

Machine Learning and Data Mining

Giorgio Medico and Carlo Gervasi

fall semester 2024

Contents

I	Part 1 : Data Mining	4
1	Introduction to Business Intelligence and Data Warehouse	5
1.1	Basic Concepts and Definitions	5
1.1.1	Authoritative Definitions	5
1.1.2	Data Warehouse Characteristics	5
1.2	BI Infrastructure and Architecture	6
1.2.1	The BI Pyramid	6
1.2.2	Platform Requirements	6
1.3	Application Fields	6
1.3.1	Commercial Applications	6
1.3.2	Service Applications	6
2	Data Processing and OLAP	7
2.1	OLAP Fundamentals	7
2.1.1	Typical Queries	7
2.1.2	OLTP vs OLAP Comparison	7
2.2	Multidimensional Analysis	7
2.2.1	Data Cube Example	7
2.2.2	OLAP Operators	8
2.3	Data Mart	8
2.3.1	Key Characteristics	8
3	ETL Process	9
3.1	Overview and Components	9
3.2	Extraction Phase	9
3.2.1	Extraction Methods	9
3.2.2	Data Source Types	9
3.3	Data Cleansing	10
3.3.1	Common Quality Issues	10
3.3.2	Cleansing Solutions	10
3.4	Transformation and Loading	11
3.4.1	Transformation Types	11
3.4.2	Loading Strategies	11
3.4.3	Loading Considerations	11
4	Data Warehouse Architectures	12
4.1	Architectural Requirements	12
4.1.1	Core Requirements	12
4.2	Architectural Models	13
4.2.1	Single-Layer Architecture	13
4.2.2	Two-Layer Architecture	13
4.2.3	Three-Layer Architecture	14
4.2.4	Architecture Comparison	14
4.3	Implementation Considerations	14
4.3.1	Technology Selection	14

4.3.2	Performance Optimization	14
4.3.3	Management and Monitoring	15
5	Dimensional Fact Model	16
5.1	Conceptual Design Approaches	16
5.1.1	Design Methodologies	16
5.2	Core Components	16
5.2.1	Facts and Measures	16
5.2.2	Dimensions and Attributes	17
5.3	Advanced Concepts	17
5.3.1	Hierarchies and Relationships	17
5.3.2	Advanced Features	17
5.4	Logical Implementation	18
5.4.1	Star Schema	18
5.4.2	Snowflake Schema	18
5.4.3	Schema Comparison	18
II	Part 2 : Machine Learning	19
6	Machine Learning	20

Part I

Part 1 : Data Mining

Chapter 1

Introduction to Business Intelligence and Data Warehouse

1.1 Basic Concepts and Definitions

Definition

Business Intelligence (BI) is a set of methodologies, processes, architectures, and technologies that:

- Transform raw data into useful information
- Support effective business strategies
- Deliver the right information to the right people, at the right time, through the right channel

1.1.1 Authoritative Definitions

According to Gartner: "Business intelligence is an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance."

According to Forrester Research: "Business Intelligence is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making."

1.1.2 Data Warehouse Characteristics

Definition

The Data Warehouse (DWH) serves as an optimized repository for decision-making and represents a specific type of Decision Support System (DSS).

Key features:

- Subject-oriented:
 - Focuses on enterprise-specific concepts
 - Examples: customers, products, sales
- Integrated and consistent:
 - Integrates data from different and heterogeneous sources
 - Provides unified view of all data

- Time variant and non-volatile:
 - Tracks and records changes over time
 - Data is static and read-only once committed
 - Retained for future reporting

1.2 BI Infrastructure and Architecture

1.2.1 The BI Pyramid

Business Intelligence is structured according to a pyramid representing different levels of data processing:

Level	Components
DECISIONS	What-if Analysis and simulation modeling
KNOWLEDGE	Data Mining and learning models
INFORMATION	OLAP and data warehouse
DATA	Operational databases and data sources

1.2.2 Platform Requirements

An effective BI platform requires:

- Ad-hoc Hardware
- Network infrastructure
- Databases
- Data Warehouse
- Front-end software (Data Visualization)

1.3 Application Fields

1.3.1 Commercial Applications

- Trade:
 - Sales and claims analyses
 - Shipment and inventory control
 - Customer care and public relations
- Financial Services:
 - Risk analysis
 - Fraud detection
 - Credit card management

1.3.2 Service Applications

- Transport: Vehicle management
- Telecommunications:
 - Customer profile analysis
 - Network performance analysis
- Healthcare: Patient admission and discharge analysis

Chapter 2

Data Processing and OLAP

2.1 OLAP Fundamentals

Definition

Online Analytical Processing (OLAP) allows users to interactively navigate data warehouse information by exploiting the multidimensional model. Data is analyzed at different levels of aggregation through subsequent OLAP operators.

2.1.1 Typical Queries

Common OLAP analytical questions include:

- Which products maximize profit?
- What is the total revenue per product category and state?
- What is the relationship between profits gained by different products?
- What is the revenue trend in the last three years?

2.1.2 OLTP vs OLAP Comparison

Feature	OLTP	OLAP
System Type	Transaction processing	Analytical processing
Data Scope	Few records per transaction	Millions of records per query
Processing Pattern	Fixed transactions	Dynamic analyses
Data Access	Read and write	Mainly read-only
Update Frequency	Continuous	Periodic
Optimization	For transaction speed	For query performance

2.2 Multidimensional Analysis

2.2.1 Data Cube Example

Here's an example of how data is organized in a multidimensional view:

Category	Type	Product	2015		2014	
			Jan	Feb	Jan	Feb
Food and Beverages	Dairy products	White milk	90	90	60	80
		Chocolate milk	60	80	70	70
		Yogurt XY	20	30	30	35
	Beverages	Cola	20	10	35	30
		Orange Juice X	50	60	60	45

2.2.2 OLAP Operators

- **Roll-up:**
 - Increases data aggregation
 - Removes detail level from hierarchy
 - Example: Product → Type → Category
- **Drill-down:**
 - Reduces data aggregation
 - Adds detail level to hierarchy
 - Example: Category → Type → Product
- **Slice-and-dice:**
 - Slice: Sets one dimension to specific value
 - Dice: Reduces data set by selection criteria
- **Pivot:**
 - Changes data presentation layout
 - Reorganizes multidimensional view
- **Drill-across:**
 - Links concepts between related cubes
 - Enables cross-cube analysis
- **Drill-through:**
 - Accesses detailed operational data
 - Connects aggregates to source data

2.3 Data Mart

Definition

A Data Mart is a subset or an aggregation of data stored in a primary data warehouse, focused on a specific business area, corporate department, or category of users.

2.3.1 Key Characteristics

- Serves as building block for incremental DWH development
- Focuses on specific business area or user group
- Provides better query performance due to smaller size
- Enables departmental data management

Note

Data Marts can be dependent (sourced from enterprise DWH) or independent (sourced directly from operational systems).

Chapter 3

ETL Process

3.1 Overview and Components

Definition

The ETL (Extraction, Transformation, Loading) process extracts, integrates, and cleans data from operational sources to feed the Data Warehouse layer. This process ensures data quality and consistency in the DWH.

3.2 Extraction Phase

3.2.1 Extraction Methods

- **Static Extraction:**
 - Initial DWH population
 - Complete snapshot of operational data
 - Used for first-time setup
- **Incremental Extraction:**
 - Regular DWH updates
 - Captures changes since last extraction
 - Based on timestamps or triggers

3.2.2 Data Source Types

- **Structured Data:**
 - Relational databases
 - Fixed format files
 - Well-defined schemas
- **Unstructured Data:**
 - No pre-defined data model
 - Variable formats
 - Requires more complex processing

3.3 Data Cleansing

3.3.1 Common Quality Issues

1. Data Duplication:

- Multiple customer registrations
- Redundant records
- Inconsistent representations

2. Missing or Incomplete Data:

- Null values
- Partial information
- Empty required fields

3. Format Issues:

- Inconsistent date formats
- Mixed number representations
- Character encoding problems

4. Value Problems:

- Invalid entries (e.g., "30th Feb 2016")
- Out-of-range values
- Logical inconsistencies

3.3.2 Cleansing Solutions

• Dictionary-based Techniques:

- Lookup tables for standardization
- Abbreviation resolution
- Format normalization

• Approximate Matching:

- Similarity functions
- Fuzzy matching
- Pattern recognition

• Custom Algorithms:

- Business rule validation
- Domain-specific checks
- Complex data verification

Note

Example of lookup table for standardization:

Short Form	Standard Form
IT	Italy
FR	France
DE	Germany
GR	Greece
ES	Spain

3.4 Transformation and Loading

3.4.1 Transformation Types

1. **Basic Conversion:**

- Data type changes
- Format standardization
- Unit conversion
- Example: Date formatting (12/11/2018 → 20181112)

2. **Data Enrichment:**

- Derived calculations
- Attribute combination
- Value enhancement
- Example: Profit = Revenue - Cost

3. **Structural Transformation:**

- Denormalization
- Key generation
- Hierarchy building
- Example: Combining customer name fields

3.4.2 Loading Strategies

- **Full Refresh:**

- Complete DWH rewrite
- Suitable for initial loads
- Ensures data consistency

- **Incremental Update:**

- Adds only changed data
- Preserves existing records
- More efficient for regular updates

3.4.3 Loading Considerations

- Performance optimization
- Data integrity maintenance
- Recovery procedures
- Load monitoring and validation

Chapter 4

Data Warehouse Architectures

4.1 Architectural Requirements

Definition

DWH architectures must satisfy specific requirements to ensure effective data management, analysis capabilities, and system maintainability.

4.1.1 Core Requirements

1. Separation

- Analytical and transactional processing kept separate
- Dedicated resources for each type of operation
- Optimized performance for both workloads

2. Scalability

- Easy hardware/software upgrades
- Handles increasing data volumes
- Accommodates growing user base
- Supports additional applications

3. Extensibility

- Integrates new technologies
- Adds functionality without redesign
- Supports evolving business needs

4. Security

- Access control and monitoring
- Data protection mechanisms
- Audit trail capabilities

5. Administrability

- Simplified management procedures
- Efficient maintenance processes
- Clear operational procedures

4.2 Architectural Models

4.2.1 Single-Layer Architecture

Definition

A simple architecture that minimizes data storage by removing redundancies, where the source layer is the only physical layer available.

- **Characteristics:**

- DWH implemented as virtual view
- Uses middleware for data access
- Minimal data duplication

- **Advantages:**

- Reduced storage requirements
- Simpler maintenance
- Always current data

- **Disadvantages:**

- Performance limitations
- No separation of workloads
- Complex query processing

4.2.2 Two-Layer Architecture

Definition

Provides clear separation between source systems and the data warehouse through distinct layers for different functionalities.

- **Layer Structure:**

- Source Layer (operational data)
- Data Staging Area (ETL processing)
- Data Warehouse Layer (analytical data)
- Analysis Layer (user access)

- **Key Features:**

- Separate physical storage
- Dedicated processing areas
- Clear functional separation

Note

The two-layer architecture represents a balance between complexity and functionality, making it a popular choice for many organizations.

4.2.3 Three-Layer Architecture

Definition

Extends the two-layer architecture by adding a reconciled layer between sources and the data warehouse, providing additional data integration capabilities.

- **Layer Components:**

1. **Source Layer:**

- Internal data sources
- External data sources
- Operational systems

2. **Reconciled Layer:**

- Integrated operational data
- Common data model
- Cleansed and standardized data

3. **Data Warehouse Layer:**

- Enterprise data warehouse
- Data marts
- Analytical databases

4.2.4 Architecture Comparison

Feature	Single-Layer	Two-Layer	Three-Layer
Complexity	Low	Medium	High
Data Redundancy	Minimal	Moderate	Higher
Performance	Limited	Good	Excellent
Maintenance	Simple	Moderate	Complex
Scalability	Limited	Good	Excellent
Data Integration	Basic	Good	Advanced

4.3 Implementation Considerations

4.3.1 Technology Selection

- Hardware infrastructure requirements
- Database management systems
- ETL tools and middleware
- Analysis and reporting tools

4.3.2 Performance Optimization

- Data partitioning strategies
- Index design and management
- Query optimization techniques
- Cache management

4.3.3 Management and Monitoring

- System health monitoring
- Performance metrics tracking
- Capacity planning
- Disaster recovery procedures

Chapter 5

Dimensional Fact Model

5.1 Conceptual Design Approaches

Definition

The DFM (Dimensional Fact Model) is a conceptual model specifically created to support data mart design. It is graphic-based and founded on the multidimensional model.

5.1.1 Design Methodologies

- **Requirement-driven Approach:**
 - Based on user analysis
 - Extracts information about facts, measures, hierarchies
 - Used when source data is complex or unavailable
- **Data-driven Approach:**
 - Based on operational source structure
 - Direct translation from source schemas
 - Used when sources are well-defined
- **Mixed Approach:**
 - Combines both methodologies
 - Balances user needs with data availability
 - Most common in practice

5.2 Core Components

5.2.1 Facts and Measures

- **Facts:**
 - Business events or processes
 - Examples: sales, orders, shipments
 - Center of analysis interest
- **Measures:**
 - Numerical properties of facts
 - Quantitative aspects
 - Examples: quantity, revenue, profit

5.2.2 Dimensions and Attributes

- **Dimensions:**
 - Analysis perspectives
 - Finite domains
 - Examples: time, product, location
- **Dimensional Attributes:**
 - Descriptive properties
 - Hierarchy components
 - Examples: product category, month name

5.3 Advanced Concepts

5.3.1 Hierarchies and Relationships

- **Standard Hierarchies:**
 - Many-to-one relationships
 - Regular structure
 - Example: Day \rightarrow Month \rightarrow Year
- **Specialized Hierarchies:**
 - Recursive relationships
 - Multiple paths
 - Optional relationships

5.3.2 Advanced Features

- **Cross-dimensional Attributes:**
 - Derived from multiple dimensions
 - Complex relationships
 - Advanced analysis capabilities
- **Convergence:**
 - Multiple paths to same attribute
 - Alternative drill-down paths
 - Complex hierarchical structures
- **Optional Elements:**
 - Conditional relationships
 - Partial hierarchies
 - Flexible structures

5.4 Logical Implementation

5.4.1 Star Schema

Definition

A denormalized schema structure where a central fact table is connected to dimension tables in a star-like pattern.

- **Components:**

- Central fact table with measures
- Surrounding dimension tables
- Simple join paths

- **Characteristics:**

- Denormalized dimensions
- Optimized for queries
- Simple to understand and use

5.4.2 Snowflake Schema

Definition

A normalized variant of the star schema where dimension tables are split into multiple related tables.

- **Structure:**

- Normalized dimensions
- Multiple related tables
- Hierarchical organization

- **Trade-offs:**

- Reduced storage space
- More complex queries
- Better data consistency

5.4.3 Schema Comparison

Aspect	Star Schema	Snowflake Schema
Structure	Simple, denormalized dimensions	Complex, normalized dimensions
Storage Space	Higher due to redundancy	Lower due to normalization
Query Performance	Better (fewer joins)	May be slower (more joins)
Maintenance	Easier to maintain	More complex maintenance
Data Integrity	Possible redundancy issues	Better referential integrity

Part II

Part 2 : Machine Learning

Chapter 6

Machine Learning