MANNHEIM
BUSINESS SCHOOL

MANAGING BIG DATA

PROF. DR. FLORIAN STAHL

UNIVERSITY OF MANNHEIM

AACSB ACCREDITED

EQUIS ACCREDITED

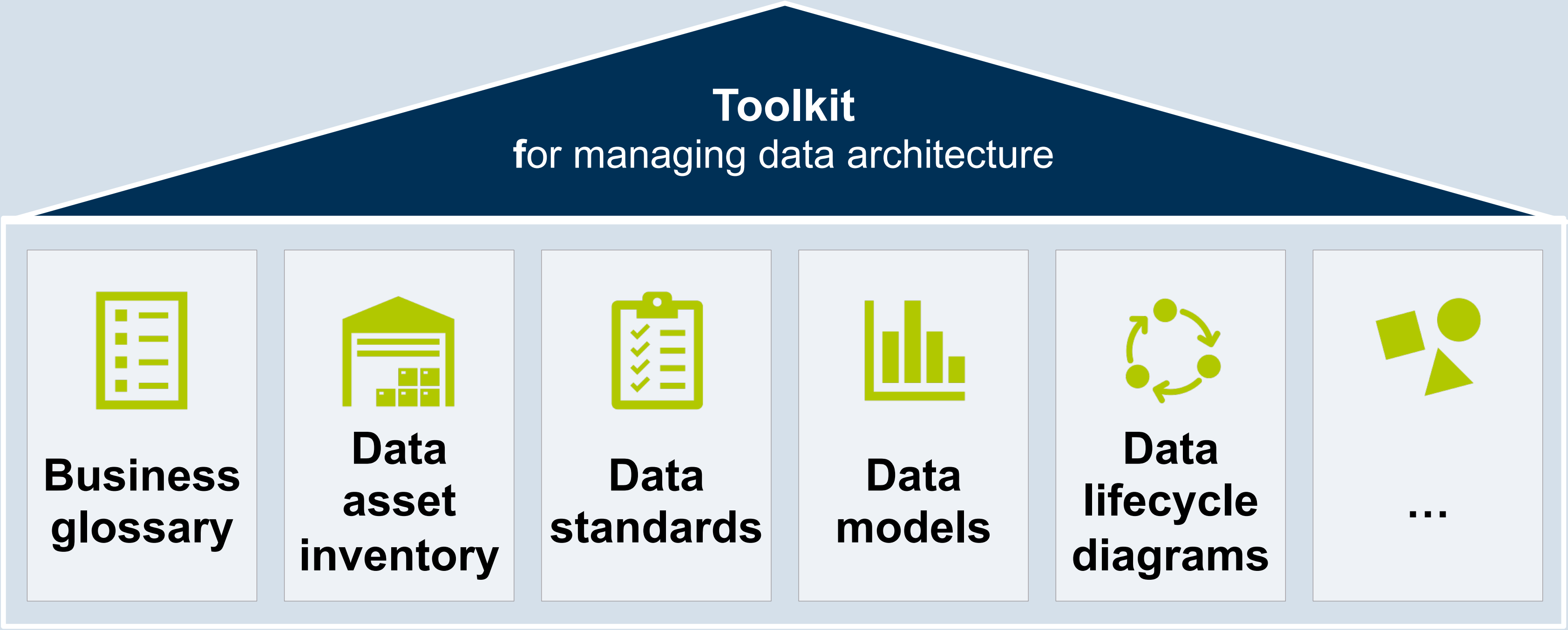
ASSOCIATION OF MBAS ACCREDITED

Managing Big Data

- Types of Data
- Data Architecture
- Master Data Management
- Data Quality
- Data Governance
- Data Security

What is Data Architecture?

Data architecture is about how data is stored and flows across the enterprise over its life cycle. The key components of managing data architecture are systems, data stores, and infrastructure.



Toolkit

for managing data architecture

- **Business glossary**
- **Data asset inventory**
- **Data standards**
- **Data models**
- **Data lifecycle diagrams**
- **…**

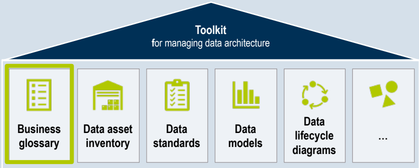
Business Glossary

A **business glossary** is a software application used to communicate and govern the organization's business concepts and terminology along with the associated definitions and relationships between those terms.

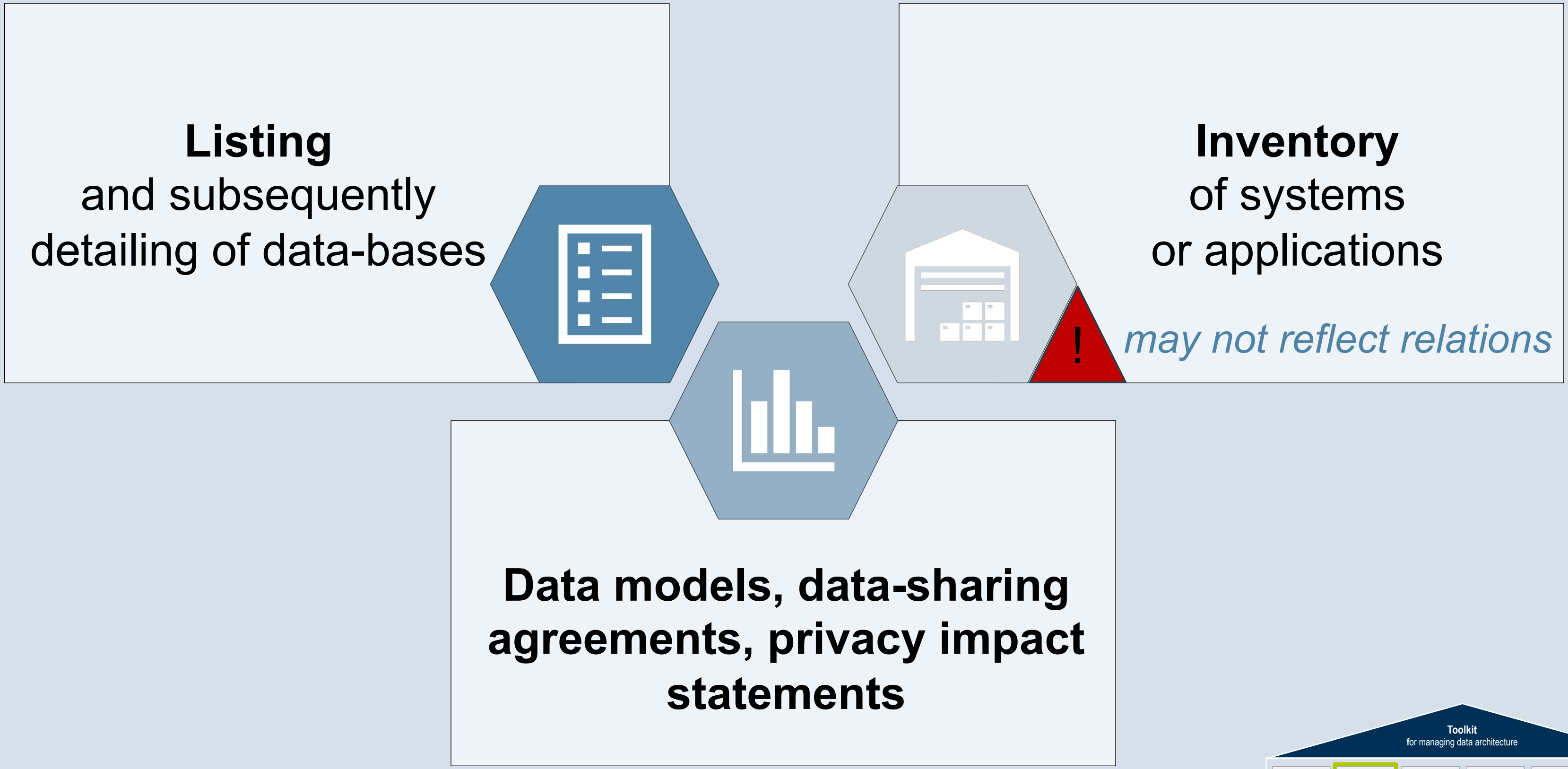
Accurate understanding of terms

The **effective use** of data

Unique definition for key business terms



Data Asset Inventory

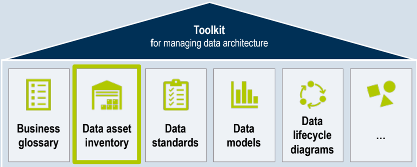


Listing
and subsequently
detailing of data-bases

Inventory
of systems
or applications

! *may not reflect relations*

Data models, data-sharing agreements, privacy impact statements



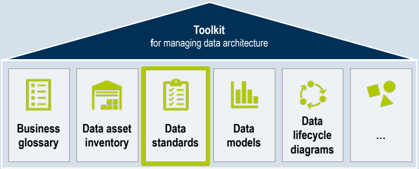
Data Standards

Detailed (or low-level) data standards might include standard terms and definitions, standard code sets, or data exchange standards.

Use **combination** of **industry** standards and **in-house** standards

Determine **data standards** to which you want to adhere

Rely on **business glossary** and **data asset inventory** while defining data standards

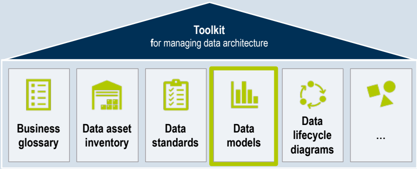


Data Models – Conceptual Data Models

A **conceptual data model** is meant to promote a common understanding of data in terms of high-level business entities and their relationships.

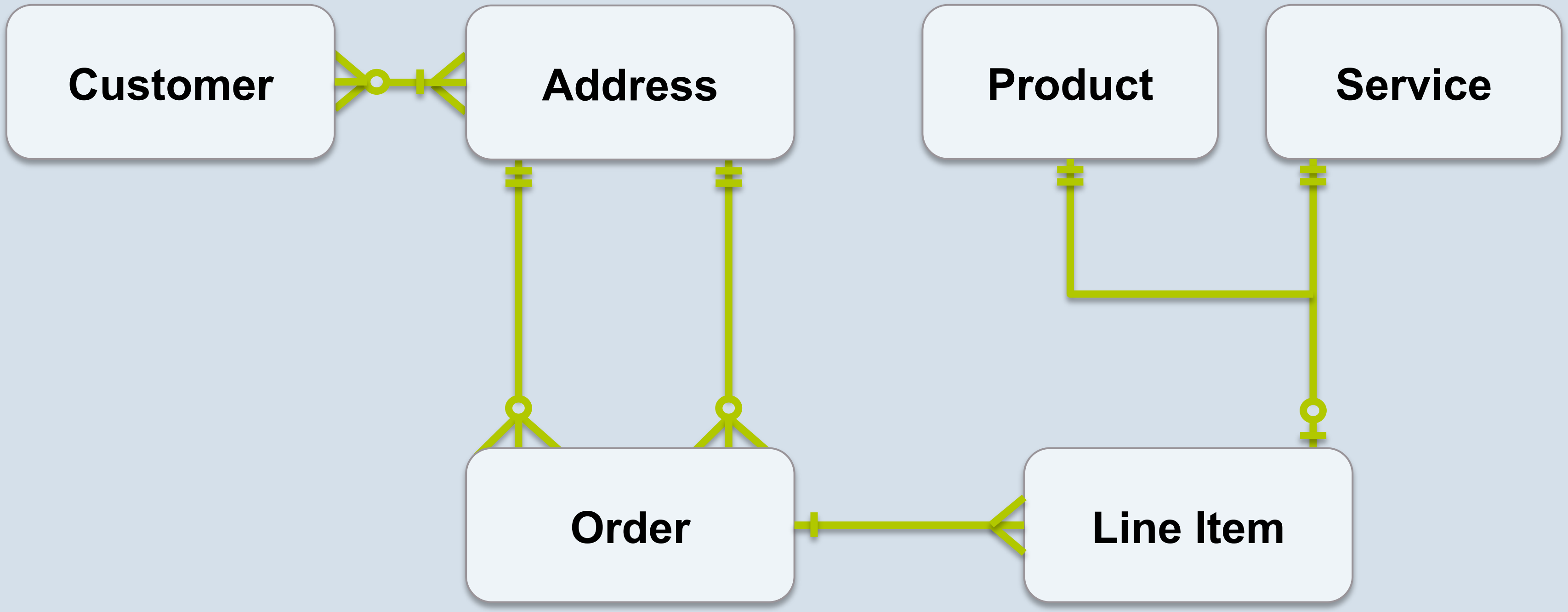
Describe and **document entities** and **attributes**

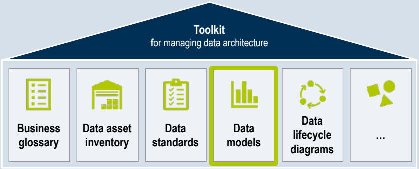
- **Widely accepted standard notations** and best practices for the creation of new models
- **Foundation for lower level logical and physical data models**
- Subject areas within conceptual data models to be created based on **business functionality**



Data Models – Conceptual Data Models

EXEMPLARY





Physical vs. Logical Data Models

Physical Data Models

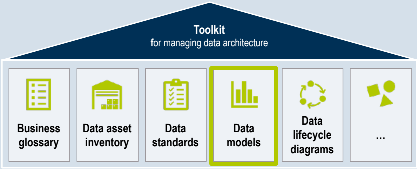
- Uses **tables, fields,** and **relations** to document how data is stored
- Relatively **common** and easily produced
- **Large database management systems** all generate physical data models based on their database

Logical Data Models

- Often used to **graphically document** low-level data requirements
- Must be separately developed
- Provide **additional context** (entity or attribute definitions etc.)
- Good way to **validate data requirements**

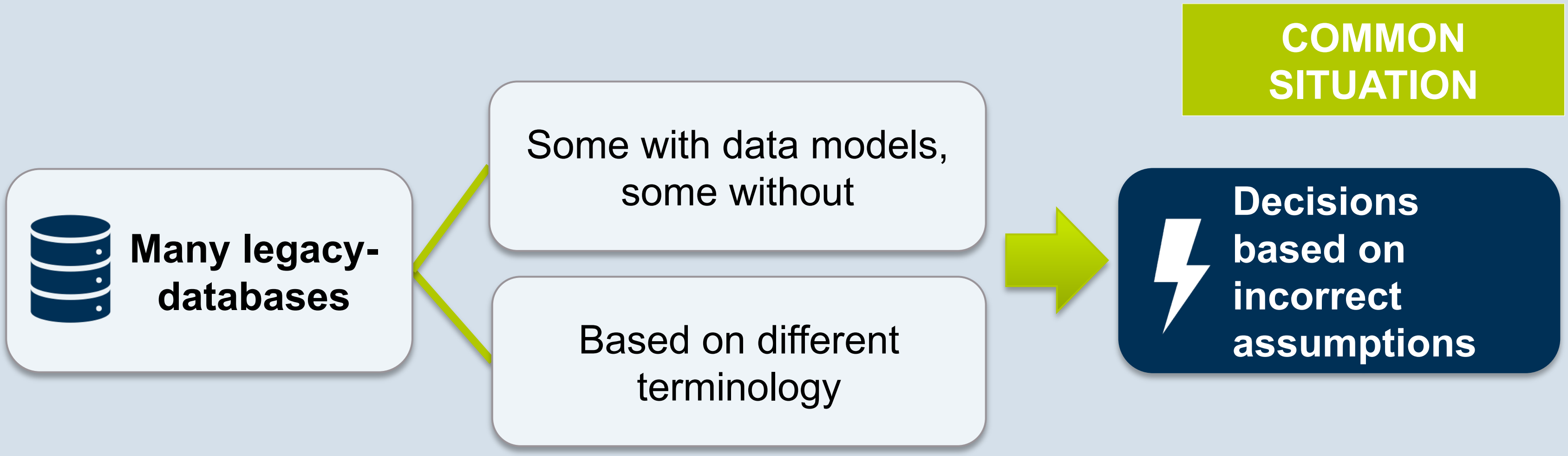
!

Key limitation:
Most physical and logical data models are **specific to a system**.



Enterprise-Level Data Models

An **enterprise-level data model** is a logical data model covering standard definition of entities and attributes across the organization.



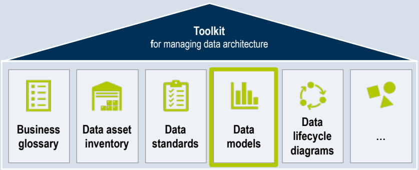
COMMON SITUATION

Many legacy-databases

Some with data models, some without

Based on different terminology

Decisions based on incorrect assumptions



Data Lifecycle Diagrams

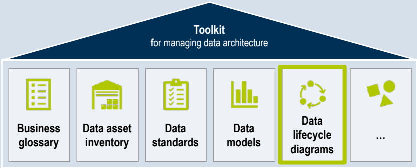
A **data lifecycle diagram (DLD)** shows how data is stored and flows across the entire enterprise.

Easily understandable
Graphical depiction of how data flows with the organization

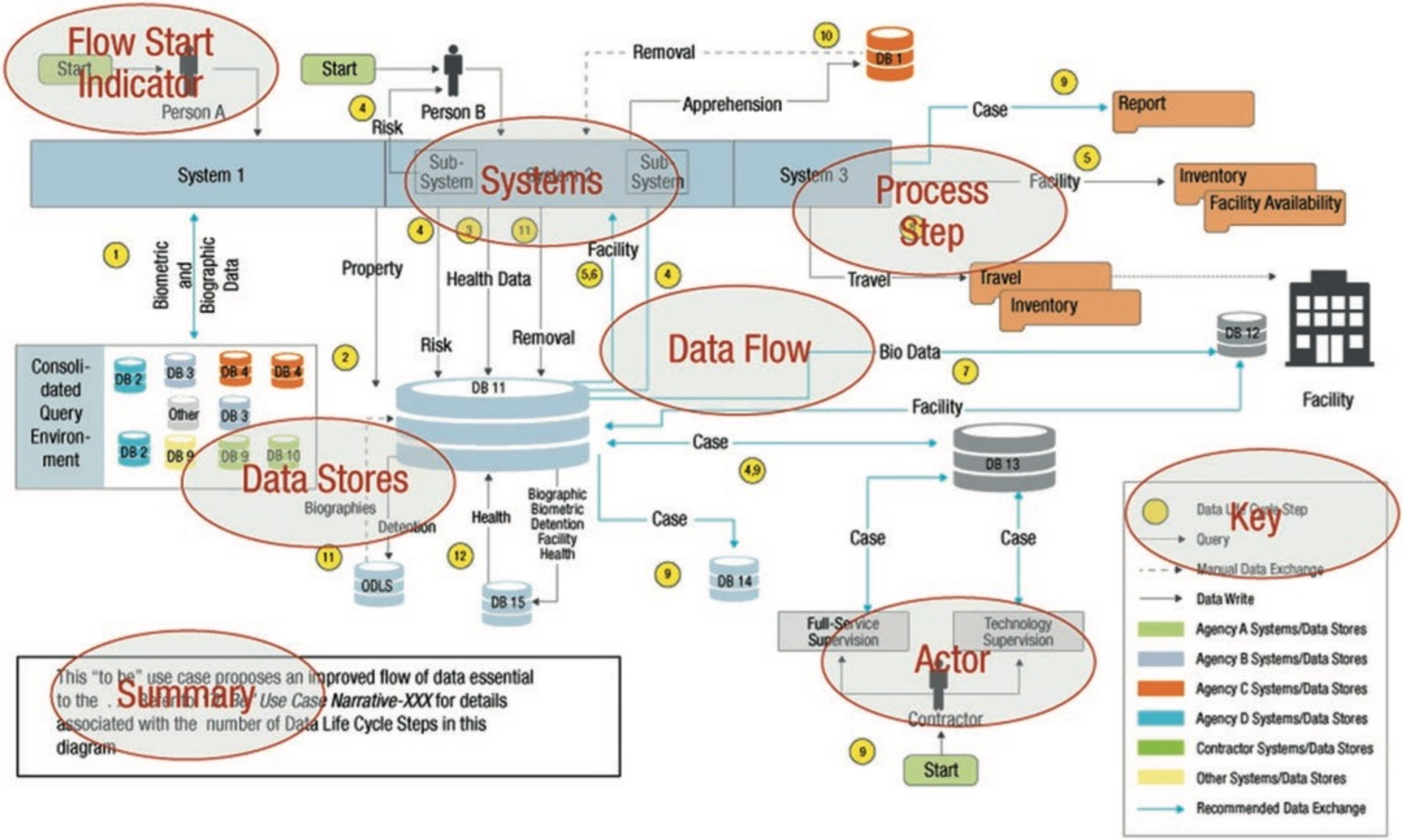
Exact **scope** of a DLD can **vary** in how it is defined

Most complete variation
Reflects the entire environment for a given type of data

Typically done for both a current as well as a desired environment



Data Lifecycle Diagrams



This "to be" use case proposes an improved flow of data essential to the . Refer to 'Use Case Narrative-XXX' for details associated with the number of Data Life Cycle Steps in this diagram

Data Lifecycle Diagrams

Flow Start Indicator

“Start” designators highlight data entry or the beginning of an independent process

Data Flow

These lines and labels show actual data flows (solid), queries (dashed), and human data exchanges (dotted)

Data Stores

Drum shapes designate where data is stored

System

Square boxes indicate applications with user interfaces

Process Step

These small, yellow numbers refer back to the diagram narrative to highlight a specific step in a process

Actor

People icons depict individuals involved in data input or exchange

Summary

A short summary of the use case; a detailed description can be found in the diagram narrative

Key

Highlights information such as different colors to indicate system and data store ownership

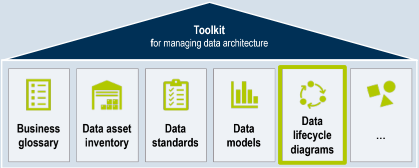
Data Lifecycle Diagrams: Applications

NORMALLY

- **Flow labels** highlight where specific data flows, is entered and re-entered, and is manually shared
- **Written narrative** accompanies each DLD and provides additional facts

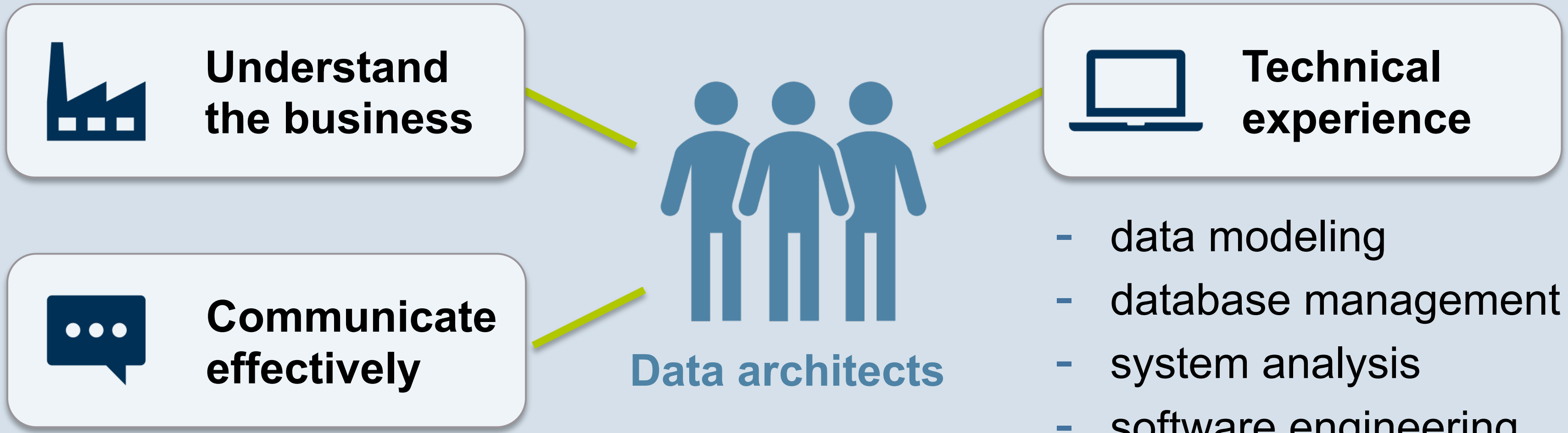
EXTENSION

- State transition overlay
- Documenting/validating a data asset inventory
- Educating new employees about data and systems
- Communicating with transitory contractors



Who are Data Architects?

Data architects guide how data is portioned into systems, stored, shared, and integrated for use, how data is standardized and organized within systems and datastores and for downstream needs.



Understand the business

Communicate effectively

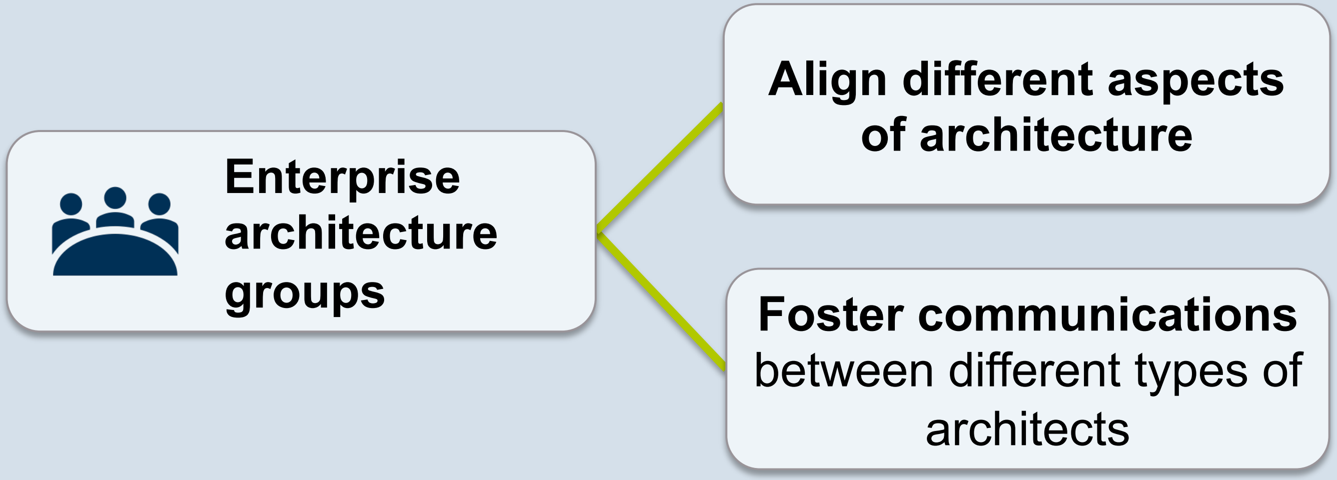
Data architects

Technical experience

- data modeling
- database management
- system analysis
- software engineering

Who is the Enterprise Architecture (EA) group?

Sometimes, all or part of data architecture is grouped with other types of architecture (business, software/application/system,…) in an **Enterprise Architecture (EA) group**.



Enterprise architecture groups

Align different aspects of architecture

Foster communications between different types of architects

Who is the Enterprise Architecture (EA) group?

Business users as important participants in **driving data architecture**

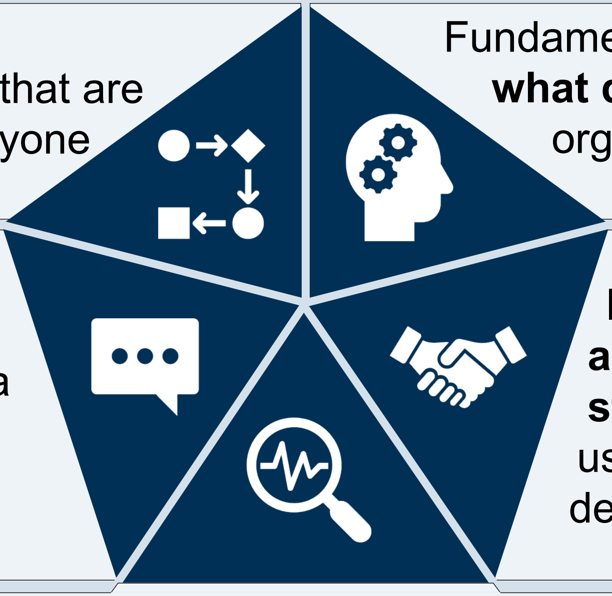
Certifications for EA

Multiple data architects within the **same organization** (given the resources required)

Need to **augment their skillsets** by working with other business and technical specialists

Working with IT to manage the software and hardware infrastructure required

Benefits of Data Architecture



Data lifecycle diagrams that are easily understood by everyone

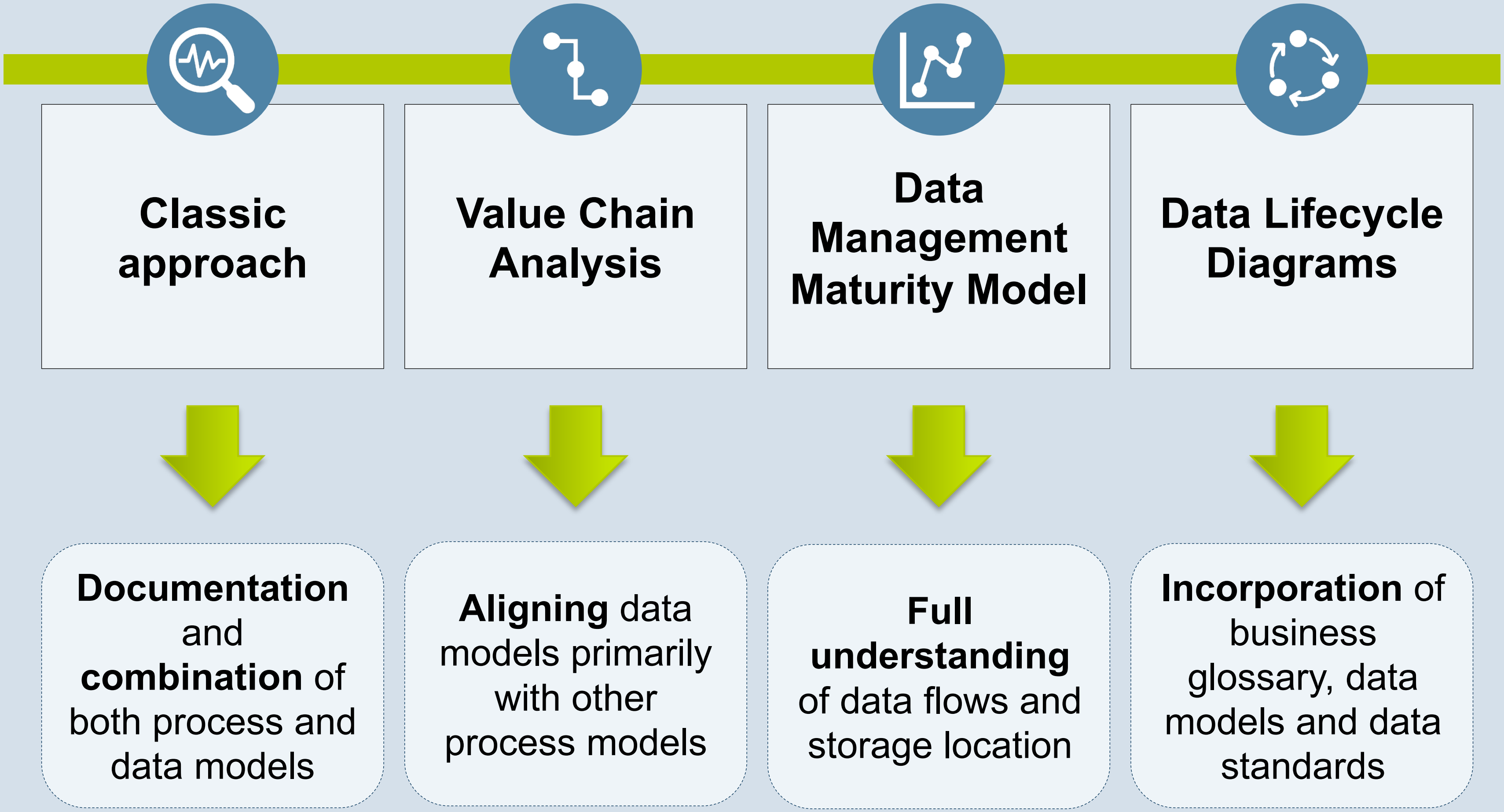
Fundamental **understanding** of **what data exists** in the organization and where it is stored

Blueprint for **communicating** how data is stored and shared for a given business process

Improved collaboration among data stakeholders using common terms and definitions for data

Understanding of (non)authoritative **data sources**, identifying redundant data stores, identifying unmet integration needs, etc.

Data Architecture Framework



Classic approach

Documentation and **combination** of both process and data models

Value Chain Analysis

Aligning data models primarily with other process models

Data Management Maturity Model

Full understanding of data flows and storage location

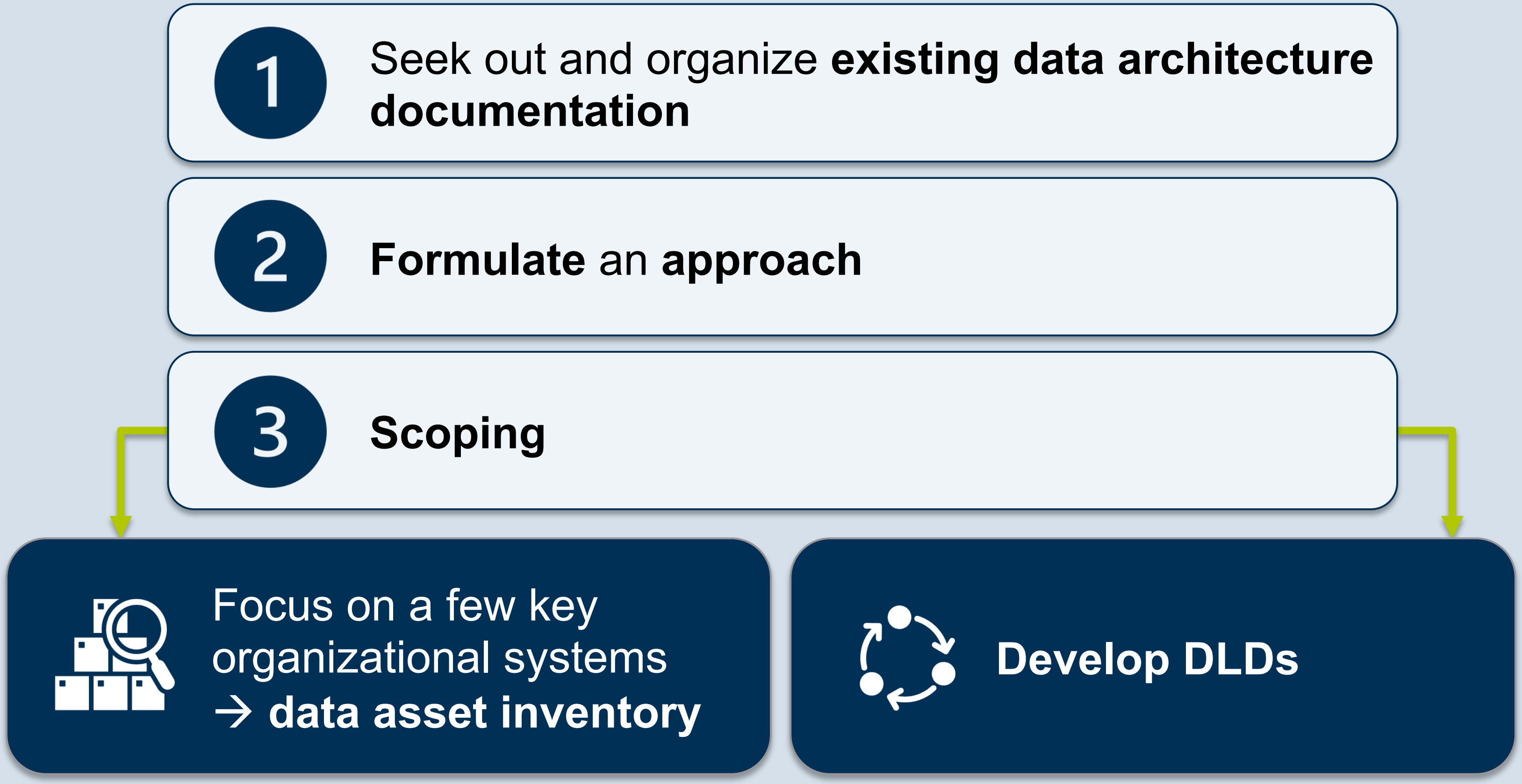
Data Lifecycle Diagrams

Incorporation of business glossary, data models and data standards

Implementing Data Architecture

- **System inventory** which could be used as a starting point to create a complementary data asset inventory
- **Data models** at the physical and the logical level
- **Agreements on subject areas** and/or **data categories**
- **System design** or **architecture documents**
- **Business glossaries** or **conceptual data models** at a more local level

Implementing Data Architecture



1 Seek out and organize **existing data architecture documentation**

2 **Formulate** an **approach**

3 **Scoping**

Focus on a few key organizational systems → **data asset inventory**

Develop DLDs

Implementing Data Architecture

! Data architecture is **not** a project-level activity. To be useful to the business, data architecture needs to be continuously maintained.

What is an inventory of our systems, applications, and data stores, and how do they interact?

What is my authoritative source for information on a particular facility?

Where is the right place to update information on a customer?

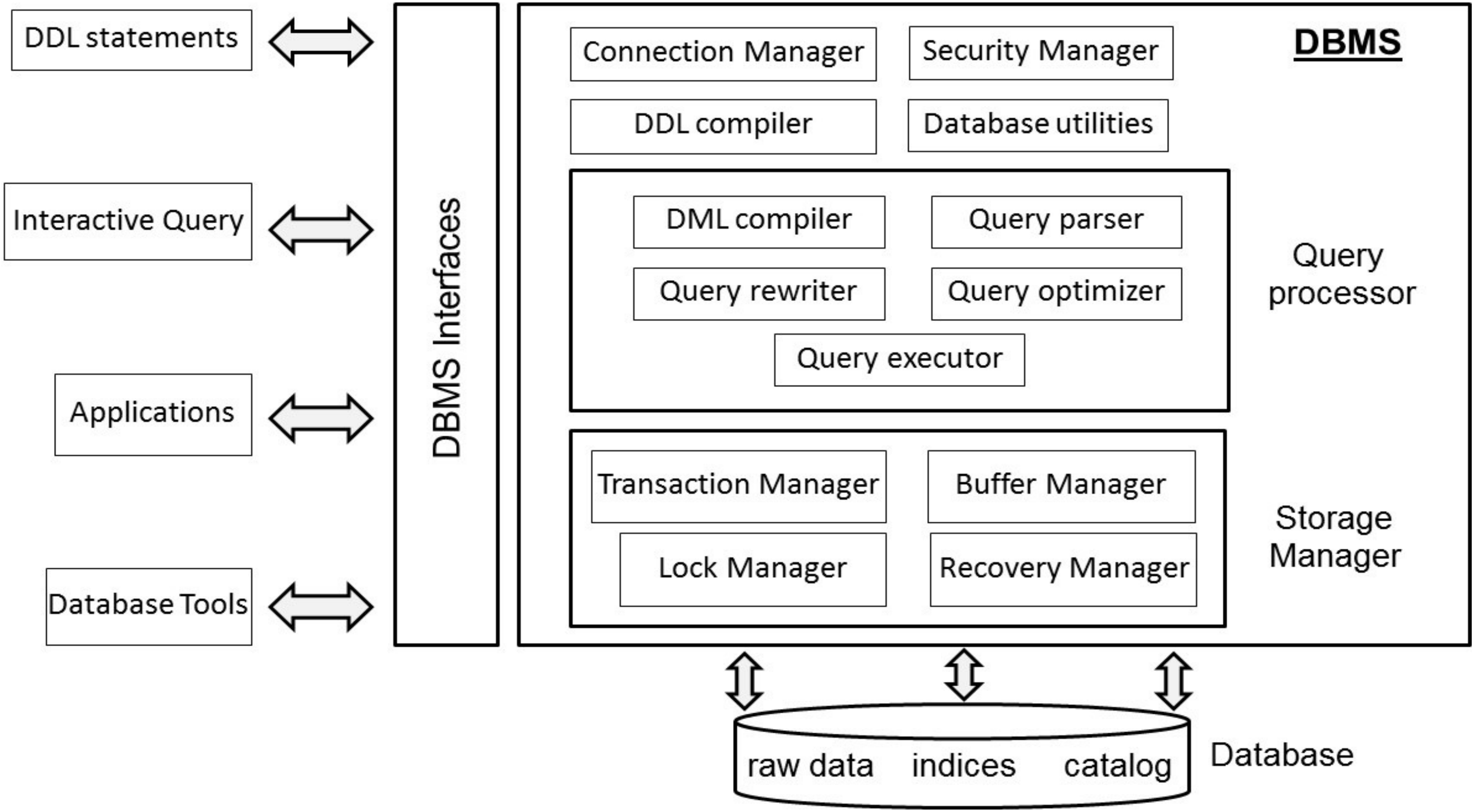
What constitutes an event or incident?

What data does our organization share with or ingest from partners?

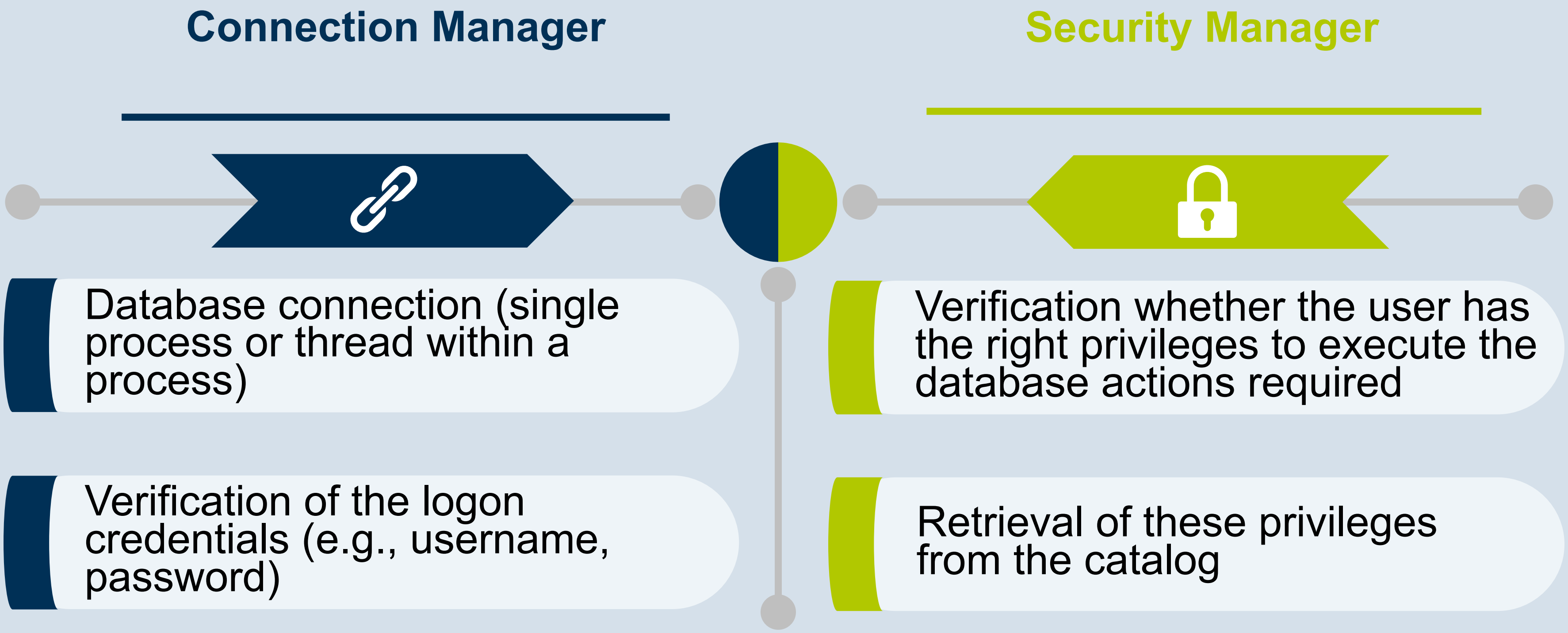
How should new data with respect to new regulations or guidelines, be integrated and stored?

Which data on several, conflicting enterprise reports is correct?

Architecture of a DBMS



Connection and Security Manager



Connection Manager

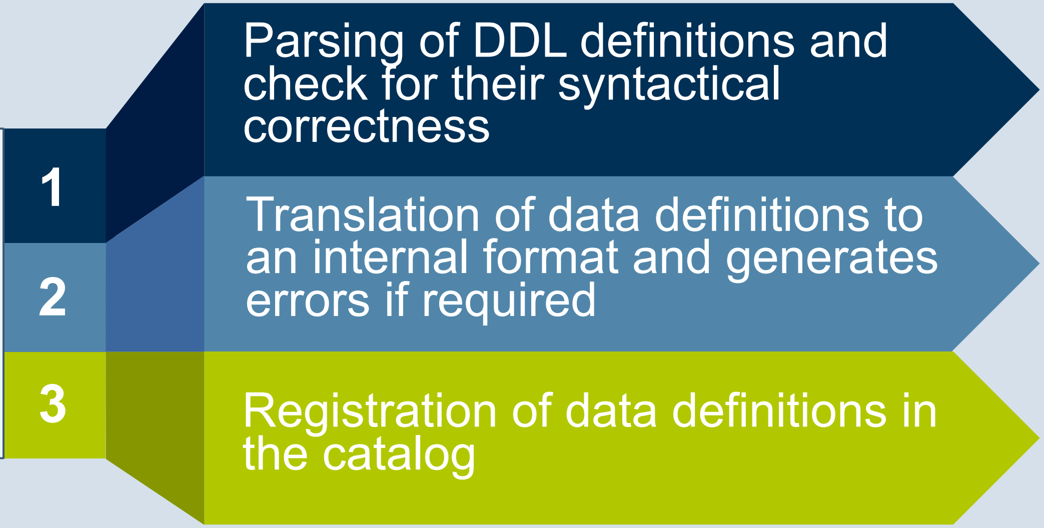
- Database connection (single process or thread within a process)
- Verification of the logon credentials (e.g., username, password)

Security Manager

- Verification whether the user has the right privileges to execute the database actions required
- Retrieval of these privileges from the catalog

DDL Compiler

- Compiles the data definitions specified in DDL
- Ideally three DDLs (internal / logical / external data model)



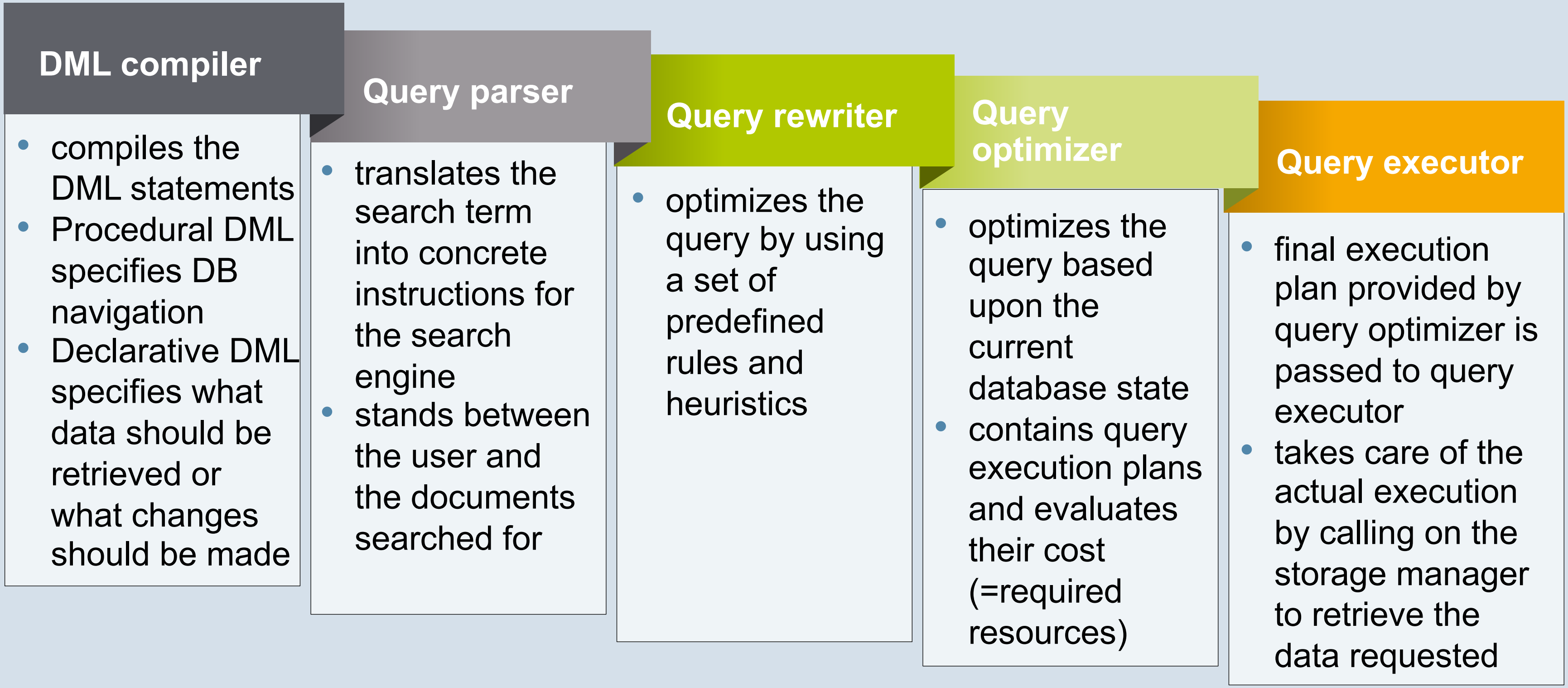
1 Parsing of DDL definitions and check for their syntactical correctness

2 Translation of data definitions to an internal format and generates errors if required

3 Registration of data definitions in the catalog

Query Processor

A **query processor** helps execute database queries such as retrieving, inserting, updating or removing data.



DML compiler

- compiles the DML statements
- Procedural DML specifies DB navigation
- Declarative DML specifies what data should be retrieved or what changes should be made

Query parser

- translates the search term into concrete instructions for the search engine
- stands between the user and the documents searched for

Query rewriter

- optimizes the query by using a set of predefined rules and heuristics

Query optimizer

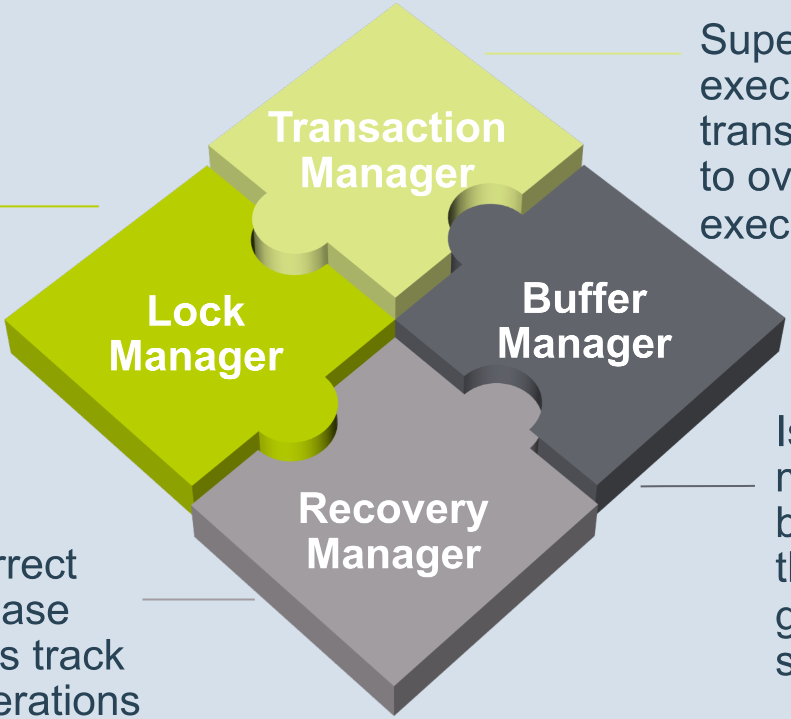
- optimizes the query based upon the current database state
- contains query execution plans and evaluates their cost (=required resources)

Query executor

- final execution plan provided by query optimizer is passed to query executor
- takes care of the actual execution by calling on the storage manager to retrieve the data requested

Storage Manager

The **Storage Manager** governs physical file access and supervises the correct and efficient storage of data.



Transaction Manager: Supervises the execution of database transactions, contributes to overall efficiency and execution performance

Lock Manager: Provides concurrency control to ensure data integrity (e.g., read lock vs. write lock)

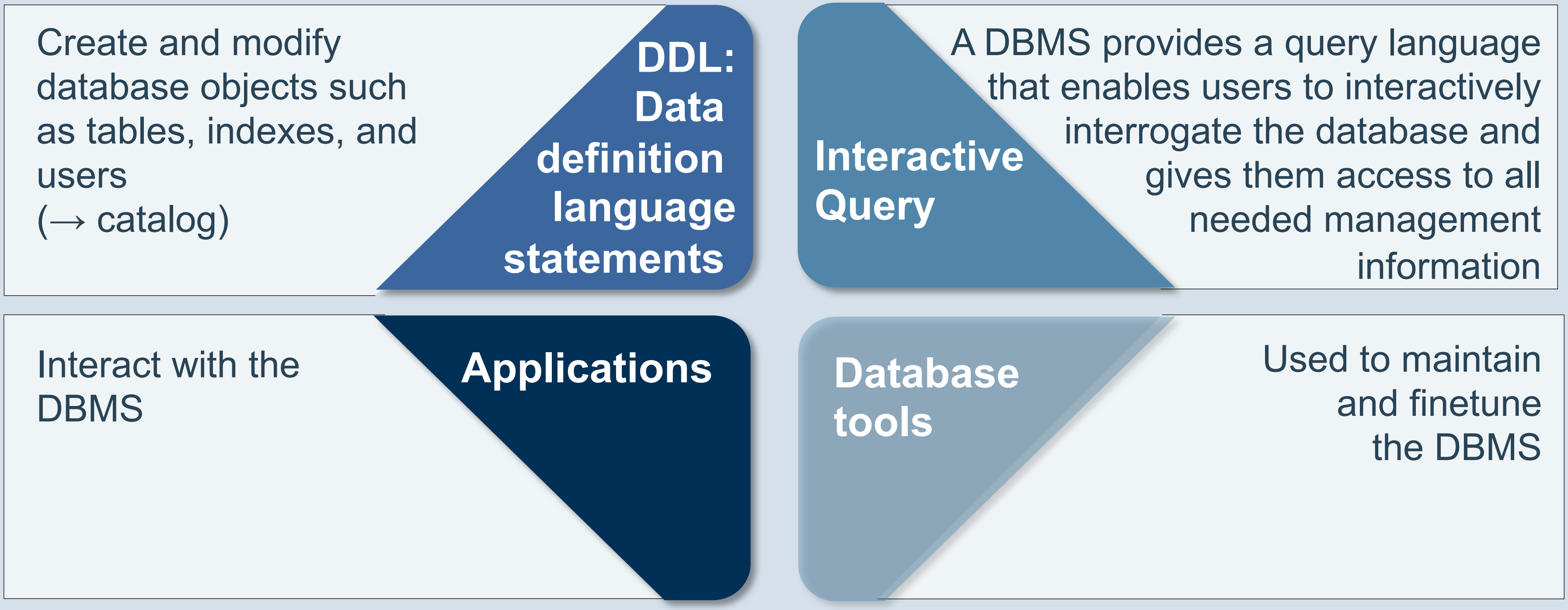
Buffer Manager: Is responsible for managing the buffer memory of the DBMS, guarantees a speedy access.

Recovery Manager: Supervises the correct execution of database transactions, keeps track of all database operations

DBM Utilities

- Loading utility
- Reorganization utility
- Performance-monitoring utilities
- User management utilities
- Backup and recovery utility

Interacting with a DBMS



DDL: Data definition language statements

Create and modify database objects such as tables, indexes, and users (→ catalog)

Interactive Query

A DBMS provides a query language that enables users to interactively interrogate the database and gives them access to all needed management information

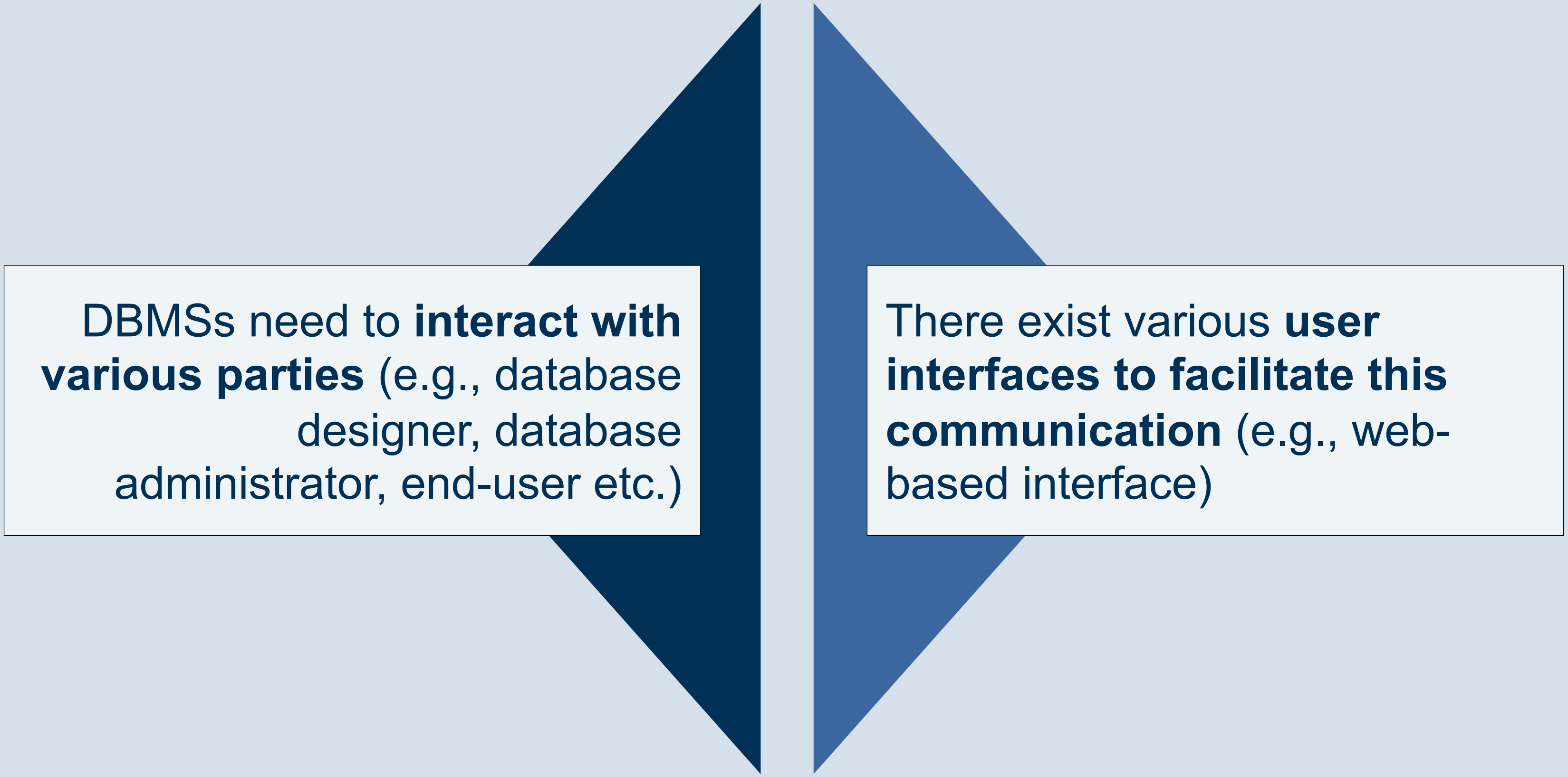
Applications

Interact with the DBMS

Database tools

Used to maintain and finetune the DBMS

DBMS Interfaces



DBMSs need to **interact with various parties** (e.g., database designer, database administrator, end-user etc.)

There exist various **user interfaces to facilitate this communication** (e.g., web-based interface)

DBMS Interfaces
Examples

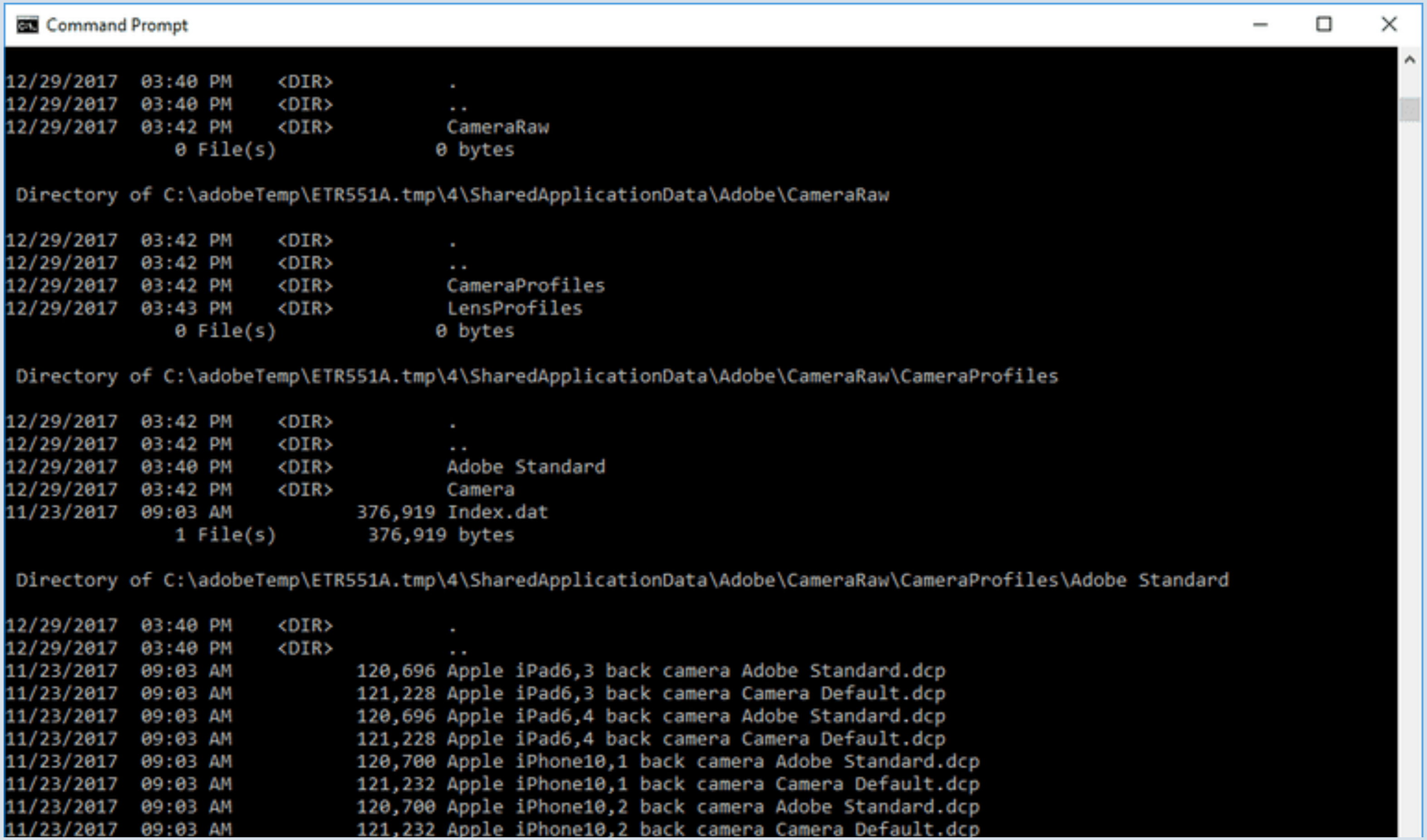
Web-based interface

Network interface

DBMS Interfaces
Examples

Command-line interface

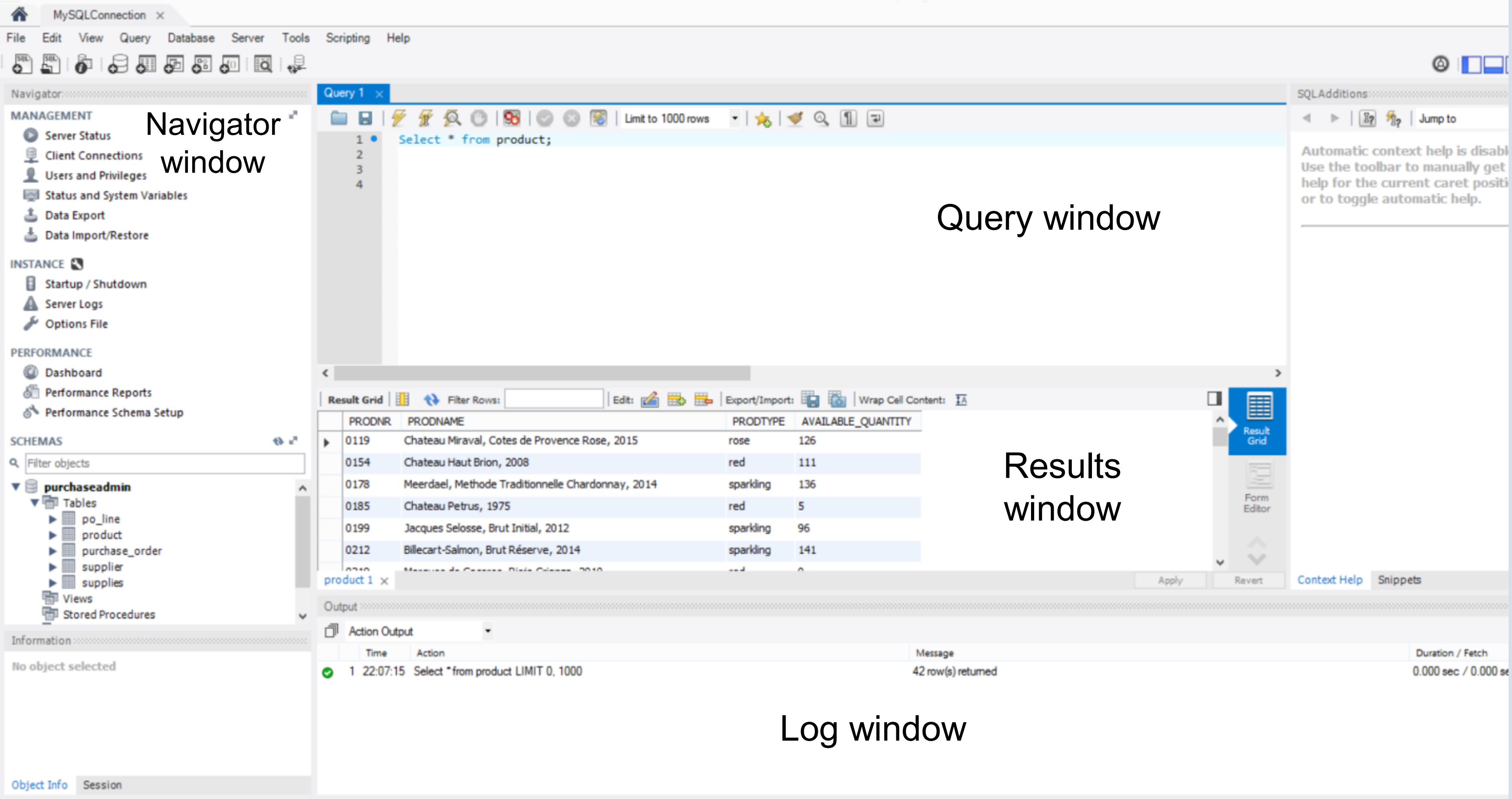
Admin interface



```
12/29/2017  03:40 PM    <DIR>          .
12/29/2017  03:40 PM    <DIR>          ..
12/29/2017  03:42 PM    <DIR>          CameraRaw
               0 File(s)              0 bytes

 Directory of C:\adobeTemp\ETR551A.tmp\4\SharedApplicationData\Adobe\CameraRaw

12/29/2017  03:42 PM    <DIR>          .
12/29/2017  03:42 PM    <DIR>          ..
12/29/2017  03:42 PM    <DIR>          CameraProfiles
12/29/2017  03:43 PM    <DIR>          LensProfiles
               0 File(s)              0 bytes

 Directory of C:\adobeTemp\ETR551A.tmp\4\SharedApplicationData\Adobe\CameraRaw\CameraProfiles

12/29/2017  03:42 PM    <DIR>          .
12/29/2017  03:42 PM    <DIR>          ..
12/29/2017  03:40 PM    <DIR>          Adobe Standard
12/29/2017  03:42 PM    <DIR>          Camera
11/23/2017  09:03 AM           376,919 Index.dat
               1 File(s)        376,919 bytes

 Directory of C:\adobeTemp\ETR551A.tmp\4\SharedApplicationData\Adobe\CameraRaw\CameraProfiles\Adobe Standard

12/29/2017  03:40 PM    <DIR>          .
12/29/2017  03:40 PM    <DIR>          ..
11/23/2017  09:03 AM           120,696 Apple iPad6,3 back camera Adobe Standard.dcp
11/23/2017  09:03 AM           121,228 Apple iPad6,3 back camera Camera Default.dcp
11/23/2017  09:03 AM           120,696 Apple iPad6,4 back camera Adobe Standard.dcp
11/23/2017  09:03 AM           121,228 Apple iPad6,4 back camera Camera Default.dcp
11/23/2017  09:03 AM           120,700 Apple iPhone10,1 back camera Adobe Standard.dcp
11/23/2017  09:03 AM           121,232 Apple iPhone10,1 back camera Camera Default.dcp
11/23/2017  09:03 AM           120,700 Apple iPhone10,2 back camera Adobe Standard.dcp
11/23/2017  09:03 AM           121,232 Apple iPhone10,2 back camera Camera Default.dcp
```

DBMS Interfaces
Examples

Graphical user interface | **Natural language interface**



Navigator window

Query window

```
Select * from product;
```

Automatic context help is disabl
Use the toolbar to manually get
help for the current caret positi
or to toggle automatic help.

Results window

PRODNR	PRODNAME	PRODTYPE	AVAILABLE_QUANTITY
0119	Chateau Miraval, Cotes de Provence Rose, 2015	rose	126
0154	Chateau Haut Brion, 2008	red	111
0178	Meerdael, Methode Traditionnelle Chardonnay, 2014	sparkling	136
0185	Chateau Petrus, 1975	red	5
0199	Jacques Selosse, Brut Initial, 2012	sparkling	96
0212	Billecart-Salmon, Brut Réserve, 2014	sparkling	141

Log window

```
   Time      Action                              Message              Duration / Fetch
1  22:07:15  Select * from product LIMIT 0, 1000 42 row(s) returned   0.000 sec / 0.000 se
```

Key Take Aways

Data Architecture

- Toolkit for managing data architecture
- Data architects and their organizational position
- Implement data architecture
- Elements of DBMS architecture