Business Intelligence

Week 2

Data Fundamentals

- Types of Data
- Data sources and Data Collection Methods
- Data quality and Data Governance
- Metadata Management

Tilahun Melak(PhD)

March, 2026

Objectives

At the end of this lecture students will be able to :

- Understand different types of data
- Identify data sources and collection methods
- Evaluate data quality dimensions
- Understand data governance principles
- Explore metadata management concepts

Introduction to Data in BI

- **Data as a Strategic Asset**
  - Data is not just raw facts but a valuable organizational resource
  - Enables competitive advantage through informed decision-making
  - Drives digital transformation and innovation
- **Data Processing in BI Systems**
  - Data is collected from multiple heterogeneous sources
  - Undergoes cleaning, transformation, and integration (ETL/ELT processes)
  - Stored in data warehouses or data lakes for analysis

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Introduction to Data in BI…

- **Challenges in Data Utilization**
  - Data silos and lack of integration
  - Poor data quality affecting analytical outcomes
  - Scalability issues with large and complex datasets
- **Importance of Data Management Practices**
  - Data governance ensures accountability and compliance
  - Data quality management improves reliability
  - Metadata management enhances discoverability and usability

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Types of Data

- **Structured Data**
  - Organized in fixed schema (rows and columns)
  - Stored in relational database management systems (RDBMS)
  - Easily queried using SQL
  - *Examples:* Banking transactions, student records, inventory databases
- **Semi-Structured Data**
  - Does not follow rigid tabular schema but contains tags or markers
  - Flexible and hierarchical in nature
  - Often stored in NoSQL databases
  - *Examples:* JSON data from web APIs, XML documents, email data

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics. *MIS Quarterly*, 36(4), 1165–1188

Types of Data

- **Unstructured Data**
  - No predefined format or structure
  - Difficult to store and analyze using traditional systems
  - Requires advanced techniques such as NLP and machine learning
  - *Examples:* Social media posts, images, videos, audio recordings

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics. *MIS Quarterly*, 36(4), 1165–1188

Structured Data

- **Definition and Characteristics**
  - Organized in fixed schema with rows and columns
  - Stored in relational databases (RDBMS)
  - Enforces data types, constraints, and relationships
  - Supports efficient querying and reporting
- **Storage and Querying**
  - Typically stored in SQL-based systems
  - Supports complex queries, joins, and transactions
  - Ideal for operational and analytical BI tasks
- **Advantages**
  - High data consistency and integrity
  - Easy to analyze with standard BI tools
  - Efficient for aggregation and reporting

Date, C. J. (2004). *An Introduction to Database Systems*. Pearson.

Structured Data…

- **Limitations**
  - Rigid schema makes adaptation difficult
  - Poor handling of multimedia or text-heavy data
  - Less suitable for modern unstructured and semi-structured sources
- **Examples in Real-World BI**
  - Banking: transaction logs and customer accounts
  - Retail: point-of-sale and inventory databases
  - Education: student enrollment and grade records

Date, C. J. (2004). *An Introduction to Database Systems*. Pearson.

Semi-Structured Data

- **Semi-structured Data**
  - Data does not follow a strict tabular schema but includes organizational tags or markers
  - Flexible and hierarchical structure
  - Often stored in NoSQL or document-oriented databases
- **Storage and Querying**
  - Supports schema-on-read rather than schema-on-write
  - Can handle variable data formats within the same dataset
  - Querying often requires specialized tools (e.g., MongoDB, Elasticsearch)
- **Advantages**
  - Flexible schema allows rapid adaptation to changing data
  - Easier integration across heterogeneous sources
  - Supports semi-structured APIs and web data

Abiteboul, S. (1997). Querying semi-structured data. ICDT Proceedings.

Semi-Structured Data…

- **Limitations**
  - Complex parsing required for analysis
  - Inconsistent formats may reduce data quality
  - Requires specialized processing frameworks
- **Examples in Real-World BI**
  - Web application logs in JSON or XML
  - Emails with structured headers and unstructured body
  - E-commerce API data streams

Abiteboul, S. (1997). Querying semi-structured data. ICDT Proceedings.

Unstructured Data

- **Unstructured Data**
  - Data with no predefined schema or structure
  - Cannot be stored easily in traditional relational databases
  - Often voluminous and complex, requiring advanced analytics
- **Storage and Processing**
  - Stored in data lakes, NoSQL systems, or object storage
  - Requires techniques such as Natural Language Processing (NLP), image/video analysis, and machine learning
  - Supports large-scale big data analytics
- **Advantages**
  - Contains rich, detailed information
  - Captures real-world phenomena beyond structured logs
  - Valuable for AI-driven predictive and prescriptive analytics

Manyika, J., et al. (2011). Big Data: The Next Frontier. McKinsey.

Unstructured Data…

- **Limitations**
  - Difficult to analyze without advanced tools
  - Requires significant computational resources
  - Data quality and consistency issues are common
- **Examples in Real-World BI**
  - Social media content (tweets, Facebook posts)
  - Customer feedback surveys in text form
  - Multimedia data: images, audio, and video recordings
  - Healthcare data: medical imaging, physician notes

Manyika, J., et al. (2011). Big Data: The Next Frontier. McKinsey.

Data Sources

- **Internal Data Sources**
  - Data generated within the organization
  - Often structured and consistent
  - *Examples:* ERP systems, CRM databases, transaction logs
- **External Data Sources**
  - Data obtained from outside the organization
  - Can be structured, semi-structured, or unstructured
  - *Examples:* Social media, government databases, market research reports

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Data Sources…

- **Machine-Generated Data**
  - Automatically created by devices, sensors, and systems
  - High volume and continuous streams
  - *Examples:* IoT sensors, server logs, network traffic data
- **Considerations for BI**
  - Assess data relevance and reliability before integration
  - Understand data update frequency and availability
  - Evaluate ethical, privacy, and regulatory considerations when sourcing external data

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

Data Collection Methods

- **Surveys and Questionnaires**
  - Direct collection from individuals or organizations
  - Can be structured or semi-structured
  - *Examples:* Customer satisfaction surveys, employee feedback forms
- **Sensors and Automated Systems**
  - Machine-generated data collected in real-time
  - Provides high-volume, continuous monitoring
  - *Examples:* IoT devices in manufacturing, weather sensors, smart meters

Groves, R. M. (2004). *Survey Errors and Survey Costs*. Wiley.

Data Collection Methods…

- **APIs and Web Scraping**
  - Collect data from external applications or websites
  - Useful for real-time and semi-structured data
  - *Examples:* Social media data extraction, stock market feeds, e-commerce product data
- **Manual Data Entry**
  - Human-driven input of data into systems
  - Can be error-prone, requires validation
  - *Examples:* Entering medical records, educational enrollment forms

Groves, R. M. (2004). *Survey Errors and Survey Costs*. Wiley.

Data Collection Methods…

- **Considerations for BI**
  - Assess accuracy, completeness, and timeliness of collected data
  - Balance automation and manual methods based on context
  - Ensure compliance with privacy regulations when collecting personal data

Groves, R. M. (2004). *Survey Errors and Survey Costs*. Wiley.

Data Integration

- **ETL (Extract, Transform, Load)**
  - Extract data from heterogeneous sources (databases, APIs, files)
  - Transform data through cleaning, normalization, and aggregation
  - Load into a centralized repository (data warehouse)
  - *Examples:* Consolidating banking transactions from multiple branches into a central warehouse
- **ELT (Extract, Load, Transform)**
  - Data is first loaded into a data lake, then transformed as needed
  - Supports big data environments and scalable processing
  - *Examples:* Storing raw log data in a data lake and transforming it for analytics on demand

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). John Wiley & Sons.

Data Integration…

- **Data Pipelines**
  - Automated workflows for moving and processing data
  - Can be batch (scheduled) or real-time (streaming)
  - *Examples:* Real-time fraud detection pipelines in financial systems
- **Data Warehousing vs Data Lakes**
  - Data Warehouse → Structured, schema-on-write, optimized for reporting
  - Data Lake → Stores raw structured, semi-structured, and unstructured data
  - *Examples:* Retail company using warehouse for sales reports and lake for customer behavior data

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). John Wiley & Sons.

Data Integration…

- **Challenges in Data Integration**
  - Data heterogeneity and incompatible formats
  - Data quality inconsistencies across sources
  - Latency and scalability issues in large systems
- **Relevance in BI**
  - Enables unified view of organizational data
  - Supports accurate and timely analytics
  - Critical for building reliable dashboards and decision-support systems

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). John Wiley & Sons.

Data Quality

- **Accuracy**
  - Degree to which data correctly represents real-world values
  - *Examples:* Correct customer addresses, accurate financial figures
- **Completeness**
  - Extent to which all required data is available
  - *Examples:* No missing fields in patient records or transaction logs

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy. *Journal of Management Information Systems*.

Data Quality…

- **Consistency**
  - Uniformity of data across different systems and datasets
  - *Examples:* Same customer ID and details across CRM and billing systems
- **Timeliness**
  - Data is up-to-date and available when needed
  - *Examples:* Real-time stock prices or current inventory levels

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy. *Journal of Management Information Systems*.

Data Quality…

- **Validity**
  - Data conforms to defined formats, rules, and constraints
  - *Examples:* Proper date formats, valid email addresses
- **Relevance in BI**
  - High-quality data leads to reliable insights and better decision-making
  - Poor data quality can result in misleading analytics and strategic errors
  - Essential for building trust in dashboards and reports

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy. *Journal of Management Information Systems*.

Data Quality…

- **Missing Data**
  - Occurs when required values are absent from datasets
  - Can lead to biased or incomplete analysis
  - *Examples:* Missing patient history in healthcare records, incomplete customer profiles
- **Duplicate Records**
  - Multiple entries representing the same entity
  - Causes overestimation and inconsistencies in reporting
  - *Examples:* Same customer registered multiple times in a CRM system

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy. *Journal of Management Information Systems*.

Data Quality…

- **Inconsistent Data**
  - Conflicting values across different systems or datasets
  - Reduces trust in BI outputs
  - *Examples:* Different addresses or account balances for the same customer in different databases
- **Outdated (Stale) Data**
  - Data that is no longer current or relevant
  - Leads to poor decision-making
  - *Examples:* Old inventory data used for supply chain decisions

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy. *Journal of Management Information Systems*.

Data Quality…

- **Invalid Data**
  - Data that does not conform to required formats or constraints
  - Often caused by poor validation mechanisms
  - *Examples:* Incorrect date formats, invalid phone numbers or IDs
- **Causes of Data Quality Issues**
  - Manual data entry errors
  - Lack of integration between systems
  - Inadequate validation and governance mechanisms

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy. *Journal of Management Information Systems*.

Data Quality…

- **Impact on Business Intelligence**
  - Produces misleading insights and unreliable dashboards
  - Reduces confidence in data-driven decisions
  - May result in financial loss or operational inefficiencies

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy. *Journal of Management Information Systems*.

Improving Data Quality

- **Data Cleaning (Cleansing)**
  - Detect and correct errors such as missing, duplicate, or inconsistent values
  - Techniques include deduplication, imputation, and outlier handling
  - *Examples:* Removing duplicate customer records; filling missing ages using statistical methods
- **Data Validation Rules**
  - Enforce constraints at the point of data entry and during processing
  - Includes format checks, range checks, and referential integrity
  - *Examples:* Email format validation; ensuring dates fall within valid ranges

Batini, C., & Scannapieco, M. (2016). *Data Quality*. Springer.

Improving Data Quality

- **Data Standardization**
  - Convert data into consistent formats and units across systems
  - Establish naming conventions and code sets
  - *Examples:* Standardizing date formats (YYYY-MM-DD); harmonizing country codes
- **Master Data Management (MDM)**
  - Create a single, authoritative source for key entities (customers, products)
  - Synchronize master records across systems
  - *Examples:* Golden customer record shared across CRM, billing, and support systems

Batini, C., & Scannapieco, M. (2016). *Data Quality*. Springer.

Improving Data Quality

- **Data Quality Monitoring & Metrics**
  - Define KPIs (accuracy %, completeness %, duplicate rate) and track over time
  - Implement dashboards and alerts for anomalies
  - *Examples:* Daily report on missing fields; alerts for sudden spikes in duplicates
- **Data Governance Integration**
  - Assign data owners and stewards responsible for quality
  - Establish policies, standards, and workflows for issue resolution
  - *Examples:* Data stewardship processes for correcting errors within SLAs

Batini, C., & Scannapieco, M. (2016). *Data Quality*. Springer.

Data Governance

- **Definition and Purpose**
  - Framework of policies, standards, and processes for managing data assets
  - Ensures data is accurate, secure, and used responsibly
  - Aligns data management with organizational objectives
- **Key Principles of Data Governance**
  - Accountability: Clear ownership of data assets
  - Transparency: Visibility into data sources and usage
  - Integrity: Maintaining data accuracy and consistency
  - Compliance: Adhering to legal and regulatory requirements

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*.

Data Governance…

- **Core Components**
  - Policies and standards for data handling
  - Roles and responsibilities (data owners, stewards, custodians)
  - Processes for data quality, security, and lifecycle management
- **Implementation Mechanisms**
  - Governance frameworks (e.g., DAMA-DMBOK)
  - Data governance committees and councils
  - Tools for data cataloging, lineage, and access control
- **Examples in Real-World BI**
  - Banking: Policies for secure handling of customer financial data
  - Healthcare: Governance of patient records to ensure privacy and accuracy
  - Government: Standardization of national statistics across agencies

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*.

Components of Data Governance

- Data Ownership
- Data Stewardship
- Data Custodianship (IT Responsibility)
- Data Policies and Standards
- Data Security and Privacy
- Compliance and Regulatory Management
- Data Lifecycle Management

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*.

Benefits of Data Governance

- Better decisions
- Compliance
- Risk reduction

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*.

Metadata Management

- Data about data
- Describes structure and meaning

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*.

Types of Metadata

- Technical
- Business
- Operational

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*.

Importance of Metadata

- Data discovery
- Governance support
- Integration facilitation

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*.

Summary

- In todays lecture we have discussed about;
  - Types of data
  - Data sources and data collection methods
  - Data quality dimensions
  - Data governance principles
  - Metadata management concepts

References

- Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.
- Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics. *MIS Quarterly*, 36(4), 1165–1188
- Date, C. J. (2004). *An Introduction to Database Systems*. Pearson.
- Abiteboul, S. (1997). Querying semi-structured data. *ICDT Proceedings*.
- Manyika, J., et al. (2011). Big Data: The Next Frontier. McKinsey.
- Groves, R. M. (2004). *Survey Errors and Survey Costs*. Wiley.
- Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). John Wiley & Sons.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy. *Journal of Management Information Systems*.
- Batini, C., & Scannapieco, M. (2016). *Data Quality*. Springer.
- Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*.