Data Governence and Data Quality

What Is Data Governance?

- Data Governance (DG) is the framework of rules, processes, and responsibilities that ensure data is:
  - Accurate
  - Secure
  - Consistent
  - Available
  - Used ethically and correctly
- It defines who can do what with which data and when.

2. Why Data Governance Matters

- Organizations need DG to:
  - Reduce data-related risks
  - Ensure compliance (GDPR, HIPAA, etc.)
  - Improve decision-making
  - Standardize data across systems
  - Increase trust in data
  - Support analytics and BI

3. Key Components of Data Governance

- Data Policies
  - Rules for how data is managed
    (e.g., data retention, access, privacy)
- Data Standards
  - Common definitions and formats
    (e.g., "customer ID must be unique")
- Data Ownership & Stewardship
- Data Owner — responsible for data assets
- Data Steward — manages data quality and processes
- Data Security & Privacy
  - Controlling access, encryption, compliance.
- Metadata Management
  - Data about data (definitions, lineage, structure)
- Data Lifecycle Management
  - Creation → Storage → Usage → Archiving → Deletion
- DATA QUALITY
  - What Is Data Quality?
  - Data Quality refers to the fitness of data for its intended use.
    Good quality data = reliable decisions.
- Dimensions of Data Quality
  - Here are the most common dimensions:

Dimension	Meaning
Accuracy	Data reflects real-world values correctly
Completeness	Required data fields are filled
Consistency	No conflicting data across systems
Timeliness	Data is up-to-date
Uniqueness	No duplicates
Validity	Follows defined formats/business rules
Integrity	Proper relationships between data

3. Causes of Poor Data Quality

- Manual data entry errors
- Inconsistent data sources
- Duplicate records
- Lack of standards
- Outdated data
- Missing values
- System migrations

4. Data Quality Processes

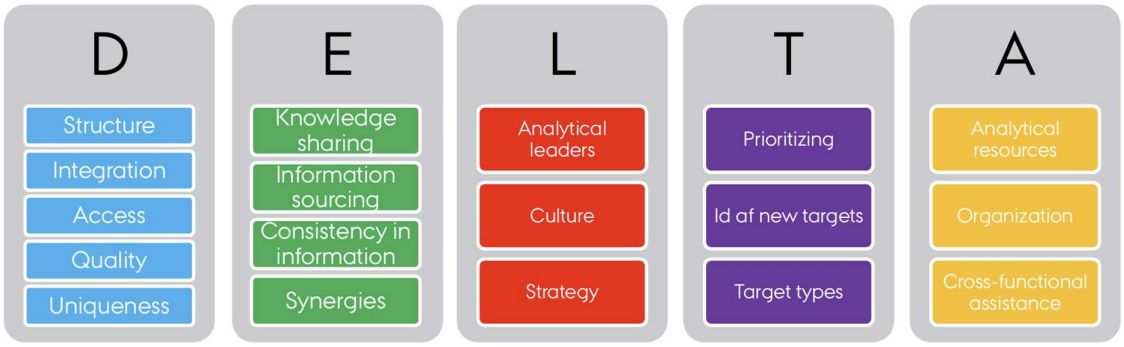
- Data Profiling
  - Analyzing data to detect issues.

- Data Cleansing
    - Fixing inaccuracies, duplicates, missing values.
- Data Matching
    - Identifying records that refer to the same entity.
- Monitoring
    - Continuous tracking of quality metrics.

HOW DATA GOVERNANCE & DATA QUALITY CONNECT

- DG sets the rules, roles, and standards.
- DQ ensures the data meets those standards.
  They work together to deliver trustworthy BI and analytics.
- Example:
  DG sets a rule → "Customer email must be valid."
  DQ checks → invalid emails → fixes errors.

Recap DELTA



T – TARGETS: KPI, DATA, DECISION

	Impaired	Localized analytics	Analytical aspiration	Analytical companies	Analytical competitors
Prioritizing	No resources	Marginally	Focused	Priority	Whole org
Id of new targets	None	Local targets	Id of opportunities	Centralized	Targets are strategic
Target types	None	Simple/functional KPIs	More org-wide KPIs	KPIs are connected to value creation	KPIs are focused on competitve advantage

How DELTA Connects to Data Governance & Data Quality

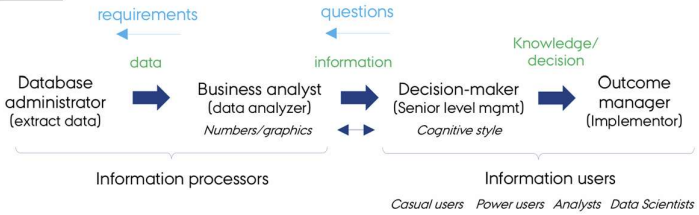
- The "D" in DELTA = Data Governance + Data Quality
  - The D (Data) element says that successful analytics requires:
    - Structured data
    - Integrated data
    - Accessible data
    - High-quality data
    - Unique, non-duplicated data
  - These points are exactly the goals of Data Governance and Data Quality.
  - Data Governance ensures:
    - Data standards
    - Definitions
    - Roles (data owners & stewards)
    - Policies for access, security, lifecycle
    - Metadata & lineage
  - Data Quality ensures:
    - Accuracy
    - Completeness
    - Consistency
    - Timeliness
    - Uniqueness
  - So the "D" pillar of DELTA cannot exist without strong Data Governance and Data Quality programs.
- "E" (Enterprise) requires governance to break silos
  - Enterprise-wide information sharing, consistency, and synergies depend on:
    - Shared definitions → *a function of data governance*
    - Standardized data formats → *data governance*
    - Consistent data across departments → *data quality*
  - Without governance, each department creates its own data rules → silos.

- “L” (Leadership) must sponsor Data Governance
    - Analytics-driven leadership supports:
        - Establishing governance councils
        - Making data quality a priority
        - Funding stewardship roles
        - Driving a data-driven culture
    - Without leadership, governance programs usually fail.
- “T” (Targets) depends on reliable data
    - Targets = KPIs + decisions + business priorities.
    - You cannot trust KPIs or decisions if:
        - Data quality is poor
        - Definitions differ across teams
        - Governance is weak
    - Example:
      If “customer churn” is defined differently in two departments, targeting churn with analytics becomes impossible.
- “A” (Analytical Resources) rely on quality-governed data
    - Analysts spend 60–80% of their time cleaning data when governance is weak.
    - Good governance reduces this, allowing analysts to:
        - Build models
        - Do BI reporting
        - Provide insights
    - High-quality + governed data → analysts work efficiently.

- Simple Summary
    - DELTA tells you what’s needed for analytics maturity, and Data Governance + Data Quality provide the foundation that makes DELTA possible — especially the Data, Enterprise, and Targets components.

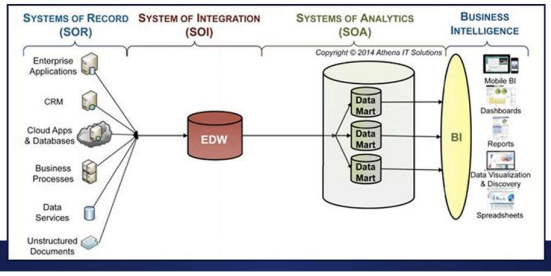
The BI Process

THE DATA JOURNEY



THE CLASSIC BI SYSTEM

Data democratization



THE BI SYSTEM

SOR – system of records

The relevant data sources for business decisions

Data should be value adding for business

SOI – system of integration

The data warehouse configuration for data captured in the SOR

Harmonizes internal and external data from the SOR to uphold the 5 Cs

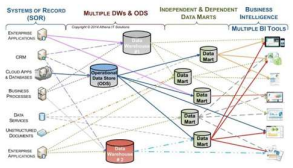
SOA – system of analytics

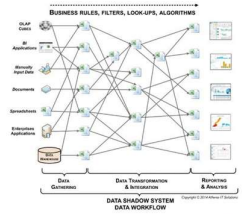
Caluculating metrics for business decisions

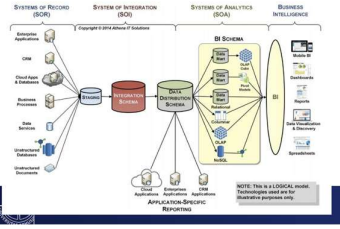
Creating data subsets matching the business decision requirements

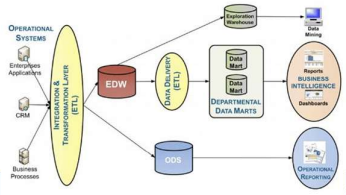
BI – business analytis

The analytic processing and presentation of data









Data Democracy

Data democratization ensures that employees have access to the right data, with a guarantee that the information they use is relevant and accurate for their specific needs, with the aim of enabling data-driven decisions to be made.

Data democratization refers to the process of enabling all members of an organization, regardless of their technical expertise, to access and analyze data. Governance in data democratization is essential to ensure data integrity and security with regulatory standards. By implementing robust governance frameworks, organizations can harness the full potential of their data assets while mitigating risks and ensuring compliance. As data continues to play a central role in organizational decision-making, effective governance will remain a critical enabler of successful data democratization.

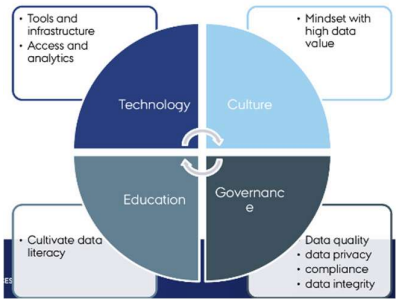
Data accessibility and usability are key concepts for democratization, where implementing strong and effective management systems allows access to data without compromising integrity, helping organizations unlock the full potential of their data assets [36].

Data security and privacy are important for building trust in data democratization. Compliance with regulations like the GDPR requires risk-based assessments to protect personal data while having a secure environment and ensuring responsible handling of the data [37].

Regulatory compliance in terms of data governance ensures organizations follow legal standards while protecting sensitive information. Embedding compliance into data governance practices removes the risks associated with data democratization, ensuring responsible and lawful data use across the organization [38].

- Data democracy means making data accessible to everyone in an organization, not just IT or analysts. It combines:
    - Centralization of data → All data in one place (e.g., data warehouse, data lake, BI platform)
    - Liberalization of use → Anyone who needs data can access it easily, safely, and quickly
- Why It Matters
    - Without data democracy:
        - Only specialists can use data
        - Decision-making is slow
        - BI becomes bottlenecked
        - Departments work with their own versions of truth
    - With data democracy:
        - More employees can make data-driven decisions
        - Teams collaborate better
        - Transparency increases
        - Innovation grows
- How Data Democracy Connects to Data Governance & Data Quality
    - You cannot democratize bad data
    - If everyone can access data, it MUST be:
        - Accurate
        - Complete
        - Consistent
        - Up-to-date
    - Data Quality is mandatory before democratizing access.

DATA ECOSYSTEM FOR DATA DEMOCRACY



- A Data Ecosystem is the combination of people, processes, technology, and policies that allows safe, efficient, and widespread use of data.
- Your points can be organized into 4 main pillars:
    - TECHNOLOGY
        - Tools and infrastructure:
            - Data warehouses, data lakes, BI tools, dashboards
            - Self-service analytics platforms
        - Access and analytics:
            - Easy access to data for employees
            - Self-service reporting and visualization
        - Technology enables democratized access, but only works if governance and quality are in place.
    - CULTURE
        - Cultivate data literacy:
        - Train employees to read, interpret, and act on data
        - Mindset with high data value:
            - Foster belief that decisions should be data-driven
            - Without culture, people won't trust or use the data, even if it's available.

- GOVERNANCE
  - Data privacy → Protect sensitive information
  - Compliance → Follow GDPR, HIPAA, or other regulations
  - Data integrity → Ensure correctness and consistency
  - Governance ensures safe democratization. Employees can access data without risking misuse or errors.
- EDUCATION & DATA QUALITY
  - Data quality → Accuracy, completeness, consistency, timeliness, uniqueness
  - Education → Training employees on data standards, definitions, and good practices
  - Good quality data + educated users = trusted decisions.
- Bottom line:
  - Data Democracy works only if you have the right technology, a data-driven culture, strong governance, and high-quality, well-understood data.

Data

- To make data truly useful for an organization, it must meet three criteria:
  - **Accessible** → Users can easily find and retrieve the data they need
  - **Usable** → Data is clean, well-structured, and understandable
  - **Valuable** → Data enables better decisions and insights
- Key Components supporting this

Component	Role
Data Infrastructure	Centralized storage (data warehouse, data lake), pipelines, cloud platforms — ensures data is available and scalable
Data Governance	Policies, standards, security, privacy, compliance, and stewardship — ensures data is reliable and safe
Modern Analytical Tools	BI platforms, dashboards, self-service analytics — makes data usable by business users
ML/AI Technology	Advanced analytics, predictive modeling, recommendation engines — unlocks new value from data

BI ARCHITECTURE

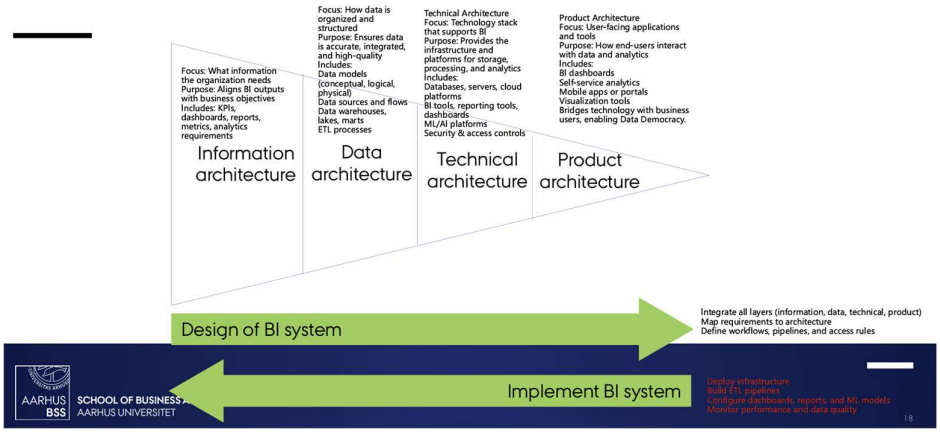
Architectural Purpose

- **A Business Intelligence Architecture is the systematic design that transforms raw data into meaningful, actionable information for the organization.**
  - Organizing Data
    - Collect, store, and structure data from multiple sources
    - Use data warehouses, lakes, or marts
    - Standardize formats and definitions
    - Supports data quality and reduces silos.
  - Democratizing Data
    - Provide self-service access to business users
    - Enable data-driven decision-making across departments
    - Ensure users can access data safely and efficiently
    - Direct link to Data Democracy.
  - Creating Business Value with Data
    - Turn data into insights, dashboards, and analytics
    - Support strategic and operational decisions
    - Enable ROI from BI investments
    - Analytics + high-quality data = business value.
  - Integration Across the Organization
    - Connect systems, databases, and applications
    - Ensure consistent and unified data definitions
    - Promote enterprise-wide collaboration
    - Strongly supported by Data Governance.
  - Securing Data Quality and Credibility (The 5 Cs)
    - Although not always explicitly named, the 5 Cs typically refer to:

Credibility	Trustworthy, accurate data
Consistency	Same definitions and formats across systems
Completeness	All required data is captured
Currency	Data is timely and up-to-date
Control	Governed access, security, compliance

    - Ensures users trust the BI outputs for decisions.
  - Systematic Design: Transforming Data into Information
    - BI architecture defines how data flows:
      - Data Collection → ETL (Extract, Transform, Load)
      - Storage → Warehouse, Lake, or Marts
      - Analysis → BI tools, dashboards, ML/AI
      - Reporting → KPIs, visualizations, decisions

Architectural Framework

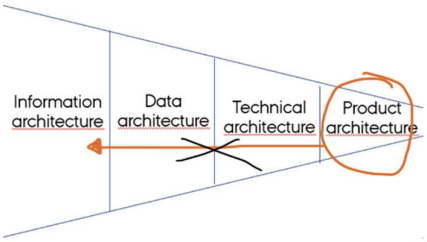


BI architecture

- Four layers of architecture
  - Information
    - Definitions, rules, regulation and management for data
  - Data
    - Requirements for data throughout the system
  - Technology
    - Requirements for hardware and software
  - Product
    - The actual choice of hardware and software

Architectural Traps

- Product Fixation
  - Focusing too much on choosing the latest BI tool or product rather than solving the business problem.
  - Mistake: Selecting software first without understanding requirements.
  - Consequence: Tool does not meet actual needs, expensive, and underutilized.
- BI Implementation Without Information Architecture
  - Skipping the planning layer that defines what information is needed.
  - Mistake: Building dashboards and reports before defining KPIs and business requirements.
  - Consequence: BI delivers irrelevant or incomplete insights.
- BI Project Not Delivered on Time or Within Budget
  - Poor planning of resources, data integration, and infrastructure.
  - Mistake: Underestimating complexity of ETL, data quality issues, or user training.
  - Consequence: Missed deadlines, cost overruns, and reduced stakeholder confidence.



- BI Solution Fails Expectations
  - End-users find dashboards confusing, incomplete, or irrelevant.
  - Mistake: Lack of engagement with business users during design.
  - Consequence: Low adoption, wasted investment.
- Blame Current Product and Select New One
  - When BI fails, organizations often blame the tool rather than architecture, governance, or process issues.
  - Mistake: Buying new software without addressing root causes.
  - Consequence: Repeating the same mistakes, increased costs, frustration.

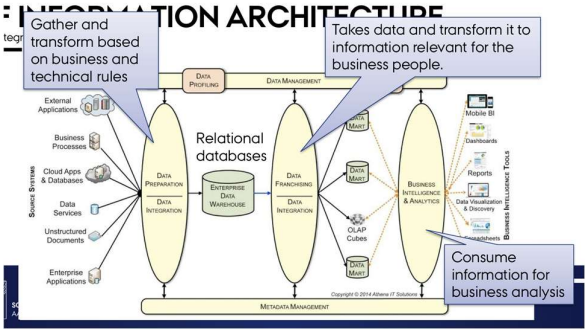
Information Architecture – Creating Data democracy

- Information Architecture (IA) is the design of how information is organized, defined, and delivered across the Business Intelligence system. It provides instructions and standards for data usage, ensuring it is trustworthy, consistent, and usable.
    - Role in Creating Data Democracy
        - Makes data accessible and usable for a wide range of users.
        - Provides rules, definitions, and procedures so users can safely interact with data.
        - Ensures that self-service analytics and democratized data access are aligned with governance.
    - Setting Up Instructions for Data
        - IA provides:
        - Definitions → Standard terminology across the organization (e.g., "customer," "revenue")
        - Procedures → How data is collected, stored, transformed, and analyzed
        - Rules & Regulations → Policies for data use, access, and compliance
        - This ensures consistent interpretation of data throughout the organization.
    - Securing Data Governance
        - IA formalizes how governance policies are applied through the BI system.
        - Ensures access controls, privacy, and compliance are embedded in every BI stage.
        - Supports auditability and accountability.
    - Ensuring Consistency and Quality Across the Data Journey
        - Tracks data from SOR → SOI → SOA → BI Applications:

Stage	Meaning	Role in IA
SOR (System of Record)	Raw operational data	Source definitions, standards
SOI (System of Insight)	Transformed, analyzed data	Ensures consistency and quality in analytics
SOA (System of Action)	Actionable outputs	Aligns with business decisions
BI Applications	Dashboards, reports, ML models	Presents trustworthy data to users

- IA ensures **data quality and governance rules are applied at every stage**, making the data reliable and consistent.

Information Architecture - Data Integration Framework



- DATA PREPARATION: From SOR to SOI
    - Definition:
      Data Preparation is the process of collecting, cleaning, transforming, and staging data from multiple sources so it can be used in a data warehouse (DW/EDW) or BI system.
    - It's often called the "rules of engagement" for data entering the warehouse.
    - Purpose
        - Ensures that only clean, consistent, and valid data enters the data warehouse
        - Applies business and technical rules to raw data
        - Supports Data Governance and Data Quality
        - Think of it as setting the "data dress code": no rule-compliant data, no entry.
    - Key Steps in Data Preparation
        - Data Gathering
            - Collect data from diverse sources:
            - Internal (ERP, CRM, operational systems)
            - External (market data, social media, third-party sources)
        - Data Transformation
            - Convert data to standard formats

        - Apply business and technical definitions
        - Enforce consistency and quality rules
    - Data Staging
        - Temporarily store data in a staging area
        - Prepare for integration into the warehouse or EDW
    - Hybrid Data Modeling
        - Create combined models from multiple sources
        - Supports complex analytics and BI reporting
    - Specialized Data Cleansing
        - Remove duplicates
        - Correct errors
        - Fill missing values
        - Ensure data integrity
- Importance
    - Takes 60–75% of BI project startup time → most BI projects fail or underperform if data prep is ignored
    - Defines core components for BI success → the rest of the BI system depends on the quality and definitions set here
    - "The system is only as good as these definitions."
- Connection to SOR → SOI

Stage	Role in Data Prep
SOR (System of Record)	Source data collection
Data Preparation	Cleaning, transforming, modeling, staging
SOI (System of Insight)	Ready-for-analysis data with consistent definitions

- Proper data preparation ensures trustworthy, usable data for analytics.

- DATA FRANCHISING
    - Data Franchising is the process of transforming raw or warehouse-level data into smaller, pre-processed, business-relevant datasets (like data marts or cubes) that can be directly used for analytics and decision-making.
    - It's about packaging data for specific business purposes, so insights are faster and easier to generate.
    - Purpose of Data Franchising
        - Create tailored datasets for specific business units or functions
        - Pre-calculate metrics, KPIs, and measures relevant to business decisions
        - Reduce repetitive work in reporting and analytics
        - Think of it as "data ready-to-use for business", instead of raw tables that need filtering and transformation every time.
    - Benefits

Speed	Reports and dashboards run faster with pre-built marts or cubes
Consistency	All users see the same measures and KPIs, reducing errors
Self-Service Analytics	Business users can explore data without IT intervention
Maintainability	BI apps are easier to manage because transformations are centralized

    - Problems Without Data Franchising
        - Every report repeats filtering, transformations, and measure creation → waste of time
        - BI software and dashboards run slowly
        - User frustration → business impatience
        - BI apps become complex to maintain
        - Limited or no self-service analytics → users depend on IT
    - Connection to BI, Governance, and Data Quality
        - Data Quality: Franchised data ensures that business-relevant metrics are accurate, complete, and consistent
        - Data Governance: Franchising enforces standard definitions and rules across all business units
        - Data Democracy: Business users can access ready-to-use datasets without deep technical skills
    - Data Franchising is a key enabler of self-service BI and faster, trustworthy decision-making.

- COMPARING DATA PREPARATION vs DATA FRANCHISING

Feature	Data Preparation	Data Franchising
Data Sources	SOR (System of Record) → Raw operational data	DW (Data Warehouse), ODS (Operational Data Store) → Pre-integrated data
Data Cleaning	Performs **cleaning of raw data** to make it accurate and usable	**Capitalizes on already cleaned data** from the warehouse
Data Conforming Process / Hierarchy	Performs **data alignment, standardization, and hierarchy building** before analytics	**Starts processing only after data is cleaned and conformed** by Data Preparation
Data Model / Dimensionality	Often **undocumented** during initial prep	**Documented**, with defined dimensions and measures for business use (DW / ODS)
Purpose	Prepare raw data for analytics; foundational step	Transform prepared data into **business-ready datasets or marts** for specific decisions
User Impact	Mostly IT / BI team responsibility	Supports **self-service analytics** and faster decision-making

Fair Data Preperation

- We litterally measure and observe on these dimensions

- Fair Data Principles

Principle	Goal	Simple Definition
F – Findable	Make it easy for both people and systems to find data.	Data should be discoverable through metadata and identifiers.
A – Accessible	Ensure authorized users can retrieve data easily.	Data should be retrievable via open, standardized protocols.
I – Interoperable	Enable data to work with other data and systems.	Data should use common standards, formats, and vocabularies.
R – Reusable	Make data ready for reuse beyond its original purpose.	Data should have clear licensing, provenance, and rich metadata.

- Examples

Not Fair
Customer ID: not unique identifier it just number if unlucky number could be somewh
Name:
Country: Is not one Format for example US, Fr and Germany
Date: not the same format
Amount: some numbers some letters
Product: No consitency some big Capital some small
Mail: Email is fair

DATASET 1 NOT FAIR

CustomerID	Name	Country	Date	Amount	Product	Email
101	John S.	US	12/10/24	45.5	Clothes	john.smith@mail.com
102	Marie D	Fr	11-11-2024	sixty	Apparel	marie.dubois@mail.fr
103	Müller	Germany	2024.11.15	75.90	clothing	hans.mueller@mail.de
104	Joao	Brasil	15/11/2024	80	Sportswear	joao.silva@mail.br
105	Jane	U.K.	13 Nov 2024	39.99	fashion	jane.doe@mail.uk

Maybe 2-3 out of 5 if you look at fair

Customer_ID: fair
Customer_name: Fair
Country_code: Fair
Purchase_date: fair
Amount_USD:
Currency_Code: it says amount in usd but then says currency code which makes no sende
Product_Gat: not fair not consistends
License: Fair

DATASET 2

Customer_ID	Customer_Name	Country_Code	Purchase_Date	Amount_USD	Currency_Code	Product_Category	Data_Source	License
CUST001	John Smith	USA	2024-12-10	45.50	**USD**	CLOTHES	POS	Internal use
CUST002	Marie Dubois	FRA	2024-11-11	60.00	**EUR**	APPAREL	OnlineStore	Internal use
CUST003	Hans Müller	DEU	2024-11-15	75.90	**EUR**	CLOTHING	CRM	Internal use
CUST004	João Silva	BRA	2024-11-15	80.00	**BRL**	SPORTSWEAR	POS	Internal use
CUST005	Jane Doe	GBR	2024-11-13	39.99	**GBP**	FASHION	OnlineStore	Internal use

DATASET 3

Fair

Customer_ID	Customer_Name	Country_Code	Purchase_Date	Amount	Currency_Code	Product_Category_Code
CUST_0001	John Smith	USA	2024-12-10	45.50	USD	CLO
CUST_0002	Marie Dubois	FRA	2024-11-10	60.00	EUR	APP
CUST_0003	Hans Müller	DEU	2024-11-15	75.90	EUR	CLO
CUST_0004	João Silva	BRA	2024-11-15	80.00	BRL	SPO
CUST_0005	Jane Doe	GBR	2024-11-13	39.99	GBP	FAS

- AIR Data and Data Democracy
    - Data Democracy = giving employees across the organization safe, easy access to data so they can make informed decisions.
    - FAIR principles ensure that this access is trustworthy, usable, and effective.
  - F – Findable → Easier Access
    - Data must be cataloged, indexed, and searchable.
    - In Data Democracy, users can quickly locate the datasets they need.
    - Example: A self-service BI portal with metadata and search functionality.
  - A – Accessible → Safe and Controlled Access
    - Data must be retrievable by authorized users through proper protocols.
    - In Data Democracy, this allows employees to access relevant data without IT bottlenecks, while maintaining security and governance.
    - Example: Role-based access to dashboards or datasets.
  - I – Interoperable → Usable Across the Organization
    - Data must follow standard formats and definitions so it can be combined with other datasets.
    - In Data Democracy, it enables cross-departmental analysis and integration.
    - Example: Standardized product codes, customer IDs, or financial measures across departments.
  - R – Reusable → Reliable and Trusted Insights
    - Data must be well-documented, clean, and prepared so it can be used repeatedly.
    - In Data Democracy, users can trust the data for decision-making without re-cleaning or re-defining it.

- Example: Pre-built data marts, cubes, or standardized KPIs for business units.

FAIR Principle	How it Enables Data Democracy
Findable	Everyone can discover the data they need, not just IT.
Accessible	Authorized users can self-serve without bottlenecks.
Interoperable	Data from different departments can be combined easily.
Reusable	Data products retain value across projects and teams.

Data Governance

- Data Governance (DG) ensures that **data remains accurate, consistent, reliable, and trustworthy** throughout its journey—from raw data in operational systems to actionable business insights in BI applications.
- It is often described as **the missing link** that connects data creation, accumulation, transformation, and consumption.
- **Core Purpose of Data Governance**
    - **Maintain data quality** across the organization
    - **Enforce consistent definitions, rules, and policies**
    - **Support BI systems and self-service analytics**
    - **Enable trust in business metrics and decisions**
- Challenges in Data Governance
    - Large company size → multiple systems, departments
    - Unstructured data (emails, documents, logs)
    - Cloud and hybrid environments
    - Big Data volumes
    - Outsourced data processing
- Where Governance Applies

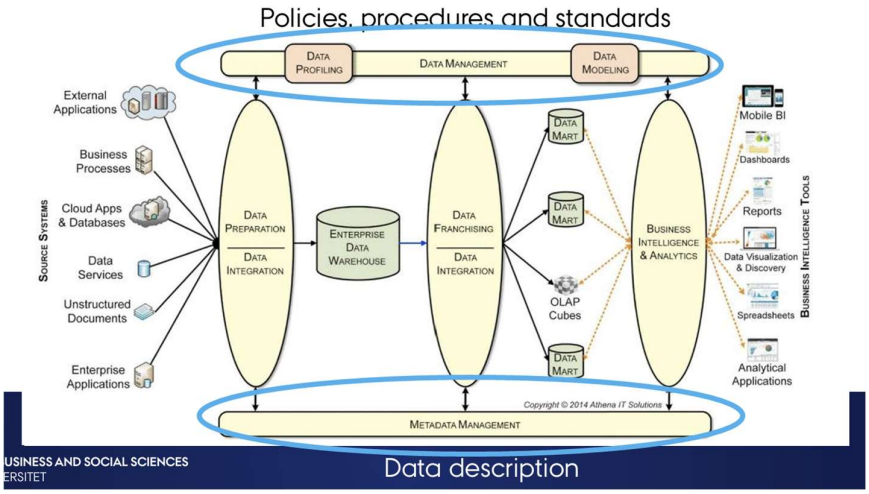
Stage	DG Role
Data Creation	Ensure proper input standards, master/reference data, and consistent definitions
Data Movement, Transformation, Integration	Monitor ETL, data lineage, workflows, transformations
Business Metrics & Data Definitions	Standardize KPIs and measures across the organization
BI Data Models & Algorithms	Ensure consistency in calculated fields, ML/AI models
Use Cases / Self-Service BI	Guide safe access, ensure metrics are reliable
Data Change Management	Track changes and updates to datasets
Monitoring Governance	Continuous auditing and compliance checks
Information Access & Delivery	Control permissions and availability of reports/dashboards
Information Consumption	Ensure users understand the data and can trust insights

- Relationship to BI & Information Architecture
    - Data Governance sits at the center of the BI ecosystem:
        - Connects **Data Creation → Transformation → Consumption**
        - Works closely with **Information Architecture and Data Integration Frameworks**
        - Implements **policies, procedures, and standards** for data handling

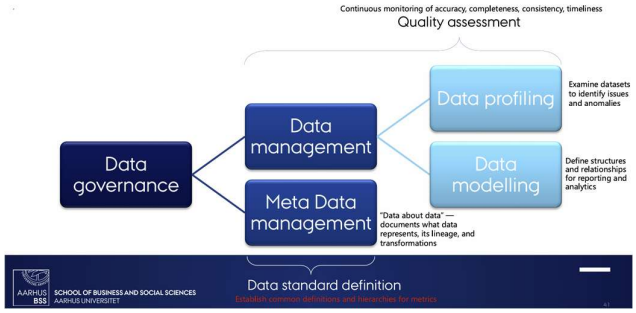
- DATA GOVERNANCE WITH INFORMATION ARCHITECTURE
    - Data Governance doesn't work in isolation—it is **embedded in the BI system** via **Information Architecture and Data Integration Frameworks**.
    - DIF – Data Integration Framework
        - Provides a **structured pipeline for moving, transforming, and integrating data** across systems.
        - Ensures that **data from diverse sources (SOR, ODS, external)** is consolidated reliably for BI.
        - Supports **Data Governance rules** like quality checks, standardization, and lineage tracking.
    - Information Architecture
        - **Defines what information is needed, how it's structured, and how it's used** across the organization.
        - Provides **instructions, standards, and documentation** for data use.
    - Policies, Procedures, and Standards
        - **These are** the rules that enforce governance **across the BI ecosystem:**
            - Policies: **Rules about access, privacy, compliance**
            - Procedures: **Steps for data entry, ETL, validation, and reporting**
            - Standards: **Naming conventions, formats, hierarchies, KPI definitions**
    - Data Description (Metadata)
        - Captures **"data about data"**
        - Documents:
            - **Definitions** → What the data represents

- **Lineage** → Where it came from and transformations applied
- **Usage** → How it is consumed in BI apps and dashboards

- Supports both **business understanding** and **IT governance** (e.g., 5 Cs, FAIR principles)



- Key Processes in Data Governance



  - Metadata Management (Core Part of DG)
    - Definition:
      Metadata is documentation of the BI system, describing:
      - What data represents → e.g., "Customer ID" means a unique customer
      - Where it came from → source systems, tables, ETL process
      - How it was transformed → cleaning, aggregation, mapping
      - What it means → business context for decision-making
    - Who uses metadata:
      - Business Users: Understand definitions to make meaningful analyses
      - IT Users: Ensure 5 Cs of data — Credible, Consistent, Complete, Current, Controlled
    - MDM –the master list of key enterprisereference data

Table 6.1 Data Standard Definition Template

METADATA	VALUE
Data Item (DI) name	The full name of the data element
DI description	A simple but unambiguous definition of the data element
DI type	Either string, integer, date/time
Data steward	The role who maintains this data element
Date published	The date this version was published as a data standard <YYYY/MM/DD>
Is part of	The parent element of the data items
Syntax	The required format of the data from the business perspective. This will include the minimum and maximum number of characters, if appropriate, and the structure of the data type or item e.g. National ID business format is NN-NNN-NNN where each N represents a digit from 0 to 9
Validation	Generic for types and specific for items. The validation rules to be applied for acceptance of data
Values	List of the acceptable values (e.g. male, female)
Default value	For any list of values, the default value to be used unless otherwise stated
Verification	Steps taken to establish the correctness of the data type or item
Comments	Additional notes
Data quality dimensions and minimum quality standards	The specific objective/subjective data quality indicators which apply to the data element (e.g. validity, completeness, usability), the metric relevant to each quality indicator and minimum measurement value

Table 6.3 <Salary> Data Standard

ATTRIBUTE	VALUE		
Data Item (DI) name	Salary		
DI description	Represents an employee's salary		
DI type (e.g. string, numeric)	Numeric		
Data steward	HR manager		
Date published	2010/04/03		
Is part of	Employee profile		
Syntax	Minimum X, maximum X numeric value		
Validation	1. No dollar sign 2. No commas		
Values	None		
Default value	Based on job title		
Verification	Should be the person's gross monthly salary		
Comments	All employees must have a salary		
Data quality dimension, metrics and minimum quality standards	DIMENSION	METRIC	MINIMUM STANDARD
	Currency	Binary value (0, 1) based on checks against job description	1 – salary must be within a given range for the job title

- Metadata Categories:
  - Data definitions
    - Your data dictionary
  - ETL source-to-target mapping
    - Documentation for examination of the ETL workflow
    - Should be within your ETL tool
  - BI applications
    - Cataloging the data accessed by BI applications, filters and queries used; workflow of data processes and data transformations

Data Profiling

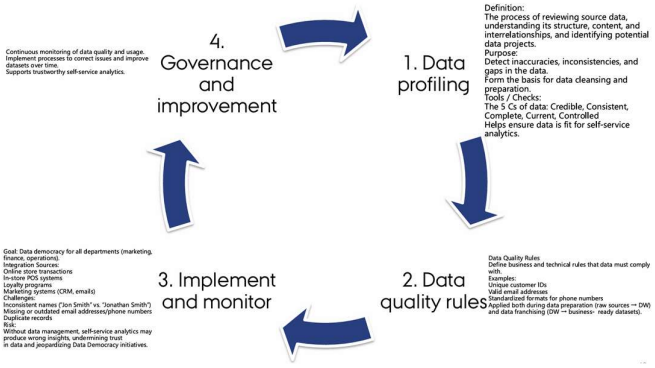
The process of **reviewing source data**, understanding its **structure, content, and interrelationships**, and identifying potential data projects.

- Purpose:
  - Detect inaccuracies, inconsistencies, and gaps in the data.
  - Form the basis for data cleansing and preparation.
- Tools / Checks:
  - The 5 Cs of data: Credible, Consistent, Complete, Current, Controlled
  - Helps ensure data is fit for self-service analytics.
- Data Profiling
  - Definition:
    - The process of reviewing source data, understanding its structure, content, and interrelationships, and identifying potential data projects.
  - Purpose:
    - Detect inaccuracies, inconsistencies, and gaps in the data.
    - Form the basis for data cleansing and preparation.
  - Tools / Checks:
    - The 5 Cs of data: Credible, Consistent, Complete, Current, Controlled
    - Helps ensure data is fit for self-service analytics.

Data profiling

Attribute	Issue Found	Example	% of Records Affected
Email	Missing values	NULL	18%
Phone Number	Invalid format	"12345"	7%
Customer ID	Duplicates	"CUST1023" repeated	3%
Birthdate	Implausible values	"1901-01-01"	0.5%
Zip Code	Mismatch with country	"94103" for Germany	2%

- Data Modeling
  - Definition:
    - The process of defining and analyzing data requirements to support business processes.
  - Purpose:
    - Design data structures, relationships, and hierarchies for reporting and analytics.
    - Ensures consistency in metrics and KPIs across BI applications.
  - Checks:
    - The 5 Cs applied during data franchising to ensure usable business-ready datasets.
- Data Management Workflow



- Implementing and Monitor

Rule	Pass Rate	Trend
DQ01	98.5%	Improving
DQ02	97.9%	Stable
DQ03	93.2%	Needs attention
DQ04	99.9%	Excellent
DQ05	95.1%	Improving

- Data Quality Rules
  - Define business and technical rules that data must comply with.
  - Examples:
    - Unique customer IDs
    - Valid email addresses
    - Standardized formats for phone numbers
  - Applied both during data preparation (raw sources → DW) and data franchising (DW → business-ready datasets).

Data Quality rules

Rule ID	Attribute	Rule Description	Validation Logic	Severity
DQ01	Email	Must not be null	`email IS NOT NULL`	High
DQ02	Email	Must match standard format	`email LIKE '%@%.%'`	High
DQ03	Phone	Must be 10 digits for US customers	`LENGTH(phone)=10`	Medium
DQ04	Customer ID	Must be unique	`COUNT(customer_id) = COUNT(DISTINCT customer_id)`	Critical
DQ05	Zip Code	Must correspond to customer country	`check_zip_country(zip, country)`	Medium

- Example

Context

A large retail company is working toward **data democracy** — giving all departments (marketing, finance, operations, etc.) access to self-service analytics using shared, governed data.

To enable this, the company builds a **Customer 360° data platform**, integrating customer data from:

Online store transactions

In-store point-of-sale (POS) systems

Loyalty programs

Marketing systems (CRM, email campaigns)

Solution:
Build a Customer 360° data platform that integrates customer data from multiple sources:
Online store transactions
In-store POS systems
Loyalty programs
Marketing systems (CRM, email campaigns)
Goal: Provide a single, unified view of the customer for analytics and decision-making.

Risk:
Poor data quality can lead to incorrect analytics results.
Users may lose trust in data, undermining Data Democracy initiatives.
BI dashboards, self-service tools, and reports could produce misleading insights.

Challenge

Different systems record customer data differently:

Inconsistent names ("Jon Smith" vs. "Jonathan Smith")

Missing email addresses

Incorrect or outdated phone numbers

Duplicate customer records

Without managing data quality, self-service analytics could lead to wrong insights — eroding trust in the data and undermining the data democracy initiative.

DATA QUALITY

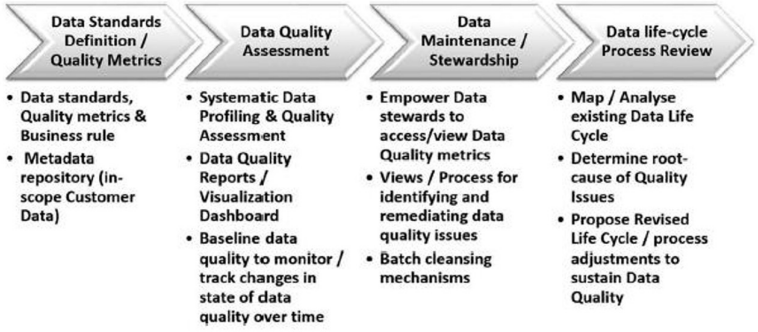


Table 6.2 Data Quality Dimensions

DIMENSIONS OF DATA QUALITY	DESCRIPTION
Accuracy	The degree to which data correctly reflects the real-world object being described
Validity	The degree to which the data conforms to a standard and business rules
Completeness	The extent to which data is not missing and is of sufficient depth and breadth. The data can be missing at multiple levels: • population – percentage of population represented • schema – attributes/tables missing • data value – missing field values
Consistency	The degree to which the data that exists in multiple locations is similarly represented and/or structured
Integrity	The degree to which data conforms to data relationship rules • Referential integrity • Uniqueness of primary key • Cardinality
Currency	The degree to which data reflects the real-world concept that it represents
Accessibility	The extent to which data is available or easily and quickly retrievable
Uniqueness	The degree to which each data record is unique
Usability	The extent to which business process(es) and/or individuals understand and are able to use this data
Relevancy	The extent to which the data is applicable to one or more business process(es) or decision(s)
Believability	The extent to which data is deemed credible by those using it

- Completeness
  - The degree to which all required data is present.

COMPLETENESS

Definition: The degree to which all required data is present.

Rating	Typical Threshold	Description	Example
High (H)	≥ 97% non-missing	Only minor missing values that don't affect usability.	0.2% missing first names
Medium (M)	90–96% non-missing	Some gaps, but analysis is still possible.	8% missing emails
Low (L)	< 90% non-missing	Significant missingness; data not reliable for analytics.	15% missing birthdates

- Validity
  - The degree to which data conforms to defined formats, ranges, or business rules.

VALIDITY

Definition: The degree to which data conforms to defined formats, ranges, or business rules.

Rating	Typical Threshold	Description	Example
High (H)	≥ 98% valid	Occasional errors, easily correctable.	0.5% invalid country codes
Medium (M)	90–97% valid	Noticeable issues; rules exist but not enforced everywhere.	3% invalid emails
Low (L)	< 90% valid	Many records violate format or business logic.	7% invalid phone numbers

- Uniqueness
  - The degree to which each record can be uniquely identified.

UNIQUENESS

Definition: The degree to which each record can be uniquely identified.

Rating	Typical Threshold	Description	Example
High (H)	≥ 99.5% unique	Duplicates rare and quickly resolvable.	0.2% duplicate customer IDs
Medium (M)	98–99.4% unique	Some duplication due to merging or entry errors.	2.5% duplicates
Low (L)	< 98% unique	Frequent duplicates causing reporting inaccuracies.	5% duplicate IDs across systems

- Consistency
    - The degree to which data is uniform across systems or conforms to reference data.

CONSISTENCY

Definition: The degree to which data is uniform across systems or conforms to reference data.

Rating	Typical Threshold	Description	Example
High (H)	≥ 98% consistent	Minimal discrepancies across sources or formats.	Country = "USA" consistently used
Medium (M)	90–97% consistent	Some differences due to format, abbreviation, or sync lag.	"US" vs. "USA"
Low (L)	< 90% consistent	Frequent mismatches, conflicting information.	"UAS" vs. "USA"; mismatched ZIP-country

- Quality Dashboard

QUALITY DASHBOARD

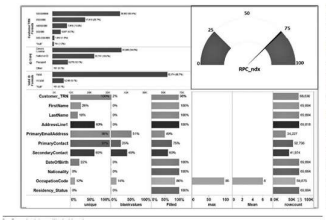


Table 6.4 Example Profile of Contact Data

VARIABLE	ROWS	FILLED (%)	MISSING (%)	UNIQUE (%)
First name	69,884	100.00	0.00	26.44
Last name	69,884	100.00	0.00	18.34
Address line 1	69,818	99.91	0.09	60.33
Address line 2	56,519	80.88	19.12	27.95
Country of residency	69,664	99.69	0.31	0.05
Nationality	69,664	99.69	0.31	0.16
Branch	69,884	100.00	0.00	0.06
Primary email address	34,227	48.98	51.02	97.93
Primary contact	52,736	75.46	24.54	97.43
Secondary contact	41,974	60.06	39.94	59.53
Customer_TRN	68,630	98.21	1.79	99.83
ID type	69,884	100.00	0.00	0.01

Data Governance Organization

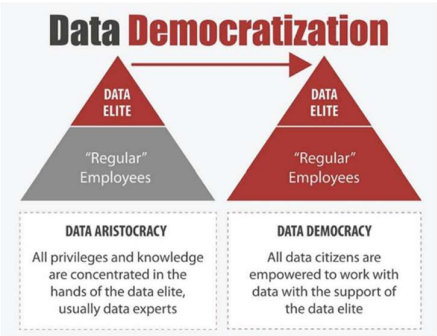
- Data Governance is primarily a business responsibility, not just an IT task. The goal is to ensure data quality, reliability, and usability across the organization.
- The Task Force
    - Purpose:
        - Oversee day-to-day governance activities
        - Ensure data quality and system integrity
    - Key Roles:

Role	Responsibility
Data Governor (Leader)	Oversees governance strategy, ensures policies are implemented
Data Owner (Data Quality)	Responsible for **accuracy, completeness, and validity** of data
Data Steward (System Quality)	Ensures **technical systems and processes** comply with governance rules

- The Business BI Committee
    - Purpose:
        - Acts like a board for the BI system and projects
        - Provides oversight, strategic guidance, and prioritization
    - Composition:
        - 6–12 business users from different departments
        - Ensures that data governance decisions align with business needs
    - Functions:
        - Approve BI projects and initiatives
        - Review and enforce data standards
        - Align governance with business objectives

Data Governance

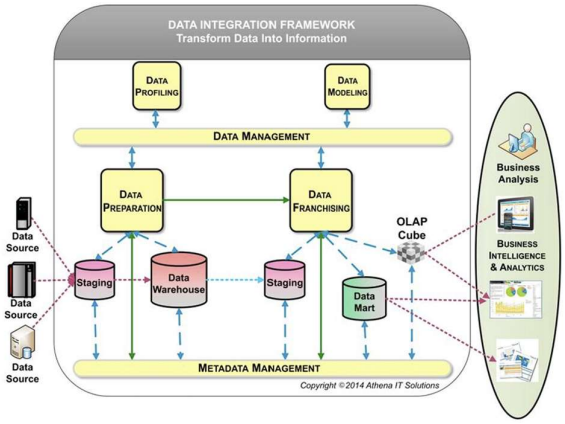
- Data Governance is an ongoing business responsibility, covering both data creation and consumption. It ensures that data is trustworthy, consistent, and fit for decision-making.
    - Constant Work on Requirements
        - Business needs evolve, so data governance is continuous.
        - Example: Adding new KPIs, business rules, or regulatory requirements.
        - Ensures the BI system remains aligned with business objectives.
    - Business Responsibility
        - Business people are accountable for data quality and governance, not just IT.
        - IT supports with infrastructure and tools, but decision-making, rules, and definitions come from business.
    - Covers Both Data Creation and Consumption
        - Data Creation: Ensures accurate, complete, and standardized data enters the system.
        - Data Consumption: Ensures users understand definitions, KPIs, and business rules when using data for reports or self-service analytics.
    - Governance Areas
        - Data Definitions: Standardized terms for consistency across systems
        - Business Rules: Policies on how data is captured, processed, and transformed
        - KPIs: Ensures metrics are defined, consistent, and reliable
    - Multiple Groups Depending on BI Project Stage
        - Different BI projects may require different governance groups:
        - Early stage → Focus on data creation and ETL rules
        - Mid stage → Focus on data integration and quality
        - Later stage → Focus on BI apps, dashboards, and consumption
- Governance must adapt to the lifecycle of BI projects within the organization.
- KEY CHALLENGES OF DATA DEMOCRACY
    - Data Democracy aims to give all employees safe, easy access to trusted data, but several challenges must be addressed:



    - Data Silos
        - Data stored in isolated systems or departments
        - Prevents a unified view of information
        - Example: Marketing has customer data separate from sales or operations
    - Access Control
        - Balancing broad access with security and privacy
        - Challenge: Providing self-service access without compromising sensitive data
    - Data Literacy
        - Users must understand, interpret, and analyze data correctly
        - Without proper literacy, misuse or misinterpretation can occur
    - Data Quality
        - Poor quality (incomplete, inconsistent, or incorrect data) undermines trust
        - Self-service analytics depends on accurate, reliable data
    - Data Governance
        - Lack of policies, standards, and oversight can lead to inconsistent definitions, metrics, and misuse
        - Governance is essential for trust, compliance, and accountability
    - Outdated Infrastructure
        - Legacy systems may struggle to handle modern BI, self-service, or big data workloads
        - Limits scalability, speed, and analytics capabilities
    - Incompatible Systems
        - Different systems may use different formats, schemas, or technologies
        - Integration is challenging, impacting data consistency and availability

DATA MANAGEMENT IN BI

- Policies, Procedures, and Standards
  - Used to stage data in relational databases (DW, EDW, or staging areas).
  - Ensures data is consistent, compliant, and controlled.
  - Closely tied to Data Governance, which provides oversight and rules.
  - Key idea: Data Management operationalizes governance rules in practice.
- Data Profiling
  - Process of analyzing source data to understand its structure, content, and relationships.
  - Helps identify issues before loading into the warehouse (e.g., missing values, duplicates, inconsistencies).
  - Supports Data Preparation by flagging problems that need correction.
  - Ensures clean, reliable data enters the analytics pipeline.
- Data Modeling
  - Defines target structures for data storage in OLAP cubes or data marts.
  - Complements data integration by specifying relationships, hierarchies, and dimensions.
  - Ensures that analytics-ready structures are consistent and optimized for BI reporting.
  - Think of this as designing the blueprint for how business users will access and analyze data.



Detailed Explanation of the Data Integration Framework
This diagram represents a Business Intelligence (BI) ecosystem, illustrating how data flows from sources through management, transformation, and analytics.

Data Sources on the left side are systems or repositories where raw data originates. Typical examples in a company include ERP and CRM systems, point-of-sale (POS) systems, web and mobile applications, as well as external datasets such as market research, social media, or third-party APIs. These sources provide heterogeneous data in different formats, structures, and quality levels, serving as the inputs for the BI system.
Staging Areas (pink cylinders) act as temporary storage where raw data is collected and pre-processed before entering the Data Warehouse. Here, data is cleaned by removing duplicates, correcting missing values, and standardizing formats, ensuring it is ready for integration. In the diagram, data from multiple sources first flows into staging areas before being passed on to the Data Warehouse.

The Data Warehouse (red cylinder) is the central repository for integrated, cleaned, and historical data. It serves as a single source of truth for the organization, supporting analytics, reporting, and decision-making. The warehouse stores standardized, historical, and structured data, which is then used as input for Data Franchising, OLAP cubes, and Data Marts.

The Data Management Layer (yellow horizontal bar) ensures data quality, governance, and consistency throughout the pipeline. It includes Data Profiling, which analyzes raw data for quality, completeness, uniqueness, and consistency, and Data Modeling, which defines target structures such as schemas, cubes, and marts, preparing the data for business analysis. This layer implements policies and standards that maintain the “five Cs” of data: Credible, Consistent, Complete, Current, Controlled.
Data Preparation (yellow box at the bottom left) involves cleansing and transforming raw source data. Processes include removing duplicates and errors, conforming data to business rules, and staging it for further processing. Cleaned and standardized data is then sent to the Data Warehouse.

Data Franchising (yellow box at the bottom right) transforms warehouse data into business-ready, tailored datasets for reporting and analytics. It supports self-service analytics, prepares Data Marts and OLAP Cubes for departments, and reduces repeated calculations for reports. Data flows from the Data Warehouse into these outputs.

Data Marts (green cylinder) are subsets of the warehouse optimized for specific business functions, such as Marketing or Finance, enabling fast queries. OLAP Cubes (gray cube) provide multidimensional representations of data, supporting complex analytics, aggregation, and slicing/dicing. Both receive data from Data Franchising and feed into BI dashboards, reports, and analysis tools. Metadata Management (yellow bar at the bottom) manages "data about data," documenting sources, transformations, definitions, and usage. It ensures traceability and transparency for both IT and business users, supporting Data Preparation and Franchising while maintaining governance, quality, and compliance.

Finally, Business Analysis / BI Applications (right side) are where end users analyze, report, and make decisions. Examples include dashboards, KPI reports, and advanced analytics such as predictive models or machine learning. These applications receive data from OLAP Cubes and Data Marts, enabling self-service analytics for departments.
Summary of Data Flow:

- Data Sources → Staging: Raw data collected and pre-processed
- Staging → Data Preparation: Cleansing and transformation
- Data Preparation → Data Warehouse: Centralized, standardized data
- Data Warehouse → Data Franchising: Tailored, business-ready datasets
- Data Franchising → Data Marts / OLAP → BI Applications: Optimized for reporting and self-service analysis
- Metadata Management: Tracks lineage, definitions, and governance throughout

Key Insights:

- End-to-End Governance: Metadata, data management, and preparation ensure trustworthy and consistent data
- Support for Data Democracy: Data Franchising, Marts, and OLAP cubes enable business users to access ready-to-use datasets
- Continuous Quality: Data profiling and modeling guarantee accurate, clean, and structured data

KEY TAKEAWAYS: BUSINESS INTELLIGENCE & DATA DEMOCRACY

- BI System = Data Supply Chain
    - BI is not just software—it's a pipeline that moves data from sources to insights, ensuring it's accurate, consistent, and actionable.
- Data Democracy Requires Architecture & Leadership
    - Giving users access to data requires planning, governance, monitoring, and leadership, not just software.
    - Self-service analytics works only when supported by rules, quality, and structured access.
- Architectural Purpose
    - The main goal of BI architecture is to generate credible data and enable Data Democracy across the organization.
- Accidental Architecture = Risky BI Design
    - Random, unplanned design of BI systems leads to inconsistent data, poor performance, and failed analytics.
    - Proper planning avoids wasted effort and frustration.
- Data Preparation & Data Franchising
    - Data Preparation: Cleans, standardizes, and stages raw data entering the warehouse
    - Data Franchising: Packages and delivers business-ready datasets for analytics
    - Together, they form the entry and exit points of the data warehouse.
- Data Governance Secures Consistency and Credibility
    - Policies, procedures, and metadata management ensure that all data is trustworthy, consistent, and compliant.
- Data Quality Assessment = KPI of the BI System
    - Measuring completeness, validity, uniqueness, and consistency ensures reliable insights.
    - Data quality is continuous, observable, and measurable, forming the heartbeat of a BI system.

Case Data Quality Assessment

Case Simulation: Data Profiling for "RetailHub Ltd."

Context:
RetailHub Ltd. is integrating customer data from online, in-store, and loyalty systems. The data governance team has performed *initial profiling* on 20,000 customer records.

Below is a sample of profiling results for 10 key fields. Students should analyze the data and rate each field's quality (High / Medium / Low) — based on **completeness**, **validity**, **uniqueness**, and **consistency**.

CASE DATA

Field Name	Description	% Missing	% Invalid	% Duplicates	Example Issues Found	Comments / Notes
Customer_ID	Unique ID for each customer	0%	0%	2.5%	"CUST1023" appears twice	Duplicate IDs from loyalty system merge
First_Name	Customer first name	0.2%	0%	N/A	-	Few missing values
Last_Name	Customer last name	0.4%	0%	N/A	-	Slightly incomplete
Email	Contact email	8%	3%	1%	"john.smith@@mail.com", missing "@"	Some invalid formats
Phone_Number	Contact phone	12%	7%	0%	"12345", "0000000000"	Missing area codes or dummy values
Date_of_Birth	Customer birth date	15%	2%	N/A	"1900-01-01", "2025-05-01"	Implausible dates
Country	Country of residence	0.5%	0.5%	N/A	"UAS" instead of "USA"	Occasional typos
Zip_Code	Postal code	6%	5%	N/A	"94103" for Germany, "AB123" missing digits	Wrong for country
Loyalty_Points	Accumulated loyalty balance	3%	0.2%	0%	Negative values in some cases	Data sync issues
Signup_Date	Date joined loyalty program	1%	0.5%	N/A	Future dates	Possible data entry errors

CASE TASKS

- Assess data quality as high, medium, low for each data string
  - Completeness (missing values), Validity (data range), Uniqueness (duplication), Consistency (uniformity)
- Set-up data quality rules

Reminder

COMPLETENESS

Definition: The degree to which all required data is present.

Rating	Typical Threshold	Description	Example
High (H)	≥ 97% non-missing	Only minor missing values that don't affect usability.	0.2% missing first names
Medium (M)	90–96% non-missing	Some gaps, but analysis is still possible.	8% missing emails
Low (L)	< 90% non-missing	Significant missingness; data not reliable for analytics.	15% missing birthdates

VALIDITY

Definition: The degree to which data conforms to defined formats, ranges, or business rules.

Rating	Typical Threshold	Description	Example
High (H)	≥ 98% valid	Occasional errors, easily correctable.	0.5% invalid country codes
Medium (M)	90–97% valid	Noticeable issues; rules exist but not enforced everywhere.	3% invalid emails
Low (L)	< 90% valid	Many records violate format or business logic.	7% invalid phone numbers

UNIQUENESS

Definition: The degree to which each record can be uniquely identified.

Rating	Typical Threshold	Description	Example
High (H)	≥ 99.5% unique	Duplicates rare and quickly resolvable.	0.2% duplicate customer IDs
Medium (M)	98–99.4% unique	Some duplication due to merging or entry errors.	2.5% duplicates
Low (L)	< 98% unique	Frequent duplicates causing reporting inaccuracies.	5% duplicate IDs across systems

CONSISTENCY

Definition: The degree to which data is uniform across systems or conforms to reference data.

Rating	Typical Threshold	Description	Example
High (H)	≥ 98% consistent	Minimal discrepancies across sources or formats.	Country = "USA" consistently used
Medium (M)	90–97% consistent	Some differences due to format, abbreviation, or sync lag.	"US" vs. "USA"
Low (L)	< 90% consistent	Frequent mismatches, conflicting information.	"UAS" vs. "USA"; mismatched ZIP-country

Solution

Access Data Quality

Field	Completeness	Validity	Uniqueness	Consistency	Instructor Notes / Justification
Customer_ID	M	H	M	M	**Duplicates (2.5%)** reduce uniqueness. Completeness perfect but some inconsistency from system merges.
First_Name	H	H	H	H	Very few missing; format consistent; low business risk.
Last_Name	H	H	H	H	Similar to First_Name — good quality overall.
Email	M	M	H	M	8% missing, 3% invalid (`@@` or missing `@`). Format issues lower validity.
Phone_Number	M	M	H	M	12% missing, 7% invalid (too short/dummy). Needs validation rule for length and numeric pattern.
Date_of_Birth	L	H	H	H	15% missing, some implausible dates (`1900`, `future`). Requires plausibility checks.
Country	H	H	H	M	Few typos (`UAS`). Good completeness, small standardization issue.
Zip_Code	M	M	H	M	Missing (6%) + invalid (5%), mismatched with country — cross-field validation needed.
Loyalty_Points	H	H	H	H	Almost all valid except rare negative values.
Signup_Date	H	H	H	H	Strong overall; only minimal invalids (future dates).

Data Quality Rules

DATA QUALITY RULES

Rule ID	Field	Rule Description	Validation Logic / Check	Severity
DQ01	Customer_ID	Must be unique	`COUNT(Customer_ID) = COUNT(DISTINCT Customer_ID)`	Critical
DQ02	Email	Must not be null	`Email IS NOT NULL`	High
DQ03	Email	Must follow standard pattern	`Email LIKE '%@%.%'`	High
DQ04	Phone_Number	Must contain 10 digits (US)	`LEN(Phone)=10 AND ISNUMERIC(Phone)=TRUE`	Medium
DQ05	Date_of_Birth	Must be within realistic range	`BETWEEN '1905-01-01' AND CURRENT_DATE - 18y`	Medium
DQ06	Zip_Code	Must match country format	`check_zip_country(zip, country)`	Medium
DQ07	Loyalty_Points	Must be non-negative	`Loyalty_Points >= 0`	Medium
DQ08	Signup_Date	Must not be in the future	`Signup_Date <= CURRENT_DATE`	High

General Data Quality Rules

General Data Quality Rules with Field Scope

Rule ID	Field / Scope	Rule Description	Validation Logic / Check	Severity	Notes / Scope Explanation
G1	All fields	Must not be null	Field IS NOT NULL	High	Applies to every column in the dataset. Ensures no missing values.
G2	All string / text fields	Must not be empty / blank	LEN(TRIM(Field)) > 0	High	Applies to text fields (e.g., First_Name, Last_Name, Email). Ensures meaningful content.
G3	All ID / key fields	Must be unique	COUNT(ID) = COUNT(DISTINCT ID)	Critical	Applies to unique identifiers (e.g., Customer_ID, Order_ID). Ensures no duplicate records.
G4	All numeric / number fields	Must be within a realistic range	Field >= Min AND Field <= Max	Medium	Applies to numbers (e.g., Age, Loyalty_Points). Detects out-of-range values.
G5	All date fields	Must not be in the future	Field <= CURRENT_DATE	High	Applies to dates (e.g., Date_of_Birth, Signup_Date). Prevents impossible future dates.
G6	Email / contact fields	Must follow standard pattern	Email LIKE '%@%.%'	High	Applies to email addresses. Ensures correct format for communications.
G7	Phone / contact fields	Must match country-specific format	LEN(Phone)=X AND ISNUMERIC(Phone)=TRUE	Medium	Applies to phone numbers. Ensures valid digits and format.
G8	Categorical / lookup fields	Must contain valid allowed values	Field IN ('A', 'B', 'C')	Medium	Applies to fields with predefined categories (e.g., Country, Status). Prevents invalid entries.
G9	Zip / postal codes	Must match country format	check_zip_country(zip, country)	Medium	Applies to postal codes. Ensures code is valid for the country.
G10	Foreign key / reference fields	Referential integrity: must exist in reference table	Field IN (SELECT ID FROM RefTable)	Critical	Applies to fields referencing other tables (e.g., Customer_ID in Orders). Ensures relationships are valid.
G11	All fields / key fields	Data consistency: duplicate records should be avoided	Check for duplicate rows on all key fields	High	Prevents repeated records across multiple columns or tables.
G12	Related fields	Logical consistency: related fields must align	IF Field1 = X THEN Field2 IN (Y,Z)	Medium	Applies to dependent fields (e.g., Country and Zip_Code). Prevents impossible combinations.
G13	All string / text fields	Standardization: values must follow consistent format	Apply formatting rules (e.g., upper-case, trimmed spaces)	Medium	Ensures consistent formatting in text fields.
G14	All numeric / number fields	Non-negative values	Field >= 0	Medium	Applies to numbers that cannot be negative (e.g., Loyalty_Points, Age).
G15	Date fields	Historical / business rules: must meet business criteria (e.g., age ≥ 18)	Date_of_Birth <= CURRENT_DATE - 18y	High	Ensures dates make sense for business logic (e.g., legal age).