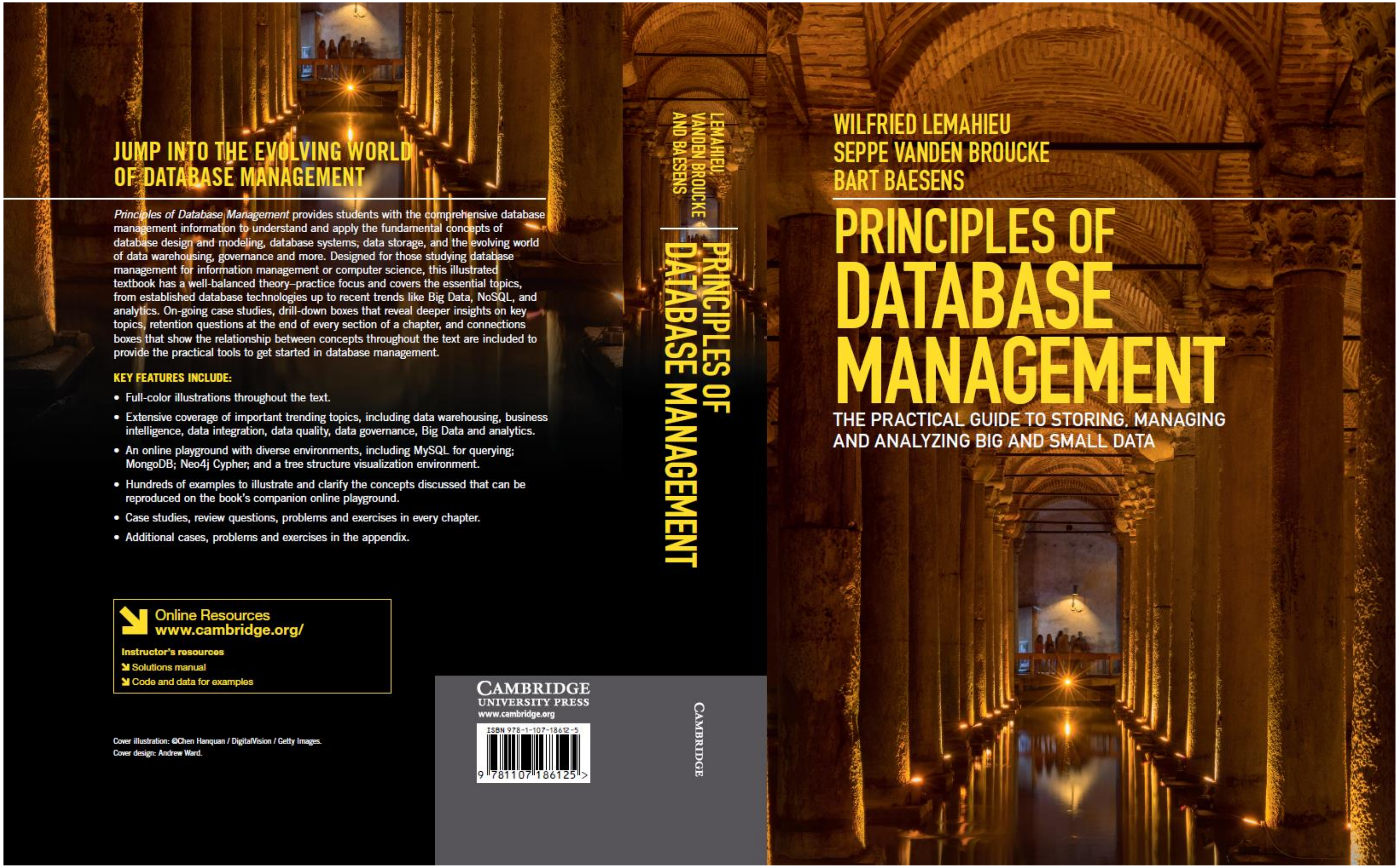
Data Integration, Data Quality and Data Governance



JUMP INTO THE EVOLVING WORLD OF DATABASE MANAGEMENT

Principles of Database Management provides students with the comprehensive database management information to understand and apply the fundamental concepts of database design and modeling, database systems, data storage, and the evolving world of data warehousing, governance and more. Designed for those studying database management for information management or computer science, this illustrated textbook has a well-balanced theory–practice focus and covers the essential topics, from established database technologies up to recent trends like Big Data, NoSQL, and analytics. On-going case studies, drill-down boxes that reveal deeper insights on key topics, retention questions at the end of every section of a chapter, and connections boxes that show the relationship between concepts throughout the text are included to provide the practical tools to get started in database management.

KEY FEATURES INCLUDE:

- Full-color illustrations throughout the text.
- Extensive coverage of important trending topics, including data warehousing, business intelligence, data integration, data quality, data governance, Big Data and analytics.
- An online playground with diverse environments, including MySQL for querying; MongoDB; Neo4j Cypher; and a tree structure visualization environment.
- Hundreds of examples to illustrate and clarify the concepts discussed that can be reproduced on the book's companion online playground.
- Case studies, review questions, problems and exercises in every chapter.
- Additional cases, problems and exercises in the appendix.

Online Resources
www.cambridge.org/

Instructor's resources
Solutions manual
Code and data for examples

Cover illustration: ©Chen Hanquan / DigitalVision / Getty Images.
Cover design: Andrew Ward.

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-18612-5

LEMAHIEU, VANDEN BROUCKE AND BAESENS

PRINCIPLES OF DATABASE MANAGEMENT

CAMBRIDGE

WILFRIED LEMAHIEU
SEPPE VANDEN BROUCKE
BART BAESENS

PRINCIPLES OF DATABASE MANAGEMENT

THE PRACTICAL GUIDE TO STORING, MANAGING AND ANALYZING BIG AND SMALL DATA

www.pdbmbook.com

Introduction

- Data and Process Integration
- Data Quality and Master Data Management
- Data Governance
- Outlook

Data and Process Integration

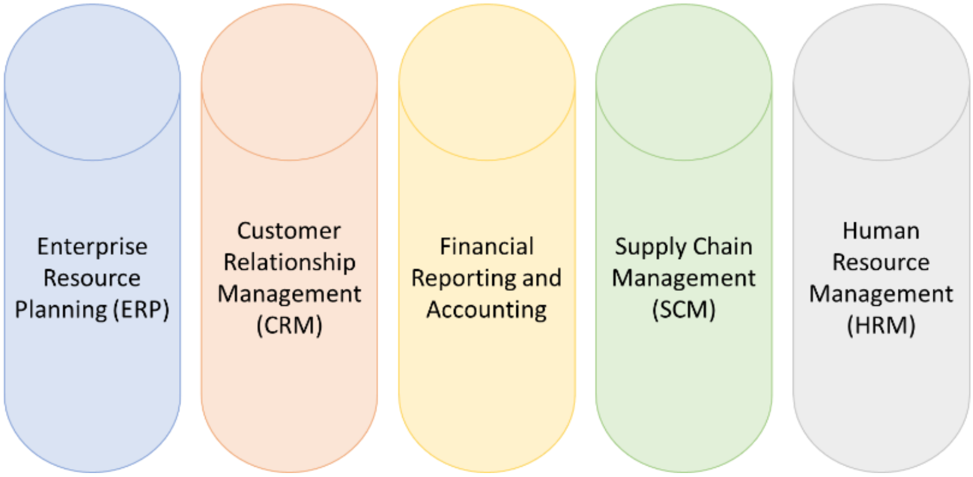
- Convergence of Analytical and Operational Data Needs
- Data Integration and Data Integration Patterns
- Data Services and Data Flows in the Context of Data and Process Integration

Convergence of Analytical and Operational Data Needs

- Data integration aims at providing a unified view and/or unified access over heterogeneous, and possibly distributed, data sources
- Process integration deals with sequencing of tasks in a business process but also governs data flows in these processes
- Both data and processes considered in data integration

Convergence of Analytical and Operational Data Needs

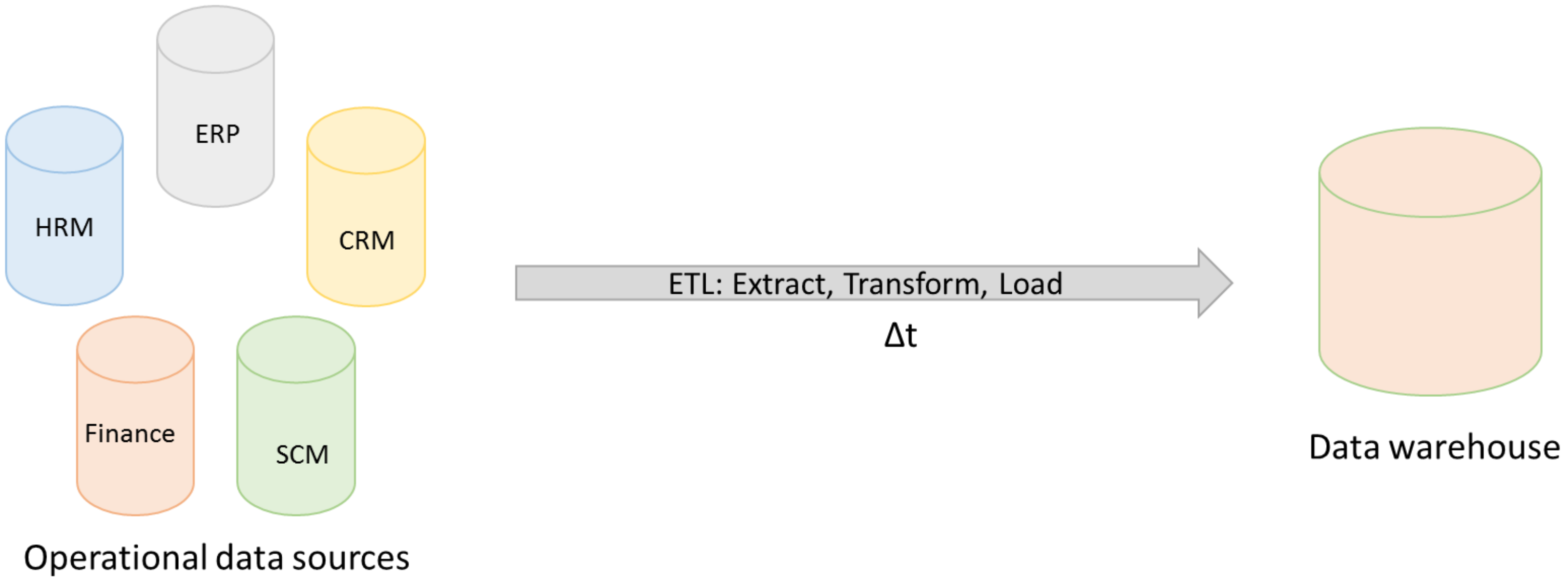
- Applications and databases traditionally organized around domains such as accounting, human resources, logistics, CRM
- Data silos aimed at operational support


Enterprise Resource Planning (ERP)
Customer Relationship Management (CRM)
Financial Reporting and Accounting
Supply Chain Management (SCM)
Human Resource Management (HRM)

- Emergence of BI and analytics triggered need to consolidate data into data warehouse

Convergence of Analytical and Operational Data Needs

- Dual data storage and processing landscape
  - operational applications: simple queries based on up to date ‘snapshot’ of the business
  - BI and analytics: complex queries based on slightly outdated data warehouse with historical, enriched and aggregated data



Convergence of Analytical and Operational Data Needs

- Convergence of operational and tactical/strategic data
- Dual focus of operational BI
  - usage of analytical techniques at the operational level
  - usage of real-time operational data combined with aggregated and historical data by tactic/strategic analytics
- Operational BI aims for low (or zero) latency

Convergence of Analytical and Operational Data Needs

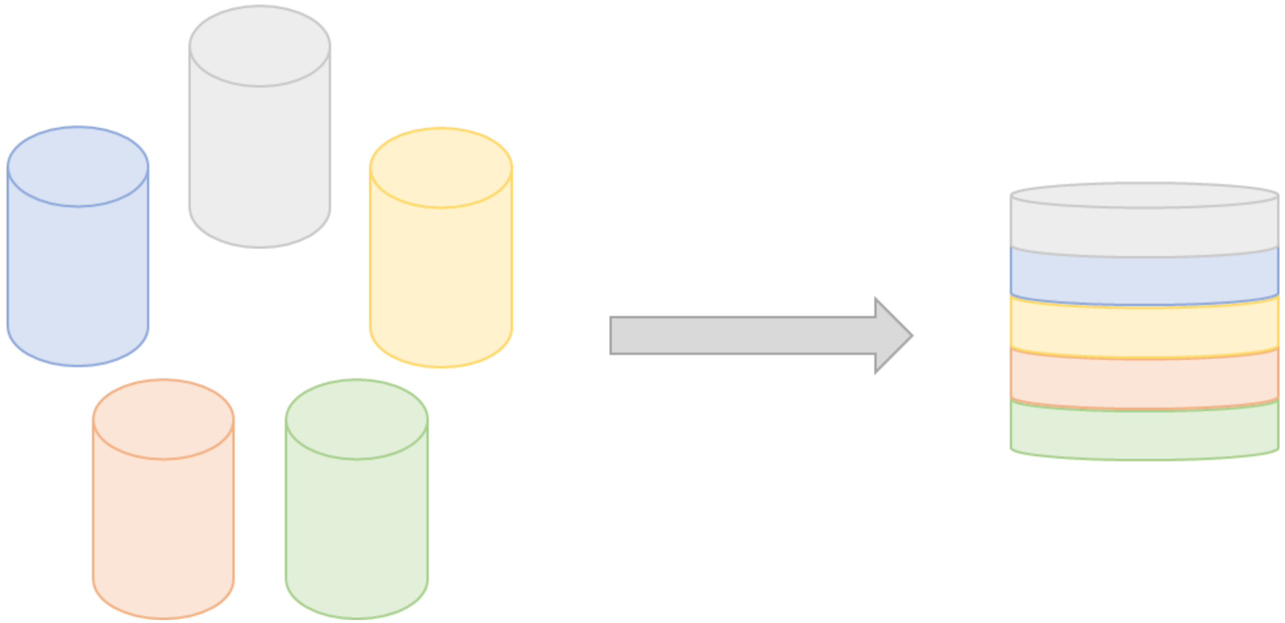
- Examples of operational BI
  - executive dashboards that monitor KPIs in real-time
  - business process/activity monitoring for timely detection of anomalies
  - real-time recommender systems (Amazon, Netflix)
- Data storage/integration challenges
  - combining traditional data types with ‘new’ types of internal and external data
  - integration of new data types

Data Integration and Data Integration Patterns

- Data integration
- Data Consolidation: Extract, Transform, Load (ETL)
- Data Federation: Enterprise Information Integration (EII)
- Data Propagation: Enterprise Application Integration (EAI)
- Data Propagation: Enterprise Data Replication (EDR)
- Changed Data Capture (CDC), Near Real Time ETL and Event Processing
- Data Virtualization
- Data as a Service and Data in the Cloud

Data Integration

- Data integration aims at providing a unified and consistent view of all data
- Extent of data integration depends on QoS
- Integration can be logical or physical

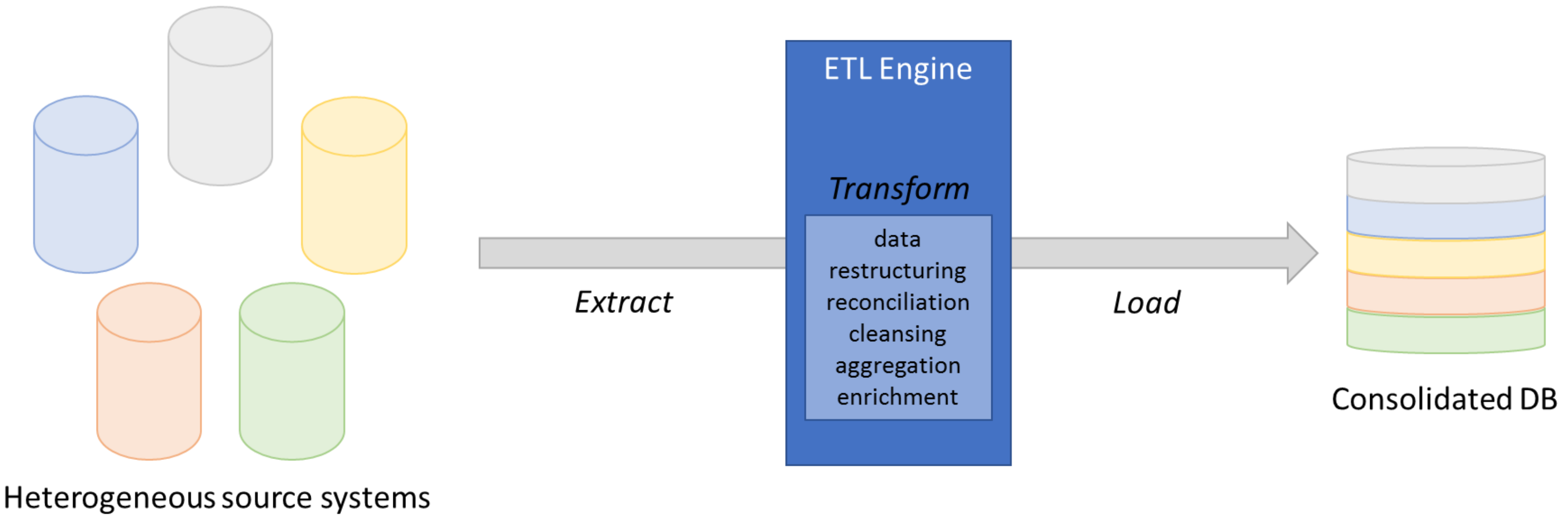


Data Consolidation: Extract, Transform, Load (ETL)

- Capture data from multiple, heterogeneous sources and integrate into a single persistent store
- ETL activities
  - extract data
  - transform data
  - load transformed data
- ETL has positive impact on data quality
- ETL induces a measure of latency and requires additional storage

Data Consolidation: Extract, Transform, Load (ETL)

- ETL variations:
  - full update or incremental refreshment strategy
  - ELT (Extract, Load, Transform): transformation directly in physical target system



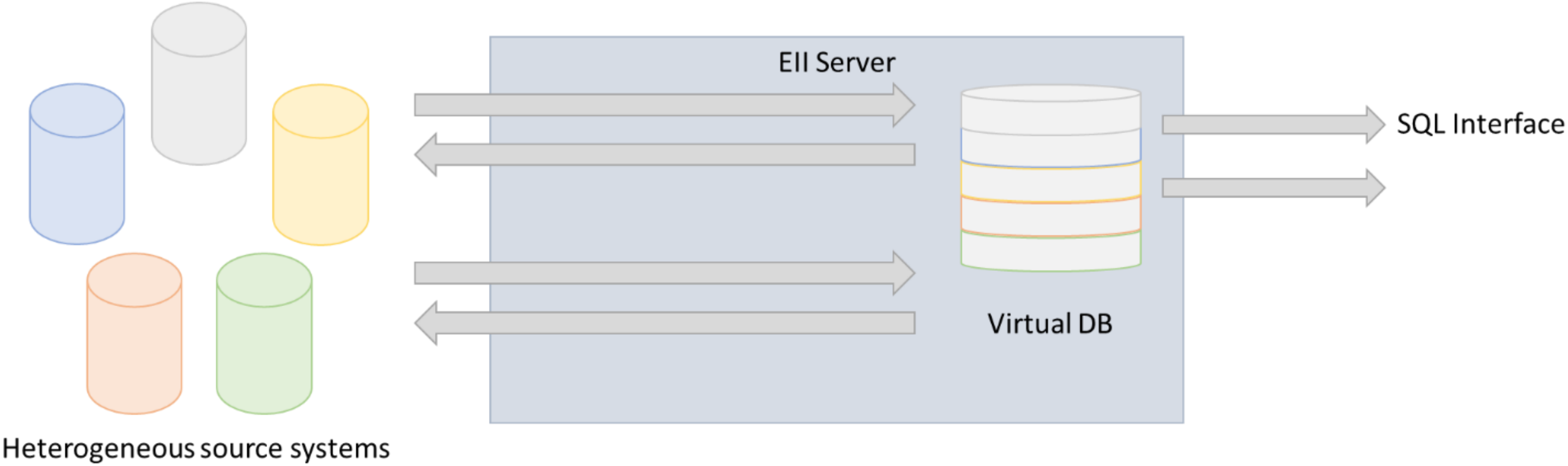
Data Consolidation: Extract, Transform, Load (ETL)

- Data lakes
  - data consolidated in native format
  - positive impact on data quality limited
  - analyzing data requires preprocessing and restructuring

Data Federation: Enterprise Information Integration (EII)

- Data federation follows a pull approach
- Example: Enterprise Information Integration (EII)
  - can be implemented by a view
  - no moving or replication of data is needed
  - enables real-time access to current data (↔ data consolidation)
  - only limited transformation and cleansing
  - read-only or write access
  - less suitable for complex queries
  - often adopted by firms as a temporary measure

Data Federation: Enterprise Information Integration (EII)



Data Federation: Enterprise Information Integration (EII)

- Performance hit since queries on the view must be translated to underlying data sources
- Operational systems may incur increased utilization rate (direct queries + queries from federation layer)
- EII solutions are limited in terms of transformation and cleansing

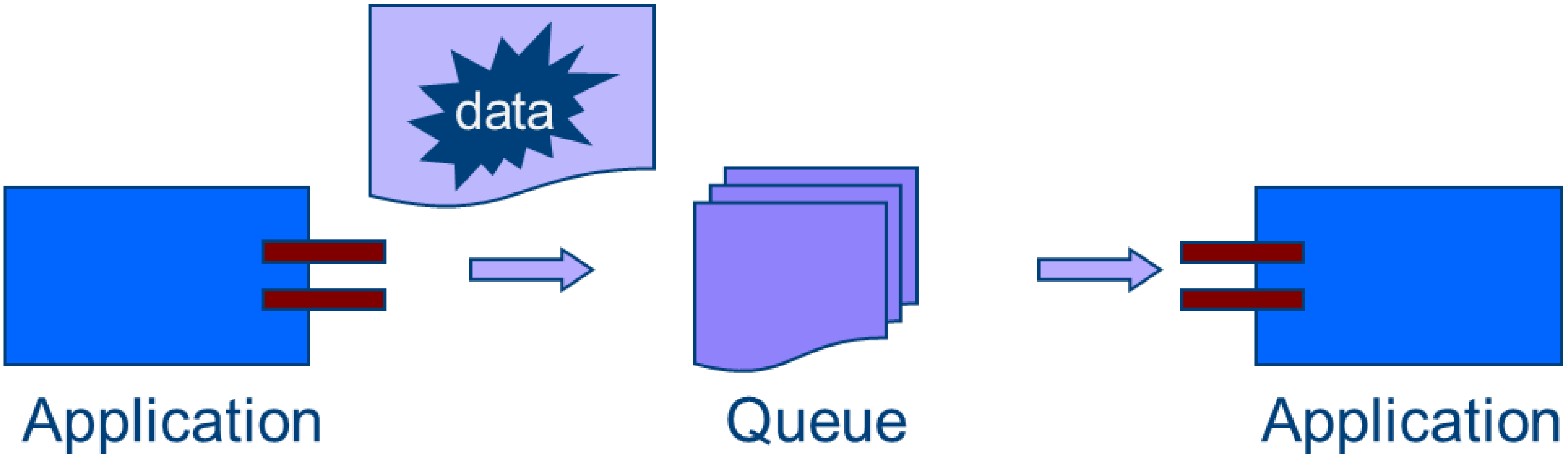
Data Propagation: Enterprise Application Integration (EAI)

- (A)synchronous propagation of updates or events in source to target system
- Two levels
  - Enterprise Application Integration (EAI): interaction between two applications
  - Enterprise Data Replication (EDR): synchronization between data stores

Data Propagation: Enterprise Application Integration (EAI)

- Enterprise Application Integration (EAI)
  - event in source application requires processing within target application
  - web services, .NET or Java interfaces, messaging middleware, etc.
  - usually involves small amounts of data being propagated from source to target application

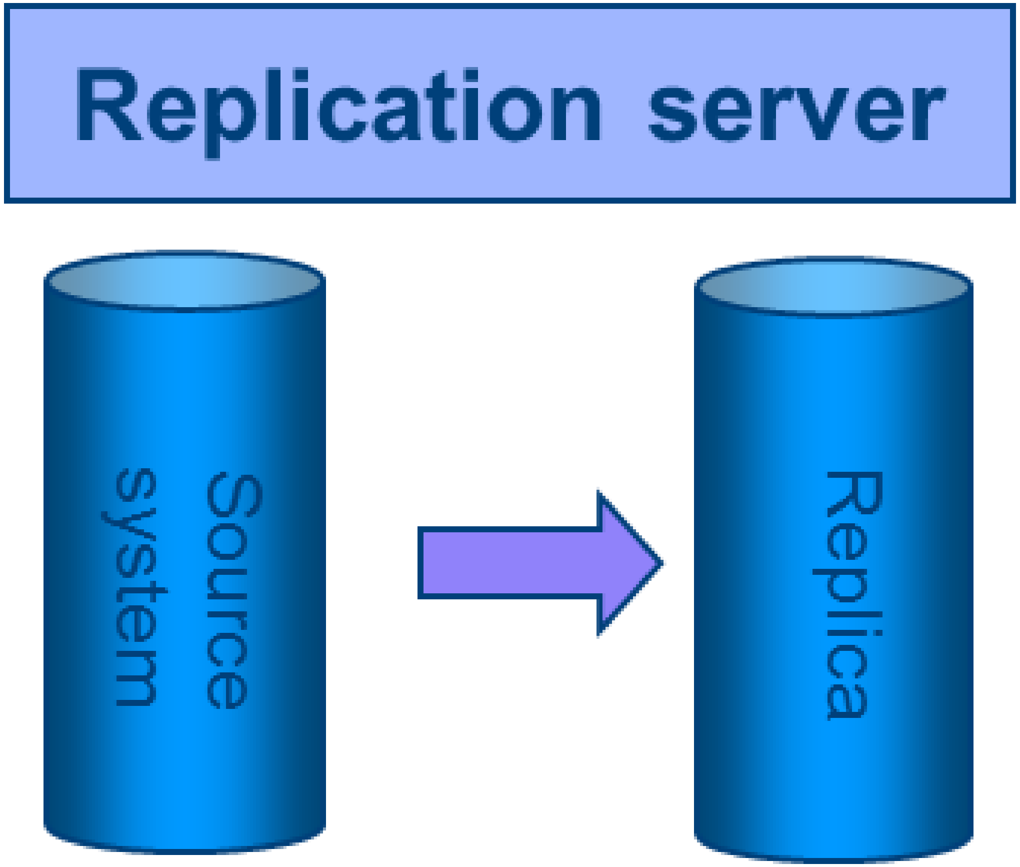
Data Propagation: Enterprise Application Integration (EAI)



Data Propagation: Enterprise Data Replication (EDR)

- Events in source system explicitly pertain to update events in data store
- Replication copies updates in source system in (near) real time to target data store
- By operating system, DBMS or replication server
- Traditionally adopted for load balancing
- Used for BI and to offload data from the source systems

Data Propagation: Enterprise Data Replication (EDR)



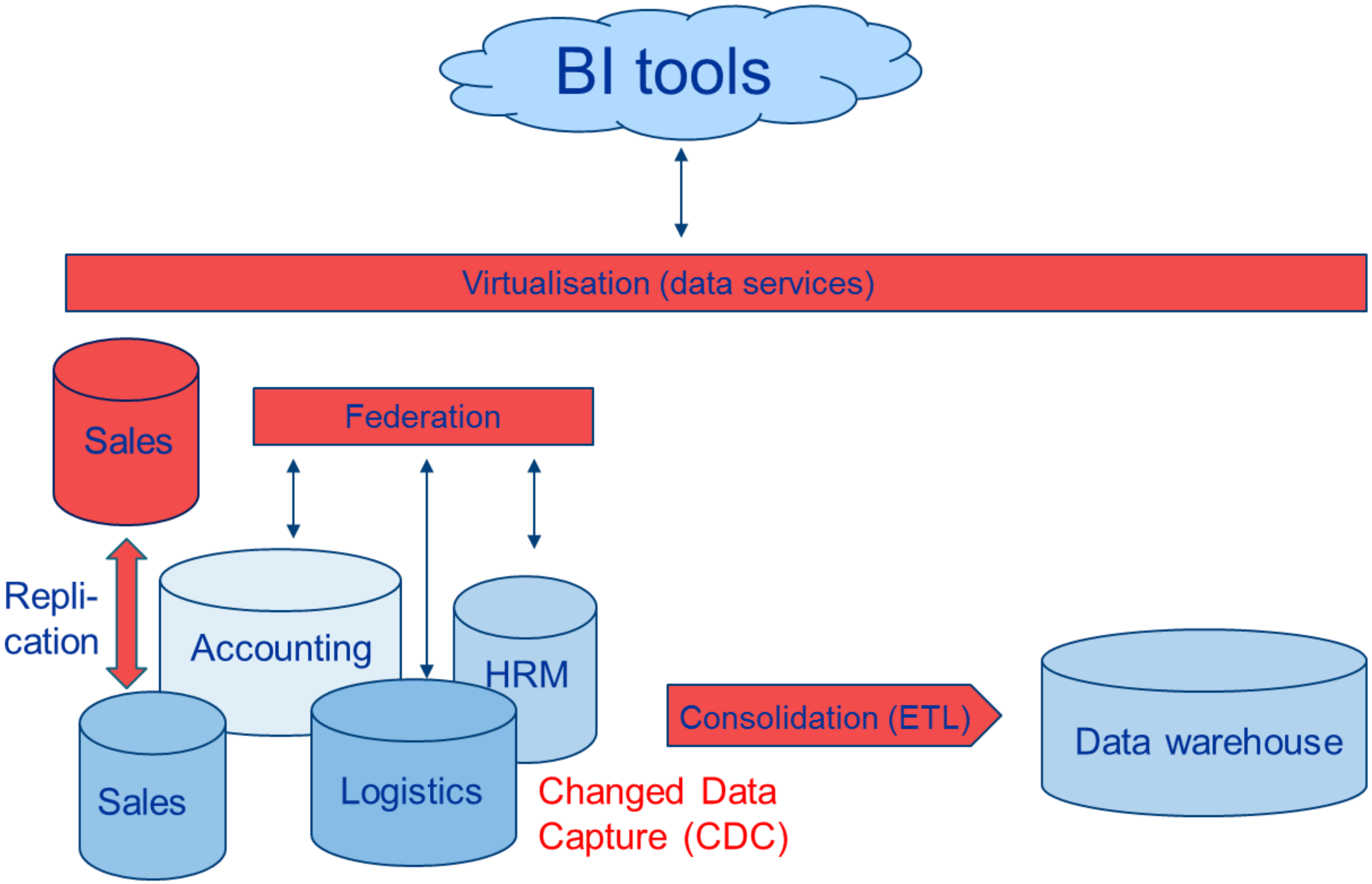
Changed Data Capture (CDC), Near Real Time ETL and Event Processing

- Changed Data Capture (CDC) adds event paradigm to ETL
- CDC can detect update events in source data and trigger ETL process (‘push’ model to ETL)
- Technically more complex but real-time capability and reduced network load
- Note: event notification pattern can also play other roles
  - Complex event processing: analytics techniques that focus on the interrelationships between events and patterns within event clouds

Data Virtualization

- Builds upon data integration patterns but isolates applications and users from the integration patterns
- ETL usually avoided: source data remains in place
- Contrary to federation (e.g., EII), virtualization does not impose a single data model
- Views can be defined and mapped top-down
- Can apply various transformations
- Views are cached transparently, and query optimization and parallelization techniques applied

Data Virtualization



Data as a Service and Data in the Cloud

- Data as a Service (DaaS): data services are offered as part of Service Oriented Architecture (SOA)
- Data services can be read-only or updatable
- Data service composition: combine data from different services into a new, composite service
- Self-service BI: data services can be composed, and then subjected to data analytics algorithms, simply by a business user dragging and dropping icons in a GUI

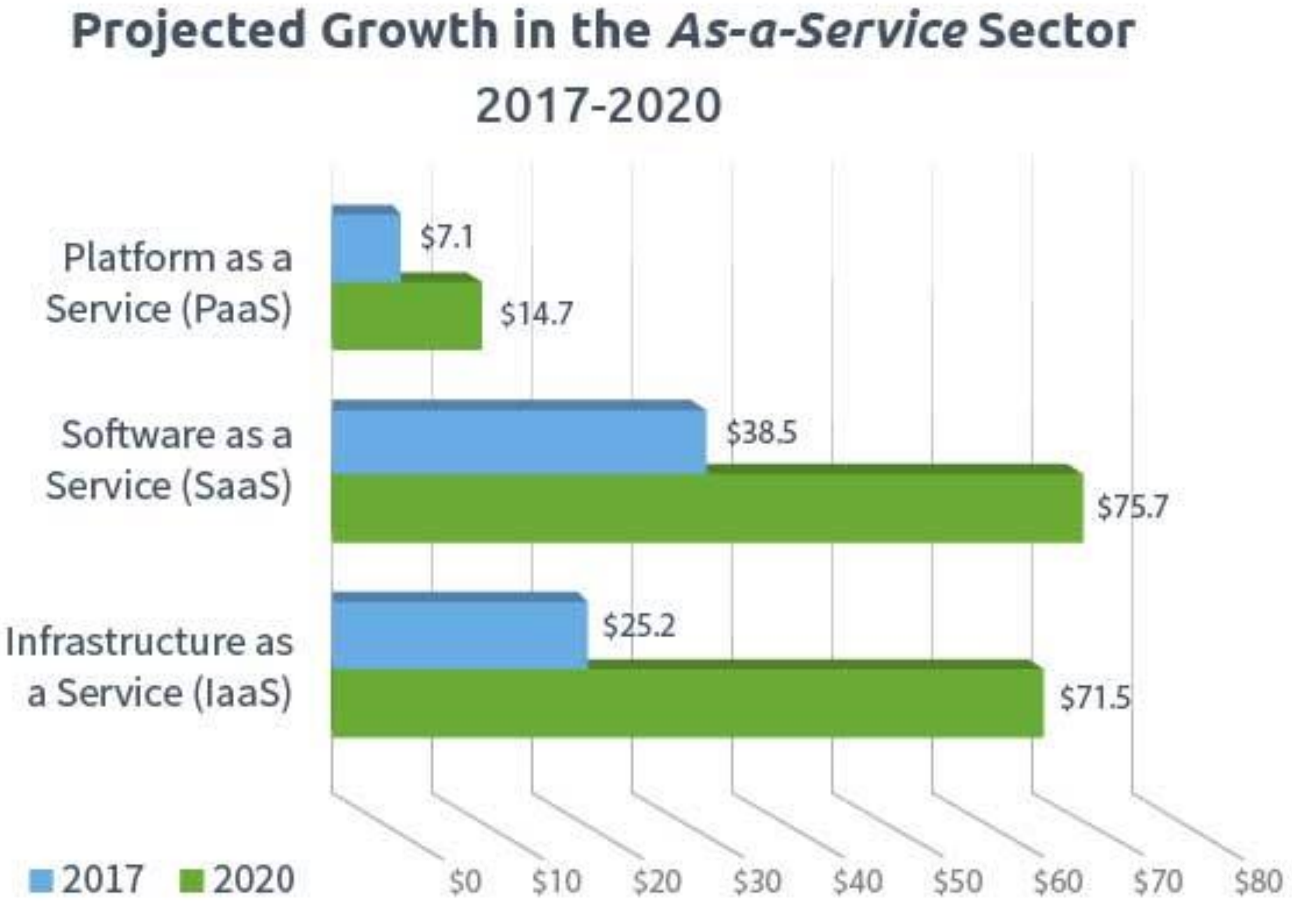
Data as a Service and Data in the Cloud

- ‘ as a service’ and ‘in the cloud’ concepts are related
- Properties of cloud computing
  - hardware, software and/or infrastructure are provided ‘on demand’ over a network
  - clouds can be public, private or hybrid
  - fading boundaries
  - converse fixed infrastructure costs, and upfront investments, into variable costs
  - risks: vendor lock-in, performance, privacy, security, accountability

Data as a Service and Data in the Cloud

- Cloud offerings
  - Software as a Service (SaaS): full applications
  - Platform as a Service (PaaS): computing platform elements
  - Infrastructure as a Service (IaaS): hardware offered as virtual machines
  - Data as a Service (DaaS): data services

Data as a Service and Data in the Cloud



Gartner, Forecast: Public Cloud Services, Worldwide, 2014-2020, 4Q16 Update, 2017.

Data Services and Data Flows in the Context of Data and Process Integration

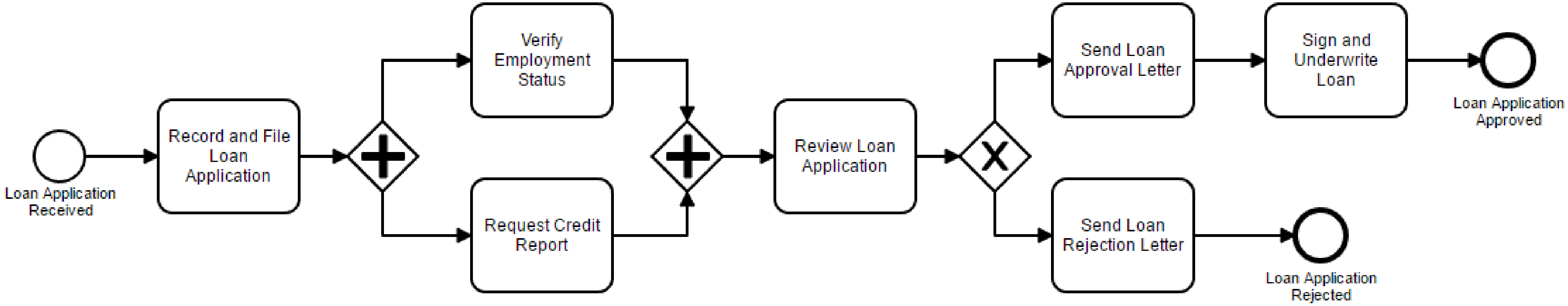
- Business Process Integration
- Patterns for Managing Sequence Dependencies and Data Dependencies in Processes
- A Unified View on Data and Process Integration

Business Process Integration

- Process integration aims at integrating business processes in an organization as much as possible
- Business process: set of tasks with a certain ordering that must be executed to reach a goal
  - Example: loan approval process
- Two perspectives:
  - control-flow: correct sequencing of tasks
  - data flow: inputs of the tasks

Business Process Integration

- Modelling of business processes is often performed using visual, flowchart-like languages BPMN, YAWL, UML Activity diagrams, etc.


Loan Application Received
Record and File Loan Application
Verify Employment Status
Request Credit Report
Review Loan Application
Send Loan Approval Letter
Sign and Underwrite Loan
Loan Application Approved
Send Loan Rejection Letter
Loan Application Rejected

Business Process Integration

- Process execution handled by process engine
- Process model translated into declarative definition of an executable process used by process engine
  - Business Process Execution Language standard (BPEL)
  - task coordination is separated from task execution
- Business processes can become quite complex
  - can consist of subprocesses
  - can span multiple organizational units
- Business processes tasks or subprocesses often offered as web services

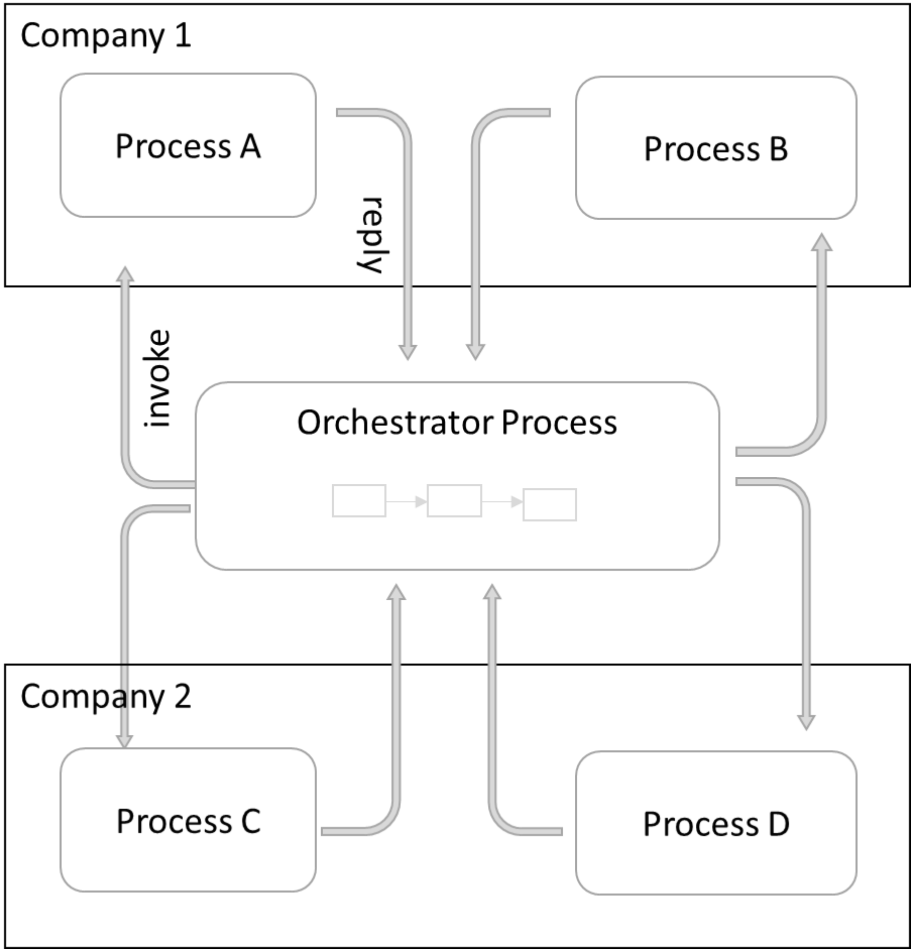
Business Process Integration

- Two types of dependencies
  - sequence dependency: execution of service B depends on the completion of the execution of service A
  - data dependency: execution of service B depends on data provided by service A

Patterns for Managing Sequence Dependencies and Data Dependencies in Processes

- Orchestration pattern
  - assumes a single centralized executable business process (orchestrator) that coordinates the interaction among different services and sub-processes
  - control flow and data flow is described at a single, central place and the orchestrator is responsible for invoking and combining the services

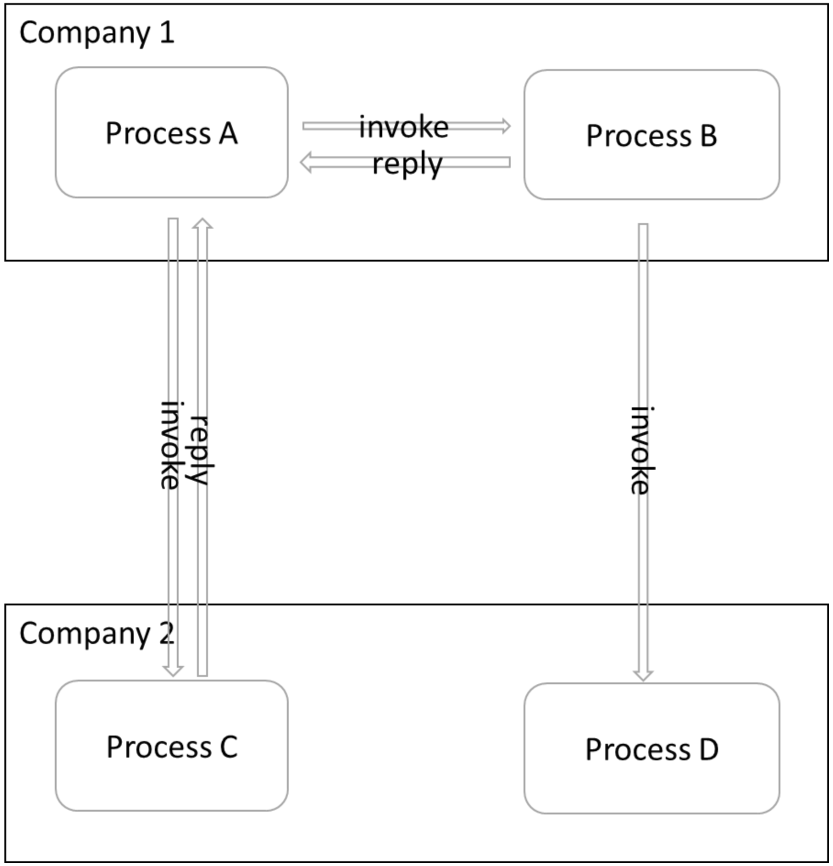
Patterns for Managing Sequence Dependencies and Data Dependencies in Processes



Patterns for Managing Sequence Dependencies and Data Dependencies in Processes

- Choreography pattern
  - relies on the participants themselves to coordinate their collaboration
  - decentralized approach where the decision logic and interactions are distributed, with no centralized point

Patterns for Managing Sequence Dependencies and Data Dependencies in Processes



Patterns for Managing Sequence Dependencies and Data Dependencies in Processes

- Combination of both often applied
- Choice of process integration pattern is made based on considerations regarding optimally managing sequence dependencies
  - – data flow then follows same pattern as control flow
- Decisions about sequence and data dependencies can be made independently
  - – some data dependencies may be satisfied using data flow at the process level, and some may be satisfied by data integration technology

Patterns for Managing Sequence Dependencies and Data Dependencies in Processes

- Data flow patterns at the process layer and data integration patterns at the data layer are complementary
- Data integration and (the data flow aspects of) process integration should be considered in a single effort

A Unified View on Data and Process Integration

- Data dependency between service A and B can be resolved in 2 ways:
  - process provides a data flow between A and B
  - service A persists the data into a data store which is also accessible to service B
- Managing data dependencies is a shared responsibility of the process and data layer
- Three types of services: workflow services, activity services, and data services
  - can correspond to actual software artefacts or used as an instrument for analysis

A Unified View on Data and Process Integration

- Workflow services
  - coordinate the control flow and data flow of a business process by triggering its respective tasks in line with the sequence constraints in the process model, and according to an orchestration or choreography pattern
  - tasks can be fully automated or human interfacing activities
  - data flow can be established by passing variables in a message or method invocation

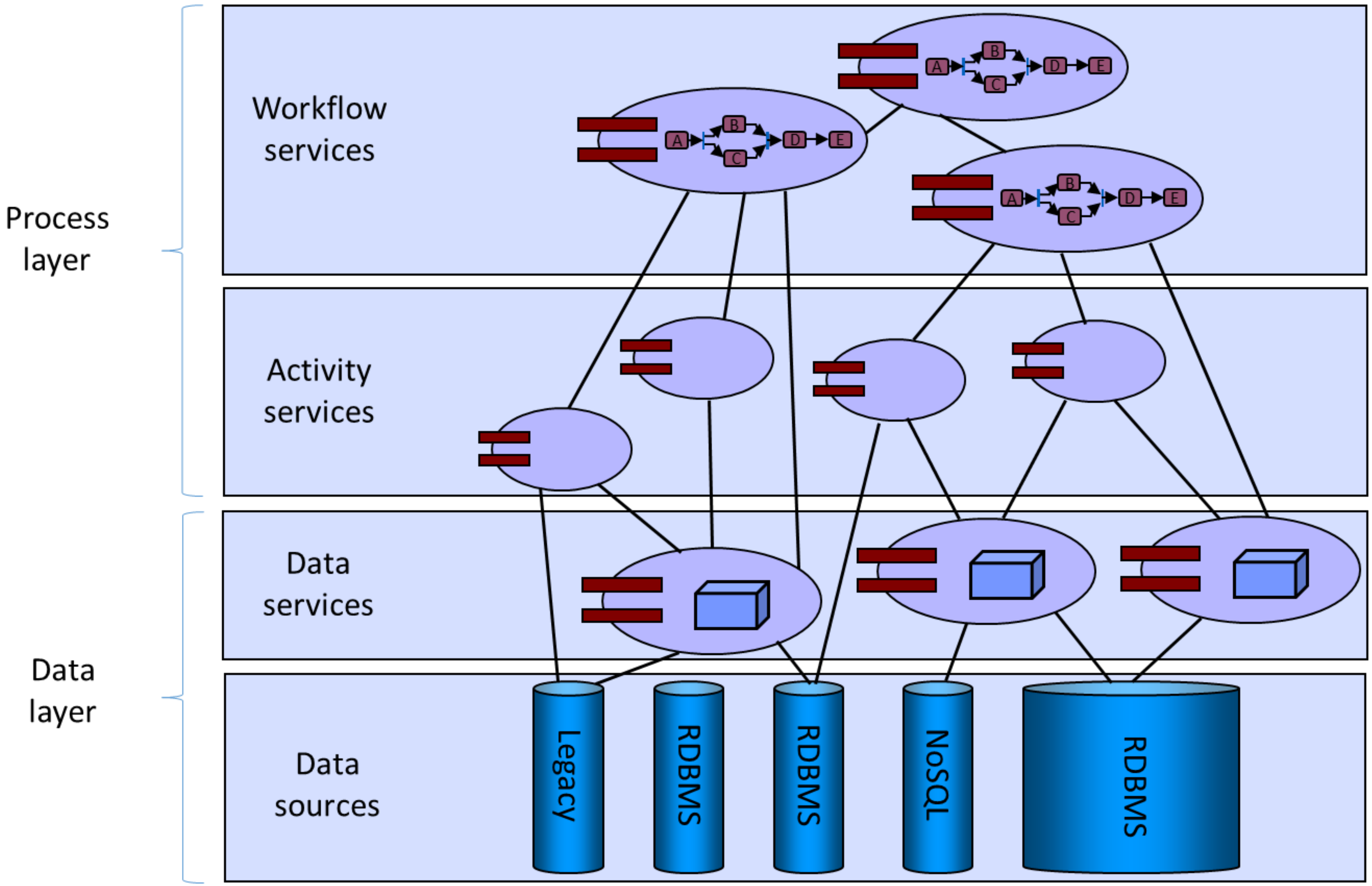
A Unified View on Data and Process Integration

- Activity services
  - perform one task in a business process
  - triggered by workflow service(s)
  - are triggered (representing the control flow) and may receive input variables (representing the data flow)
  - may return result to the workflow service or alter business state
  - may interact with data services to retrieve business state not provided in input variables

A Unified View on Data and Process Integration

- Data services
  - provide access to the business data
  - CRUDS functionality: Create, Read, Update, Delete and Search
  - read-only or not
  - unified access to the underlying data and realized using data integration patterns

A Unified View on Data and Process Integration



A Unified View on Data and Process Integration

- Data services can be realized according to different data integration patterns
  - federation provides real time, comprehensive data about business state
  - if extensive transformation, aggregation and/or cleansing are needed, or performance is an issue, it is better to use consolidation
  - ff only performance is a criterion without the need for transformation/cleansing or historical data, replication can be used

A Unified View on Data and Process Integration

- Data services perspective and process perspective should be combined to provide activity services with necessary input data
- Balance between input through data flow and through data layer is context dependent
- Sometimes, all necessary input data will be provided as part of the triggering of the activity service (comfort data)
  - Trade-off: more comfort data implies less dependence on data layer but increases risk of outdated data

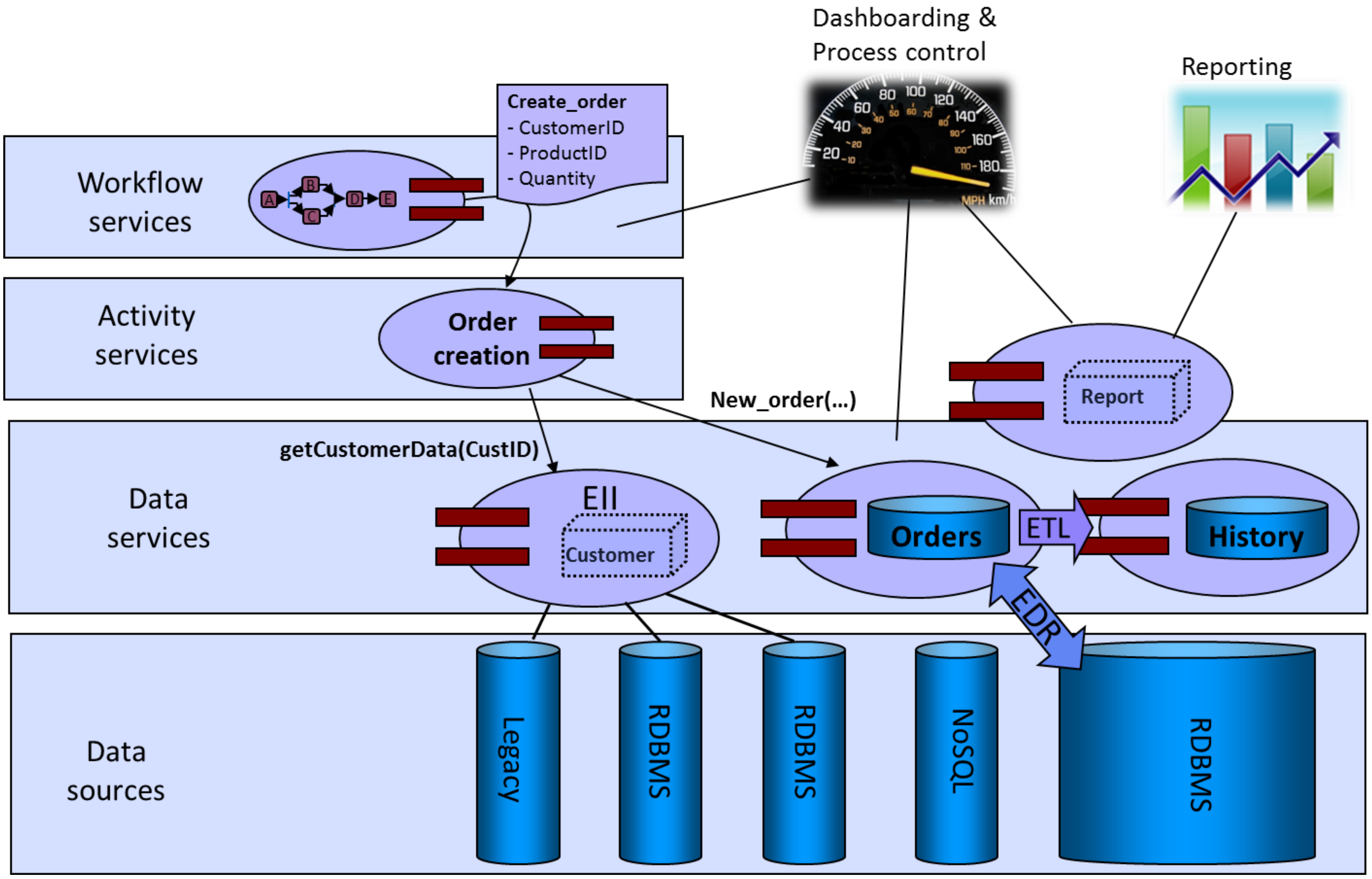
A Unified View on Data and Process Integration

- Data lineage refers to the whole trajectory followed by a data item, from its origin, possibly over respective transformations and aggregations, until it is ultimately being used or processed
- Take data integration patterns at data layer level, and data flow at business processes level into account to see the whole picture regarding data lineage and assess impact on data quality

A Unified View on Data and Process Integration

- Event data (when was an order created ? what is the order quantity) can be safely passed as data flow, as these data will never change
- Business state data (what is the customer’s current address ? what is the current stock ?) is safer to retrieve through the data layer when needed

A Unified View on Data and Process Integration



A Unified View on Data and Process Integration

- Most SOA enabled data integration suites provide different data related infrastructure services:
  - data profiling services
  - data cleansing services
  - data enrichment services
  - data transformation services
  - data event services
  - data auditing services
  - metadata services

Searching Unstructured Data and Enterprise Search

- Principles of Full Text Search
- Indexing Full Text Documents
- Web Search Engines
- Enterprise Search

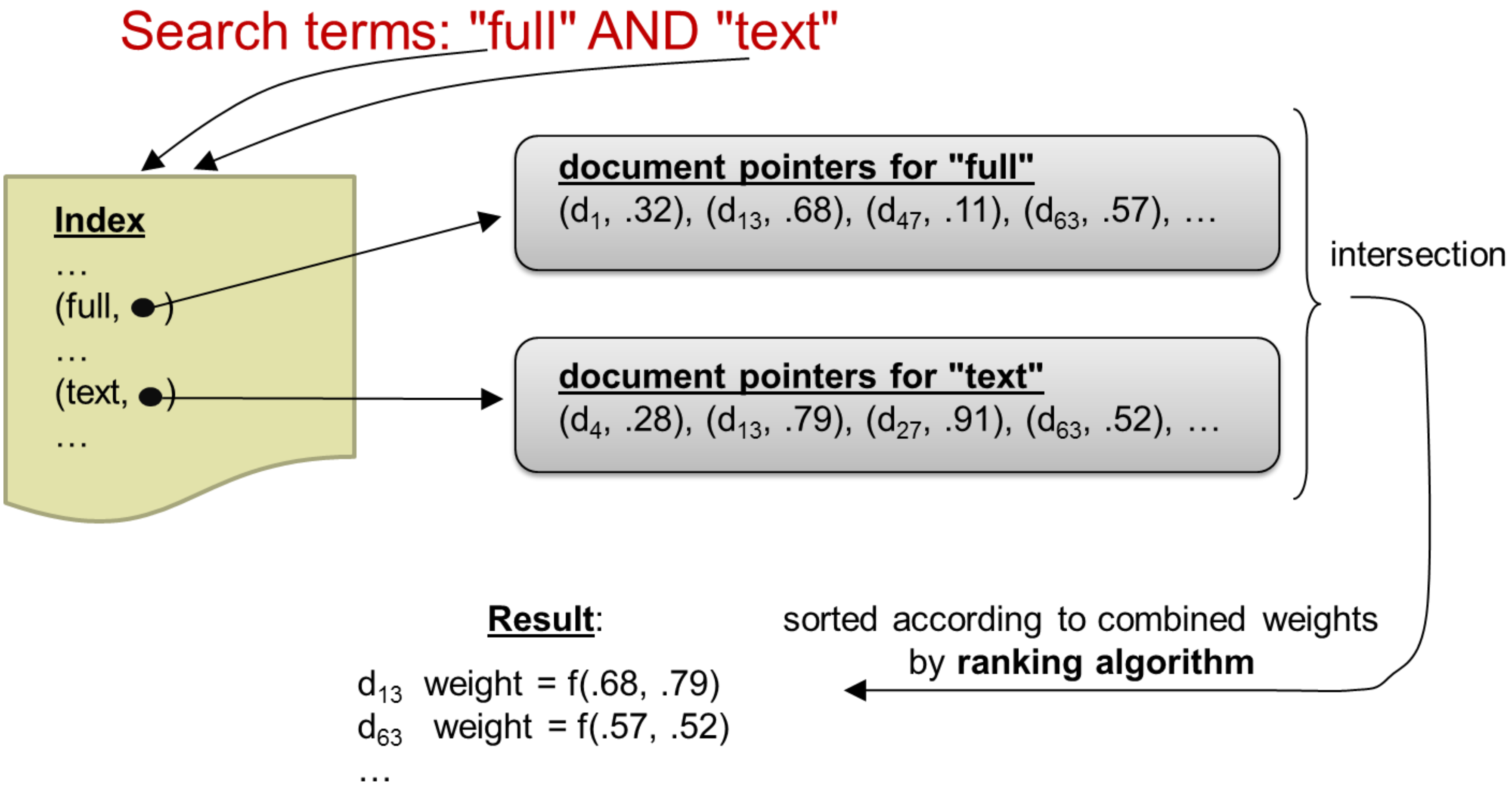
Principles of Full Text Search

- Structured data: can be described according to a formal logical data model
- Unstructured data: no finer grained components in a text document that can be interpreted in a meaningful way
- Idea of full text search is that individual text documents can be selected from a collection of documents according to the presence of a single (or combination of) search term(s)
- Additional criteria: proximity and absence
- Relevance can be measured by the frequency with which the search term(s) occur(s)

Indexing Full Text Documents

- Inverted index for indexing full text documents
  - document collection is parsed upfront, with only relevant terms being withheld
  - index entry is created for every individual search term consisting of (search term, list pointer) pairs, with the list pointer referring to a list of document pointers
  - for search term t_i the list is typically of this format: $[(d_{i1}, w_{i1}), \ldots (d_{in}, w_{in})]$. A list item (d_{ij}, w_{ij}) contains a document pointer d_{ij} and a weight w_{ij} denoting how important term t_i is to document j.
  - most search engines contain a lexicon, which maintains some statistics per search term
- Full text search then comes down to searching the index

Indexing Full Text Documents



Search terms: "full" AND "text"

Index

…

(full, ●)

…

(text, ●)

…

document pointers for "full"

(d_1, .32), (d_{13}, .68), (d_{47}, .11), (d_{63}, .57), …

document pointers for "text"

(d_4, .28), (d_{13}, .79), (d_{27}, .91), (d_{63}, .52), …

intersection

sorted according to combined weights by **ranking algorithm**

Result:

d_{13} weight = f(.68, .79)

d_{63} weight = f(.57, .52)

…

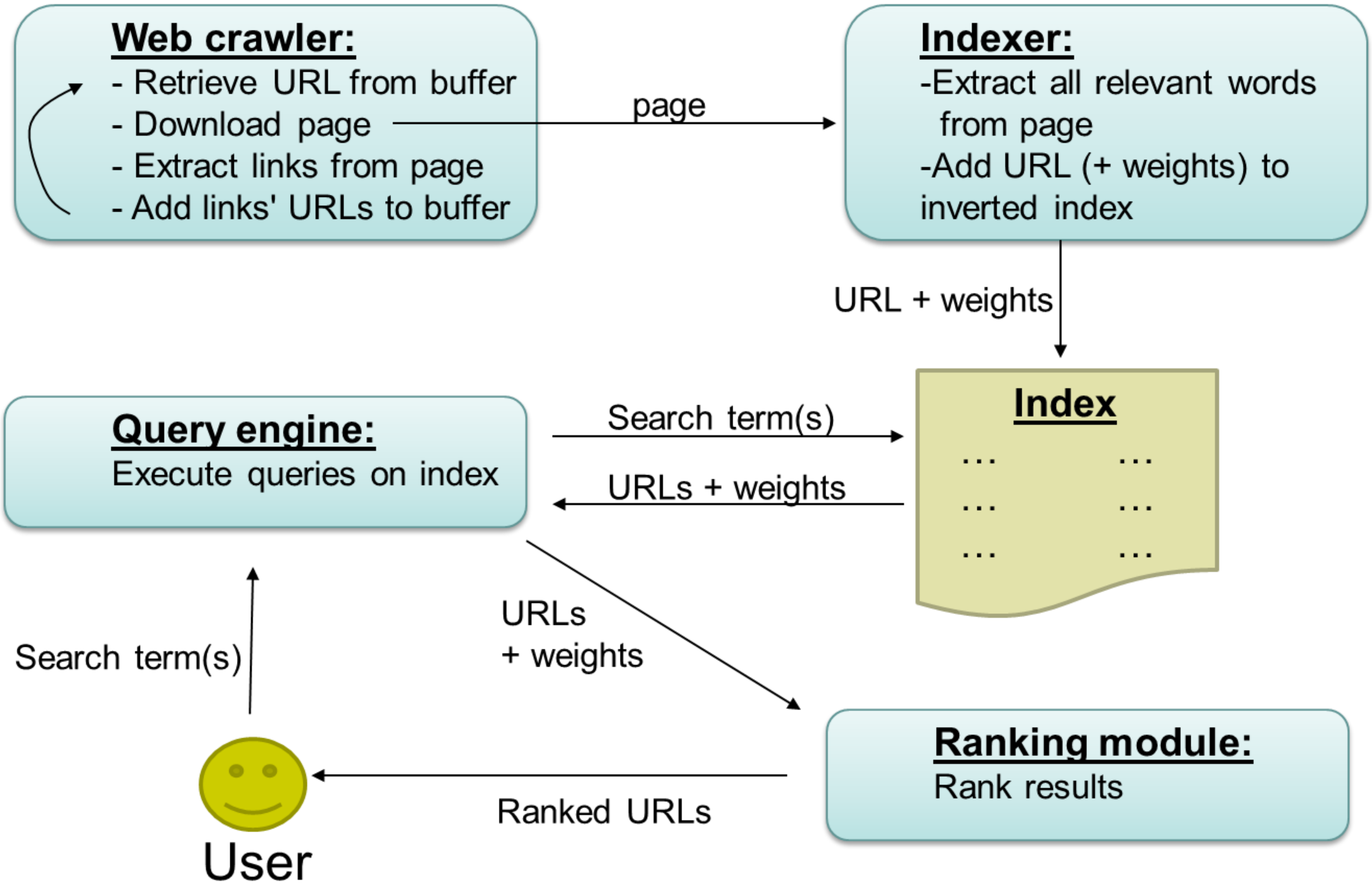
Indexing Full Text Documents

- Extensions
  - thesaurus
  - proximity
  - fuzzy logic or similarity measures
  - text mining
  - document metadata

Web Search Engines

- Web crawler (web spider): retrieves web pages, extracts their links and adds these URLs to a buffer that contains the links to pages yet to be visited
- Indexer: extracts all relevant terms from the page and updates the inverted index
  - each relevant term corresponds to an index entry, referring to a list with (d_{ij}, w_{ij}) pairs, with d_{ij} the web page's URL and w_{ij} the weight
- Ranking module: sorts the result set

Web Search Engines



Web crawler:
- Retrieve URL from buffer
- Download page
- Extract links from page
- Add links' URLs to buffer

Indexer:
-Extract all relevant words from page
-Add URL (+ weights) to inverted index

Query engine:
Execute queries on index

Ranking module:
Rank results

Enterprise Search

- Enterprise search: practice of making content stemming from various distributed data sources in an organization searchable
- Apache Lucene
  - information retrieval from text by offering indexing and searching capabilities
- ELK stack
  - Elasticsearch: adds additional APIs, distributed search support, grouping and aggregation in queries, and allows to store documents in JSON
  - Logstash: tool to collect and process data to store it into a backend
  - Kibana:web based analytics, visualization and search interface

Data Quality and Master Data Management

- Data integration is related to data quality
- Data quality can be defined as “fitness for use”
- Example data quality dimensions
  - data accuracy
  - data completeness
  - data consistency
  - data accessibility
  - data timeliness

Data Quality and Master Data Management

- Data integration can both improve and hamper data quality
  - E.g., environments where different integration approaches have been combined, leading to a jungle of systems
- Master data management (MDM): series of processes, policies, standards, and tools to help organizations to define and provide a single point of reference for all data that is “mastered”
  - provide a trusted, single version of the truth
  - focus on unifying company-wide reference data types

Data Quality and Master Data Management

- Setting up an MDM initiative involves many steps and tools, including data source identification, mapping out the systems architecture, constructing data transformation, cleansing and normalization rules, providing data storage capabilities, monitoring and governance facilities, …
- A key element is a centrally governed data model and metadata repository
- Data integration approaches can be used as a method to achieve maturity in master data management

Data Governance

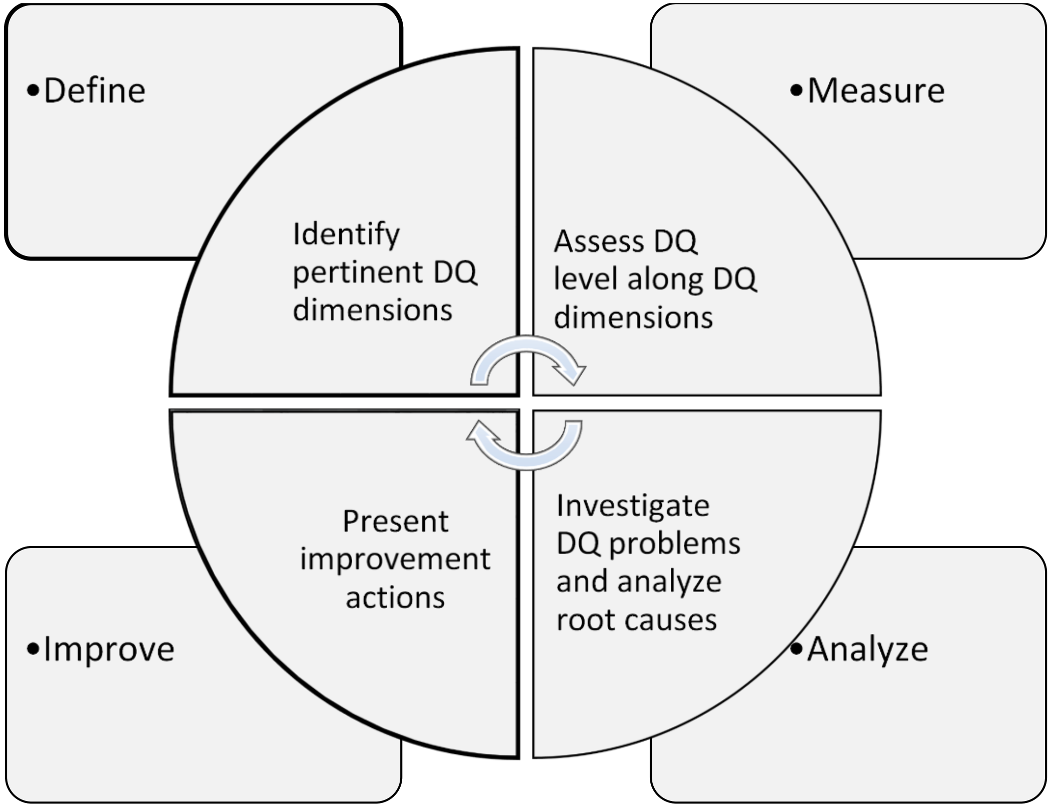
- Basic Ideas
- Total Data Quality Management (TQDM)
- Capability Maturity Model Integration (CMMI)
- Data Management Body of Knowledge (DMBOK)
- Control Objectives for Information and Related Technology (COBIT)
- Information Technology Infrastructure Library (ITIL)

Basic Ideas

- Organizations are increasingly implementing company-wide data governance initiatives to govern and oversee data quality and data integration
- Aim of data governance is to set up a company-wide controlled and supported approach towards data quality, accompanied by data quality management processes
- Manage data as an asset rather than a liability
- Different frameworks and standards have been introduced for data governance

Total Data Quality Management (TQDM)

- Wang, 1998



- Define
- Measure
- Analyze
- Improve

Identify pertinent DQ dimensions

Assess DQ level along DQ dimensions

Investigate DQ problems and analyze root causes

Present improvement actions

Capability Maturity Model Integration (CMMI)

- Geared towards the improvement of business processes
- Developed at Carnegie Mellon University (CMU)
- CMMI defines the maturity of a process by 5 levels
- Likewise, the Data Management Maturity Model also applies 5 levels of maturity to the governance of data, its quality and infrastructure:
  - Level 1 performed: emphasis is on data repair
  - Level 2 managed: there is awareness of the importance of managing data
  - Level 3 defined: data is treated as a critical asset for successful performance
  - Level 4 measured: data is treated as a source of competitive advantage and seen as a strategic asset
  - Level 5 optimized: data is seen as critical to survival in a dynamic market

Data Management Body of Knowledge (DMBOK)

- Overseen by DAMA International (the Data Management Association) and lists best practices towards data quality management, metadata management, data warehousing, data integration, and data governance
- Currently in its second version

Control Objectives for Information and Related Technology (COBIT)

- Created by the international professional association ISACA for IT management and governance
- Describes a series of implementable control sets and organizes them in a logical framework
- Goal is to link business goals to IT goals, starting from business requirements and mapping these to IT requirements, and hence provide metrics and maturity models to measure the effectiveness of these IT goals
- Comprehensive framework encompassing much more than just data governance

Information Technology Infrastructure Library (ITIL)

- Set of detailed practices for IT service management that focuses on aligning IT services with the needs and requirements of business
- Published in 5 volumes, each of which covers a different IT service management lifecycle stage
- Encompasses much more governance than just data quality and integration aspects

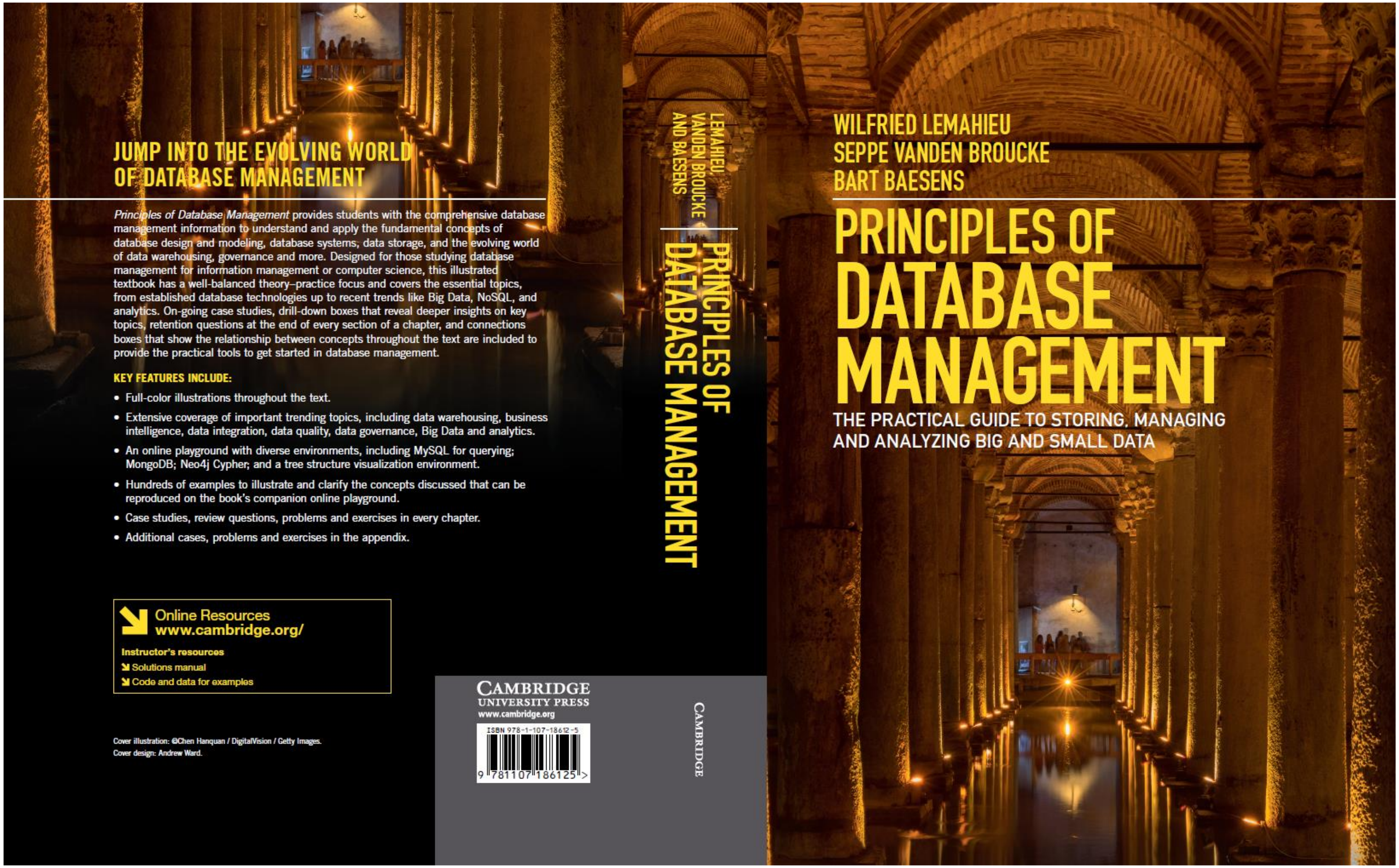
Outlook

- Many vendors, and cloud providers are trying to offer ways to handle the data integration issue in a world where companies are either moving their data to the cloud or are shifting to a Big Data environment
  - Sqoop and Flume for Hadoop
  - Apache Kylin
  - Google Cloud Dataflow and BigQuery ETL
  - Amazon Redshift
  - Amazon Relational Database Service (RDS)

Conclusions

- Data and Process Integration
- Data Quality and Master Data Management
- Data Governance
- Outlook

More information?



JUMP INTO THE EVOLVING WORLD
OF DATABASE MANAGEMENT

Principles of Database Management provides students with the comprehensive database management information to understand and apply the fundamental concepts of database design and modeling, database systems, data storage, and the evolving world of data warehousing, governance and more. Designed for those studying database management for information management or computer science, this illustrated textbook has a well-balanced theory–practice focus and covers the essential topics, from established database technologies up to recent trends like Big Data, NoSQL, and analytics. On-going case studies, drill-down boxes that reveal deeper insights on key topics, retention questions at the end of every section of a chapter, and connections boxes that show the relationship between concepts throughout the text are included to provide the practical tools to get started in database management.

KEY FEATURES INCLUDE:

- Full-color illustrations throughout the text.
- Extensive coverage of important trending topics, including data warehousing, business intelligence, data integration, data quality, data governance, Big Data and analytics.
- An online playground with diverse environments, including MySQL for querying; MongoDB; Neo4j Cypher; and a tree structure visualization environment.
- Hundreds of examples to illustrate and clarify the concepts discussed that can be reproduced on the book's companion online playground.
- Case studies, review questions, problems and exercises in every chapter.
- Additional cases, problems and exercises in the appendix.

Online Resources
www.cambridge.org/

Instructor's resources

- Solutions manual
- Code and data for examples

Cover illustration: ©Chen Hanquan / DigitalVision / Getty Images.
Cover design: Andrew Ward.

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-107-18612-5

LEMAHIEU
VANDEN BROUCKE
AND BAESENS

PRINCIPLES OF
DATABASE MANAGEMENT

CAMBRIDGE

WILFRIED LEMAHIEU
SEPPE VANDEN BROUCKE
BART BAESENS

PRINCIPLES OF
DATABASE
MANAGEMENT

THE PRACTICAL GUIDE TO STORING, MANAGING
AND ANALYZING BIG AND SMALL DATA

www.pdbmbook.com