Information Integration
Chapter 2. Federated Databases

SIA & SDBIS

2.1 Data Integration
Concept, Process, Architectures

- Data Integration Concept.
- Data Integration Process.
- Data Integration Architectures and Strategies.

Data Integration Concept
(Some) Definitions

- "... a set of techniques that enable building systems geared for flexible *sharing* and *integration* of data across multiple *autonomous* data providers" [1, 1]
- "... a set of procedures, techniques, and technologies used to design and build *processes* that extract, restructure, move, and load data in either operational or analytic data stores either in real time or in batch mode." [2, 3]

Data Integration Concept
(Other related) Definitions

- “The practice associated with managing data that *travels* between applications, data stores, systems, and organizations is traditionally called data integration (DAMAinternational,2009)” [3, 7]

- “... about the *consolidation* of data, but it is the *movement*, not the persistence that is the focus. Data interface refers to an application written to implement the movement of data between systems. ” [3, 7]

Types of Data Integration (Goal oriented)

- Transactional data integration.
- Analytical (Business intelligence) data integration.

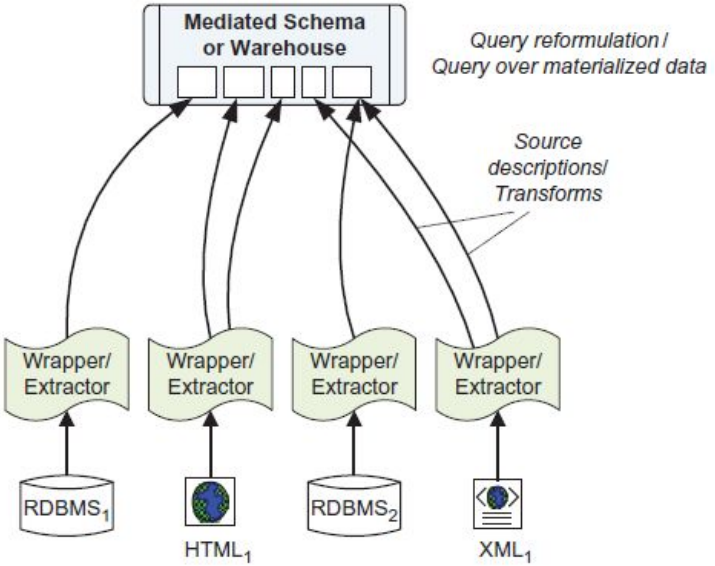
Data Integration Problems and Challenges

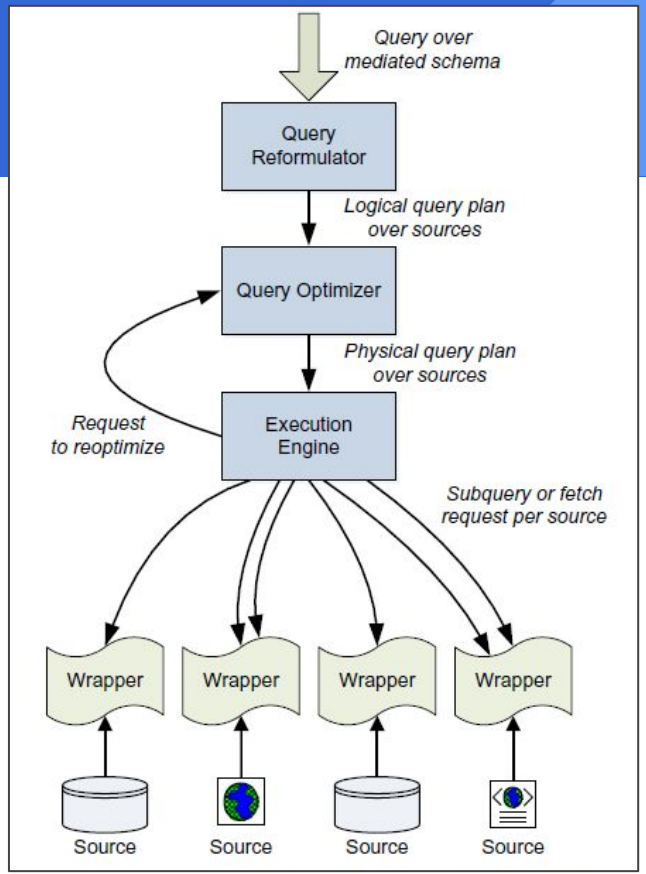
- Format: heterogeneous data format/data models
- Accessibility and autonomy: networking, access drivers, web-enabling
- Synchronous/Asynchronous Systems
- Scope: Operational/Transactional (OLTP) vs. Analytical (OLAP.BI)
- Complexity: number of sources, data format (in)compatibilities
- Source Data Query Language/Procedures (to extract data)

Data Integration Process Perspectives/Approaches

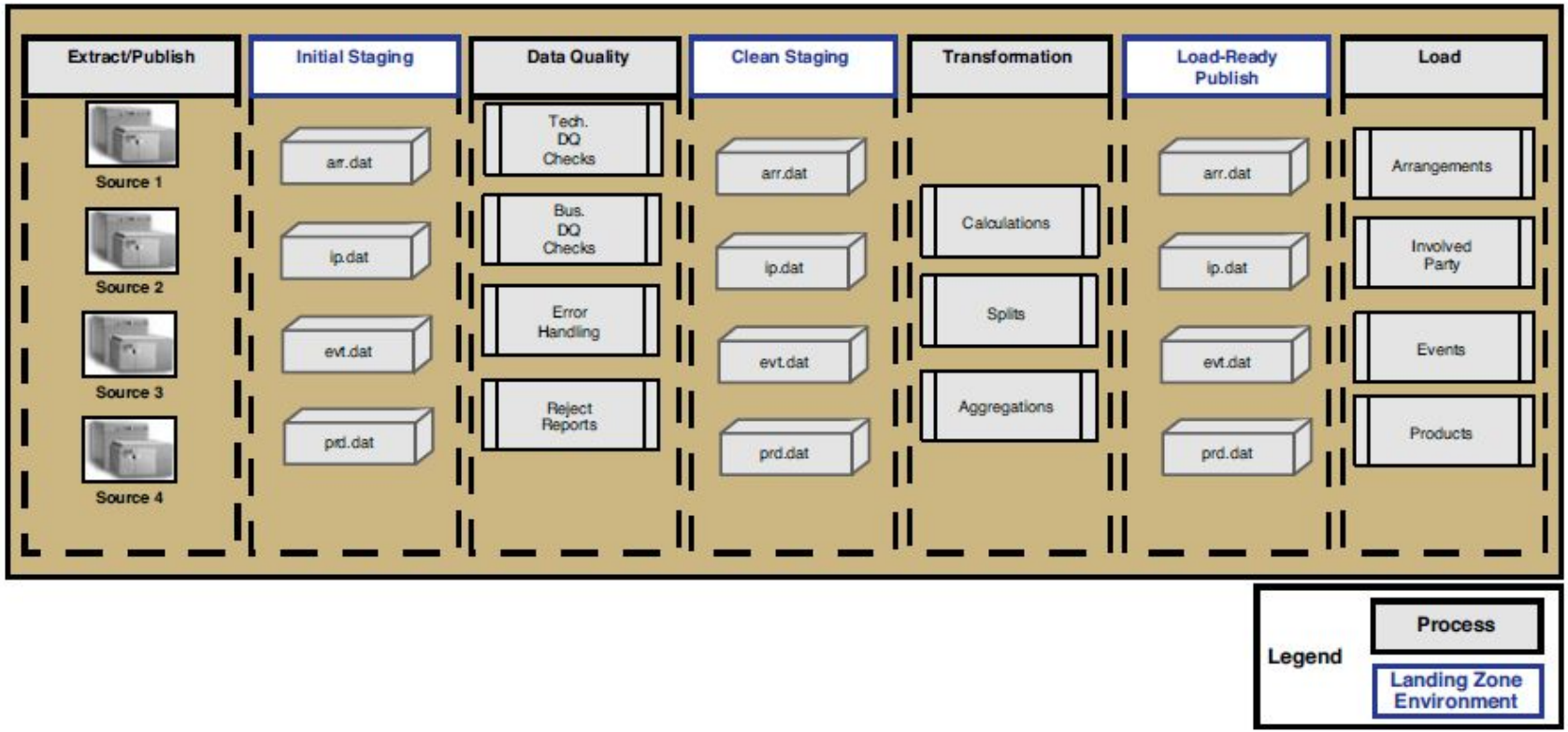
- Database(d) perspective: virtual database.
- Datawarehouse-ing perspective: ETL Extract-Transform-Load.
- Data-in-motion perspective.

Database(d) *perspective* [1, 10..14] (virtual database)

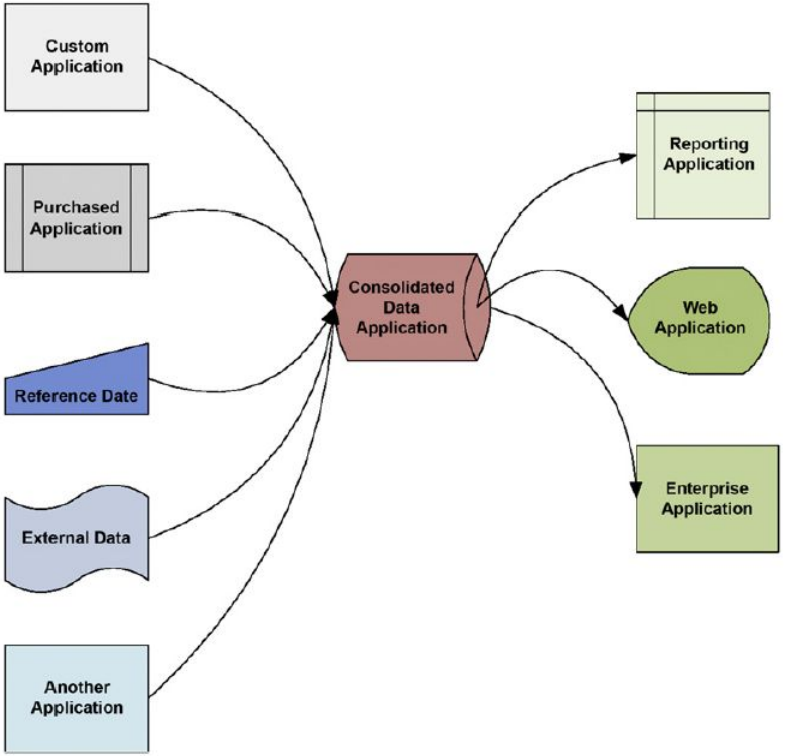


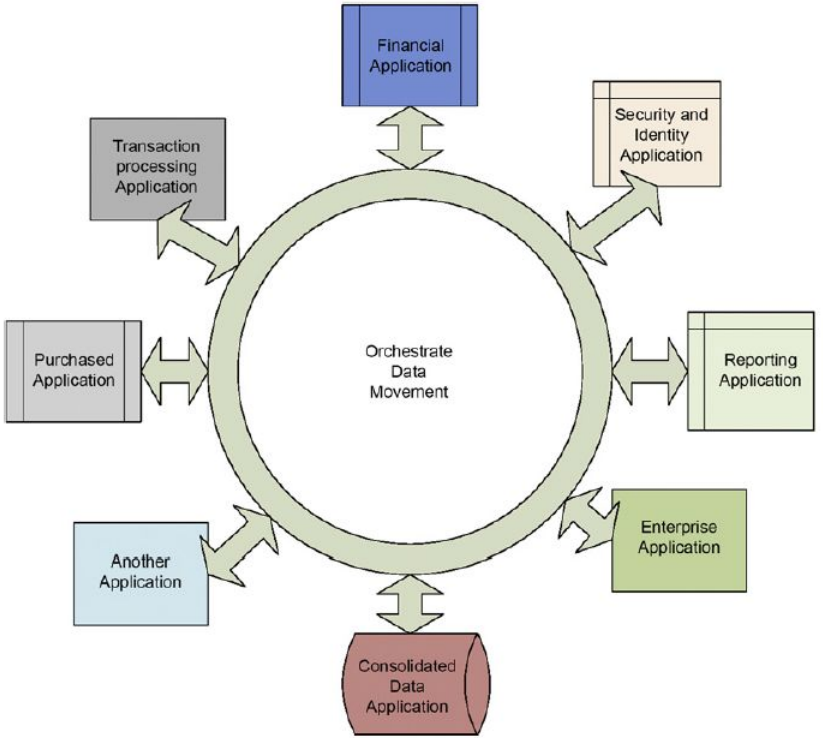


Data Warehousing perspective [2, 20]



Data “in motion” perspective [3, 7]





Data Integration Process
Common Activities

- Data Acquisition: Access/Extract Data Process (from Data Sources):
  - Data Type *Mapping* (Format Matching).
- Data Quality Process:
  - Data *Integrity Checking*
    - to detect:
      - inconsistent Data
      - missing Data
      - invalid Data
    - resolution
      - *Data Filtering*
      - *Data Enhancement.*

Data Integration Process
Common Activities

- Data Transformation Process:
  - Data Matching:
    - Structured Type/Schema *Mapping* (Format Matching);
    - *Entity Resolution*:
      - entity mediation;
      - entity merging;
      - reference reconciliation (reference matching);
  - Data Consolidation:
    - calculation/derivation;
    - splitting;
    - aggregation.

DI Architectures and Strategies

- DI Architectural Types
- DI Components

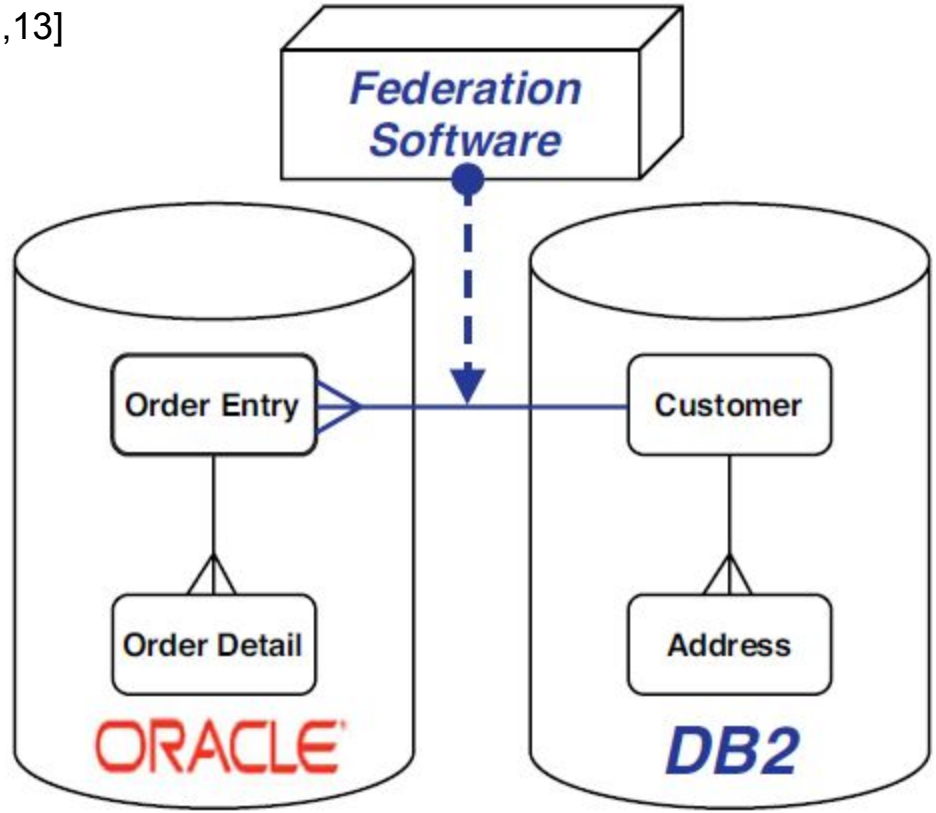
DI Architectural Strategies

- Data Integration Approaches [1]
  - Data Warehousing approach
  - Virtual Integration
- Architectural Integration Patterns [2]:
  - EAI (Enterprise Application Integration),
  - Federation (and virtualization),
  - SOA (Service Oriented Architecture),
  - ETL (Extract Transform Load)
- Data “In Motion” Integration Categories [3]:
  - Batch Data Integration
  - Real-time Data Integration
  - Data Virtualization

Federation Architecture [2,13]

- Federation: disparate data(bases) integration into a unified logical structure.
- Federation Software.

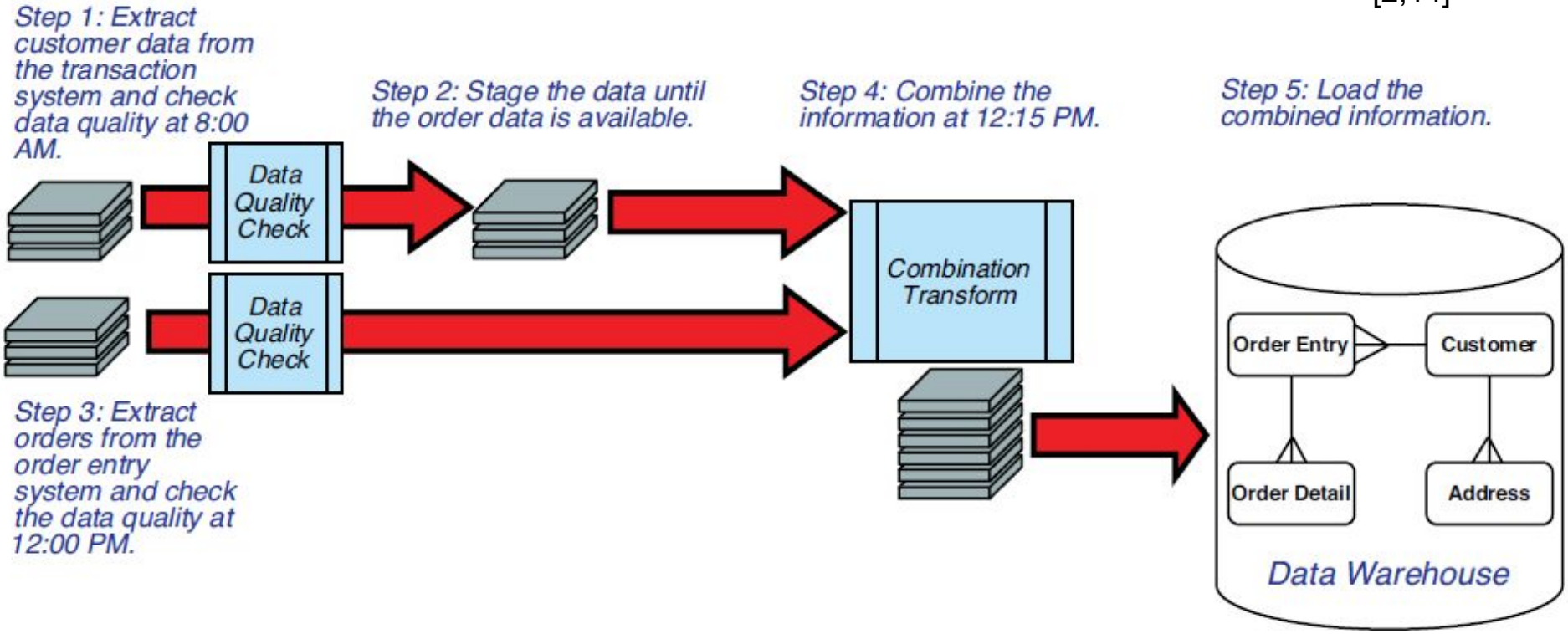
[2,13]


Federation Software
Order Entry
Order Detail
Customer
Address
ORACLE
DB2

ETL-DataWarehouse Architecture [2,14]

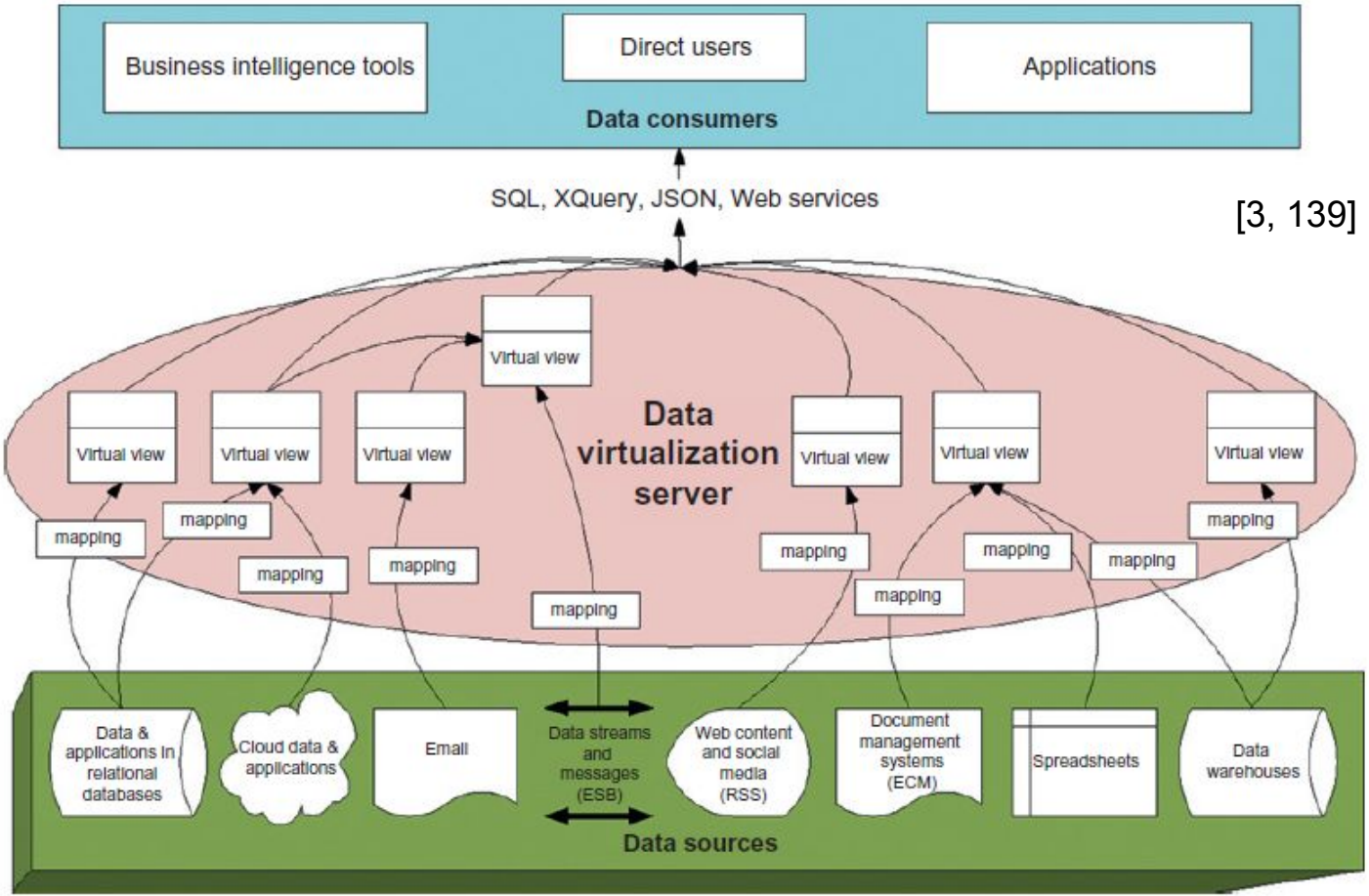
- Collection and aggregation of transactional data.
- Analytics and Business Intelligence target.
- ETL Software.

[2,14]


Step 1: Extract customer data from the transaction system and check data quality at 8:00 AM.
Step 2: Stage the data until the order data is available.
Step 4: Combine the information at 12:15 PM.
Step 5: Load the combined information.
Data Quality Check
Data Quality Check
Combination Transform
Step 3: Extract orders from the order entry system and check the data quality at 12:00 PM.
Order Entry
Customer
Order Detail
Address
Data Warehouse

Data Virtualization Architecture [3,139]

- Federation (combined with Real-time) sub-category.
- Virtual Database as common view and single access point.
- Data Virtualization Software.


Business intelligence tools
Direct users
Applications
Data consumers
SQL, XQuery, JSON, Web services
Virtual view
Virtual view
Virtual view
Virtual view
Data virtualization server
Virtual view
Virtual view
Virtual view
mapping
mapping
mapping
mapping
mapping
mapping
mapping
mapping
mapping
mapping
Data & applications in relational databases
Cloud data & applications
Email
Data streams and messages (ESB)
Web content and social media (RSS)
Document management systems (ECM)
Spreadsheets
Data warehouses
Data sources

[3, 139]

Data Integration Architecture
Components [1, 10]

- Data Sources
- Data Wrappers, Data Extractors
- Mediated Schema, Integration Platforms, Integration Schema
- Transformation Processors:
  - Source descriptors
  - Schema mappings
  - Transformations
- Data Warehouses
- Virtual Databases

Data Integration and DATA ENGINEERING

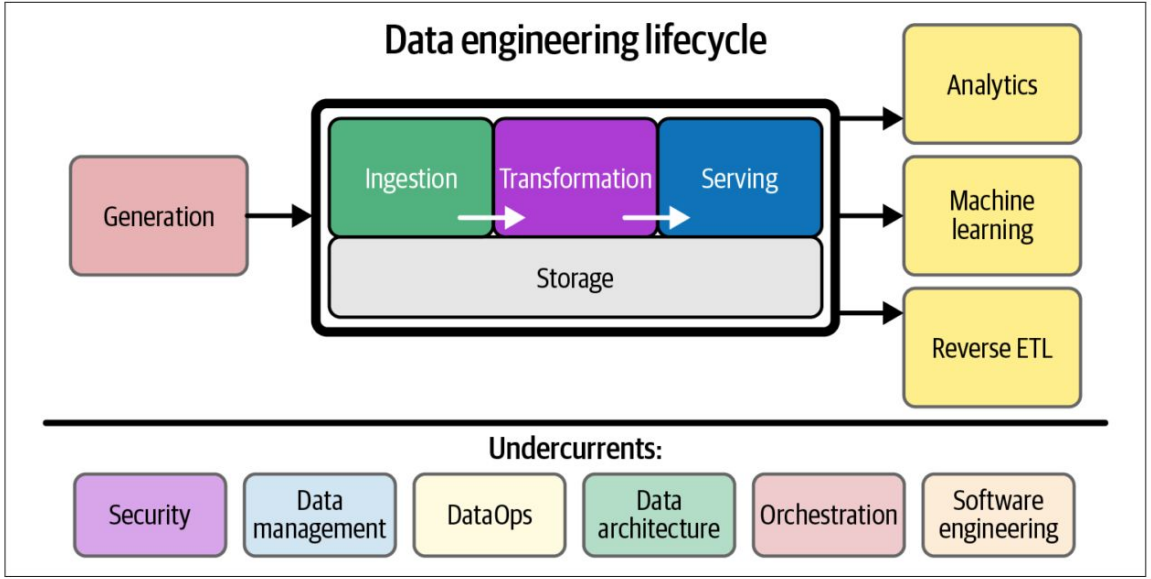
- **Data engineering** definition [4, 4]

*"**Data engineering** is the development, implementation, and maintenance of systems and processes that take in raw data and produce high-quality, consistent information that supports downstream use cases, such as analysis and machine learning. Data engineering is the intersection of security, data management, DataOps, **data architecture**, orchestration, and software engineering. A data engineer manages the data engineering lifecycle, beginning with getting data from source systems and ending with serving data for use cases, such as analysis or machine learning."* [Reis&Housley, 2022]

DATA ENGINEERING Lifecycle

Data Engineering lifecycle is a part of Data Lifecycle. [4, 5]

(Reis&Housley, 2022)

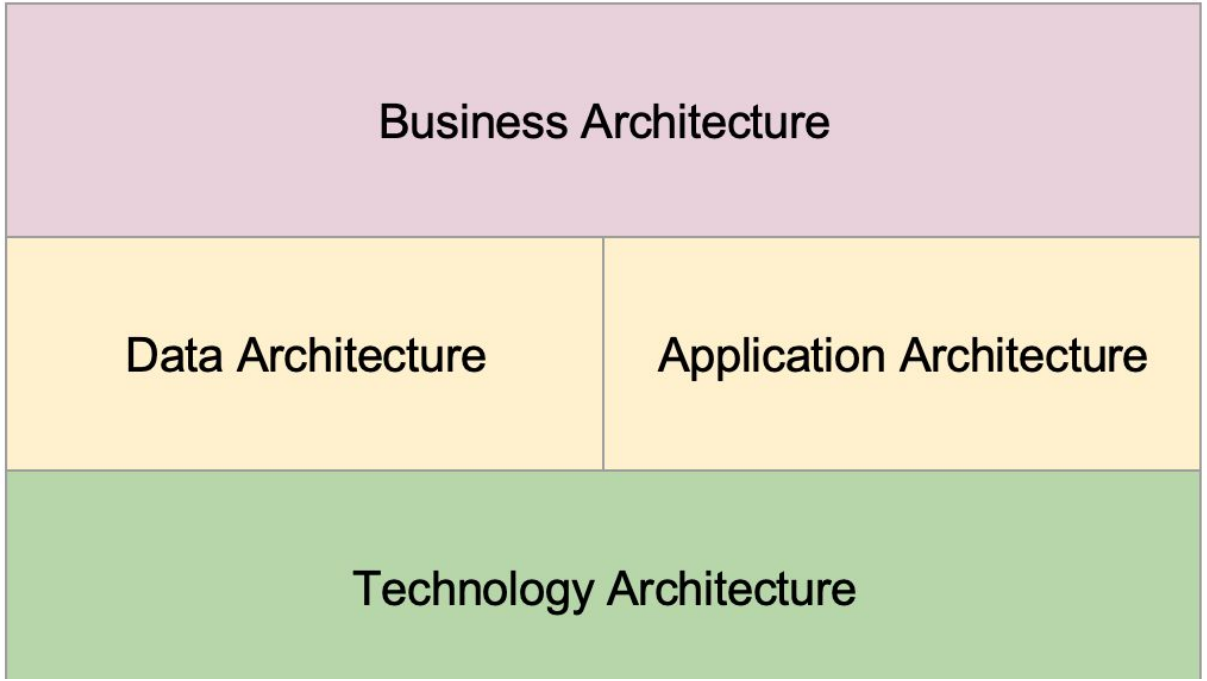


DATA ENGINEERING Lifecycle

- Data Engineering lifecycle is a part of Data Lifecycle.
  - **Generation**: source systems.
  - **Storage**: persistence support across the entire data engineering cycle (ingestion, transformation, serving).
  - **Ingestion**:
  - **Transformation**: data adjusting from initial (source) form to be used downstream.
  - **Serving data**: providing to downstream consumers such as Analytics (Business Intelligence, Operational Analytics, Embedded Analytics) or Machine Learning tools.

Data Architecture: domain of Enterprise Architecture

TOGAF Architectural Model [5]



Data Architecture: domain of Enterprise Architecture

- TOGAF Architectural Model [5]
  - **Business Architecture** refers to enterprise strategy and key business processes.
  - **Data Architecture** refers to logical and physical data assets and data management.
  - **Application Architecture** refers to application (and services) interactions, relationships and business process support.
  - **Technology Architecture** refers to logical and hardware infrastructure to support data and application architectures.

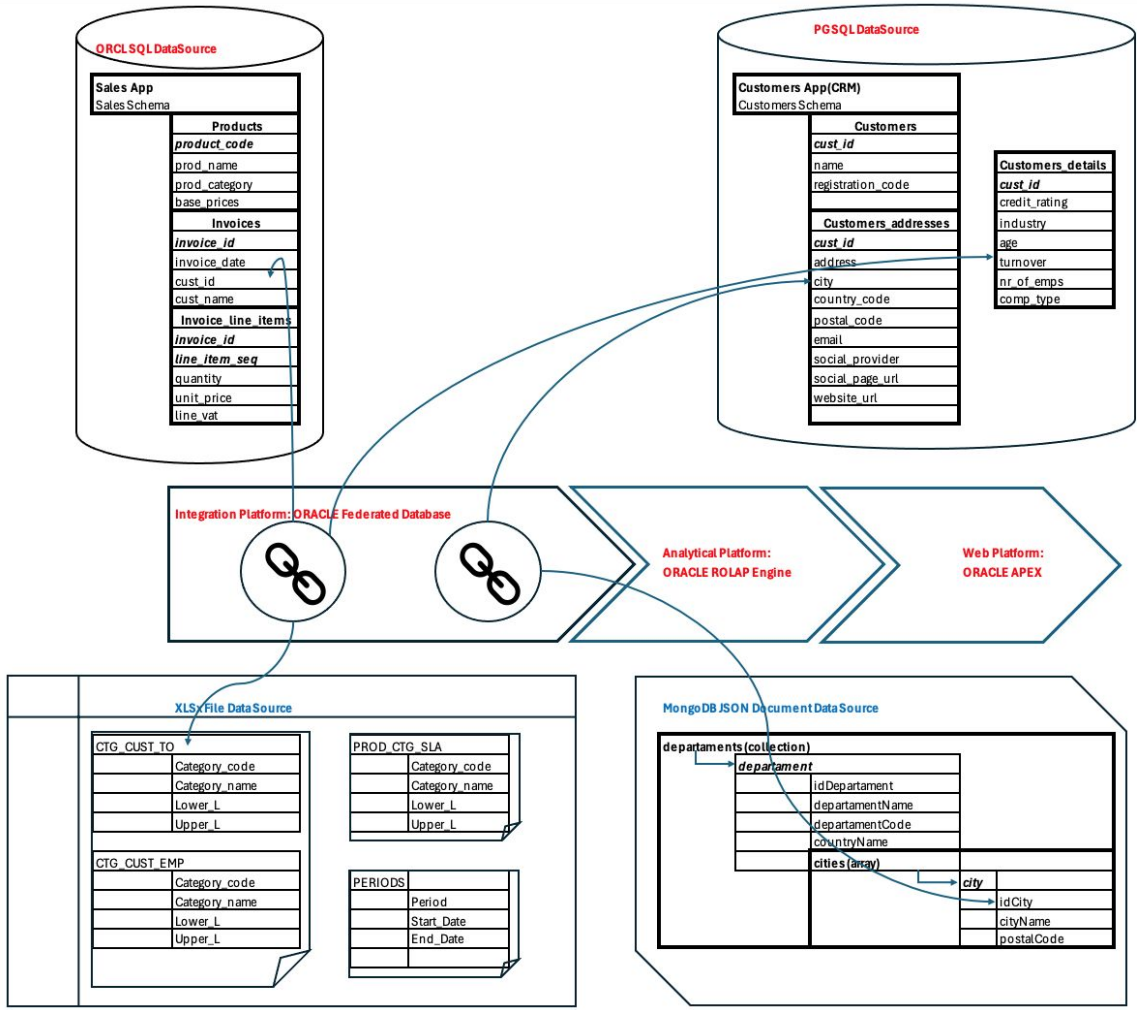
Data Architecture

- DA definition [4, 75]:

*"**Data architecture** is the design of systems to support the evolving data needs of an enterprise, achieved by flexible and reversible decisions reached through a careful evaluation of trade-offs."* (Reis&Housley, 2022)

Data Integration and DATA ENGINEERING

- Data engineering(DE) could be considered a more advanced perspective/approach than Data Integration:
  - broader in scope: covering more data management activities;
  - more comprehensible concerning Data Architecture
- Data Integration(DI) and Data Engineering overlap, but DI is more focused on the ingestion and transformation phases of DE lifecycle, and is less concerned on storage and analytics.


ORCL SQL DataSource
Sales App
Sales Schema
Products
product_code
prod_name
prod_category
base_prices
Invoices
invoice_id
Invoice_date
cust_id
cust_name
Invoice_line_items
invoice_id
line_item_seq
quantity
unit_price
line_vat
PGSQL DataSource
Customers App(CRM)
Customers Schema
Customers
cust_id
name
registration_code
Customers_addresses
cust_id
address
city
country_code
postal_code
email
social_provider
social_page_url
website_url
Customers_details
cust_id
credit_rating
industry
age
turnover
nr_of_emps
comp_type
Integration Platform: ORACLE Federated Database
Analytical Platform:
ORACLE ROLAP Engine
Web Platform:
ORACLE APEX
XLSx File Data Source
CTG_CUST_TO
Category_code
Category_name
Lower_L
Upper_L
CTG_CUST_EMP
Category_code
Category_name
Lower_L
Upper_L
PROD_CTG_SLA
Category_code
Category_name
Lower_L
Upper_L
PERIODS
Period
Start_Date
End_Date
MongoDB JSON Document Data Source
departaments (collection)
departament
idDepartament
departamentName
departamentCode
countryName
cities (array)
city
idCity
cityName
postalCode

References

- 1. AnHai Doan, Alon Halevy, Zachary Ives, *Principles of Data Integration*, 2012 Elsevier, Inc.
- 2. Anthony Giordano, *Data integration : blueprint and modeling techniques for a scalable and sustainable architecture*, 2010, Pearson Education, Inc.
- 3. April Reeve, *Managing Data in Motion Data Integration Best Practice Techniques and Technologies*, 2013 Elsevier, Inc.
- 4. Joe Reis and Matt Housley, *Fundamentals of Data Engineering*, O'Reilly Media, Inc., 2022
- 5. TOGAF, *Digital Edition of the TOGAF Standard, ADM Techniques, 2. Architecture Principles*, [online] Available at: https://pubs.opengroup.org/togaf-standard/adm-techniques/chap02.html [Accessed 17.05.2023].