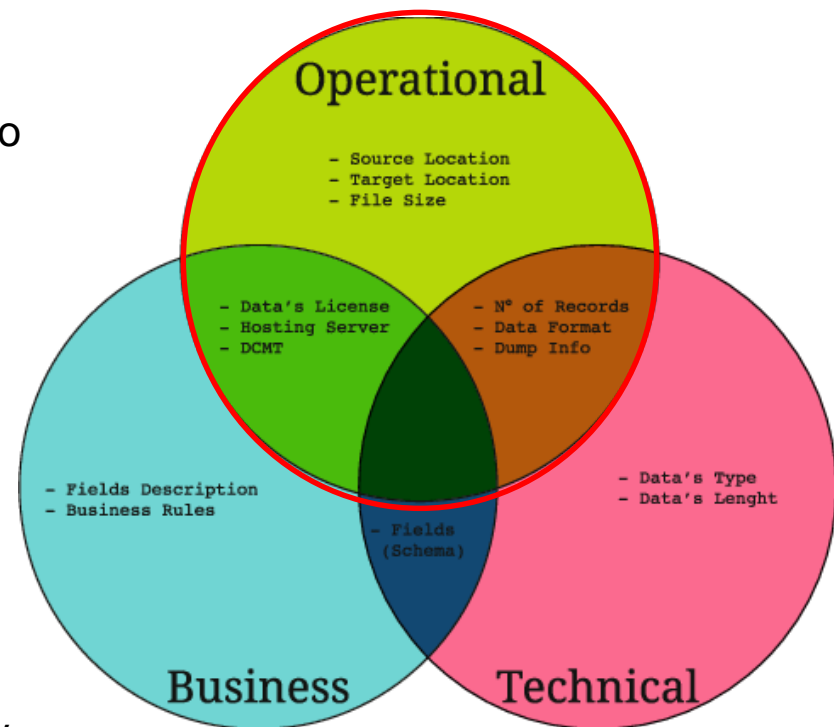
- **2) Technical Metadata** includes structural information such as primary and foreign key attributes and indices, data types and values.
 - ❑ gives information about the **structure** of data, where it **resides**, and other **technical** details related to finding data in its native database.
 - ❑ mainly used by **software tools** to understand and process data
 - ❑ stores data **mapping** and **transformations** from source systems to the data warehouse
 - ❑ illustrates information related to **system functions** and **metadata**



Difference between Business and Technical Metadata

Business Metadata	Technical Metadata
Business terms and definitions for tables and columns	Physical table and column names
Subject area names	Data mapping and transformation logic
Query and report definitions	Source systems information
Report mappings	Foreign keys and indexes
Data Steward information	ETL process names

- **3) Operational Metadata** offers a connection between the metadata repository and the data warehouse
 - ❑ Allows adding physical database columns to the data warehouse tables
 - ❑ includes currency of data and data lineage and enables easier use for **business** and **technical** consumers.
 - Currency of data means whether the data is active, archived, or purged
 - Lineage of data means the history of data migrated and transformation applied on it.
 - ❑ **Benefits** of operational metadata
 - ❑ referenced at a row level of granularity in DW (unlike in meta data repository)
 - ❑ provides a detailed row level explanation of actual information content



Data Warehouse users can **use metadata** in a variety of situations to build, maintain and manage the system

- **Finding Data**

- ☐ Data in warehouses may reach terabyte levels
- ☐ Metadata serves as a roadmap during the development of a DW, making the process of finding a relevant object considerably easier
- ☐ Search data is faster due to the small size of metadata

- **Using Data**

- ☐ helps utilise large sums of data present in the data warehouse without using the actual dataset.
- ☐ gives information about the retrieval, structure, terminology, and regulations governing the data warehouse

- **Metadata**
 - ✓ Introduction to Metadata and Metadata Examples
 - ✓ Data Warehouse and Metadata
 - ✓ Metadata Repository and Category
 - ✓ **Challenges of Metadata Management**
- **Data Mart**
 - ✓ Introduction to Data Mart
 - ✓ Types of Data Mart
 - ✓ Inmon and Kimball's Data Warehouses
 - ✓ Implementing Data Mart
 - ✓ Advantage/ Disadvantage of Data Mart
 - ✓ Use Cases Example

Consider the following questions:

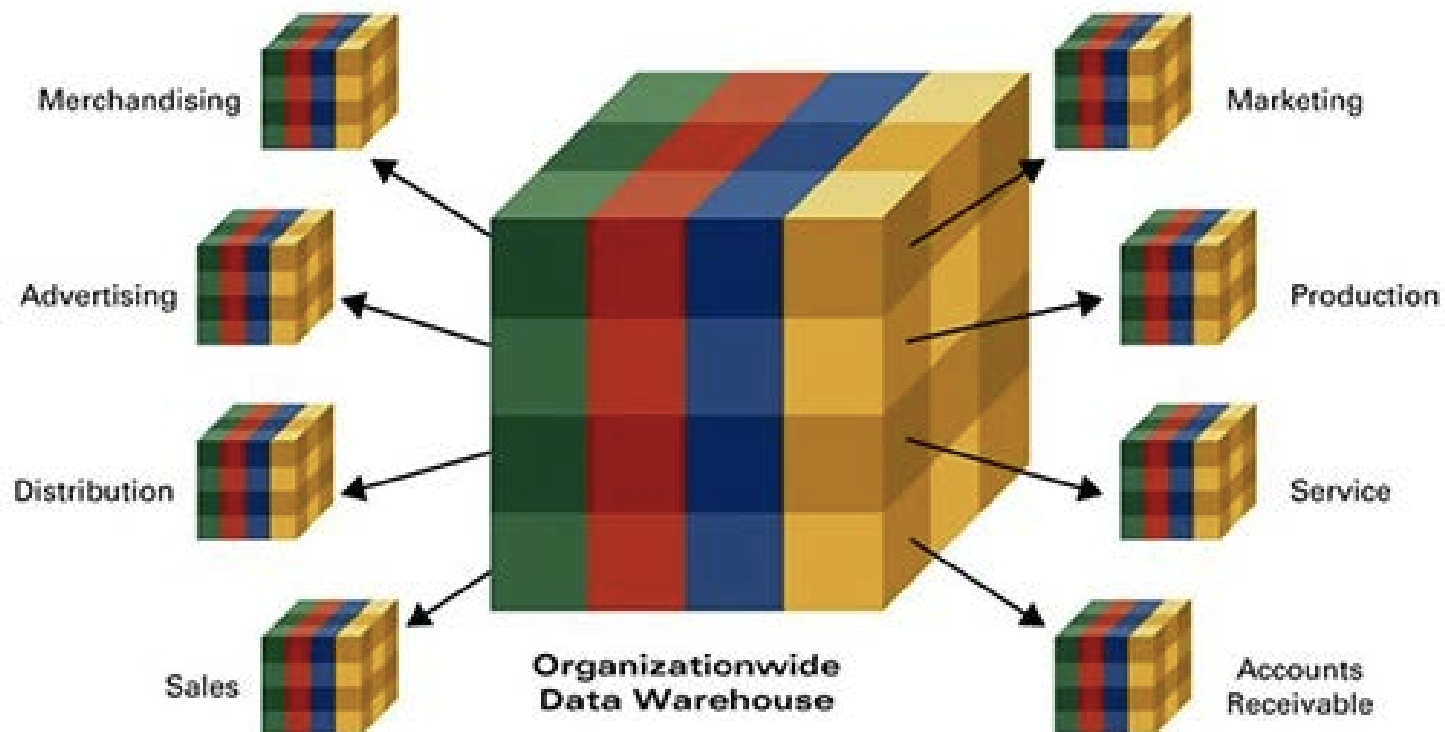
- **How do you ensure** that you are exploiting the metadata you are collecting to the **fullest, possible extent**?
- **How do you make sure** that your metadata is **easily accessible** and **effectively used** across your organisation?
- **How do you ensure** that it is kept **up-to-date** so that new metadata about new data is incorporated?

A few challenges of managing metadata:

- **Disparate information sources**
 - Wide **variety** of sources, in a big organisation, metadata is **scattered** across the organisation
 - Significant portion of every organisation's vital data resides **outside** of its databases
 - Hard to maintain **consistent** and **easy-to-understand** format
- **Enforcing business rules for metadata**
 - Creating a context of enforceable **business rules** around the metadata is an important aspect of maintaining data integrity and usability.
 - Data repositories do not help you understand the **relationships** around the data
 - Clarifying the **dependencies** associated with data is challenging
- **Effective communication**
 - Information about **how to use data** is hard to find or hard to use
 - There are no industry-wide accepted **standards**
 - Clear **communication** is vital to leveraging metadata

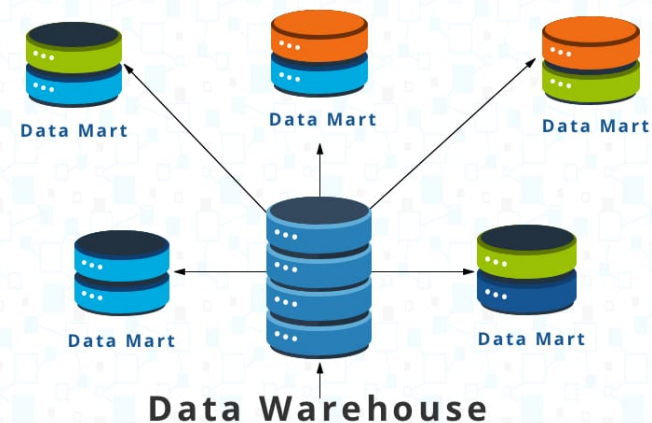
- **Metadata**
 - ✓ Introduction to Metadata and Metadata Examples
 - ✓ Data Warehouse and Metadata
 - ✓ Metadata Repository and Category
 - ✓ Challenges of Metadata Management
- **Data Mart**
 - ✓ **Introduction to Data Mart**
 - ✓ Types of Data Mart
 - ✓ Inmon and Kimball's Data Warehouses
 - ✓ Implementing Data Mart
 - ✓ Advantage/ Disadvantage of Data Mart
 - ✓ Use Cases Example

Data Mart Overview



What is data mart?

- A data mart is a **subset** of a data warehouse oriented to a specific business line.
 - Focuses on particular business domain as marketing or sales.
- Data marts contain repositories of summarised data collected for analysis on a **specific section or unit** within an organisation, for example, the **sales department**.
- A data mart can be created from an existing data warehouse - the top-down approach



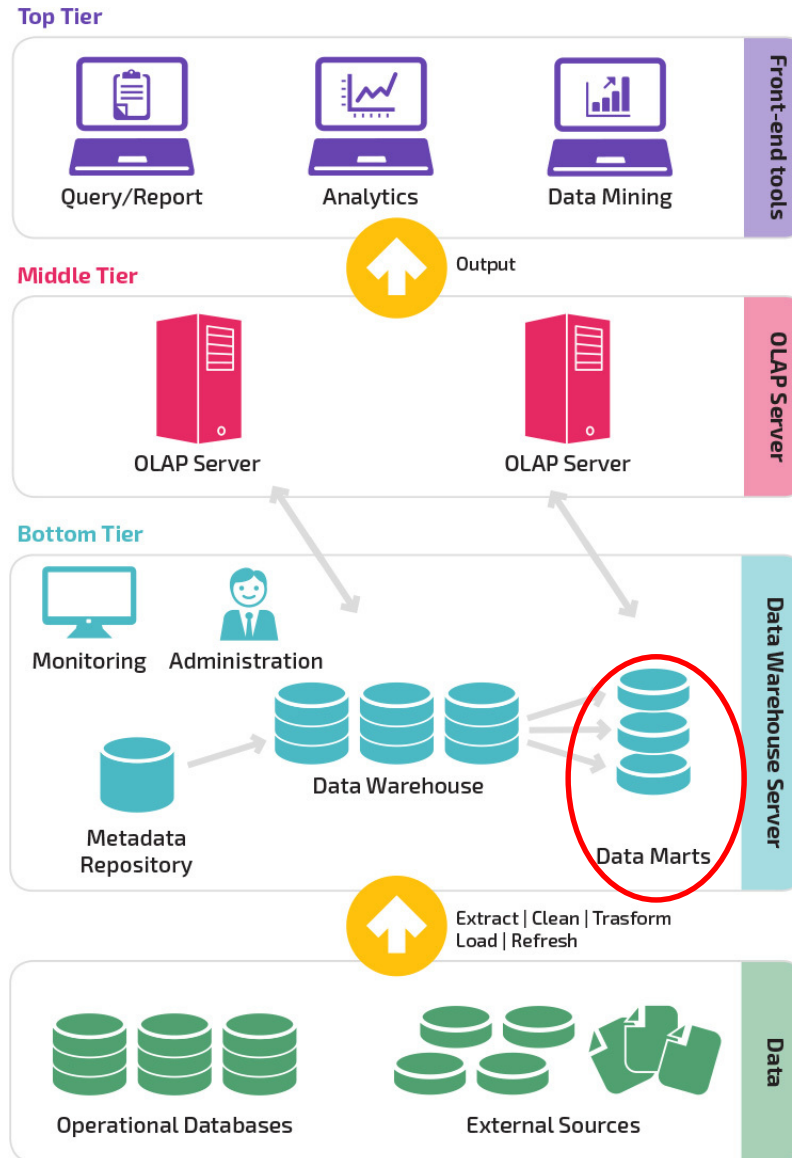
Consider the below reasons to build a data mart

- If you want to partition the data with a set of user access control strategy.
- If a particular department wants to see the query results much faster instead of scanning huge Data Warehouse data.
- If a department wants data to be built on other hardware (or) software platforms.
- If a department wants data to be designed in a manner that is suitable for its tools.

Why do we need Data Mart?

- Data Mart focuses only on functioning of **particular** department of an organisation.
- The data mart can **improve** the **response time** of users due to the reduction of data.
- It provides **easy access** to frequently requested data.
- It is **easier** to **implement** when compare to corporate data warehouse. (The cost of implementing Data Mart is certainly lower)
- Compared to Data Warehouse, a data mart is **agile**. In case of change in model, data mart can be built **quicker** due to a smaller size.
- Data is partitioned and allows very **granular** access control privileges.

DW Architecture and DataMart



- Dedicated **single subject matter**
- Focuses in on the subject matter by **consolidating** and **integrating** information from **various sources**.
- Usually dedicated for a **specific** business function or purpose.
- Built using a dimensional model with **star schema**. This allows data marts to have **multidimensional** analytical capabilities.

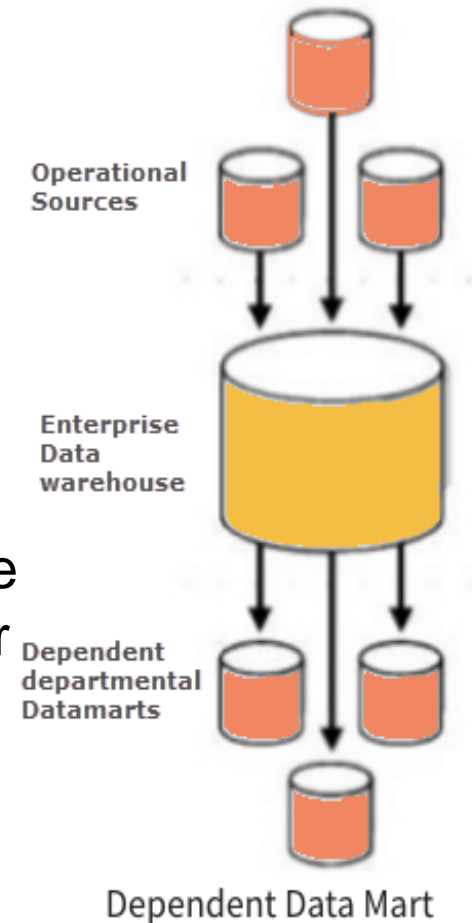
- **Metadata**
 - ✓ Introduction to Metadata and Metadata Examples
 - ✓ Data Warehouse and Metadata
 - ✓ Metadata Repository and Category
 - ✓ Challenges of Metadata Management
- **Data Mart**
 - ✓ Introduction to Data Mart
 - ✓ **Types of Data Mart**
 - ✓ Inmon and Kimball's Data Warehouses
 - ✓ Implementing Data Mart
 - ✓ Advantage/ Disadvantage of Data Mart
 - ✓ Use Cases Example

There are three main types of data mart:

- **Dependent:** Dependent data marts are created by drawing data directly from operational, external or both sources.
- **Independent:** Independent data mart is created without the use of a central data warehouse.
- **Hybrid:** This type of data marts can take data from data warehouses or operational systems.

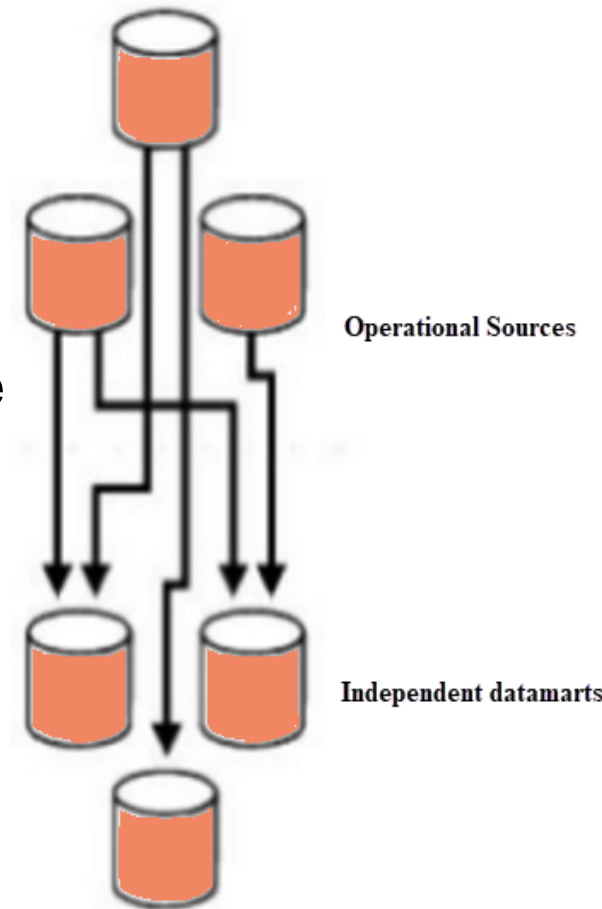
Dependent Data Mart

- A dependent data mart allows sourcing organisation's data from a single Data Warehouse. It is one of the data mart example which offers the benefit of centralisation. If you need to develop one or more physical data marts, then you need to configure them as dependent data marts.
- Dependent Data Mart in data warehouse can be built in **two different ways**. Either where a user can access both the data mart and data warehouse, depending on need, or where access is limited only to the data mart.



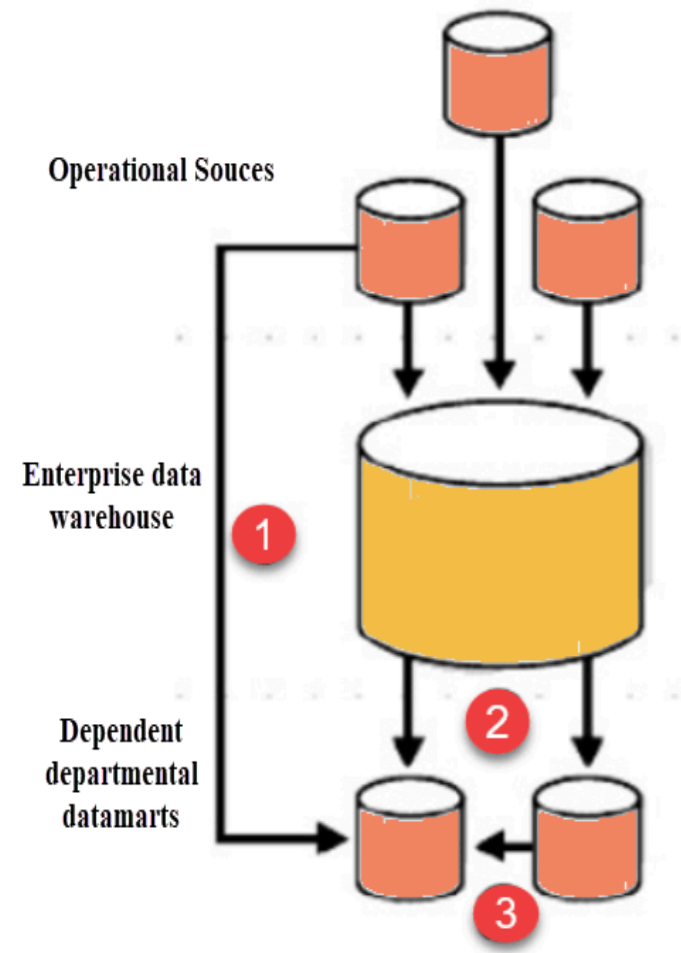
Independent Data Mart

- An independent data mart is created **without** the use of central Data Warehouse. This kind of Data Mart is an ideal option for smaller groups within an organisation.
- An independent data mart has neither a relationship with the enterprise data warehouse nor with any other data mart. In Independent data mart, the data is input separately, and its analyses are also performed autonomously.
- Implementation of independent data marts is antithetical to the motivation for building a data warehouse. First of all, you need a consistent, centralised store of enterprise data which can be analysed by multiple users with different interests who want widely varying information.



Hybrid Data Mart

- A hybrid Data Mart combines input from sources apart from Data Warehouse. This could be helpful when you want ad-hoc integration, like after a new group or product is added to the organisation.
- It is the best data mart example suited for multiple database environments and fast implementation turnaround for any organisation. It also requires least data cleansing effort. Hybrid Data Mart also supports large storage structures, and it is best suited for flexible for smaller data-centric applications.



Hybrid Data Mart

Independent Data Mart

Pros:

- Easier and faster to implement
- More flexible

Cons:

- It is easy to form an information island (because it disengages from the data warehouse, when multiple independent data marts grow to a certain scale, enterprises will only add some information islands because there is no centralised data warehouse coordination.)
- Not a true enterprise-wide solution and can become very costly over time as more and more are added.
- They do not provide the historical depth of a true data warehouse.

Dependent Data Mart

Pros:

- Generally considered a better solution than independent marts. (because dependent marts use the warehouse as their foundation)
- There is no historical limit to the data
- Improve data analysis quality
- Data integrity is ensured.

Cons:

- Take longer and more expensive to implement.

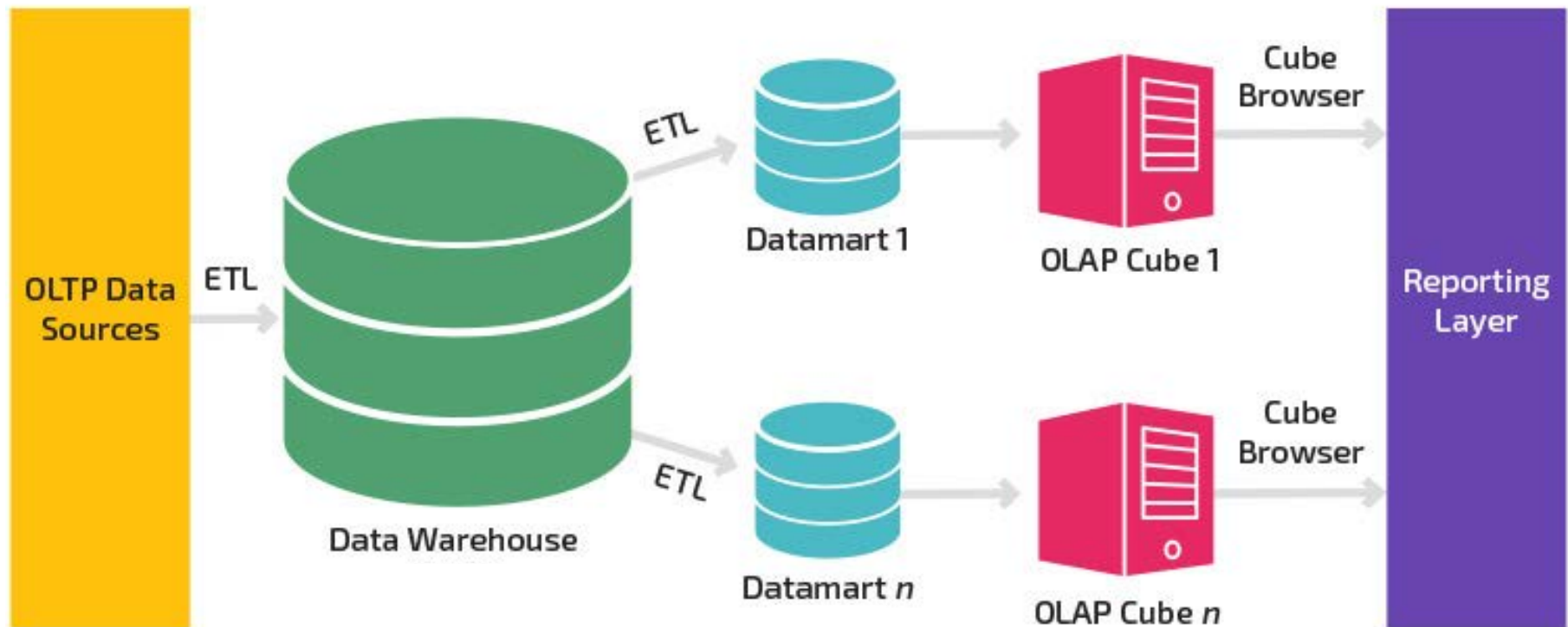
- **Metadata**
 - ✓ Introduction to Metadata and Metadata Examples
 - ✓ Data Warehouse and Metadata
 - ✓ Metadata Repository and Category
 - ✓ Challenges of Metadata Management
- **Data Mart**
 - ✓ Introduction to Data Mart
 - ✓ Types of Data Mart
 - ✓ **Inmon and Kimball's Data Warehouses**
 - ✓ Implementing Data Mart
 - ✓ Advantage/ Disadvantage of Data Mart
 - ✓ Use Cases Example

Bill Inmon and Ralph Kimball are two outstanding pioneers in the field of data warehouse. They have different views on how data warehouses should be designed from the organisation's perspective.

- **Bill Inmon's** approach favours a **top-down** design in which the data warehouse is the centralised data repository and the most important component of an organisation's data systems.
- The **Inmon** approach first builds the centralised corporate data model, and the data warehouse is seen as the physical representation of this model. Dimensional data marts related to specific business lines can be created from the data warehouse when they are needed.

Inmon Model for Data Warehouses

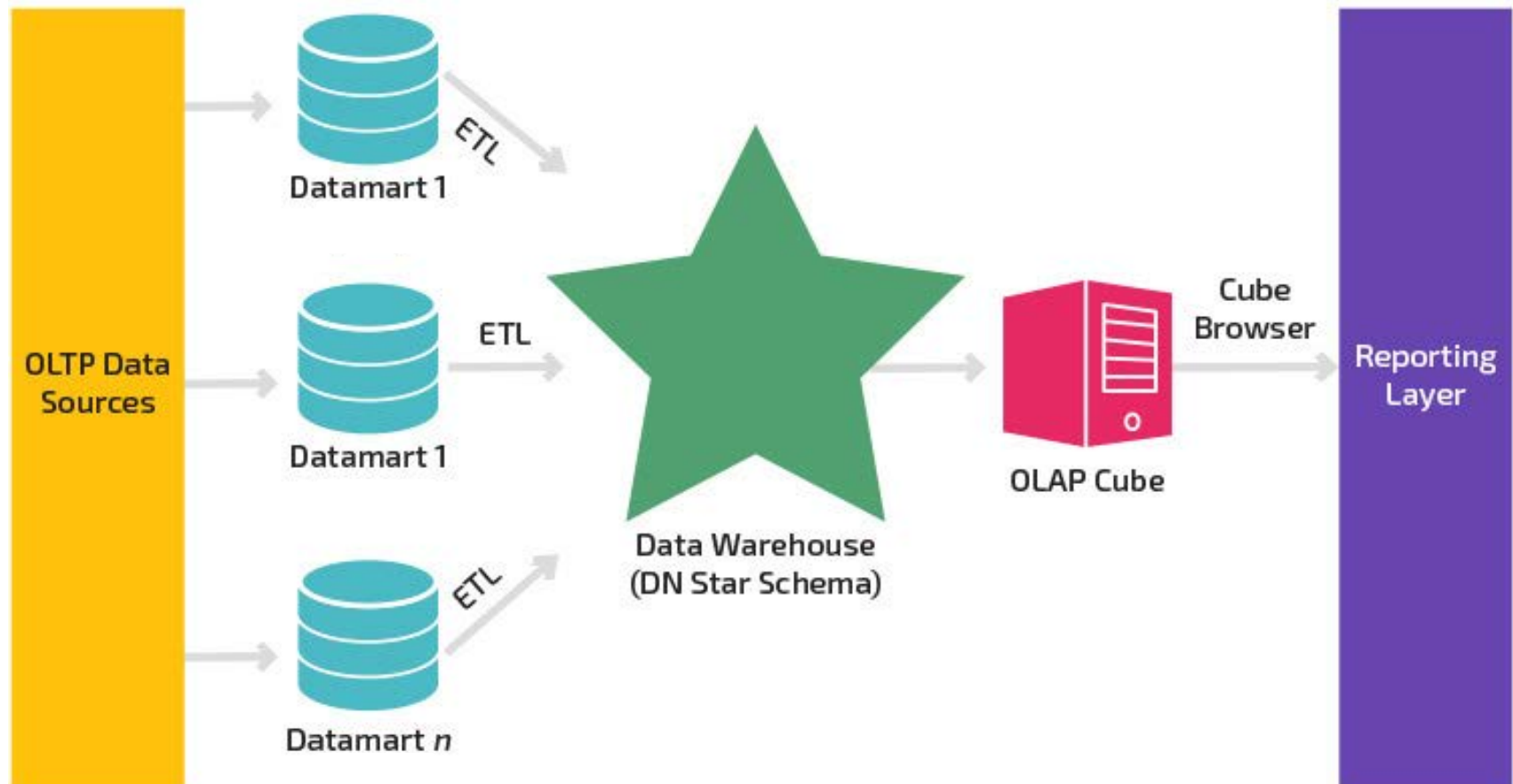
Inmon Model



- In the Inmon model, data in the data warehouse is integrated, meaning the Data Warehouse is the source of the data that ends up in the different Data Marts. This ensures data integrity and consistency across the organisation.
- **Ralph Kimball's** Data Warehouse design starts with the most important business processes. In this approach, an organisation creates **Data Marts** that aggregate relevant data around subject-specific areas. The Data Warehouse is the combination of the organisation's individual data marts.

Kimball's Model for Data Warehouses

Kimball Model

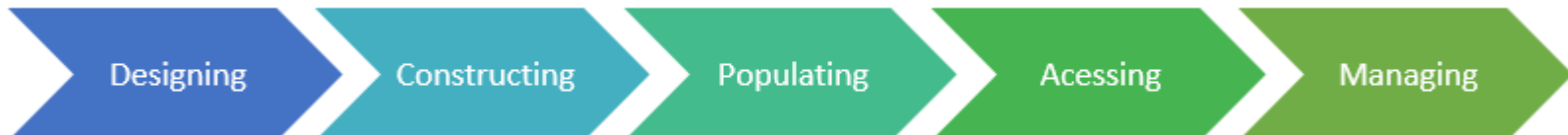


- With the **Kimball** approach, the Data Warehouse is the conglomerate of a number of Data Marts. This is in contrast to **Inmon's** approach, which creates Data Marts based on information in the warehouse. **Kimball** said “the data warehouse is nothing more than the union of all Data Marts.”

- **Metadata**
 - ✓ Introduction to Metadata and Metadata Examples
 - ✓ Data Warehouse and Metadata
 - ✓ Metadata Repository and Category
 - ✓ Challenges of Metadata Management
- **Data Mart**
 - ✓ Introduction to Data Mart
 - ✓ Types of Data Mart
 - ✓ Inmon and Kimball's Data Warehouses
 - ✓ **Implementing Data Mart**
 - ✓ Advantage/Disadvantage of Data Mart
 - ✓ Use Cases Example

Steps of Implementing a Data Mart

- **Designing**
 - Gathering the business & technical requirements and Identifying data sources.
 - Selecting the appropriate subset of data.
 - Designing the logical and physical structure of the Data Mart.
- **Constructing**
 - Implementing the physical database designed in the earlier phase, e.g., table, indexes, views, etc.
- **Populating**
 - Source data to target data mapping
 - Extraction of source data
 - Cleaning and transformation operations on the data
 - Loading data into the Data Mart
 - Creating and storing metadata



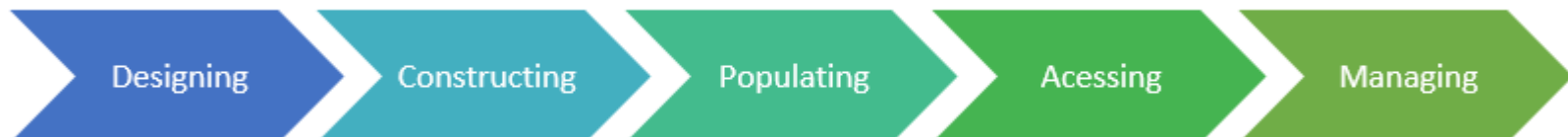
Steps of Implementing a Data Mart

- **Accessing**

- Querying data, creating reports, charts.
- Set up and maintain database structures
- Set up API and interfaces if required.

- **Managing**

- Ongoing user access management.
- System optimisations and fine-tuning to achieve the enhanced performance.
- Adding and managing fresh data into the Data Mart.
- Planning recovery scenarios and ensure system availability in the case when the system fails.



- **Managing** big data and gaining valuable business insights
- **Efficient access** - A data mart is a time-saving solution for accessing a specific set of data for business intelligence
- **Inexpensive data warehouse alternative** - Data marts can be an inexpensive alternative to developing an enterprise data warehouse, where required data sets are smaller.
- **Improve data warehouse performance** - Significantly can reduce analytics processing
- **Data maintenance** - Different departments can own and control their data.

Disadvantage of a Data Mart

- Many enterprises create too many disparate and **unrelated** Data Marts without much benefit.
 - ❑ Creating too many **data marts** become cumbersome sometimes.
- Data Mart cannot provide company-wide data analysis as their **data set is limited**.
 - ❑ Data Mart stores the data related only to specific function
- Data Marts are meant for **small business** needs. Increasing the size of data marts will **decrease** its **performance**.

Comparison with Data Warehouse

	Data Mart	Data Warehouse
Focus	A single subject or functional organisation area	Enterprise-wide repository of disparate data sources
Data Sources	Relatively few sources linked to one line of business	Many external and internal sources from different areas of an organisation
Size	Less than 100 GB	100 GB minimum but often in the range of terabytes for large organisations
Normalization	No preference between a normalised and denormalised structure	Modern warehouses are mostly denormalised for quicker data querying and read performance
Decision Types	Tactical decisions pertaining to particular business lines and ways of doing things	Strategic decisions that affect the entire enterprise

Comparison with Data Warehouse

	Data Mart	Data Warehouse
Cost	typically from \$10,000 upwards	Varies but often greater than \$100,000; for cloud solutions costs can be dramatically lower as organisations pay per use
Setup Time	3-6 months	At least a year for on-premise warehouses; cloud data warehouses are much quicker to set up
Data Held	Typically summarised data	Raw data, metadata, and summary data

- **Metadata**
 - ✓ Introduction to Metadata and Metadata Examples
 - ✓ Data Warehouse and Metadata
 - ✓ Metadata Repository and Category
 - ✓ Challenges of Metadata Management
- **Data Mart**
 - ✓ Introduction to Data Mart
 - ✓ Types of Data Mart
 - ✓ Inmon and Kimball's Data Warehouses
 - ✓ Implementing Data Mart
 - ✓ Advantage/Disadvantage of Data Mart
 - ✓ **Use Cases Example**

Data Mart

- Marketing analysis and reporting favor a Data Mart approach because these activities are typically performed in **a specialised business unit**, and do not require enterprise-wide data.
- A financial analyst can use a finance Data Mart to carry out **financial reporting**.

Data Warehouse

- A company considering an expansion needs to incorporate data from a variety of data sources across the organisation to come to an informed decision. This requires a data warehouse that aggregates data from sales, marketing, store management, customer loyalty, supply chains, etc.
- Many factors drive profitability at an insurance company. An insurance company reporting on its profits needs a centralised data warehouse to combine information from its claims department, sales, customer demographics, investments, and other areas.

References and Readings

- [What is Metadata](#)
- [Data Mart Tutorial](#)
- [Data Mart or Data Warehouse](#)



Copyright Notice

Material used in this recording may have been reproduced and communicated to you by or on behalf of **The University of Western Australia** in accordance with section 113P of the *Copyright Act 1968*.

Unless stated otherwise, all teaching and learning materials provided to you by the University are protected under the Copyright Act and is for your personal use only. This material must not be shared or distributed without the permission of the University and the copyright owner/s.