SARACUS CONSULTING

Einführung in Metadatenkataloge

Lukas Hestermann und Matthias Lohmann

Agenda

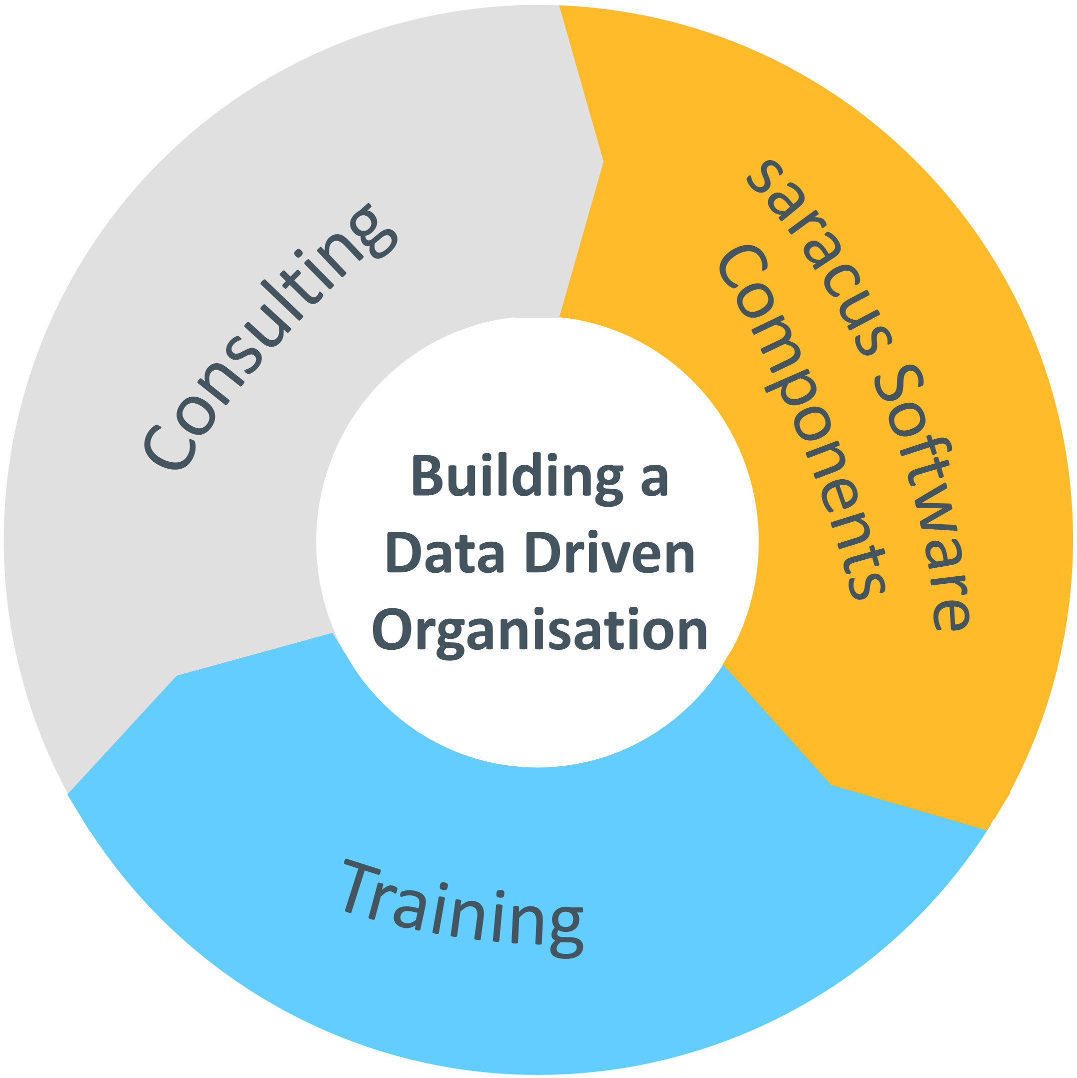
1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
5. Metadata Strategy und Data Catalog-Einführung

Vorstellungsrunde

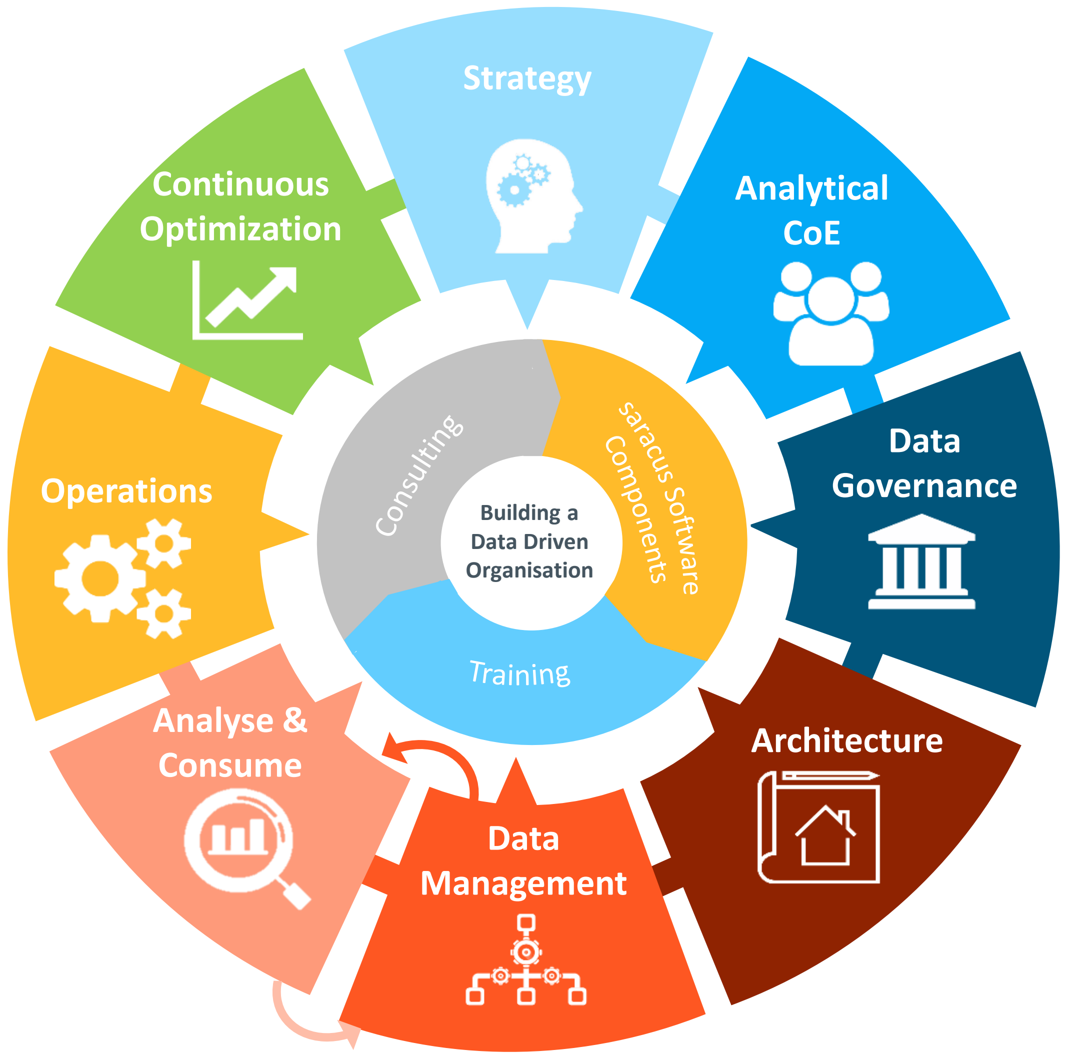
Kurze Vorstellungsrunde:

- Beschreibung Ihres Aufgabengebietes/Tätigkeiten
- Kenntnisse im Metadaten- und Datenkatalog-Umfeld
- Erwartungen an das Seminar

saracus Focus and Mission



saracus Framework: Building a Data Driven Organisation

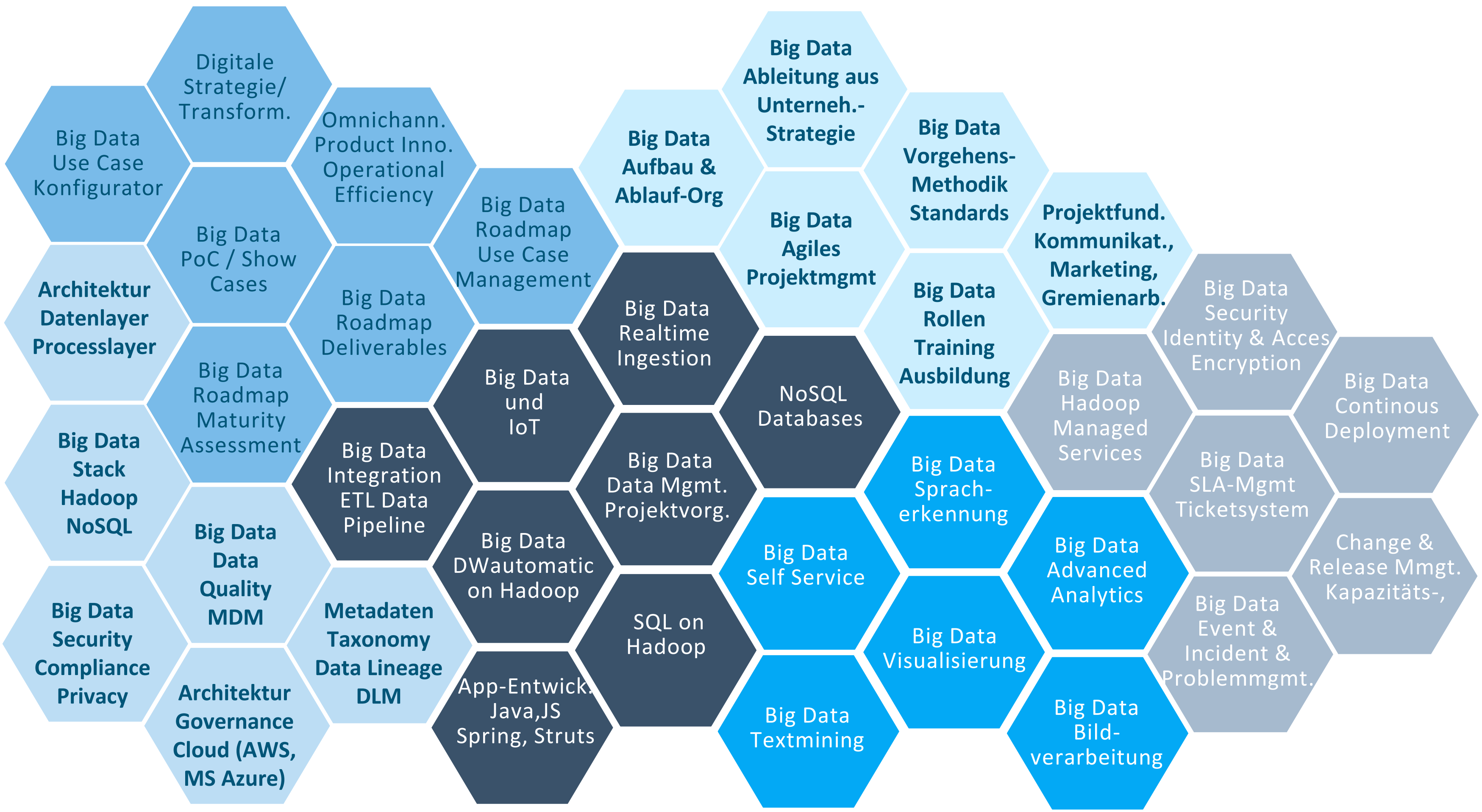


SARACUS CONSULTING

saracus: Consulting Services:
BI, DWH, Big Data Engineering, Advanced Analytics

SARACUS CONSULTING

saracus consulting Produkte
Big Data Analytics, Daten und Technologie



Strategy | Big Data Architektur / Governance | Big Data Management Center of Ex. (CoE) | Data Management | Analysis | Operations

SARACUS CONSULTING

Seminar-Angebot Big Data academy

Strategie und Methodik

- Entwicklung einer DWH-Strategie
- Agiles Projekt-management
- Analytical Architectur
- Planung, Einführung u. Betreiben eines BICC
- Review und Redesign Data Warehouse

Integration und Data Warehousing

- Kompaktwissen Data Warehouse
- Daten-Design in einer komplexen Daten-architektur
- Design u. Implementierung der ETL-Prozesse
- Versionierung als Kernproblem der Bewirtschaftung
- Datenqualität im DWH
- Vorgehen ETL-Tool Auswahl
- DWH Governance: Daten als Ressource

Big Data Engineering

- Einführung in Big Data und Hadoop
- Hadoop Administrator Training
- Fast SQL auf Hadoop
- Hadoop Developer-Training
- Real Time Stream Processing
- Data Engineer Training

Business Intelligence

- Vorgehen und Verfahren der Informationsbedarfs-Analyse
- Grundlagen der dimensionalen Modellierung im DWH
- Spezialfragen der dimensionalen Modellierung
- Design Techniken für Dashboards u. Scorecards

Data Science und Advanced Analytics

- Data Scientist Grundlagen
- Predictive Analytics (Statistische Methoden, Neuronale Netze, Decision Trees)
- Integration von Social Analytics, Web Analytics und Business Intelligence
- Deep Learning Neuronale Netze

Level:

- 100 (Einführungsseminar)
- 200 (Fortgeschrittene)
- 300 (Experten)

Und los geht’s!

Agenda

1. Grundlagen Metadaten
   – Was sind Metadaten
   – Warum Metadaten
   – Quellen von Metadaten
   – Metadaten Standards
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
5. Metadata Strategy und Data Catalog-Einführung

COURSE CURRICULUM

Purpose:
To define metadata and show its importance to the organization.

Outcome:

- Understand the definition of metadata
- Distinguish the various types of metadata
- Recognize the importance & relevance of metadata

Agenda

1. Grundlagen Metadaten
   - Was sind Metadaten
   - Warum Metadaten
   - Quellen von Metadaten
   - Metadaten Standards
2. Metadaten Management und Data Catalogs
3. Informatica EDC Präsentation
4. Funktionalitäten von Data Catalogs
5. Data Catalogs: Ausgewählte Themen
6. Metadata Strategy und Data Catalog-Einführung

What is Metadata

THE DEFINITION OF METADATA

What is Metadata

Metadata is
Data In Context

Metadata is the „Who, What, Where, Why, When & How" of Data

Who	What	Where	Why	When	How
Who created this data?	What is the business definition of this data element?	Where is this data stored?	Why are we storing this data?	When was this data created?	How is this data formatted? (character, numeric, etc.)
Who is the Steward of this data?	What are the business rules for this data?	Where did this data come from?	What is its usage & purpose?	When was this data last updated?	How many databases or data sources store this data?
Who is using this data?	What is the security level or privacy level of this data?	Where is this data used & shared?	What are the business drivers for using this data?	How long should it be stored?	
Who "owns" this data?	What is the abbreviation or acronym for this data element?	Where is the backup for this data?		When does it need to be purged/deleted?	
Who is regulating or auditing this data?	What are the technical naming standards for database implementation?	Are there regional privacy or security policies that regulate this data?			

Metadata Places Information in Context

Context drives relevance & importance

–> drives action. “You’re at 140.”

Metadata
Street address or PO Box?
Are you mailing something or visiting me?
Will I get my mail delivered?

Address

IQ

Metadata
What test was used?
When was this test taken?
Will this get me into Mensa?

140
Data

Metadata
Is this in beats per minute?
Am I exercising or sitting?
Am I healthy?

Heart Rate

Weight

Metadata
Is this pounds or kilograms?
How does this relate to my BMI?
Do I need to go on a diet?

We use Metadata all of the time – creating context

- We use metadata all of the time to put information in context.
- The human brain naturally creates associations to understand new information.
  - Core characteristics & properties
  - Relationships to other information
  - Timing and historical context
- We are all bombarded with a massive amount of data entering our brains each day – metadata helps us *put data into context and determine what is important & why*.
- Organizations face the same issue with their data. With massive volumes of data, metadata helps categorize information, put it in context, and determine what is important & why.

Data vs. Metadata

Customer

First Name	Last Name	Company	City	Year Purchased
Joe	Smith	Komputers R Us	New York	1970
Mary	Jones	The Lord's Store	London	1999
Proful	Bishwal	The Lady's Store	Mumbai	1998
Ming	Lee	My Favorite Store	Beijing	2001

Metadata

Data

Data vs. Metadata

Customer

STR01	STR02	TXT123	TXT127	DT01
Joe	Smith	Komputers R Us	New York	1970
Mary	Jones	The Lord's Store	London	1999
Proful	Bishwal	The Lady's Store	Mumbai	1998
Ming	Lee	My Favorite Store	Beijing	2001

Metadata

Data

Metadata Adds Context & Definition

Customer

First Name	Last Name	Company	City	Year Purchased
Joe	Smith	Komputers R Us	New York	1970
Mary	Jones	The Lord's Store	London	1999
Proful	Bishwal	The Lady's Store	Mumbai	1998
Ming	Lee	My Favorite Store	Beijing	2001

Is this the city where the customer lives or where the store is located?

Definition	Last Name represents the surname or family name of an individual.
Business Rules	In the Chinese market, family name is listed first in salutations.
Format	VARCHAR(30)
Abbreviation	LNAME
Required	YES
Etc.	Numerous technical & business metadata including security, privacy, nullability, primary key, etc.

Technical & Business Metadata

- Technical Metadata describes the structure, format, and rules for storing data
- Business Metadata describes the business definitions, rules, and context for data.
- Data represents actual instances (e.g. John Smith)

Technical Metadata

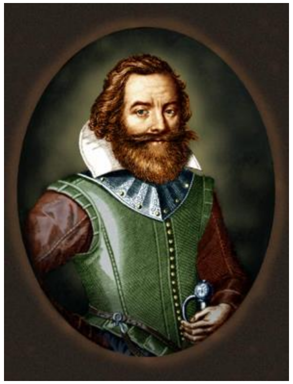
```
CREATE TABLE EMPLOYEE (
    employee_id       INTEGER NOT NULL,
    department_id     INTEGER NOT NULL,
    employee_fname    VARCHAR(50) NULL,
    employee_lname    VARCHAR(50) NULL,
    employee_ssn      CHAR(9) NULL);

CREATE TABLE CUSTOMER (
    customer_id       INTEGER NOT NULL,
    customer_name     VARCHAR(50) NULL,
    customer_address VARCHAR(150) NULL,
    customer_city     VARCHAR(50) NULL,
    customer_state    CHAR(2) NULL,
    customer_zip      CHAR(9) NULL);
```

Business Metadata

Term	Definition
Employee	An employee is an individual who currently works for the organization or who has been recently employed within the past 6 months.
Customer	A customer is a person or organization who has purchased from the organization within the past 2 years and has an active loyalty card or maintenance contract.

Data



John Smith

Business vs Technical Metadata

„A customer is a person or organization who purchases a product or service from…"

„CUST_LNM is VARCHAR (30) on the Oracle database CustDB1."

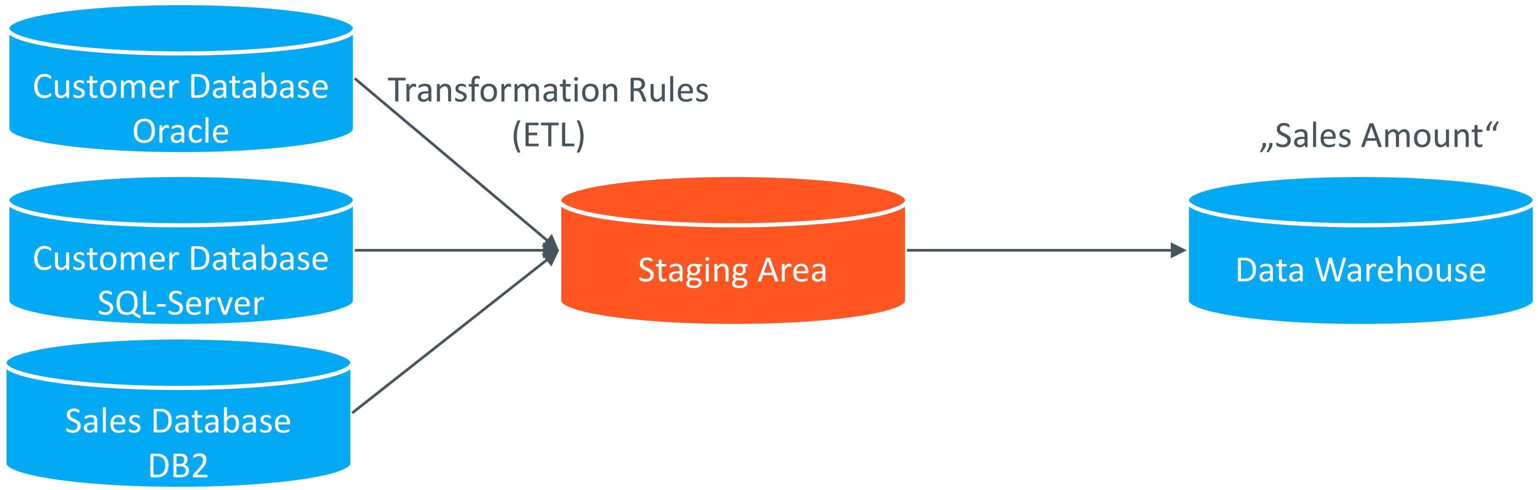
Business Metadata	Technical Metadata
• Definitions & Glossary • Data Steward • Organization • Privacy Level • Security Level • Acronyms & Abbreviations • Business Rules • Etc.	• Column structure of a database table • Data Type & Length (e.g. VARCHAR(20)) • Domains • Standard abbreviations (e.g. CUSTOMER -> CUST) • Nullability • Keys (primary, foreign, alternate, etc.) • Validation Rules • Data Movement Rules • Permissions • Etc.

Relationship Metadata

SHOWING HOW INFORMATION INTERRELATES

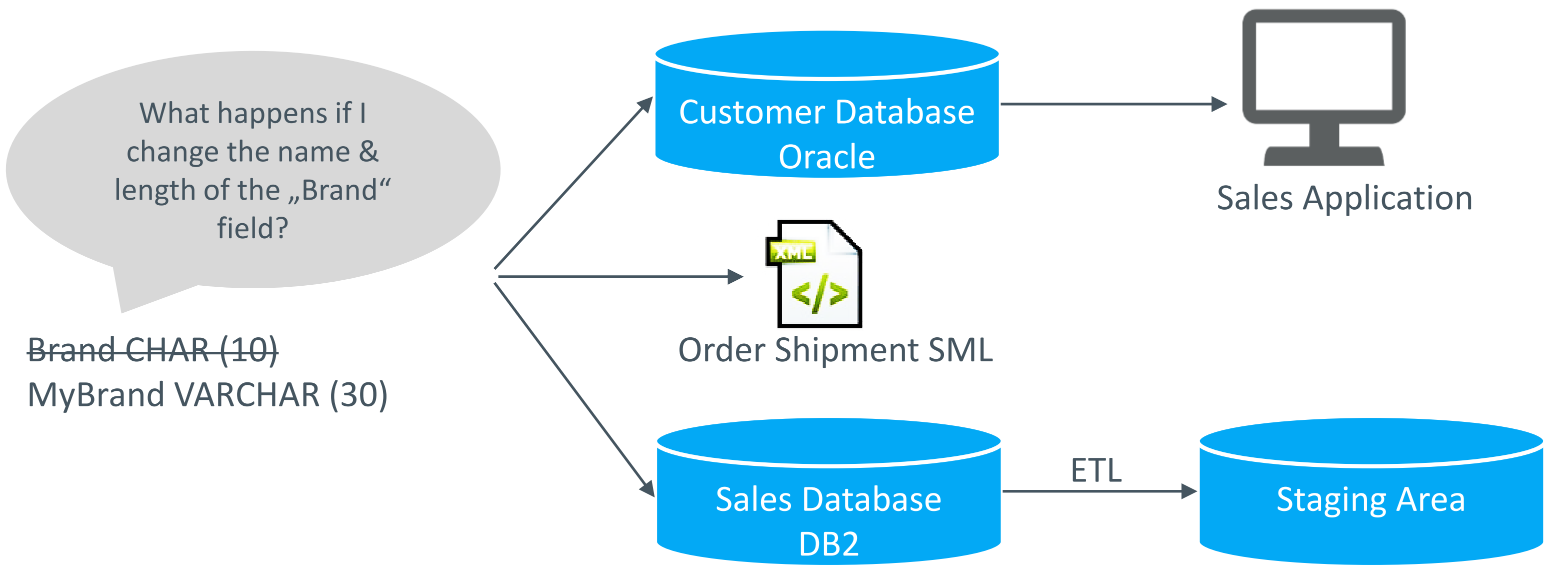
Data Lineage

- **Data Lineage** shows the source to target mapping, or provenance for information.
- For example, to understand how "Sales Amount" in a data warehouse is calculated, it is necessary to understand where the data came from and how it was manipulated along the way.



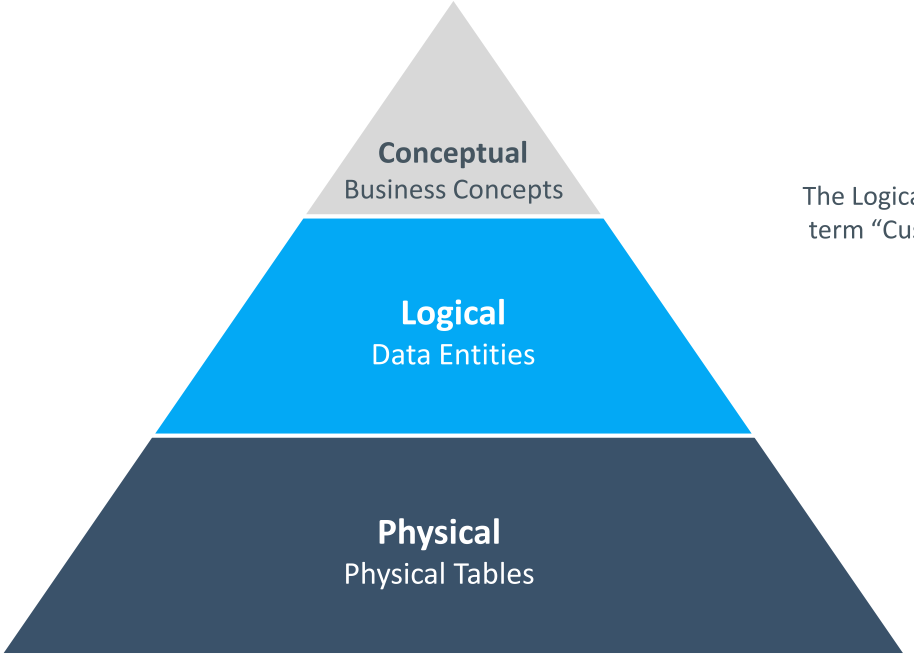
Impact Analysis & Where Used

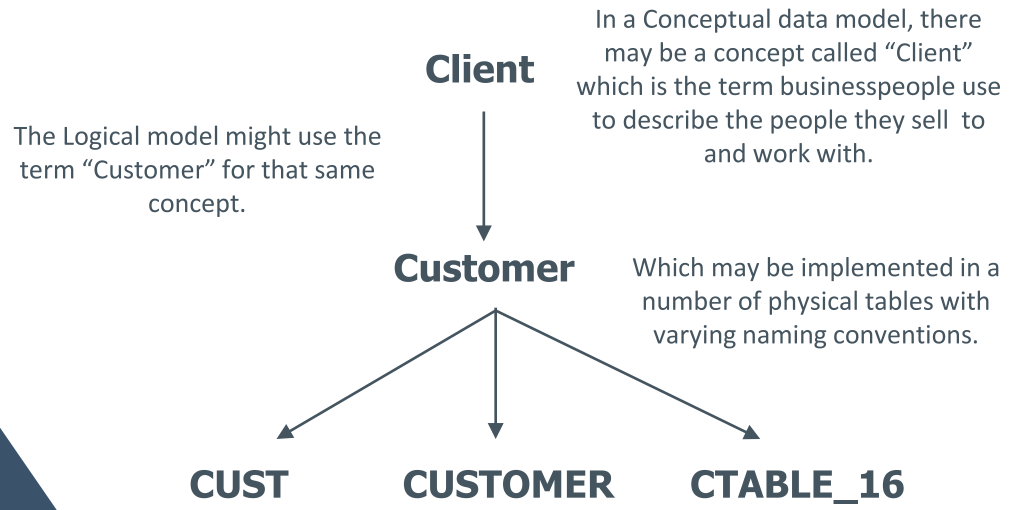
- Impact Analysis shows the relationship between a piece of metadata and other sources that rely on that metadata to assess the impact of a potential change.
- For example, if I change the length & name of a field, what other systems that are referencing that field will be affected?



Design Layer Relationships

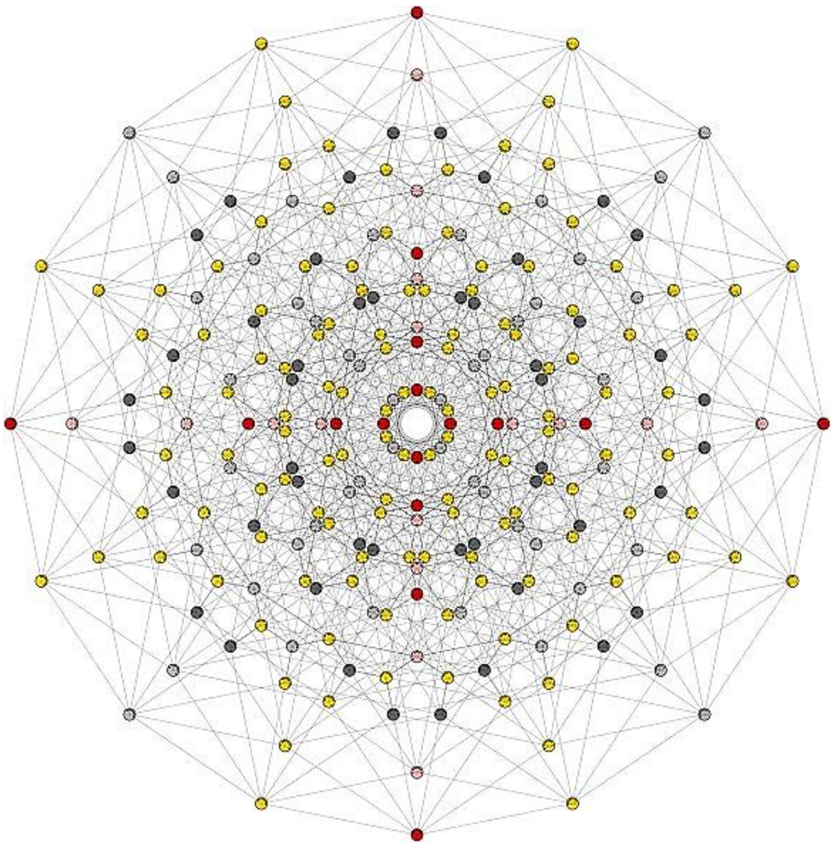
- In a data model, for example, there are several design layers that describe a given data concept.





Graph Relationships

- Graph databases are ideal for analyzing metadata relationships between objects and finding patterns in those relationships.
- Suitable for saving metadata



Types of Metadata

Structural Metadata

- Where
- How

Formatting & Storage

Descriptive Metadata

- Who
- What
- When
- Why
- Where

Description, Usage, & Context

Relationship Metadata

- Data Lineage
- Impact Analysis
- Design Layers
- Graph Patterns

Linkage

Context Helps Prioritize Importance & Relevance

- By placing data in contact, metadata helps determine relevance and importance.
  - The number is 140 – *what does it mean and why do I care?*
  - John Smith – *what is his relationship to me and why do I care?*
  - I have 200 Oracle databases – *what information is stored within them and why do I care?*
  - This customer information is 2 years old – *is it still relevant?*
  - What's the definition of customer? Are we talking about prospects or current, existing customers? – *how are we going to use this customer information?*
  - What does the database field "TR_01" contain? - *Is it important?*
  - If I delete this field - *what other systems would be affected?*
  - The figure "total sales" on this report – how was it calculated? How many sources were summarized? – *is it meaningful?*
  - We just found a correlation between total income and length of time unemployed – how are we calculating total income? What is the definition of unemployed – does volunteer or part-time work count as employment? – *Can I trust this analysis?*
  - Etc.....

Summary

- Metadata provides Data in Context
  - Who, What, Where, When & How of Data
  - Characteristics, Relationships, and Timeliness are key aspects
- Numerous types of Metadata exist
  - Business & Technical
  - Structural, Descriptive, Relationship
- Relationships are critical
  - Lineage
  - Where Used
  - Semantic
  - Graph Patterns

Agenda

1. Grundlagen Metadaten
   - Was sind Metadaten
   - Warum Metadaten
   - Quellen von Metadaten
   - Metadaten Standards
2. Metadaten Management und Data Catalogs
3. Informatica EDC Präsentation
4. Funktionalitäten von Data Catalogs
5. Data Catalogs: Ausgewählte Themen
6. Metadata Strategy und Data Catalog-Einführung

METADATAS IMPORTANCE TO THE BUSINESS

COURSE CURRICULUM

Purpose:

To show metadata's importance to the business.

Outcome:

- Understand how metadata management affects business results.
- Provide examples of metadata's affect on business operations
- Show metadata's value for IT and data management projects

Agenda; The Business Value of Metadata

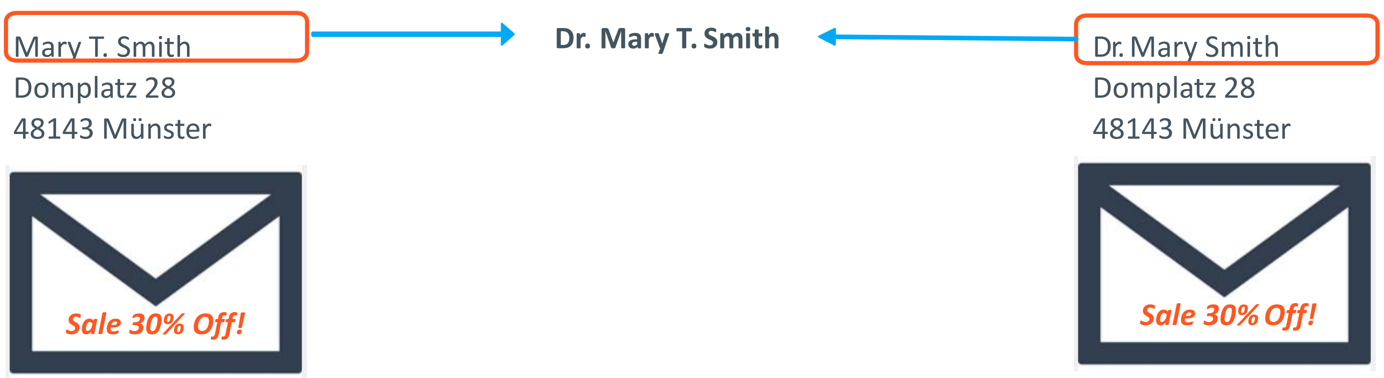
1. Wasted Costs
2. Brand Damage
3. Financial Reporting & Audit
4. Big Data Analytics
5. Efficiencies – Reuse
6. Agility & Change Management
7. Data Governance

METADATA'S IMPORTANCE TO THE BUSINESS

AFFECTING THE BOTTOM LINE

Wasted Costs – Duplicate Mailings

- Has this ever happened to you? You receive two advertisements from the same company announcing an upcoming sale.
- Don’t they know you are the same person?
- Not only have they wasted money sending two mailings, you might be less likely to go to that sale (brand damage).



Common Structural Metadata can help

- This data duplication can occur due to poor metadata.
- In this case, the structural metadata was not consistent across data sources, making it difficult to properly match data instance (e.g. Mary T. Smith vs. Dr. Mary Smith)
- Common, Standardized structural metadata helps improve data quality & data duplication issues.

Common Metadata

Salutation	First Name	Middle Initial	Last Name
Dr.	Mary	T	Smith
Mr.	Marco	S	DiPietro
Etc.			

Metadata
Data Source #1

First Name	Middle Initial	Last Name
Mary	T	Smith
Marco	S	DiPietro
Etc.		

Metadata
Data Source #2

Salutation	First Name	Last Name
Dr.	Mary	Smith
Mr.	Marco	DiPietro
Etc.		

Brand Damage – Knowing the customer

- Has this ever happened to you? You have a credit card with Company X.
  - On the same day that you receive your bill, you also receive an advertisement to sign up for the credit card.
  - And it's at a better rate than you currently have!
- Don't they know you already have a credit card account with them?
- You're not very happy with your credit card company. You certainly don't feel like a valued customer and start looking around for another offer.
  - Not only did they waste money sending mailings to the wrong target audience.
  - Larger damage has occurred in the brand reputation and customer sentiment.

Descriptive Metadata can help

- This type of miscommunication can occur due to poor descriptive metadata.
- For example, perhaps the marketing department requested:
  - **“Let’s send an ad campaign to our customer list”**

A customer is an individual with an active credit card account.

A customer is an individual or household who has engaged with our organization in the past year.

Billing “Customer” Database

Marketing “Customer” Database

Poor Metadata Management can be Expensive

On average organizations waste 15-18% of their budgets dealing with data problems.

Source: Experian

56% of UK marketing organizations say managing data quality is a ‘significant challenge’ .

Source: UK Marketing Today

The US economy loses $3.1 trillion a year due to poor data quality .

Source: Artemis Ventures

In the US, 6.9 billion pieces of mail are undeliverable annually because of address issues .

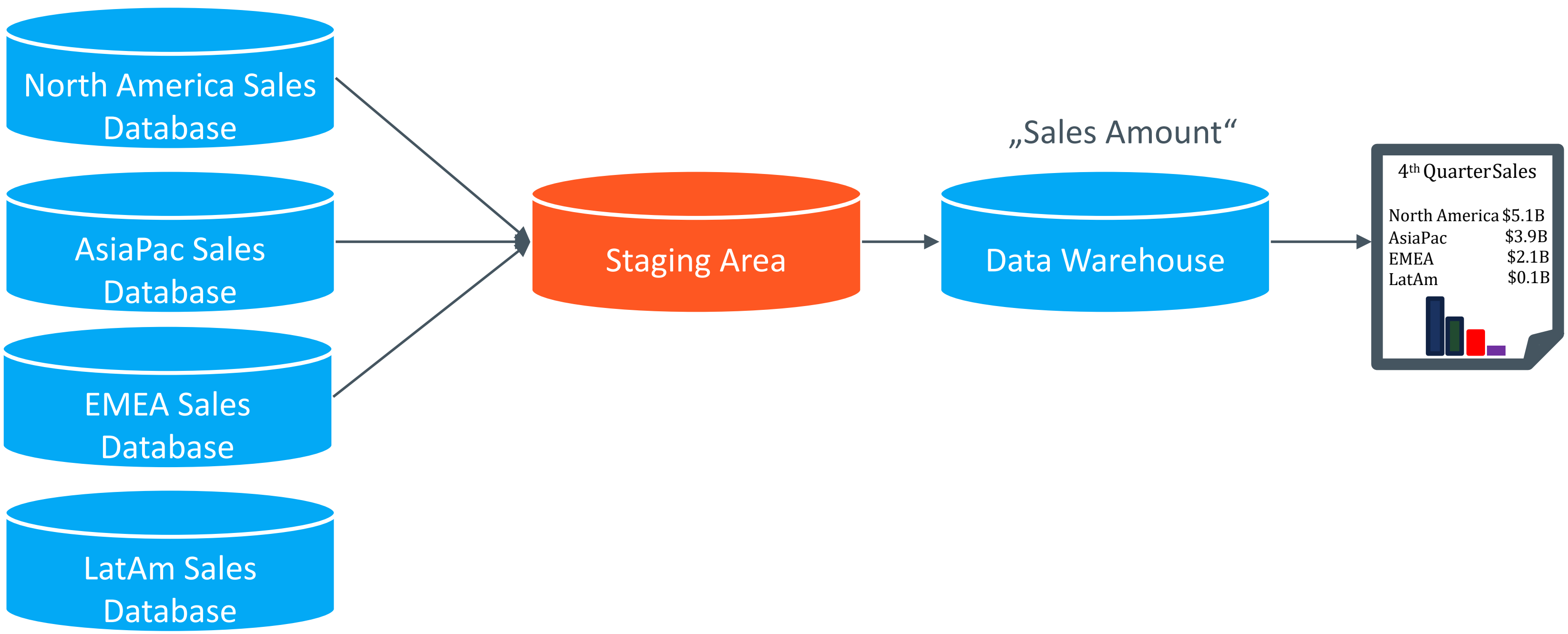
Source: US Postal Service

A very Expensive Example - NASA

- On September 23, 1999 NASA lost the $125 million Mars Climate Orbiter spacecraft after a 286-day journey to Mars.
- Missing Metadata was the culprit
  - Thruster data was sent in English units of pound-seconds (lbf s) instead of Metric units of newton-seconds (N s)
- This metadata inconsistency caused thrusters to fire incorrectly, sending the craft off course – about 90km.
- In addition to the cost of the orbiter were:
  - Brand and Reputational Damage
  - Lost Opportunities for research on the Martian atmosphere & climate

Audit & Tracebility

- Reporting errors at an international retail chain spurred an internal audit to evaluate how financial figures were calculated.
- Because this company had good metadata tracking and lineage, they were easily able to show how information was sourced & manipulated to create key reports.



Big Data Analytics

- Modern advances in data analytics & big data storage provide a wealth of opportunities
  - But the analytics are only as good as the quality of the underlying data
  - Metadata is critical – where did the data come from? What was its intended purpose? What are the units of measure? What is the definition of key terms?
- Good data analysis is based on good data. Good data requires good metadata.

Metadata is Critical for Big Data Analytics & BI

The absence of commonly understood and shared metadata and data definitions and the lack of data governance are cited as the main impediments to the success of Data Lakes.

Source: Radiant Advisors

Many data scientists and BI professionals spend an estimated 50 – 90% of their time cleaning and reformatting data to make it fit for purpose.

Source: DataCenterjournal.com

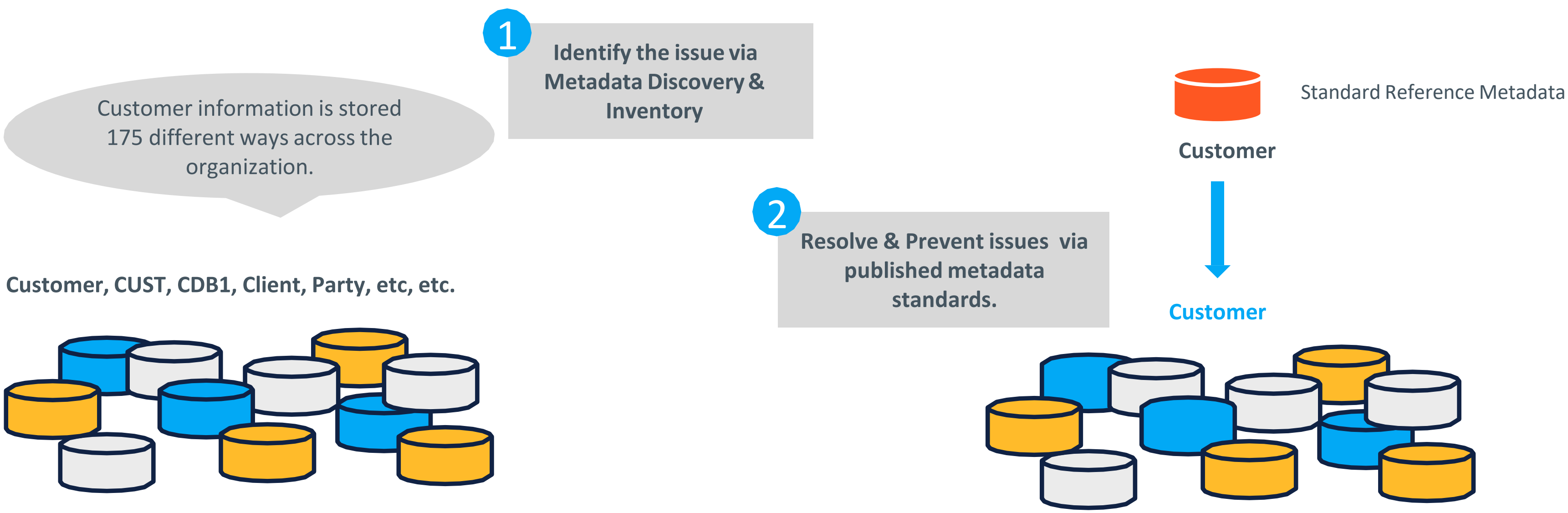
71% of interviewees surveyed in larger global organizations expect data- driven digitization to help their business grow. But…

- 70% say the biggest barrier is finding the right data
- 62% cite inconsistent data.

Source: Stibo Systems

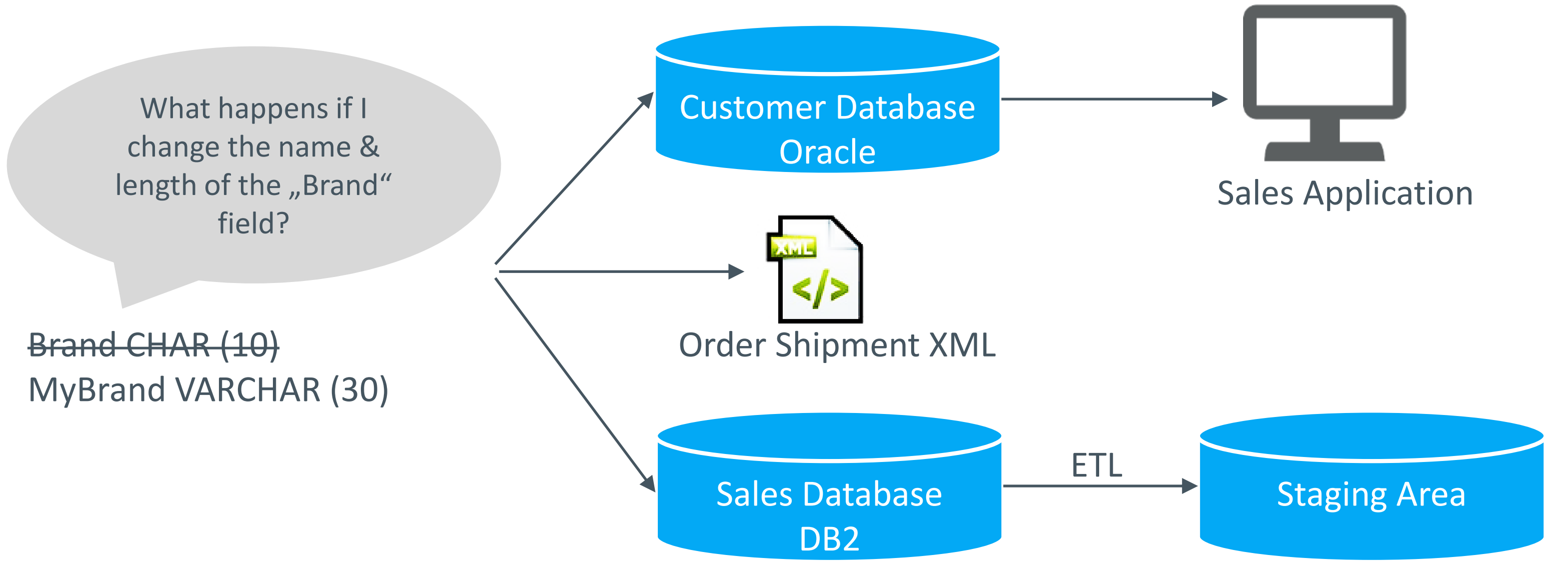
Efficiencies & Reuse

- Metadata Management can help rationalize data storage throughout the organization, leading to significant efficiencies & cost reduction.



Agility & Change Management

- Metadata management provides an inventory & roadmap of data assets & their interrelationships with other data, systems, and application.
- With this roadmap in place, it is easier to assess the impact of a proposed change, significantly reducing development and maintenance time -> *driving agility & responsiveness*



Data Governance

- Data Governance is the process of managing and improving data for the benefit of all stakeholders. Metadata is key to an effective data governance program.
- Key data governance artifacts that are metadata-driven include:
  - Business Glossary: Defining the business metadata definitions for critical data elements.
  - Data Stewardship: Aligning data stewardship or ownership roles to key data objects.
  - Data Standards: Metadata standards provide governance and rules for current and future development, based on both business and IT input.
  - Privacy & Security: Identification of privacy and security levels for business data is a key aspect of data governance can be managed with metadata definitions.
  - Traceability & Audit: Understanding of how data is used across the organization, how key financial figures are calculated, etc.
- Metadata management is a key foundation for any Data Governance initiative.

Common Definitions and owner assignment : Business Glossary Definitions for core Data Elements

What do we need to govern?	What is the benefit?	Use case	Why do we need to govern?
Common Definitions and owner assignment Core Data Elements	Defining the business terms is crucially important as confusion and conflict around data usage centres on what data means. A succinct definition should describe the term using the business language as well as the business purpose of the term. Synonyms and tpnyms should identified and eliminated. The benefits can be summarised as: • Imporved fact-based decision making • Reduced cost of change • Increased speed to market • Reduced confusion among the information consumers.	Quite a number of organisations suffer from the inconsistent view of a very basic measurement: the number of customer information reported by various Business Units. One report can state that the company has 1 million customers, another can say 500,000 and yet another states 1,5 million. This might be as a result of each department´s unique definition of the term „customer“: marketing may define „customer“ as a household versus an individual: finance may define „customer“ by the number of accounts; and while operations may define „customer“ by the number of customer-to-product relationships.	Misunderstandings caused by incorrect interpretation of enterprise information increase the risk of expensive errors. The complete business costs of ambiguous information are even more dramatic. Productivity plummets when executives in the boardroom or employees at the coal face waste time searching for or misinterpreting poorly named data.

Consequences of non-governing

Increased risk of inaccurate regulatory reporting, erroneous business performance reporting, reduced customer satisfaction, increased complexity, unreliable data

Common Definitions and owner assignment for Micro Services

<table>
<tr><th>What do we need to govern?</th><th>What is the benefit?</th><th>Use case</th><th>Why do we need to govern?</th></tr>
<tr><td>Common Definitions and owner assignment
Mirco services</td><td>Business driven owner assignment for micro services supports
• Consistent data and computing
• Reuse where it is necessary to have a single source of truth
• Higher Transparency regarding IT assets and therefore improves faster reporting
• Encapsulation of domain internal knowledge
• Respecting the bounded context of domain which is a core success factor for implementing micro services</td><td>• One typical example is calculating tariffs.Maybe it´s a good idea to have different services for calculating different products. But calculating different prices for the same product will cause trouble.
• Knowing which business data is stored in which micro service is also much easier if the services have a clear owner assignment</td><td>It`s unlikely that without any governance the development of mirco services in different teams and organisations leads to a consistent service architecture.</td></tr>
</table>

Consequences of non-governing

Risk of inconsistent data and computing finally leads to inconsistent customer communication and compromises omni-channel strategy.

Group Metrics/KPIs governance

What do we need to govern?	Benefit	Use case	Why do we need to govern?
Group Metrics/KPIs	The misunderstanding of a metric (i.e. calculated / derived information based on multiple attributes) or key performance indicator (KPI) can lead to implementation delays, lost staff productivity, lost business opportunities and in the worst case, poor or incorrect business decisions	A good retail example is: What is "Net Sales Revenue"? Does it include taxes, commissions, shipping or cost of goods sold? Is there a common, well documented understanding across the organisation use case: Definition of "new business" is changing from BU to BU and causing inconsistent and inaccurate reporting. Consequently it is adversely impacting the decisions and actions regarding the business performance.	To be able to accurately measure the business performance and the value of the current and potential customers, we need well defined, consistently calculated measures. Otherwise our business decisions will be based on false premise.

Consequences of non-governing

Increased risk of inaccurate regulatory reporting, erroneous business performance reporting

Summary

- Effective Metadata Management has a positive effect on business operations and efficiency
  - Cost Reduction, Efficiencies, and Reuse
  - Brand Reputation
  - Financial Reporting & Audit
  - Big Data Analytics
  - Agility & Change Management
  - Data Governance

Agenda

1. Grundlagen Metadaten
   - Was sind Metadaten
   - Warum Metadaten
   - Quellen von Metadaten
   - Metadaten Standards
2. Metadaten Management und Data Catalogs
3. Informatica EDC Präsentation
4. Funktionalitäten von Data Catalogs
5. Data Catalogs: Ausgewählte Themen
6. Metadata Strategy und Data Catalog-Einführung

COURSE CURRICULUM

Purpose:
To outline the various sources of metadata across and beyond the organization.

Outcome:

- Learn the various sources of metadata.
- Understand how these sources interrelate within
- and beyond the organization.

SOURCES OF METADATA

METADATA IS EVERYWHERE

Sources & Types of Metadata

There are many Sources and Types of Metadata

- Relational databases
- Data Models
- Text Documents
- XML
- Open Data
- Internet of Things (IoT)
- Photos / Images
- Social Media
- COBOL Copybooks

- Application Code
- Data Transformation / ETL Tools
- Spreadsheets
- Data Quality Tools
- Business Process Models
- Business Intelligence (BI) Tools
- ERP, CRM, and Packed Applications
- Big Data platforms
- Etc.... ***there are many more***

Technical & Business Metadata

- The technical structure of a relational database is defined by **DDL (data definition language)**. It describes the structure / schema for how data is stored in a database.
- A **Glossary or Data Dictionary** generally stores the business metadata.

Technical Metadata

```
CREATE TABLE EMPLOYEE (
    employee_id        INTEGER NOT NULL,
    department_id      INTEGER NOT NULL,
    employee_fname     VARCHAR(50) NULL,
    employee_lname     VARCHAR(50) NULL,
    employee_ssn       CHAR(9) NULL);

CREATE TABLE CUSTOMER (
    customer_id        INTEGER NOT NULL,
    customer_name      VARCHAR(50) NULL,
    customer_address VARCHAR(150) NULL,
    customer_city      VARCHAR(50) NULL,
    customer_state     CHAR(2) NULL,
    customer_zip       CHAR(9) NULL);
```

Business Metadata

Term	Definition
Employee	An employee is an individual who currently works for the organization or who has been recently employed within the past 6 months.
Customer	A customer is a person or organization who has purchased from the organization within the past 2 years and has an active loyalty card or maintenance contract.

Data

John Smith

Relation Database Metadata

Customer

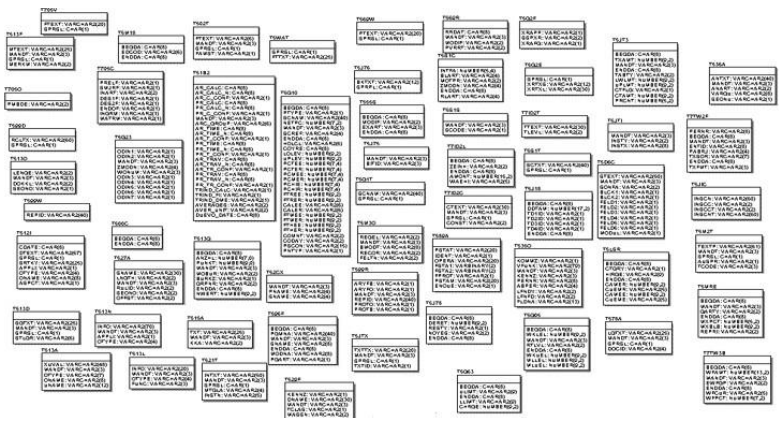
First Name	Last Name	Company	City	Year Purchased
Joe	Smith	Komputers R Us	New York	1970
Mary	Jones	The Lord's Store	London	1999
Proful	Bishwal	The Lady's Store	Mumbai	1998
Ming	Lee	My Favorite Store	Beijing	2001

Definition	Last Name represents the surname or family name of an individual.
Business Rules	In the Chinese market, family name is listed first in salutations.
Format	VARCHAR(30)
Abbreviation	LNAME
Required	YES
Etc.	Numerous technical & business metadata including security, privacy, nullability, primary key, etc.

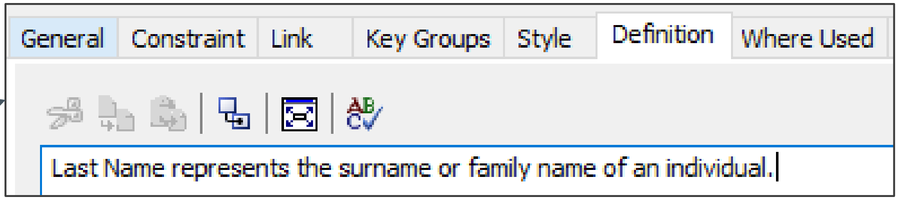
Data Models are a good Source of Metadata

- Data Models are another good source of both business & technical metadata for relational databases.
- They store structural metadata as well as business rules & definitions.

Technical Metadata



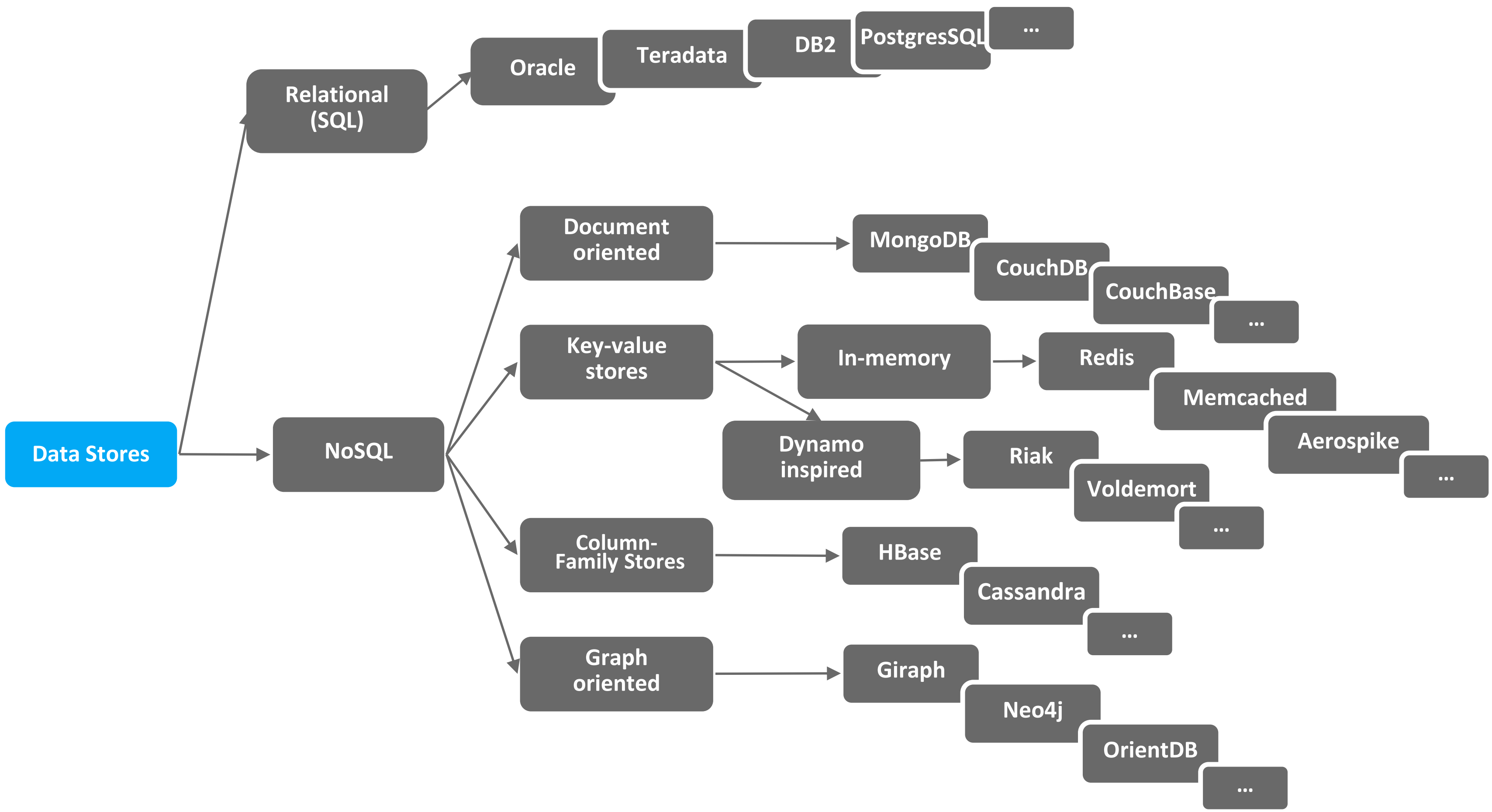
Business Metadata



Customer

Customer_ID	CHAR(18) NOT NULL
First Name	CHAR(18) NOT NULL
Last Name	CHAR(18) NOT NULL
City	CHAR(18) NULL
Date Purchased	CHAR(18) NULL

Data Store Types



NoSQL Metadata – Key Value Databases

- NoSQL Databases are often optimal solutions for flexibility & performance in certain scenarios.
  - One common NoSQL database is a key-value pair database (e.g. Redis, Oracle NoSQL, etc.)
  - They can support extremely high volumes of records & state changes per second through distributed processing and distributed storage.
  - Use cases include: Managing user sessions in web applications, online gaming, online shopping carts, etc.
- While they clearly have their strengths, metadata management is not one of them.
  - Metadata for NoSQL databases is typically minimal or non-existent.
  - The structure & metadata is generally determined by the application code

Key	Value
1839047	John Doe, Prepaid, 40.00
9287320	01/01/2008, 50.00, Green

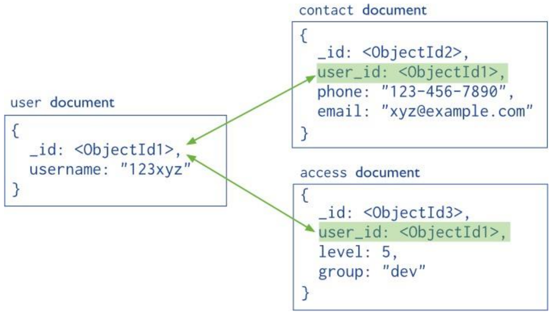
NoSQL Metadata – Document Databases

- Document databases are popular ways to store unstructured information in a flexible way (e.g. multimedia, social media posts, etc.)
- Each Collection can contain numerous Documents which could all contain different fields.

{type: "Artifact",
medium: "Ceramic"
country: "China",
}

{type: "Book",
title: "Ancient China"
country: "China",
}

- Some data modeling can be done, allowing metadata to be stored in data modeling tools.



*

* Example from docs.mongodb.com

Big Data Platform Metadata

- Big Data platforms (e.g. Hadoop-based) are typically based on system of files (HDFS)
- As a result, the detailed structure that is found in a relational database platform does not exist
- Metadata still exists for these platforms.
- Technical Metadata
  - Tree structure of HDFS directories
  - Directory and file attributes (ownership, permissions, quotas, replication factor, etc.)
  - Metadata about logical data sets (e.g. format, statistics, etc.)
  - Data ingest & transformation lineage
- Business Metadata
  - Description of file
  - Tags
  - There are components that allow you to add structure within the Hadoop ecosystem (e.g. Hive)

```
data/dfs/name
├── current
│ ├── VERSION
│ ├── edits_0000000000000000001-0000000000000000007
│ ├── edits_0000000000000000008-0000000000000000015
│ ├── edits_0000000000000000016-0000000000000000022
│ ├── edits_0000000000000000023-0000000000000000029
│ ├── edits_0000000000000000030-0000000000000000030
│ ├── edits_0000000000000000031-0000000000000000031
│ ├── edits_inprogress_0000000000000000032
│ ├── fsimage_0000000000000000030
│ ├── fsimage_0000000000000000030.md5
│ ├── fsimage_0000000000000000031
│ ├── fsimage_0000000000000000031.md5
│ └── seen_txid
└── in_use.lock
```

CoBol Copybook Metadata

- **What is a COBOL Copybook?** – In COBOL, a copybook file is used to define data elements that can be referenced by many programs
- **What is COBOL Copybook Metadata?** – structure, definition

```
01  STUDENT.
  20  ID                               PIC 9(8).
  20  FIRST_NAME                       PIC X(32).
  20  LAST_NAME                        PIC X(32).
*                                      YYYYMMDD
  20  DATE_OF_BIRTH                    PIC S9(8) COMP.
  20  NUMOF_COURSES                    PIC 9(4) COMP.
  20  NUMOF_BOOKS                      PIC 9(4) COMP.
  20  COURSES.
    25 COURSE OCCURS 8 TIMES DEPENDING ON NUMOF_COURSES.
      30  COURSE_ID                    PIC 9(8).
      30  COURSE_TITLE                 PIC X(48).
      30  INSTRUCTOR_ID                PIC 9(8).
      30  NUMOF_ASSIGNMENTS            PIC 9(4) COMP.
      30  ASSIGNMENTS OCCURS 4 TIMES DEPENDING ON NUMOF_ASSIGNMENTS.
        40  ASSIGNMENT_TYPE            PIC X(12).
        40  ASSIGNMENT_TITLE           PIC X(48).
*                                      YYYYMMDD
        40  DUE_DATE                   PIC S9(8) COMP.
        40  GRADE                      PIC S9V9.
  20  BOOKS.
    25  BOOK OCCURS 1 TO 5 TIMES DEPENDING ON NUMOF_BOOKS.
      30  ISBN                         PIC X(10).
*                                      YYYYMMDD
      30  RETURN_DATE                  PIC 9(8) COMP.
```

Metadata
Describes structure & format of data

SARACUS CONSULTING

XML Metadata

- **What is XML?** – (Extensible Markup Language) is used to store and transport data. It's often a complement to HTML, which is used to format the data.
- **What is XML Metadata?** – Similar to DDL, an XML Schema (XSD) defines the structure & format of data

Metadata

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="shiporder">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="orderperson" type="xs:string"/>
      <xs:element name="shipto">
        <xs:complexType>
          <xs:sequence>
            <xs:element name="name" type="xs:string"/>
            <xs:element name="address" type="xs:string"/>
            <xs:element name="city" type="xs:string"/>
            <xs:element name="country" type="xs:string"/>
          </xs:sequence>
        </xs:complexType>
      </xs:element>
          </xs:sequence>
    <xs:attribute name="orderid" type="xs:string"
    use="required"/>
  </xs:complexType>
</xs:element>
</xs:schema>
```

XSD

Data

```xml
<?xml version="1.0"?>
<shipto>
  <name>John Smith</name>
  <address>123 Main ST</address>
  <city>Boise</city>
  <country>USA</country>
</shipto>
```

XML

Data

Order Shipment

Ship to:
John Smith
123 Main ST
Boise
USA

…………………………
…………………………
…………………………

Json Metadata

- **What is JSON?** – (JavaScript Object Notation) is a minimal, readable format for structuring data. It is used primarily to transmit data between a server and web application, as an alternative to XML.
- **What is JSON Metadata?** – structure, definition

For example, assume we have a JSON based product catalog. This catalog has a product which has an id, a brand, a price, and an optional set of tags.

Data

```
{
"id": 127849,
"brand": "Super Cooler",
"price": 12.50,
"tags": ["camping", "sports"]
}
```

Example Product in the API

Context Needed (i.e. Metadata)

- Can the ID contain letters?
- What is a brand?
- Is a price required?
- Etc.

Metadata

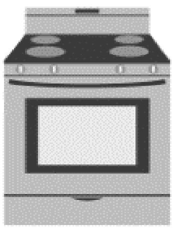
```
{
"$schema": "http://json-schema.org/draft-04/schema#",
"title": "Product",
"description": "A retail product from Acme's online catalog",
"type": "object",
"properties": {
  "id": {
    "description": "The unique identifier for a product",
    "type": "integer"
  },
  "brand": {
    "description": "The brand name of the product as shown in the online catalogue",
    "type": "string"
  },
  "price": {
    "type": "number",
  },
  "tags": {
    "type": "array",
    "items": {
      "type": "string"
    },
    "minItems": 1,
    }
},
"required": ["id", "brand", "price"]
    }
```

JSON Schema

SARACUS CONSULTING

IoT Metadata

- **What is the IoT?** – The Internet of Things (IoT) is a network of physical devices that are able to share data over a network.
- **What is IoT Metadata?** – Metadata is necessary to provide context around the readings generated by IoT devices, e.g. units of measure, type of measurement, etc. .



```
<timestamp value='2016-03-07T16:24:30'>
      <numeric name='Temperature' value='140' unit='F'/>
</timestamp>
```

```
<timestamp value='2016-03-07T16:24:30'>
      <numeric name='HeartRate' value='140' unit='BPM'/>
</timestamp>
```

140

Is **140** the temperature of my stove or my max heart rate on my run?

Image Metadata

- Metadata is critical for locating images online, as well as identifying copyright information, etc.
- Some information is system-generated, while other is user-defined.

Technical Metadata (Embedded in Photo)

Camera:	Apple iPhone 6 Plus
Lens:	iPhone 6 Plus back camera 4.15mm f/2.2 Shot at 4.2 mm Digital Zoom: 5.006134969×
Exposure:	Auto exposure, Program AE, 1/7,937 sec, f/2.2, ISO 32
Flash:	Auto, Did not fire
Date:	April 13, 2016 5:35:53PM (timezone not specified) (1 month, 11 days, 14 hours, 14 minutes, 46 seconds ago, assuming image timezone of US Pacific)
File:	3,264 × 2,448 JPEG (8.0 megapixels) 800,782 bytes (782 kilobytes)

Descriptive Metadata (User Defined)

Title	DATAVERSITY EDW 2016 San Diego
Keywords	EDW 2016, San Diego, Bay Photos
Location	San Diego

Administrative Metadata (User Defined)

Author	Donna Burbank
Copyright	None
Licensing	None

Social Media Metadata

- Metadata from Social Media, such as Twitter, can help identify trend and sentiment analysis, for example.

Author

Text/Content of Tweet

Hash Tag

# of Retweets

Embedded Metadata

Date/Time
Location
Language
Device Used
ID of Tweet
Etc.

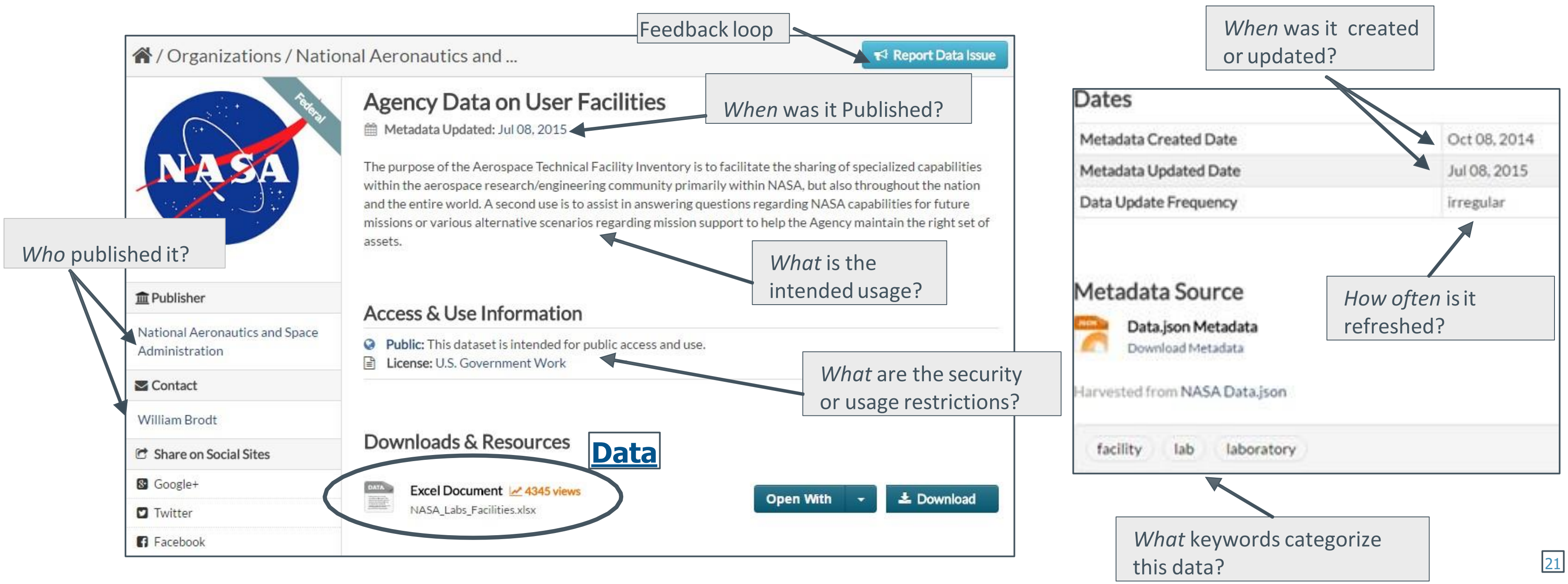
Open Data Metadata

- **What is a Open Data?** – Open Data is data that can be freely used and redistributed by anyone

Many governmental organizations, for example, provide open data sets

Open Data Metadata

- **What is Open Data Metadata?** – Metadata provides the context that makes open data usable and credible.



Sample JSON Metadata file for Open Data

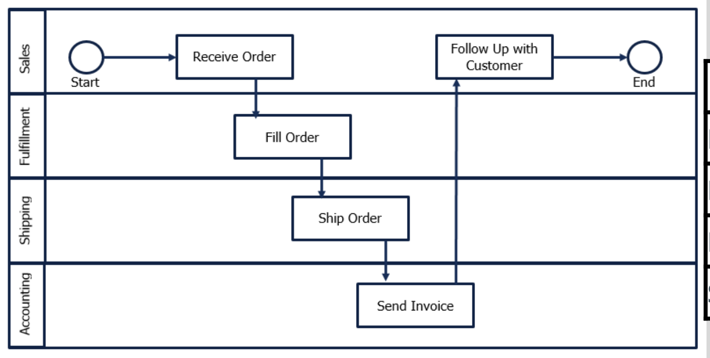
{"@type": "dcat:Dataset", "_id":
{"$oid": "55942a79c63a7fe59b497552"}, "accessLevel": "public", "accrualPeriodicity":
"irregular", "bureauCode": ["026:00"], "contactPoint": {"@type": "vcard:Contact", "fn":
"William Brodt", "hasEmail": "mailto:wbrodt@nasa.gov"},
"description": "The purpose of the Aerospace Technical Facility Inventory is to facilitate the sharing of specialized capabilities within the aerospace research/engineering community primarily within NASA, but also throughout the nation and the entire world. A second use is to assist in answering questions regarding NASA capabilities for future missions or various alternative scenarios regarding mission support to help the Agency maintain the right set of assets.",
"distribution": [{"@type": "dcat:Distribution", "downloadURL":
"http://open.nasa.gov/datasets/NASA_Labs_Facilities.xlsx",
"mediaType": "application/vnd.ms-excel"}],
"identifier": "NASA-0000061", "keyword": ["Lab", "Laboratory", "Facility"], "language": ["en-US"],
"license": "http://www.usa.gov/publicdomain/label/1.0/",
"modified": "2014-06-05",
"programCode": ["026:000"],
"publisher": {"@type": "org:Organization", "name": "National Aeronautics and Space Administration"},
"references": ["https://nrpi.hq.nasa.gov/ATFI/", https://nrpi.hq.nasa.gov/ATFI/URLLinks.cfm"],
"spatial": "United States",
"title": "Agency Data on User Facilities"}

This same information can be downloaded as a JSON file.

Business Process Model Metadata

- Business Process Models describe key activities within the organization.
- Linking these processes to the data that is Created, Updated, or Deleted (CRUD) is important to understanding data usage.

Business Process Model



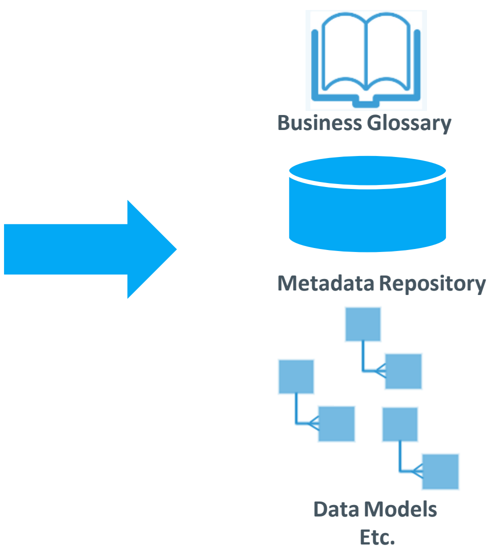
CRUD Matrix

	Customer	Order	Account	Invoice	Product
Receive Customer Order	R	C	C, R		
Process Customer Order	C,R,U		R,U		R
Fill Order	R,U		R,U		R,U
Send Invoice	R,U		R,U	C	

Human Metadata

- Much business metadata and the history of the business exists in employee’s heads.
- It is important to capture this metadata in an electronic format for sharing with others.
- Avoid the dreaded “I just know”

Part Number is what used to be called Component Number before the acquisition.



Speadsheets Metadata

- Spreadsheets, like the documents describe earlier, have important metadata properties.

Business Term	Abbreviation	Definition
After Action Review	AAR	Team recap after every activity to share learning & improve best practices.
Activty Based Costing	ABD	Costs are allocated to products via cost drivers linked to various categories linked to the costs of manufacturing.
Component Number	C/N	Unique identifier associated with a given design for manufacture within ACME Corp.
Manufacturing Change Order	MCO	A change order used to make a manufacturing change. This typically does not involve a design change to the item.
Part Number	P/N	Unique identifier associated with a given design for manufacture within ACME Corp.
Etc.		...

Metadata

Properties

Size	9.02KB
Title	ACME Corp Business Terms
Tags	terms, acronyms
Comments	Business Terms for ACME C...
Template	
Status	Production
Categories	Manufacturing
Subject	Glossary
Hyperlink Base	Add text
Company	ACME Corporation

Related Dates

Last Modified	Today, 12:34 PM
Created	Today, 12:31 PM
Last Printed	

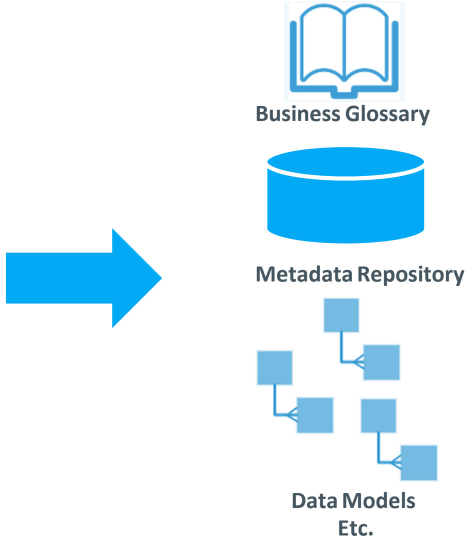
Related People

Manager	Joe Smith

Metadata in Spreadsheets

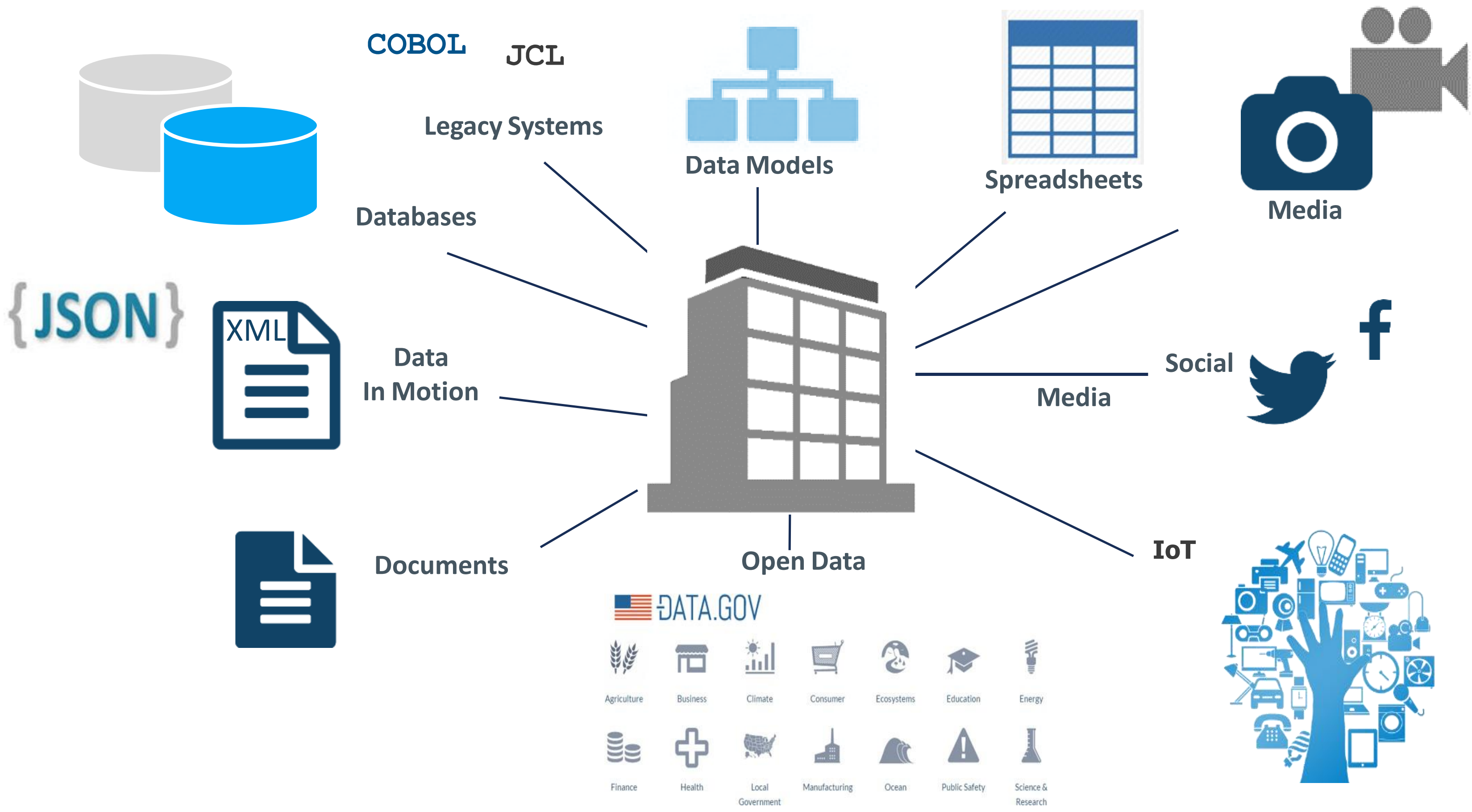
- These spreadsheets commonly contain critical metadata that should be shared across the organization (similar to the “human metadata”).

Business Term	Abbreviation	Definition
After Action Review	AAR	Team recap after every activity to share learning & improve best practices.
Activty Based Costing	ABD	Costs are allocated to products via cost drivers linked to various categories linked to the costs of manufacturing.
Component Number	C/N	Unique identifier associated with a given design for manufacture within ACME Corp.
Manufacturing Change Order	MCO	A change order used to make a manufacturing change. This typically does not involve a design change to the item.
Part Number	P/N	Unique identifier associated with a given design for manufacture within ACME Corp.
Etc.		...



Metadata across & beyond the Organization

- Metadata exists in many sources across & beyond the organization.



Summary

- There are many Sources and Types of Metadata
  - Relational databases
  - Data Models
  - Text Documents
  - XML
  - Open Data
  - Internet of Things (Iot)
  - Photos / Images
  - Social Media
  - COBOL Copybooks
  - Etc.
- A Consolidated View of metadata is a valuable asset to the organization

Agenda

1. Grundlagen Metadaten
   - Was sind Metadaten
   - Warum Metadaten
   - Quellen von Metadaten
   - Metadaten Standards
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
5. Metadata Strategy und Data Catalog-Einführung

COURSE CURRICULUM

Purpose:

To provide an overview of metamodels and metadata industry standards.

Outcome:

- Learn the definition, importance and usage of a metamodel
- Understand metadata registries
- Identify key industry standards

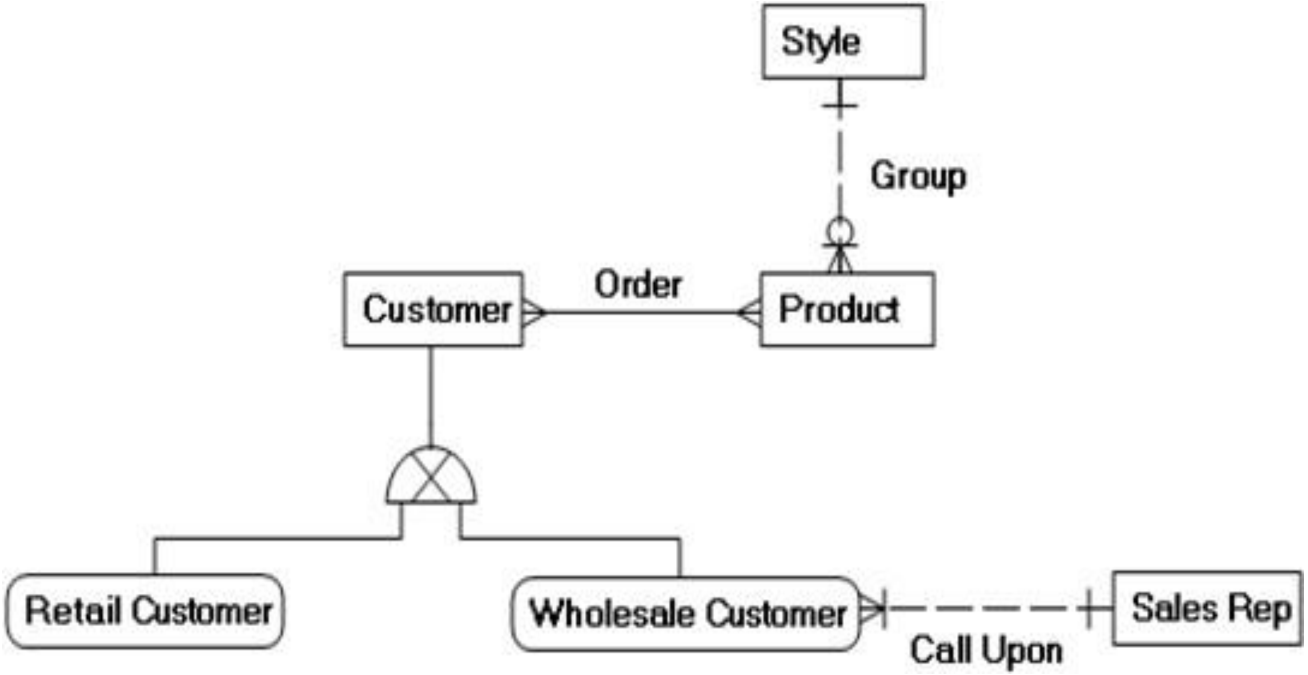
Agenda: Metamodels & Metadata Standards

1. What are Metamodels?
2. The Importance of Metamodels
3. What are Metadata Registries?
4. Metadata Standards

DEFINITION & OVERVIEW

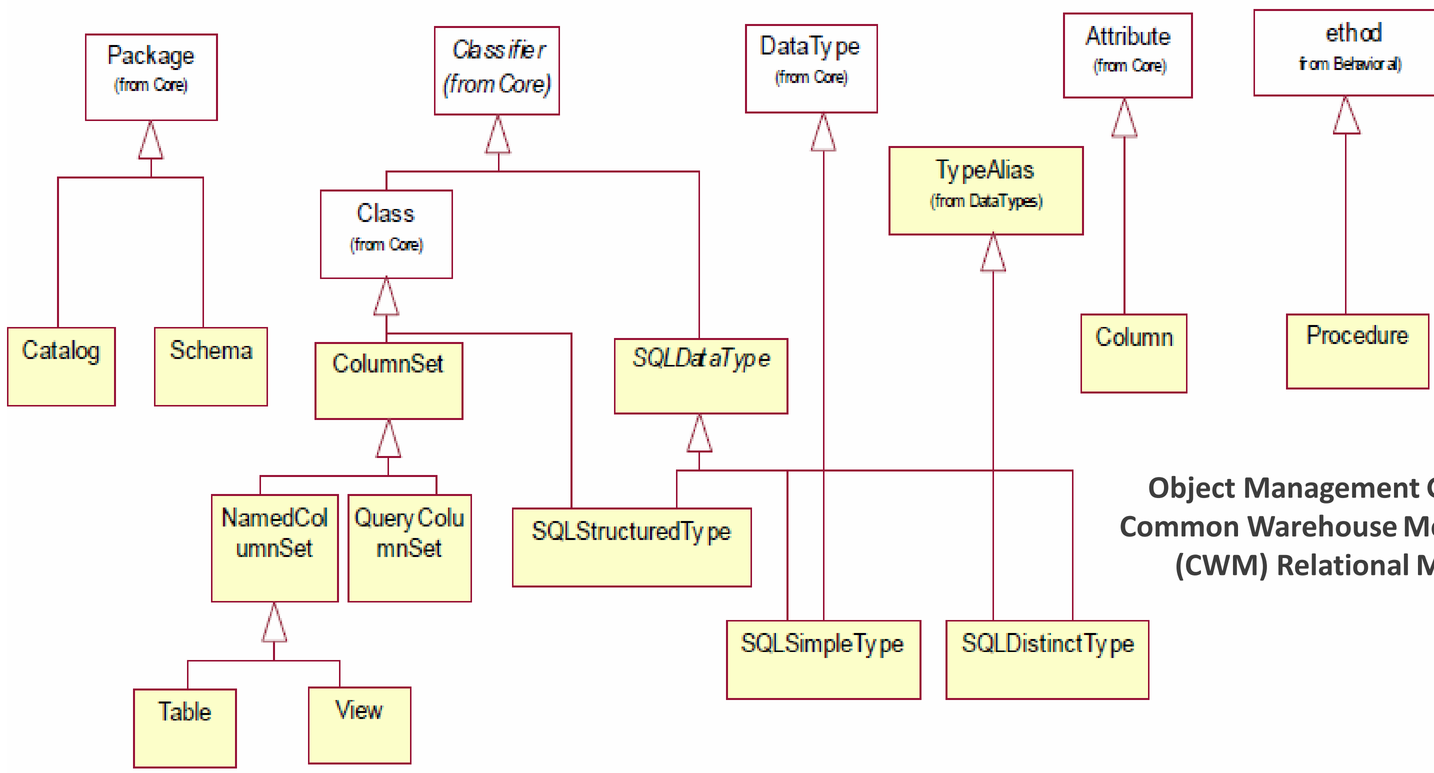
A Data Model describes a the Data of a Business

- Many of use are familiar with data models, which describes the core business entities and their associated relationships, definitions and business rules.



A Metamodel is a Data Model for Metadata

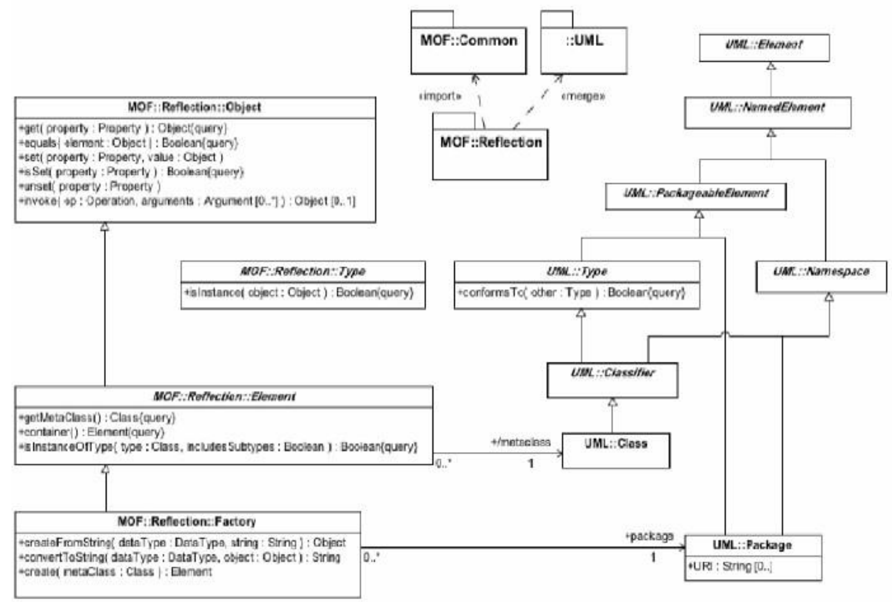
- Think of a Metamodel as a data model for metadata, which describes the core metadata objects, and their relationships and associated business rules.



Object Management Group's Common Warehouse Metamodel (CWM) Relational Model

A Meta-Metamodel is the Framework for the Metamodel

- A Meta-Metamodel provides the framework for defining models for metadata. It defines a common abstract syntax for defining metamodels.
- For example, the MOF (Meta Object Facility) from the OMG (Object Management Group) is a meta-metamodel for the CWM (Common Warehouse Metamodel)
- Basically, it provides a common way to express the models, which is important for interchange between systems.
- The actual metadata is stored in some sort of database persistence mechanism, e.g. Metadata Repository database



Meta Levels

- The Object Management Group (OMG) uses the following levels to describe their architecture.

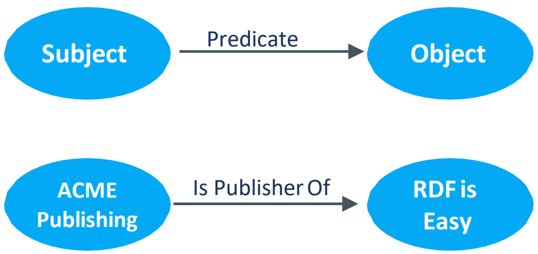
Meta Level	Term	Example
M3	Meta-Metamodel	MOF
M2	Metamodel	CWM
M1	Metadata	Metadata from a data warehouse
M0	Data	Data from a data warehouse

Most of your work will be done here.

- Other standards and tools use similar levels, e.g.
  - M2 – a model describing how the metadata is stored
  - M3 – the framework/method you use to store & share the metadata (e.g. XML)

The Semantic Web & RDF

- The RDF (Resource Description Framework) model from the World Wide Web Consortium (W3C) provides a way to link resources on the web (people, places, things). It provides a common framework for applications to share information without losing meaning.
  - Search Engines
  - Exchanging data between datasets
  - Sharing information with applications / APIs
  - Building social networks
  - Etc.



- The goal is to move from a web of documents to a web of data.
- The Framework is a simple way to express relationships between resources.
  - IRIs (International Resource Identifiers) (e.g. URI) identify resources
  - Simple triples relate objects together in the format: <subject> <predicate> <object>
  - These relationships create a connected Graph
  - There are several serialization formats, with RDF XML being a common one. For example:
    - Turtle is a human-friendly format
    - RDF/XML
    - JSON-LD
  - Schemas define the vocabularies used to describe the objects
    - Dublin Core and Schema.org are described further in the Standards section

Creating a Web of Data


```
"@context": "http://schema.org", *
"location": {
    "@type": "Place",
    "name": "Sheraton San Diego Hotel & Marina",
    "address": {
      "@type": "PostalAddress",
     "streetAddress": "1380 Harbor Island Drive",
      "addressLocality": "San Diego",
      "addressRegion": "CA",
      "postalCode": "92101"
    },
     "telephone" : "+1-877-734-2726",
     "image":
"http://edw2016.dataversity.net/uploads/ConfSiteAssets/
72/image/sheraton.jpg",
     "url":"http://edw2016.dataversity.net/travel.cfm"
```

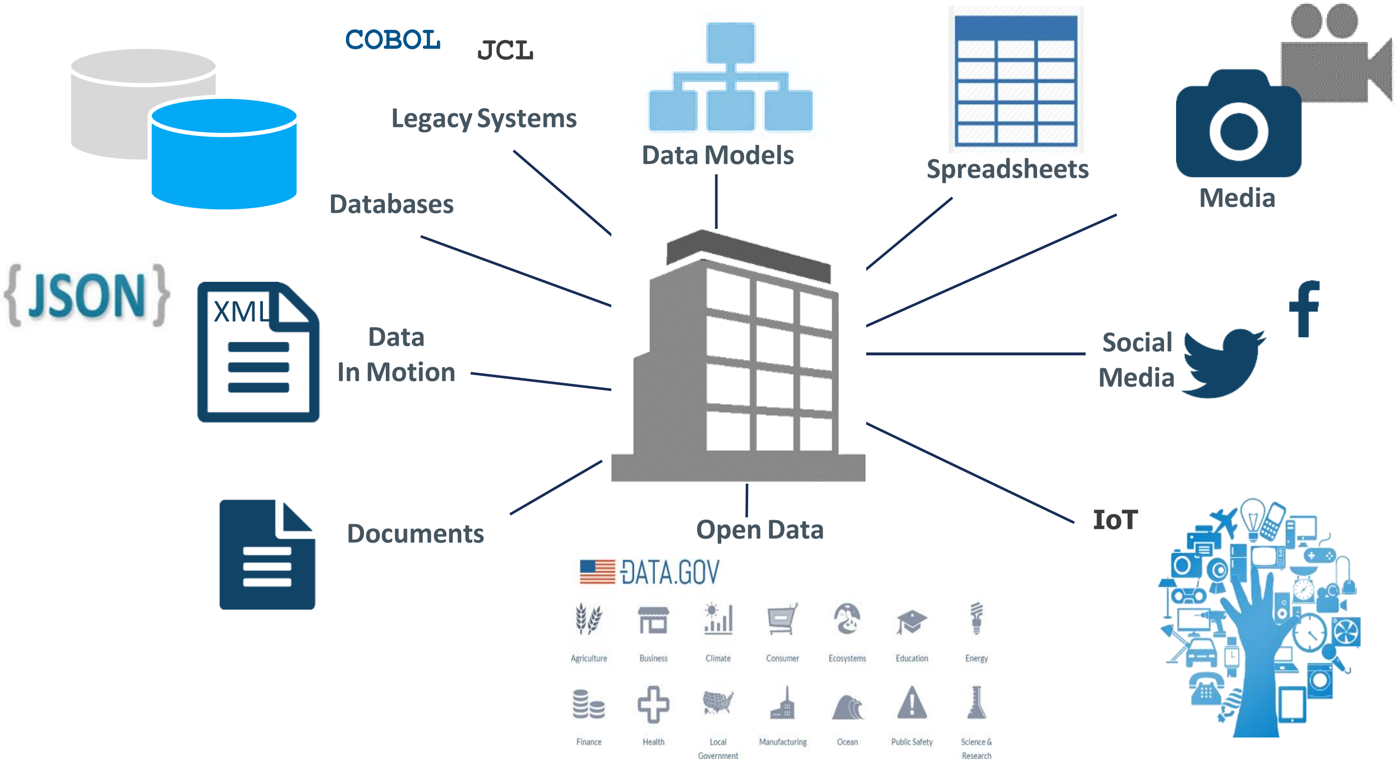
@type: Place

Sheraton San Diego Hotel & Marina
1380 Harbor Island Drive
San Diego, California 92101 USA

```
"@context": "http://schema.org",
"location": {
   "@type": "Place",
   "name": "Sheraton San Diego Hotel & Marina",
   "address": {
     "@type": "PostalAddress",
    "streetAddress": "1380 Harbor Island Drive",
     "addressLocality": "San Diego",
     "addressRegion": "CA",
     "postalCode": "92101"
   },
    "telephone" : "+1-877-734-2726",
    "image": "http://mysite.com/edw16photo.jpg",
    "url":"http://mysite.com/myphotos"
 },
```

The importance of Metamodels

- Each type of Metadata requires its own Metamodel for Storage.
  - Some have overlapping model objects (e.g. Relational Databases & Data Models both contain TABLEs)
  - Metadata Standards help rationalize metadata across within & across multiple sources



SHARING DATA CONSISTENTLY

Metadata Registries

- **What is a Metadata Registry?** Metadata registries are used whenever data must be used consistently between a group of organizations or within an organization (e.g. via XML or other data exchange). For example:
  - Health care research organizations
  - Governmental organizations
  - Data warehouse teams
- Registries typically contain both the definitions of elements, as well as the structural constraints (e.g. data types)

Metadata Registry – An Example

- This is an example of a Metadata Registry for sharing information between European (EU) Institutions.

MDR Metadata Registry

MDR > *Home*

Welcome to the Metadata Registry (MDR). The Metadata Registry registers and maintains definition data (metadata elements, named authority lists, schemas, etc.) used by the different European Institutions involved in the legal decision making process gathered in the Interinstitutional Metadata Maintenance Committee (IMMC) and by the Publications Office of the EU in its production and dissemination process.

The following definition data are maintained in the Metadata Registry:

- Named Authority Lists (Common Authority Tables/Value lists)
- IMMC Core metadata exchange protocol
- European Legislation Identifier (ELI)
- OP Core metadata element set
- EuroVoc thesaurus and alignments (SKOS/XML distributions)
- Common Data Model (CDM) - Ontology of the CELLAR (content and metadata repository)
- OJEEP (Official Journal Electronic Exchange Protocol)
- Style sheets for presentation

Other reference data maintained at the Publications Office of the EU:

- Formex (Formalized Exchange of Electronic Publications)

The Metadata Registry is maintained by the Publications Office of the EU.

The following NAL's are maintained in the Metadata Registry:

1. Released versions

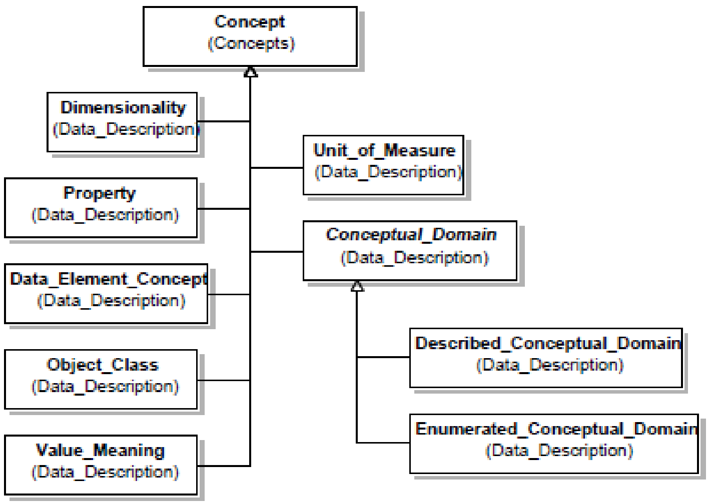
- Address types
- Administrative territoral units
- Administrative territoral units types
- Case reports
- Case statuses
- Concept statuses
- Continents
- Corporate bodies
- Countries
- Currencies
- Dataset statuses
- Dataset types
- Data themes
- Distribution types
- Documentation types
- EU budget amount statuses
- EU budget stages
- EU budget statuses
- EU programmes
- Events *(under review)*
- File types
- Formations of the Court
- Frequencies
- Honorific
- Human sexes
- Interinstitutional procedures
- Label types

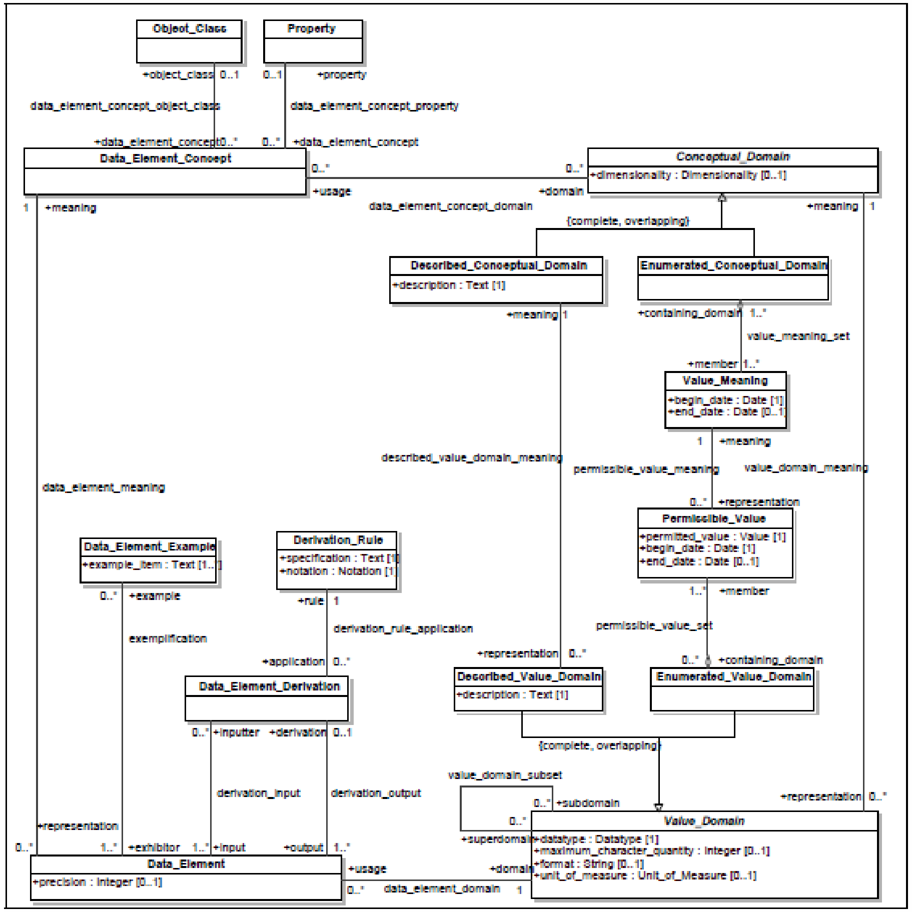
ISO / IEC11179 Metadata Registry Standard

- To help ensure consistency The International Organization for Standardization (ISO) has published standards for a metadata registry The ISO/IEC11179 Metadata Standard for data exchange
  - It provides for the attributes of data elements and associated metadata to be specified and registered as metadata items in a metadata registry. (i.e. Metamodel)
  - This provides a common way for organizations to share information via metadata registries
- This is a multi-part standards that specifies the following:
  - Part 1: Framework
  - Part 2: Classification
  - Part 3: Registry metamodel and basic attributes
  - Part 4: Formulation of data definitions
  - Part 5: Naming and identification principles
  - Part 6: Registration
  - Part 7: Datasets

Sample ISO Metamodel

- The following are examples of the Data Description metamodel in the ISO/IEC11179-3 Metadata Standard.



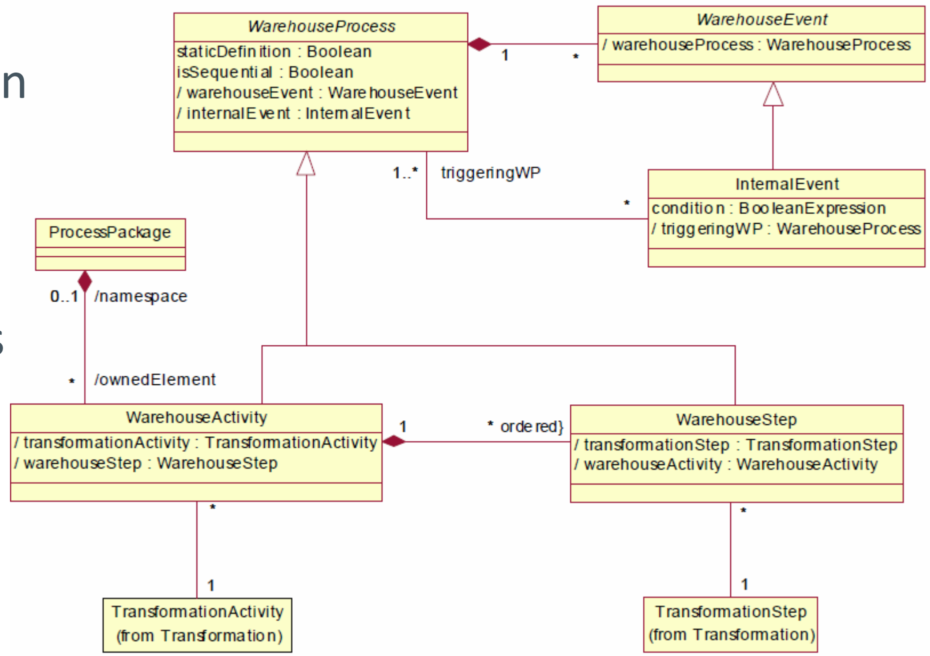


PROVIDING A COMMON INTERCHANGE

- Metadata Standards provide a common way of sharing & integrating metadata from various sources
- There are a number of Industry Standards focusing on various subject areas. For example:
  - Data Warehousing: Common Warehouse Metamodel (CWM)
  - Open Data: Common Core Metadata Schema
  - Publications, Media, Library Info: Dublin Core
  - IoT Data: A standard does not currently exist, although there is much discussion in the industry
  - Geospatial Data: ISO 19115 standards exist for sharing geospatial metadata internationally
  - Etc. – There are many more for specific data subject areas
- In addition, a number of tools provide their own metamodels & standards. While the downside is their proprietary nature, the benefit is that they often store a wider array of metadata properties.
  - Metadata Repositories
  - Data Modeling Tools
  - Etc.

Data Warehousing - CWM

- The Object Management Group (OMG) has defined the Common Warehouse Metamodel (CWM) to share metadata regarding common warehouse object.
- There are a number of packages including:
  - Relational
  - Multidimensional
  - XML
  - OLAP
  - Data Mining
  - Warehouse Process
  - Warehouse Operation
  - Etc.

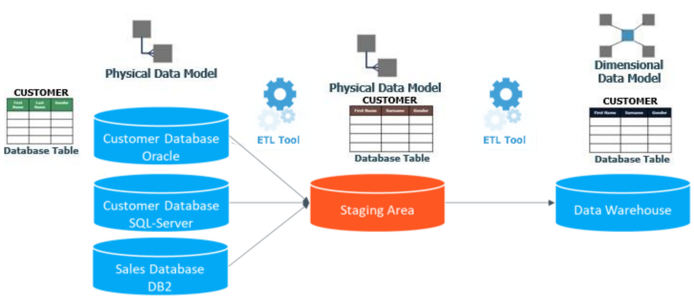


OMG's Common Warehouse Metamodel (CWM) Warehouse Process Model

Industry Standard vs. Tool Specific Metamodels

- An industry standard metamodel has the benefit of providing common interchange between a variety of tools & sources.
- A tool-specific model has the limitation of being proprietary to a vendor, but often provides a wider array of information about that source.

Industry Standard –
Breadth of Sources & Reuse



- Table Name
- Columns
- Column Names
- Data Type
- Description

Tool-Specific –
Depth & Detail

- Table Name
- Columns
- Column Names
- Data Type
- Nullability
- Primary Key
- Domain
- Description
- Display Color
- Subject Area
- Etc.

Open Data Standards – Project Open Data

- The US Government provides standards for Open Data sharing through a published Metadata Schema

Project Open Data

Project Open Data Metadata Schema

"Common Core" Required Fields

The following "common core" fields are required, to be used to describe each entry:

(Consult the 'Further Metadata Field Guidance' section lower in the page to learn more about the use of each element, including the range of valid entries where appropriate. Consult the schema maps to find the equivalent DCAT, Schema.org, and CKAN fields.)

Field	Label	Definition
title	Title	Human-readable name of the asset. Should be in plain English and include sufficient detail to facilitate search and discovery.
description	Description	Human-readable description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the asset is of interest.
keyword	Tags	Tags (or keywords) help users discover your dataset; please include terms that would be used by technical and non-technical users.
modified	Last Update	Most recent date on which the dataset was changed, updated or modified.
publisher	Publisher	The publishing entity.
contactPoint	Contact Name	Contact person's name for the asset.
mbox	Contact Email	Contact person's email address.
identifier	Unique Identifier	A unique identifier for the dataset or API as maintained within an Agency catalog or database.
accessLevel	Public Access Level	The degree to which this dataset **could** be made publicly-available, *regardless of whether it has been made available.* Choices: public (Data asset is or could be made publicly available to all without restrictions), restricted public (Data asset is available under certain use restrictions), or non-public (Data asset is not available to members of the public)

Sample JSON Metadata File for Open Data

- SAMPLE JSON METADATA FILE FOR OPEN DATA

{"@type": "dcat:Dataset", "_id":
{"$oid": "55942a79c63a7fe59b497552"}, "accessLevel": "public", "accrualPeriodicity":
"irregular", "bureauCode": ["026:00"], "contactPoint": {"@type": "vcard:Contact", "fn":
"William Brodt", "hasEmail": "mailto:wbrodt@nasa.gov"},
"description": "The purpose of the Aerospace Technical Facility Inventory is to facilitate the sharing of specialized capabilities within the aerospace research/engineering community primarily within NASA, but also throughout the nation and the entire world. A second use is to assist in answering questions regarding NASA capabilities for future missions or various alternative scenarios regarding mission support to help the Agency maintain the right set of assets.",
"distribution": [{"@type": "dcat:Distribution", "downloadURL":
"http://open.nasa.gov/datasets/NASA_Labs_Facilities.xlsx",
"mediaType": "application/vnd.ms-excel"}],
"identifier": "NASA-0000061", "keyword": ["Lab", "Laboratory", "Facility"], "language": ["en-US"],
"license": "http://www.usa.gov/publicdomain/label/1.0/",
"modified": "2014-06-05",
"programCode": ["026:000"],
"publisher": {"@type": "org:Organization", "name": "National Aeronautics and Space Administration"},
"references": ["https://nrpi.hq.nasa.gov/ATFI/", https://nrpi.hq.nasa.gov/ATFI/URLLinks.cfm"],
"spatial": "United States",
"title": "Agency Data on User Facilities"}

DUBLIN CORE METADATA INITIATIVE

METADATA
Dublin Core® Metadata Initiative
INNOVATION

- The Dublin Core Metadata Initiative provides a common metadata standards for resources such as media, library books, etc.
- It defines standards for information such as:

Title Creator
Subject
Description
Publisher
Contributor
Date
Type

Format
Identifier
Source
Language
Relation
Coverage
Rights

- Resources can be described using:
  - Text
  - HTML
  - XML
  - RDF XML

Sample Metadata

Format="video/mpeg; 5 minutes"
Language="en"
Publisher="Kats Online, LLC"
Title="My Favorite Cat Video"
Subject="Cats"
Description="A short video of a black cat playing with string."

Schema.org

- **Schema.org** is a vocabulary that webmasters can use to mark-up Web pages for the Semantic Web, so that search engines understand what the pages are about .
  - Created by a group of search providers (e.g. Google, Microsoft, Yahoo and Yandex).
  - Vocabularies are developed by an open community process
    - Through GitHub (https://github.com/schemaorg/schemaorg)
    - Using the public-schemaorg@w3.org mailing list
- The schemas are a set of 'types', each associated with a set of properties. The types are arranged in a hierarchy. There are currently over 570 types, including:
  - Creative works
  - Organization
  - Person
  - Place, LocalBusiness, Restaurant
  - Product, Offer, AggregateOffer
  - Etc.
- There are also extensions for particular industries such as:
  - auto.schema.org
  - health-lifesci.schema.org

Schema.org Schema

Schema.org core schema

This is the RDFa representation of the schema.org schema, the underlying representation of the schema.org vocabulary.

It is represented in a form based on W3C RDF/RDFS. We encourage proposals for schema.org improvements to be expressed in this same style. For Discussion please use the W3C Web schemas group.

See datamodel for more details.

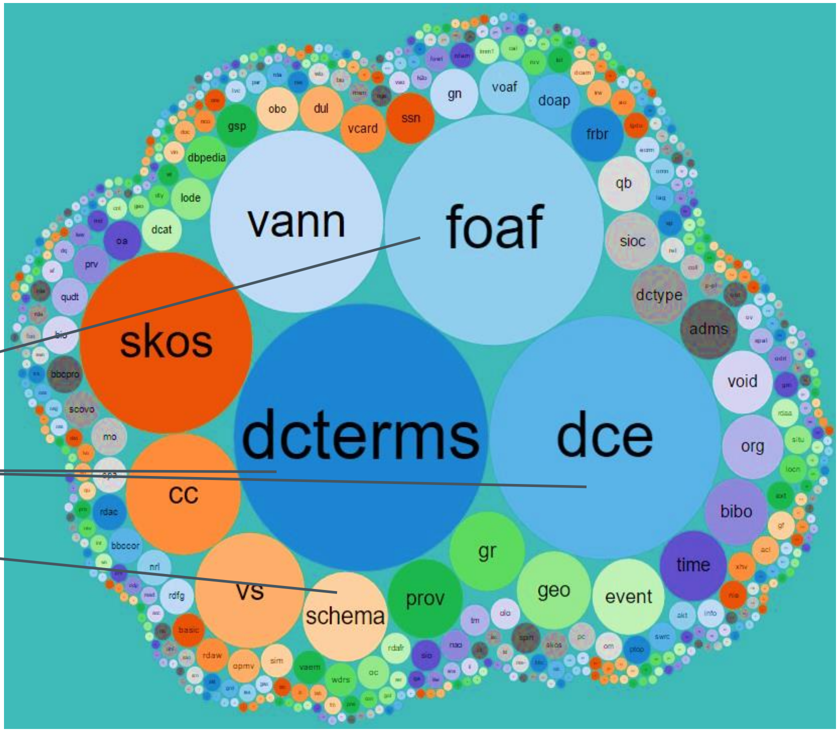
Note: the style of RDFa used here may change in the future. To see the substantive content of the schema, view the HTML source markup. We use a simple subset of RDFa for syntax, including prefixes that are declared in the RDFa initial context.

Thing
The most generic type of item.
CreativeWork
The most generic kind of creative work, including books, movies, photographs, software programs, etc.
Subclass of: Thing
Source: rNews
WebPage
A web page. Every web page is implicitly assumed to be declared to be of type WebPage, so the various properties about that webpage, such as <code>breadcrumb</code> may be used. We recommend explicit declaration if these properties are specified, but if they are found outside of an itemscope, they will be assumed to be about the page.
Subclass of: CreativeWork
AboutPage
Web page type: About page.
Subclass of: WebPage
Organization
An organization such as a school, NGO, corporation, club, etc.
Subclass of: Thing
Place
Entities that have a somewhat fixed, physical extension.
Subclass of: Thing
LocalBusiness
A particular physical business or branch of an organization. Examples of LocalBusiness include a restaurant, a particular branch of a restaurant chain, a branch of a bank, a medical practice, a club, a bowling alley, etc.
Subclass of: Organization
Subclass of: Place
MedicalOrganization
A medical organization (physical or not), such as hospital, institution or clinic.
Subclass of: Organization
Dentist
A dentist.
Subclass of: MedicalOrganization
Subclass of: schema:ProfessionalService

There are many other Common Schemas & Vocabularies

- The Dublin Core and Schema.org are two popular schemas, but many more exist for particular subject areas, industries, etc.
- The Linked Open Vocabularies site (LOV) provides a helpful listing

551 Vocabularies in LOV



Friend of a Friend

Dublin Core

Schema.org

Summary

- A metamodel provides a common format & structure for storing metadata
- A meta-metamodel provides a common syntax and way of expressing a metamodel
- A metadata registry provides common metadata definitions for sharing metadata between and within organizations
- Metadata standards provide common metamodels for specific subject areas, tools, and industries

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
   – Was sind Data Catalogs
   – Alation Präsentation - Überblick
   – Informatica EDC Präsentation
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
5. Metadata Strategy und Data Catalog-Einführung

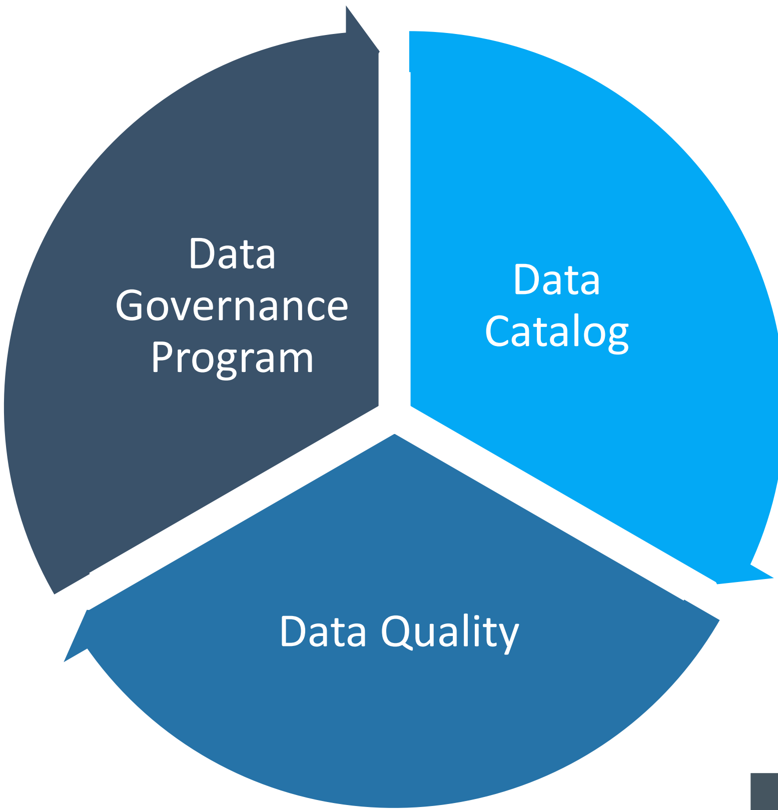
Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
   - Was sind Data Catalogs
   - Alation Präsentation - Überblick
   - Informatica EDC Präsentation
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
5. Metadata Strategy und Data Catalog-Einführung

Komponenten eines Data Governance Frameworks

Data Security Management

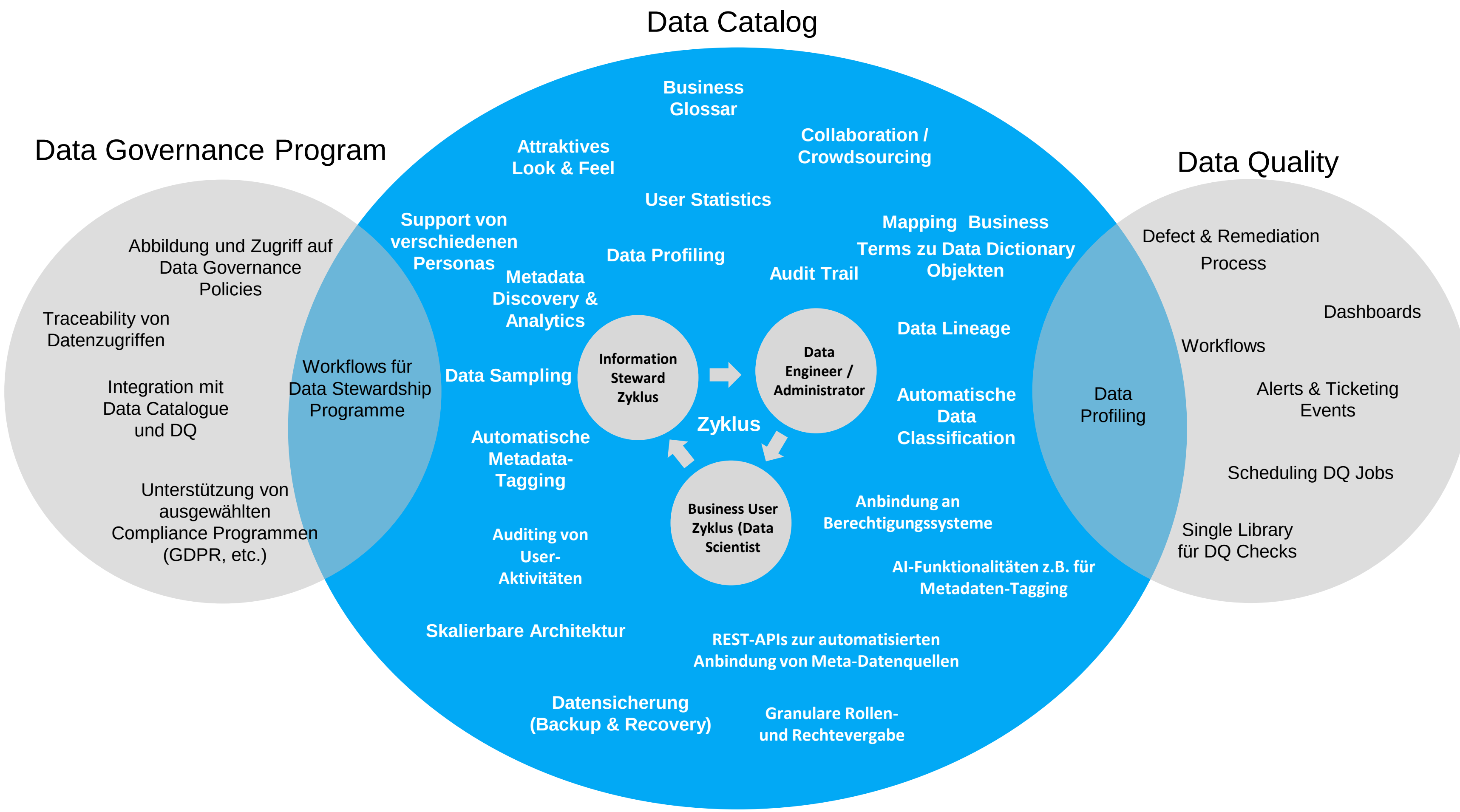
Master Data Management



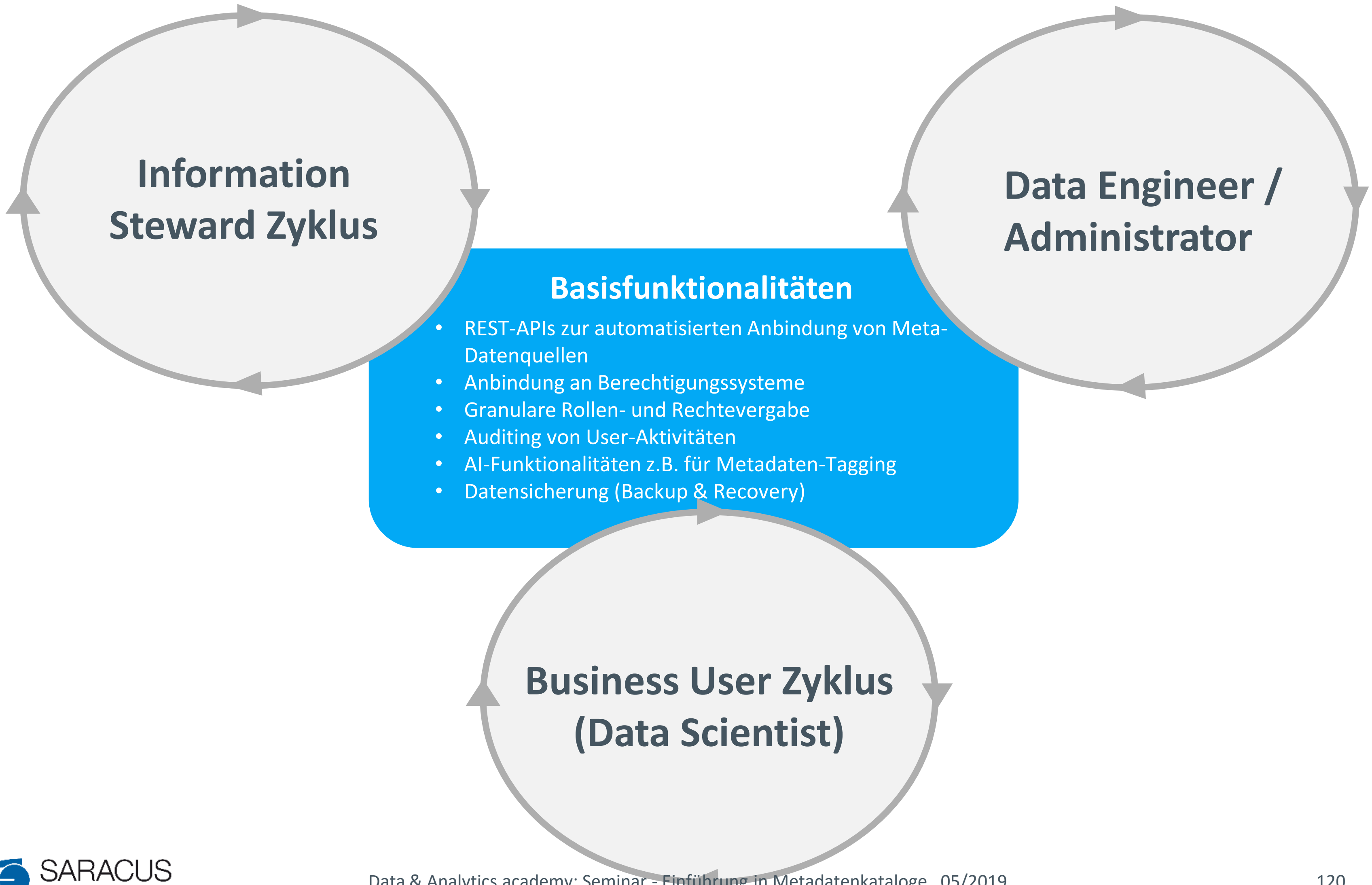
Data Management / ETL

Data Lifecycle Management

Data Catalog: Funktionalitäten
Abgrenzung zu Data Governance Program and DQ



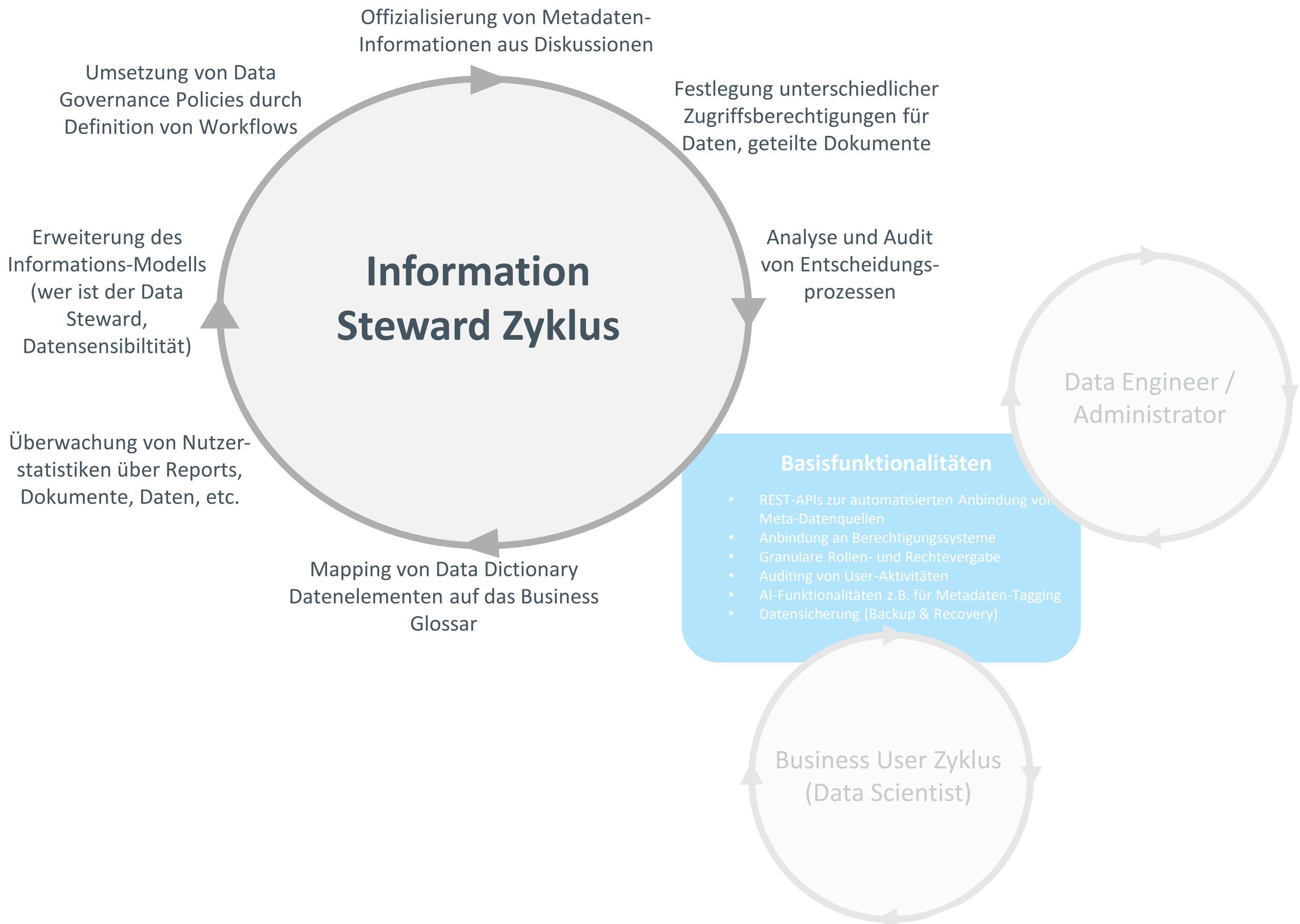
Aktivitäten nach Rollen im Datenportalkontext
Basis für Ableitung von Toolkriterien



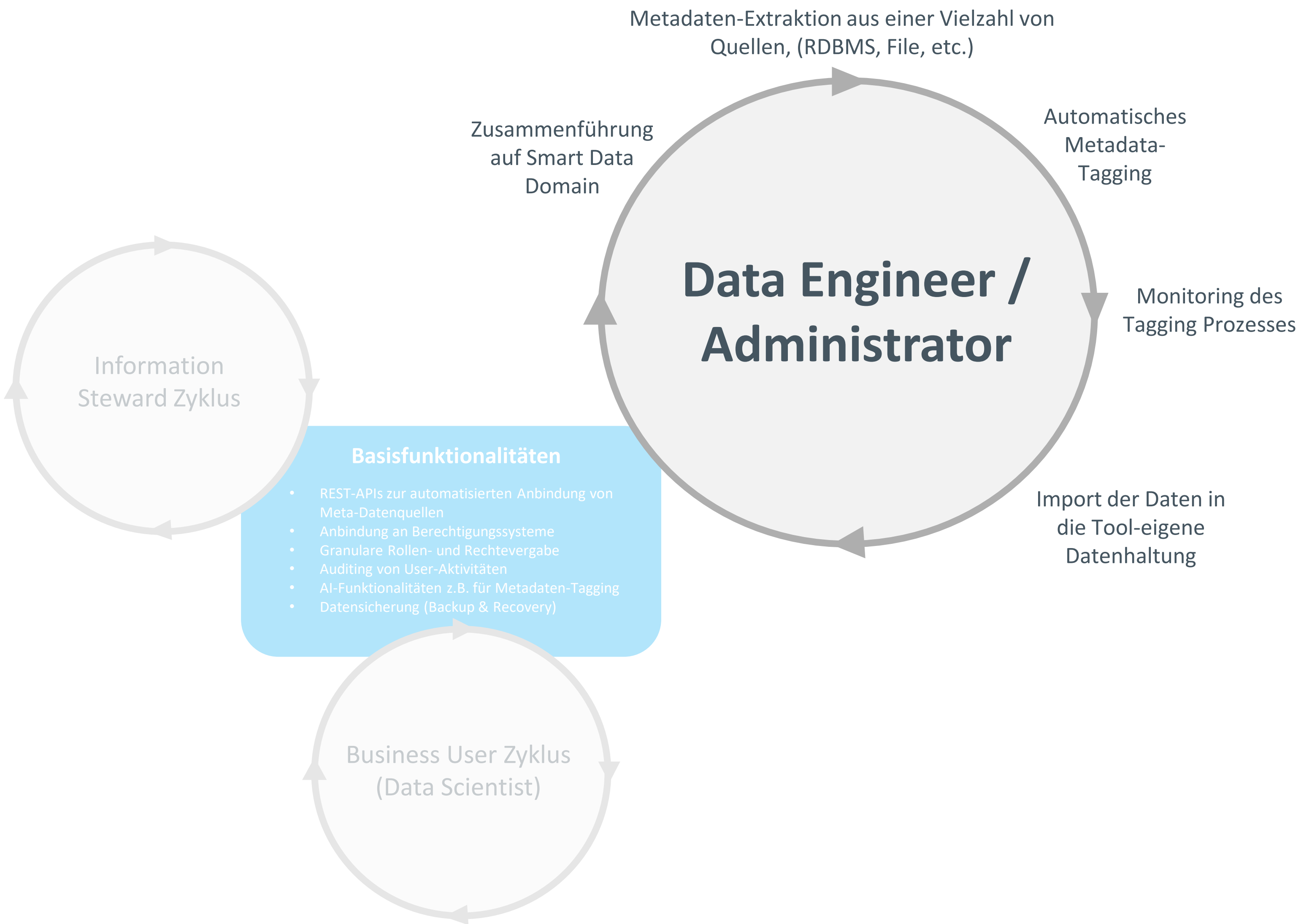
Basisfunktionalitäten

- REST-APIs zur automatisierten Anbindung von Meta-Datenquellen
- Anbindung an Berechtigungssysteme
- Granulare Rollen- und Rechtevergabe
- Auditing von User-Aktivitäten
- AI-Funktionalitäten z.B. für Metadaten-Tagging
- Datensicherung (Backup & Recovery)

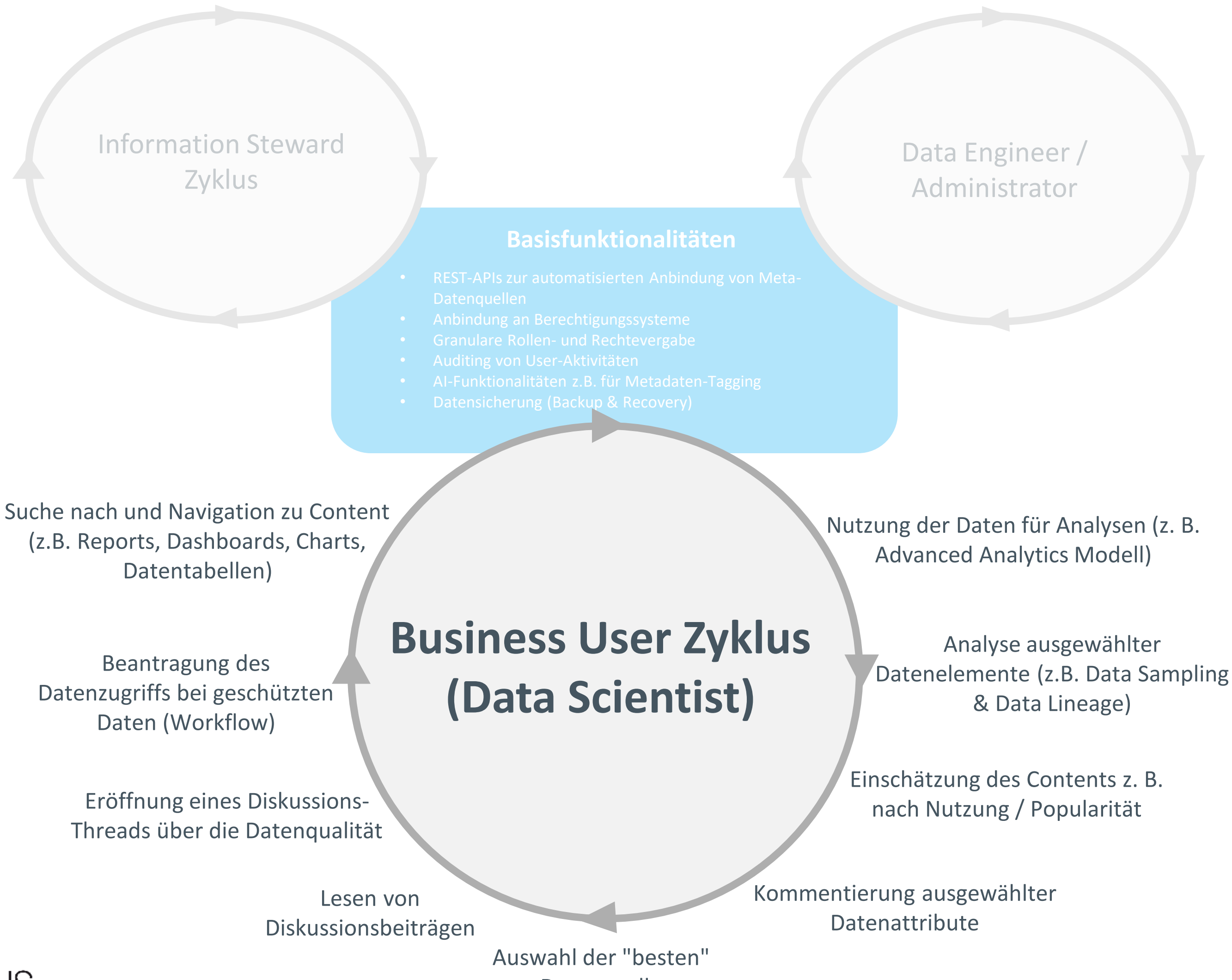
Aktivitäten nach Rollen im Datenportalkontext
Basis für Ableitung von Toolkriterien



Aktivitäten nach Rollen im Datenportalkontext
Basis für Ableitung von Toolkriterien



Aktivitäten nach Rollen im Datenportalkontext
Basis für Ableitung von Toolkriterien



Typen von Data Catalogs

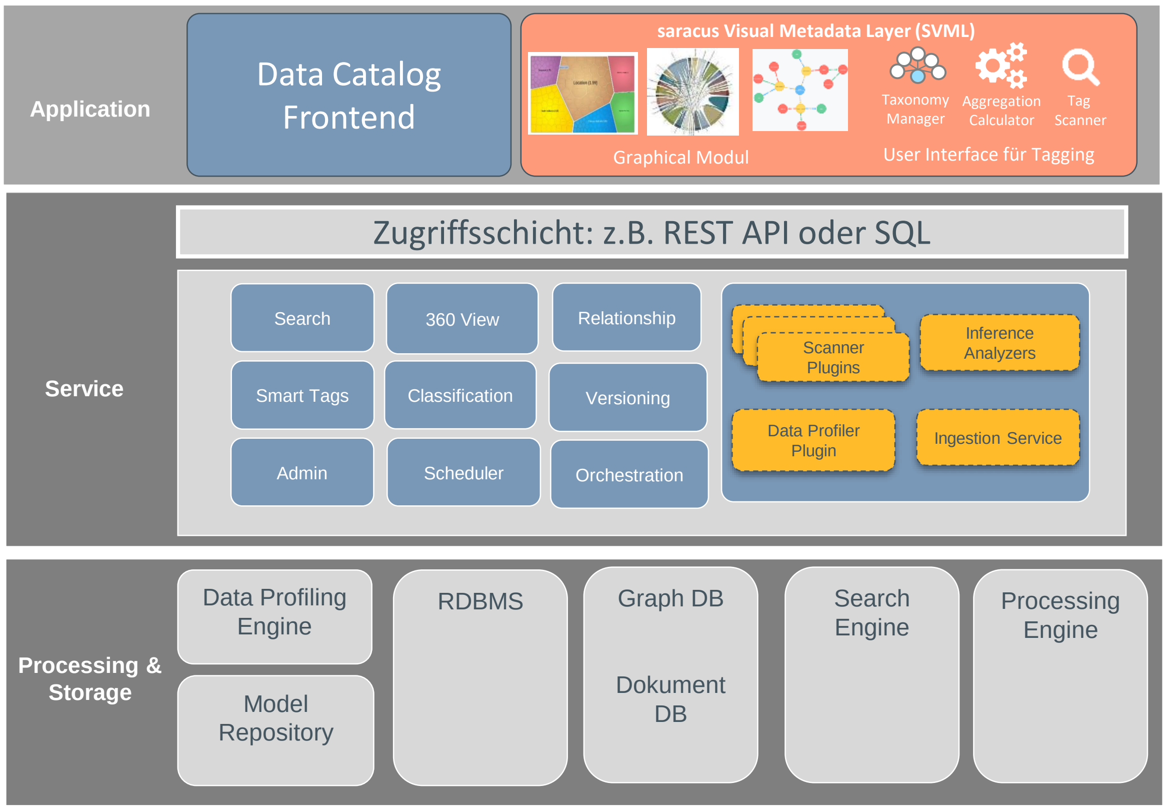
Catalog Type	Catalog Characteristics
Standalone	•Catalog of data sets and operations •Supports data set search and evaluation •Seamless user experience requires high level of interoperability
Integrated with Data Preparation	•Catalog of data sets and operations in a tool that includes extensive data preparation features and functions •Seamless user experience for finding, evaluating, and preparing data •Requires high level of interoperability with analysis tools
Integrated with Data Analysis	•Catalog of data sets in a tool that includes extensive data analysis and visualization features and functions •May catalog operations and support basic data preparation •Seamless user experience to find and analyze data •Requires high level of interoperability with data preparation tools when advanced data preparation capability is needed
Data integration tools and data lake management solutions that embed data cataloging as a feature within a broader solution	Data integration tools and data lake management solutions that embed data cataloging as a feature within a broader solution (see "Magic Quadrant for Data Integration Tools").
Fully Integrated Solution	•Catalog of data sets and operations in a tool that includes extensive features and functions for data preparation, analysis, visualization, governance, and security •Seamless user experience throughout the analytics lifecycle •Interoperability becomes important in organizations where multiple data preparation and/or analysis tools are used

So far not available

What`s In An Information Catalog?

- Metadata Repositories
- Data sources
- Business glossary
  - Disparate and trusted data names
  - Semantic Frameworks
- Data classifications
- Policies
- Rules to enforce policies
- Schemas
- Communities
- Collaborations
- Raw / In-progress / trusted data sets
- Data profiles
- Ingestion workflows
- Refinery workflows
- BI/Analytical artifacts
- Roles, e.g. data stewards, data owners, data experts, curators
- Asset membership
- Consumers data marketplace
- Data provisioning workflows
- Metadata lineage
- Impact Analysis
- …

Typical Architecture



Data Catalog Tools

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
   – Was sind Data Catalogs
   – Alation Präsentation - Überblick
   – Informatica EDC Präsentation
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
5. Metadata Strategy und Data Catalog-Einführung

SARACUS CONSULTING

Demo: Alation

Alation Summary

Main features:

- Metadata Catalog coming with connectors to many different Database Systems
  - DB2, Hive, Oracle, SQLServer, Teradata, …
- Compose: SQL-client that integrates the metadata information while typing queries
- Easy to use web interface with advanced search capabilities
- Wiki like structure for documenting business metadata that can be linked to technical objects
- Strong focus on crowd sourcing and collaboration features

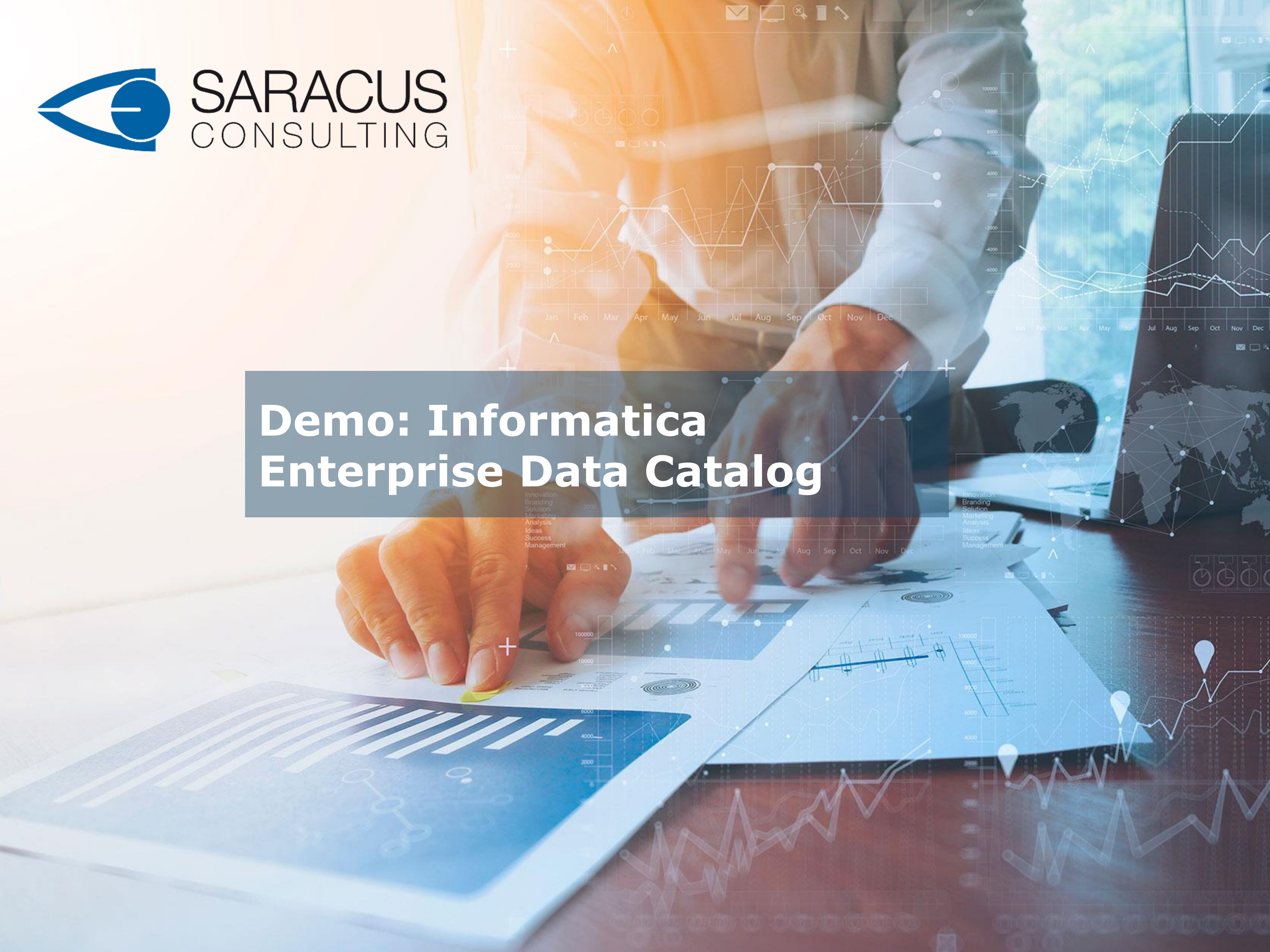
Technical aspects

- Metadata storage is based on RDBMS and Search engine
- REST-API for Metadata Extraction and editing

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
   - Was sind Data Catalogs
   - Alation Präsentation - Überblick
   - Informatica EDC Präsentation
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
5. Metadata Strategy und Data Catalog-Einführung

SARACUS
CONSULTING



Demo: Informatica Enterprise Data Catalog

EDC Summary

Main features:

- Strong AI features for Data Classification
- Good data profiling feature
- Integrates Data Governance Tool (Axon) and Self Service Tool (EDL) from the Informatica Suite
- Strong Lineage capabilities (but requires Informatica Powercenter)

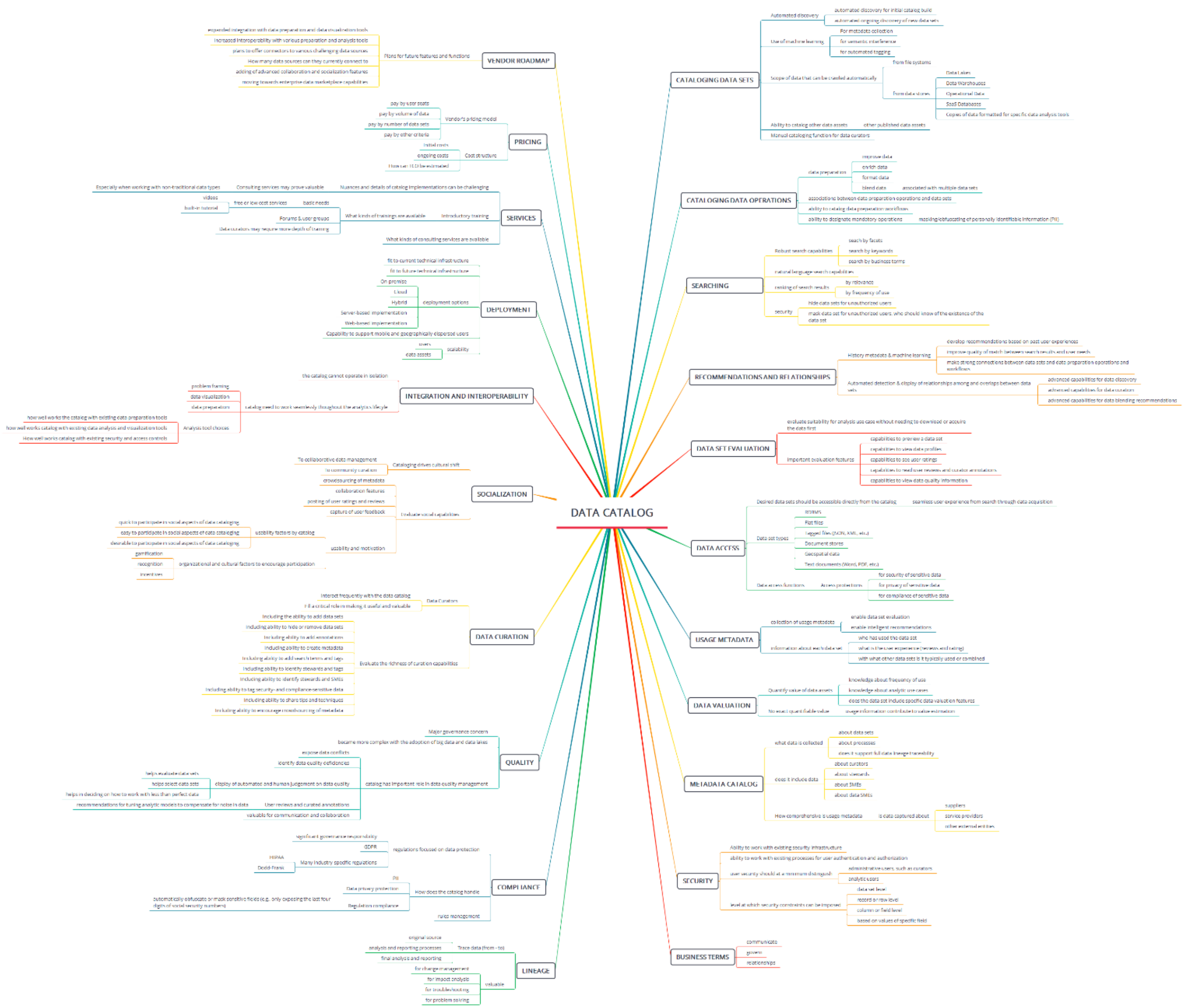
Technical aspects

- Metadata storage is based on RDBMS Big Data Systems (Solr, Hbase, …)
- REST-API for Metadata Extraction and editing (well documented)
- Java API

Agenda

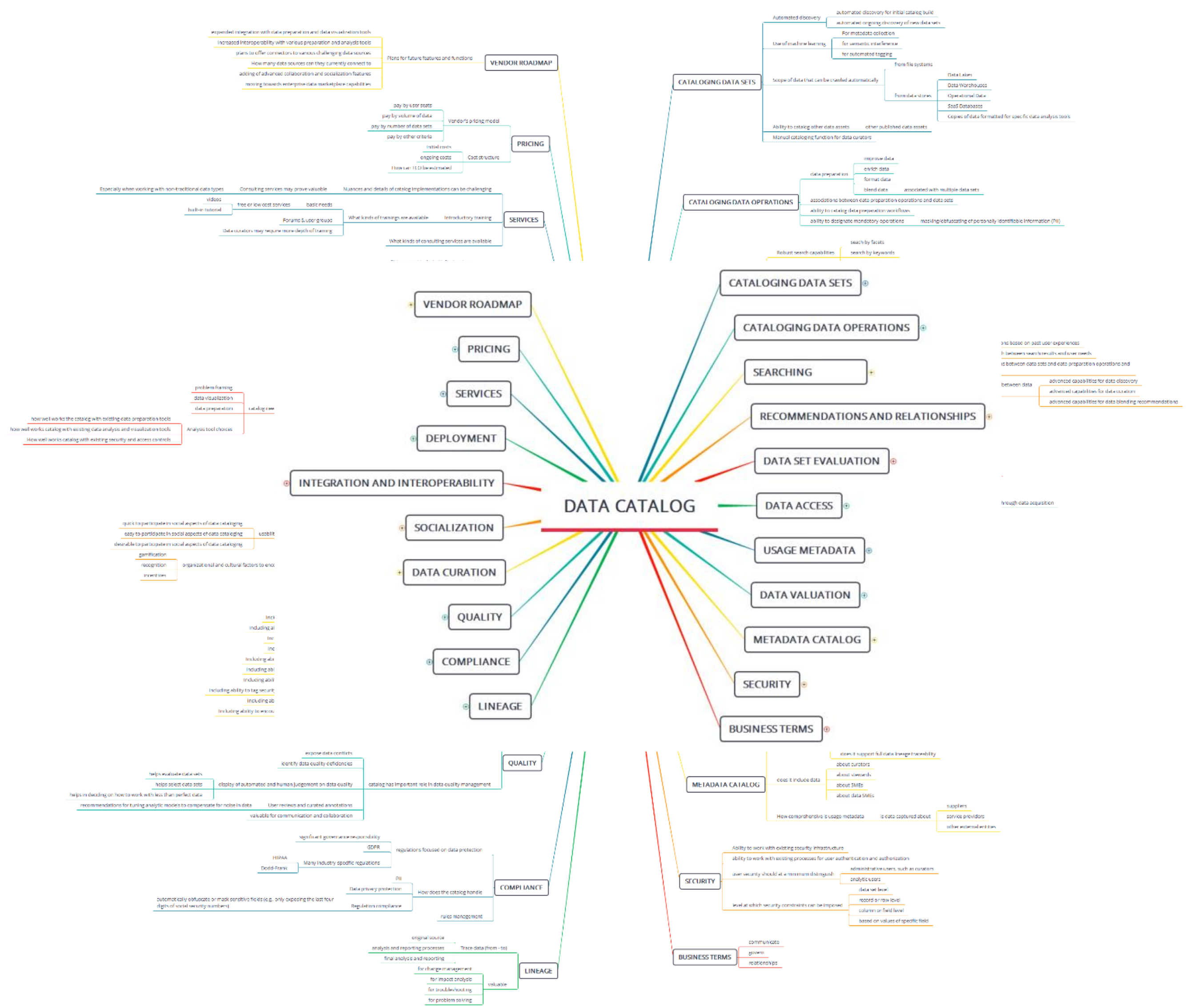
1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
   - Cataloging Data Sets
   - Searching
   - Data Set Evaluations
   - Data Access
   - Usage Metadata
   - Recommendations
   - Compliance
   - Lineage
   - Integration
   - zusätzliche Toolkriterien
4. Data Catalogs: Ausgewählte Themen
5. Metadata Strategy und Data Catalog-Einführung

Übersicht



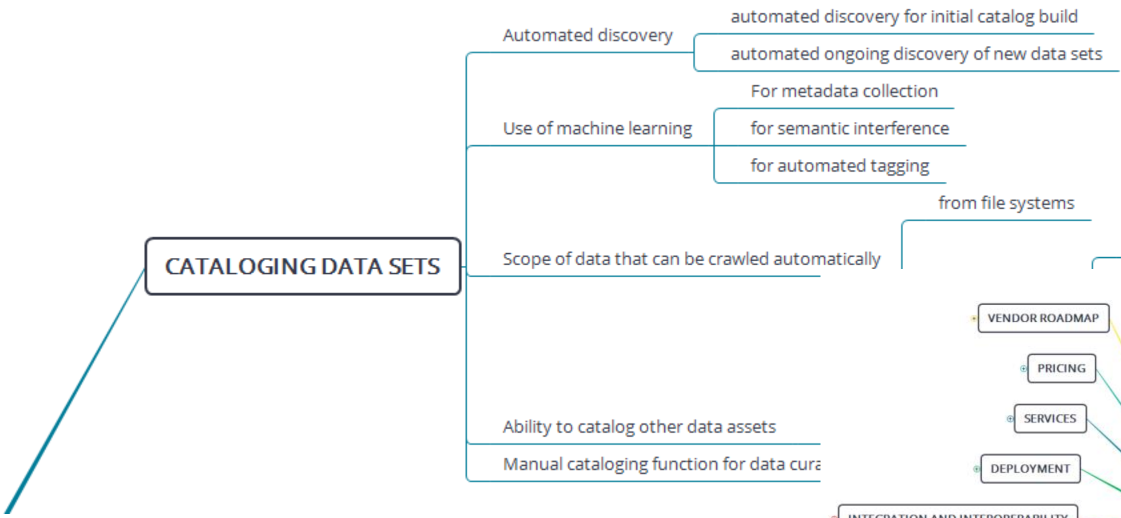
DATA CATALOG
VENDOR ROADMAP
PRICING
SERVICES
DEPLOYMENT
INTEGRATION AND INTEROPERABILITY
SOCIALIZATION
DATA CURATION
QUALITY
COMPLIANCE
LINEAGE
CATALOGING DATA SETS
CATALOGING DATA OPERATIONS
SEARCHING
RECOMMENDATIONS AND RELATIONSHIPS
DATA SET EVALUATION
DATA ACCESS
USAGE METADATA
DATA VALUATION
METADATA CATALOG
SECURITY
BUSINESS TERMS

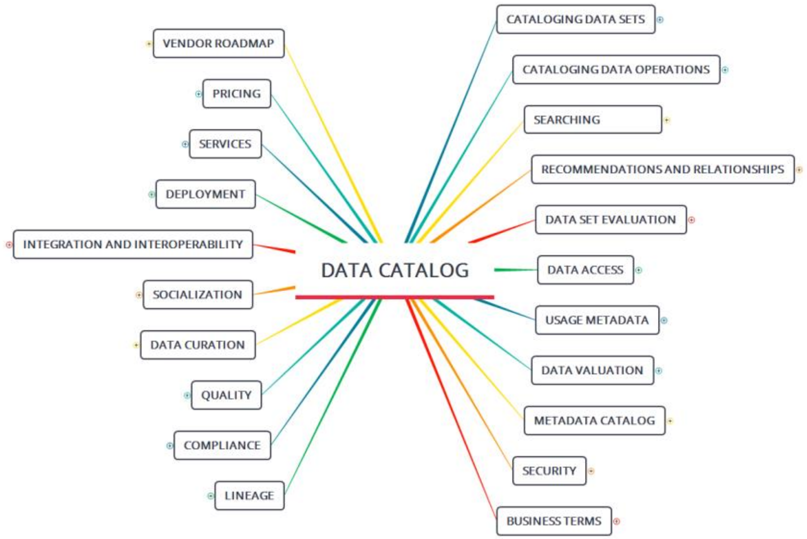
Übersicht



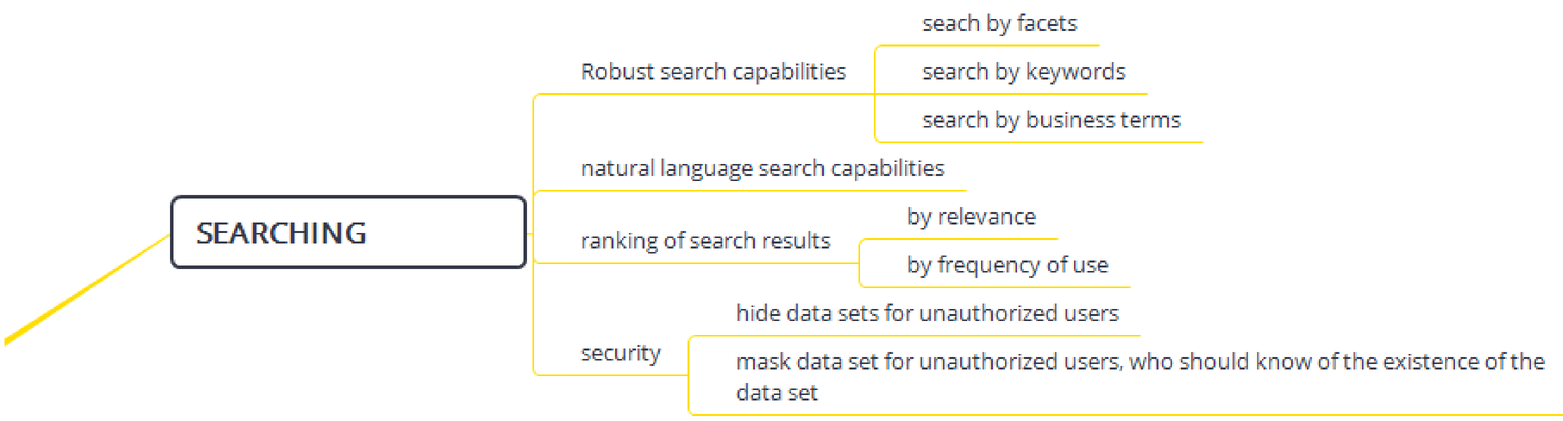
DATA CATALOG
VENDOR ROADMAP
PRICING
SERVICES
DEPLOYMENT
INTEGRATION AND INTEROPERABILITY
SOCIALIZATION
DATA CURATION
QUALITY
COMPLIANCE
LINEAGE
CATALOGING DATA SETS
CATALOGING DATA OPERATIONS
SEARCHING
RECOMMENDATIONS AND RELATIONSHIPS
DATA SET EVALUATION
DATA ACCESS
USAGE METADATA
DATA VALUATION
METADATA CATALOG
SECURITY
BUSINESS TERMS

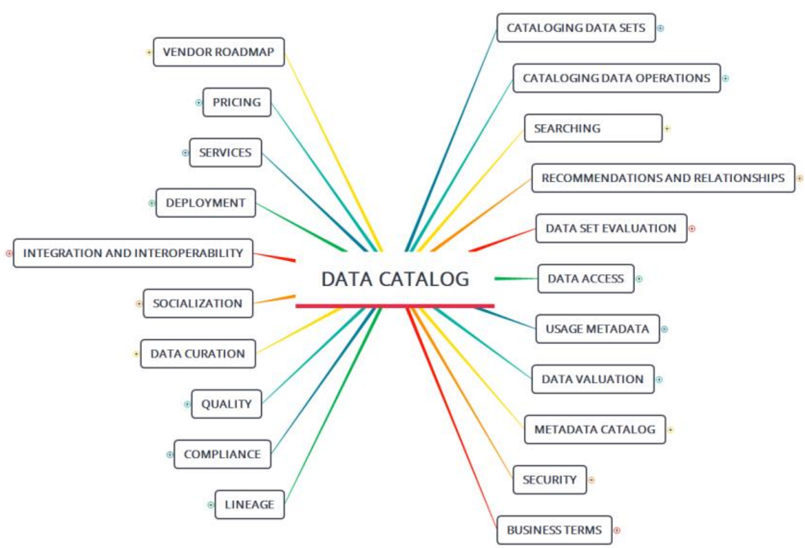
Übersicht: Cataloging Data Sets





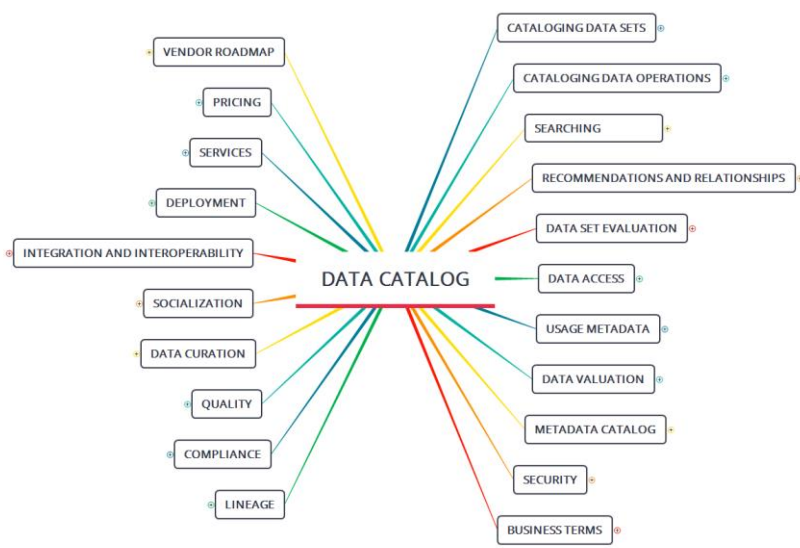
Übersicht: Searching





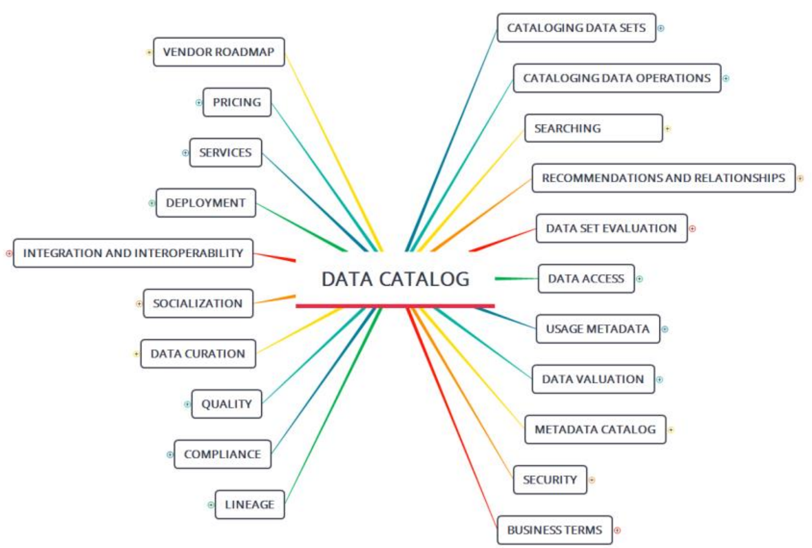
Übersicht: Data Set Evaluation

- DATA SET EVALUATION
  - evaluate suitability for analysis use case without needing to download or acquire the data first
  - important evaluation features
    - capabilities to preview a data set
    - capabilities to view data profiles
    - capabilities to see user ratings
    - capabilities to read user reviews and curator annotations
    - capabilities to view data quality information

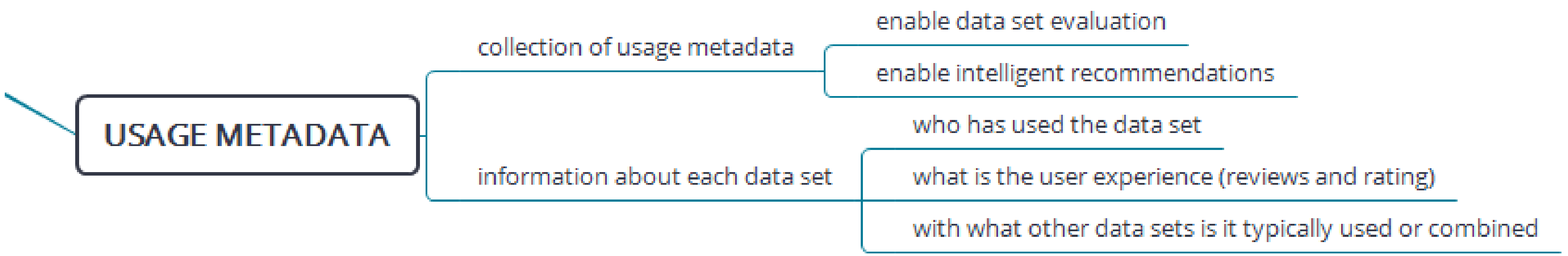


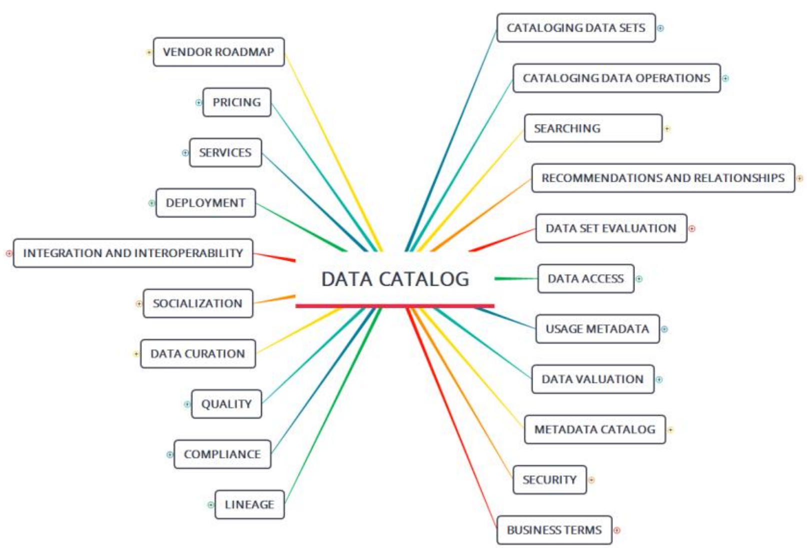
Übersicht: Data Access

- DATA ACCESS
  - Desired data sets should be accessible directly from the catalog
    - seamless user experience from search through data acquisition
  - Data set types
    - RDBMS
    - Flat files
    - Tagged files (JSON, XML, etc.)
    - Document stores
    - Geospatial data
    - Text documents (Word, PDF, etc.)
  - Data access functions
    - Access protections
      - for security of sensitive data
      - for privacy of sensitive data
      - for compliance of sensitive data

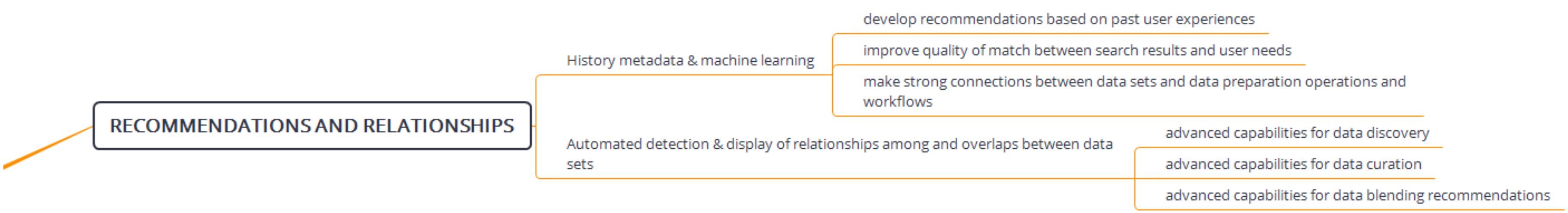


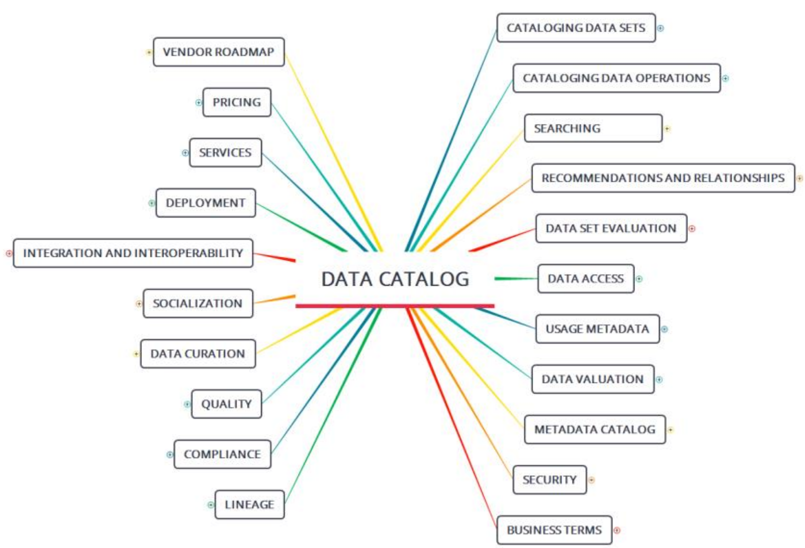
Übersicht: Usage Metadata



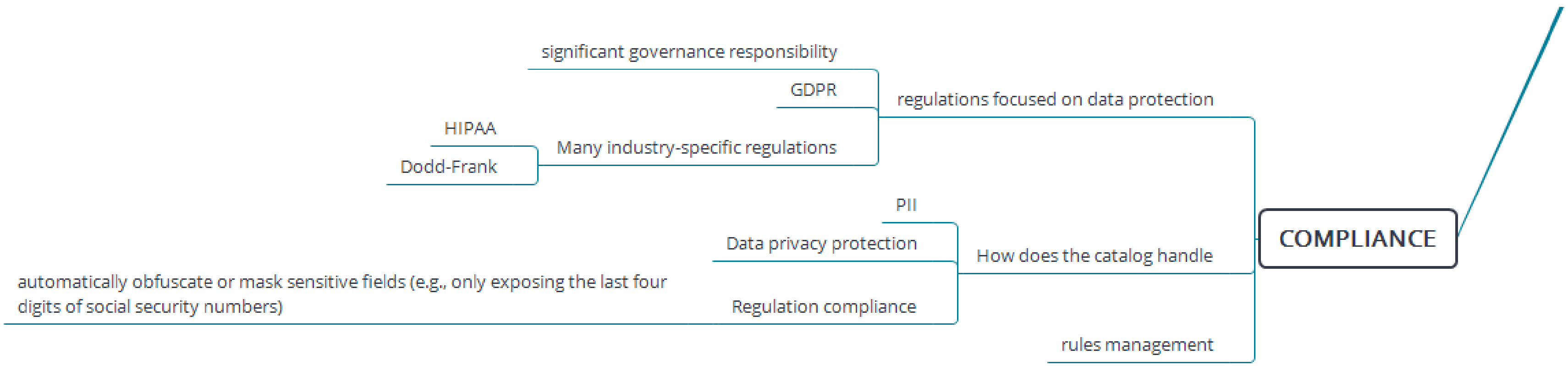


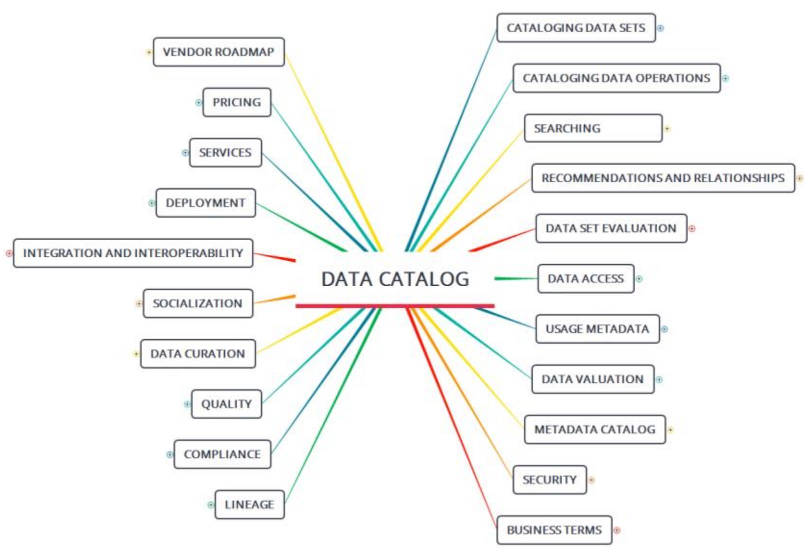
Übersicht: Recommendations and Relationships



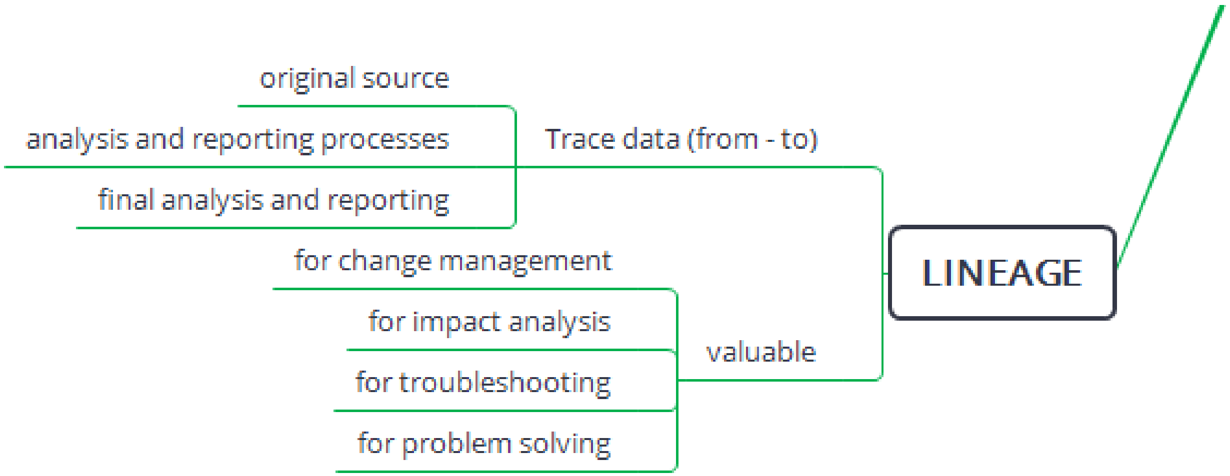


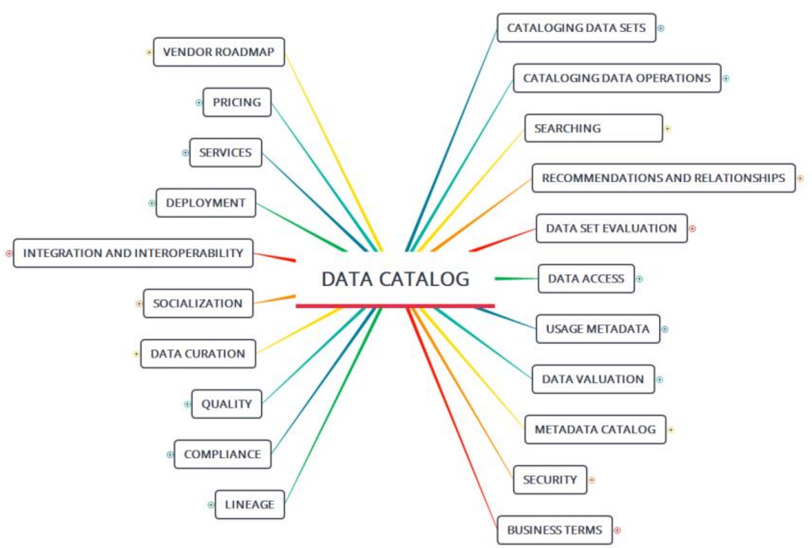
Übersicht: Compliance



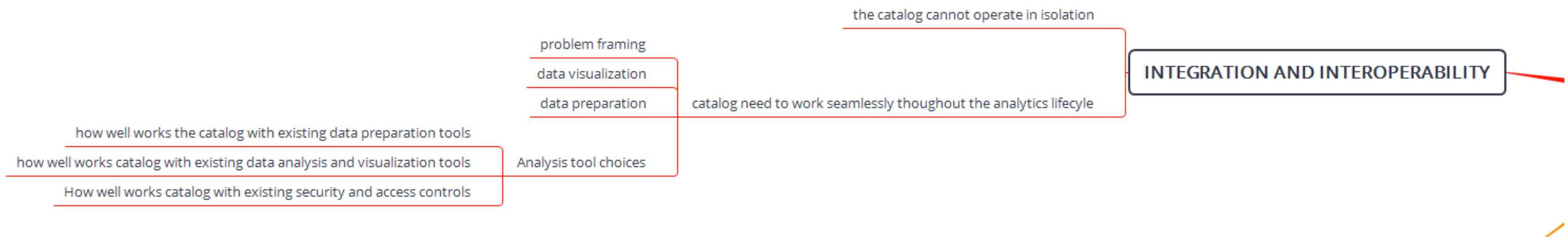


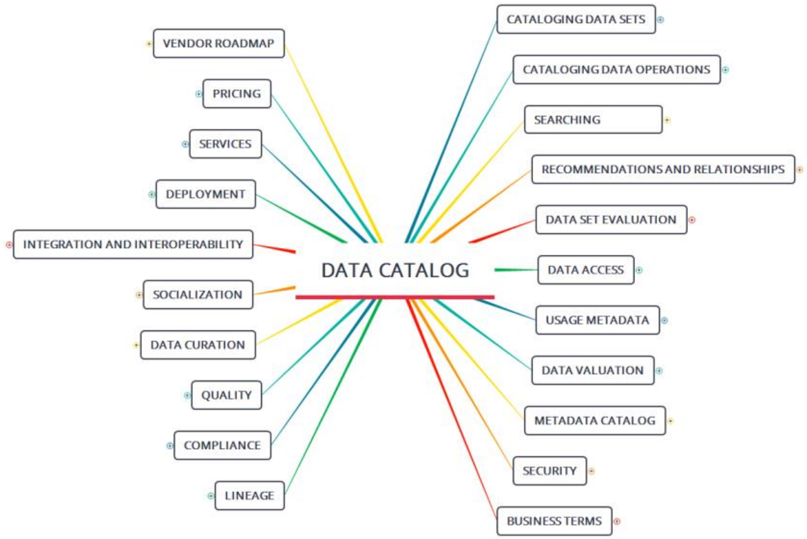
Übersicht: Lineage



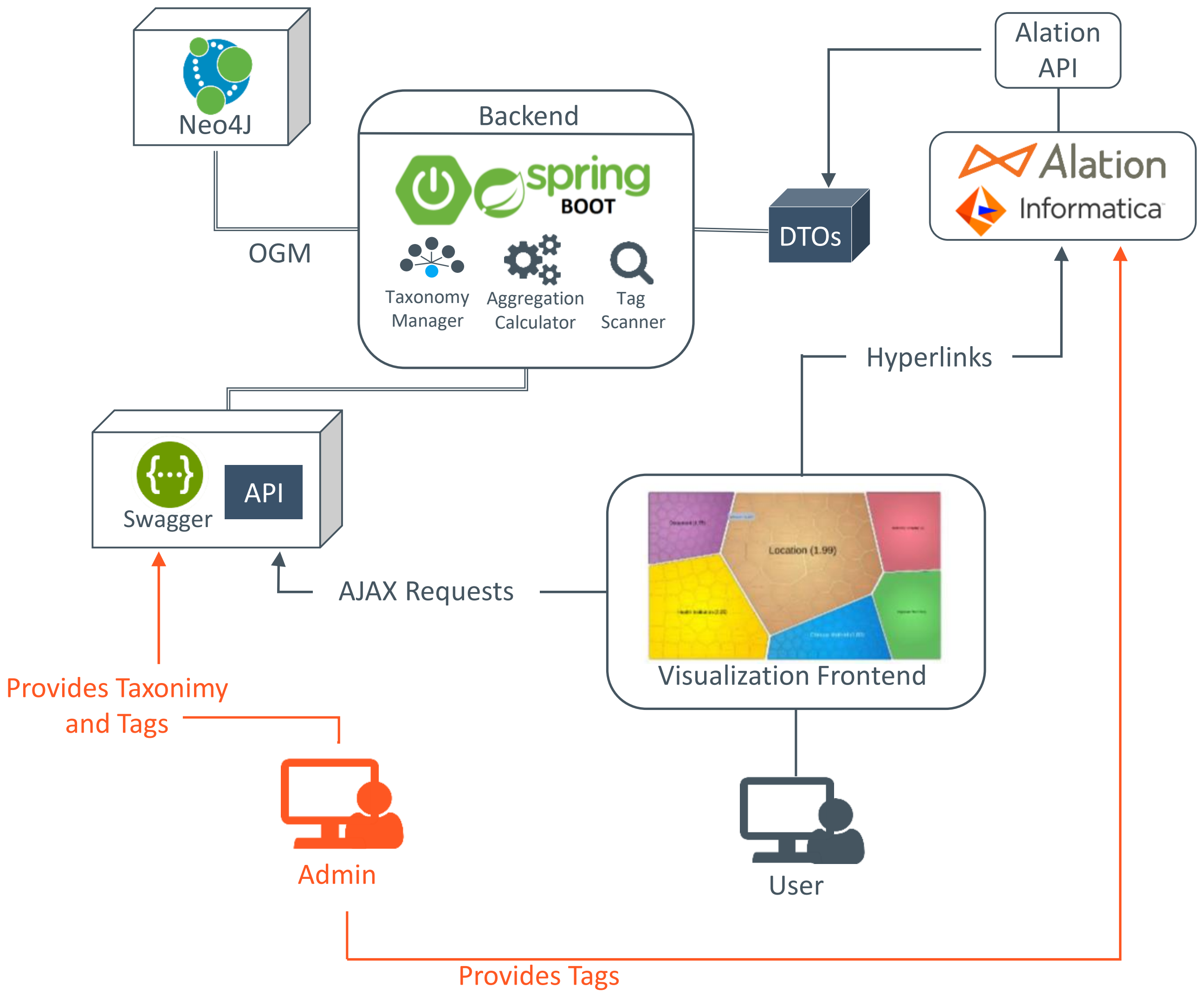


Übersicht: Integration and Interoperability





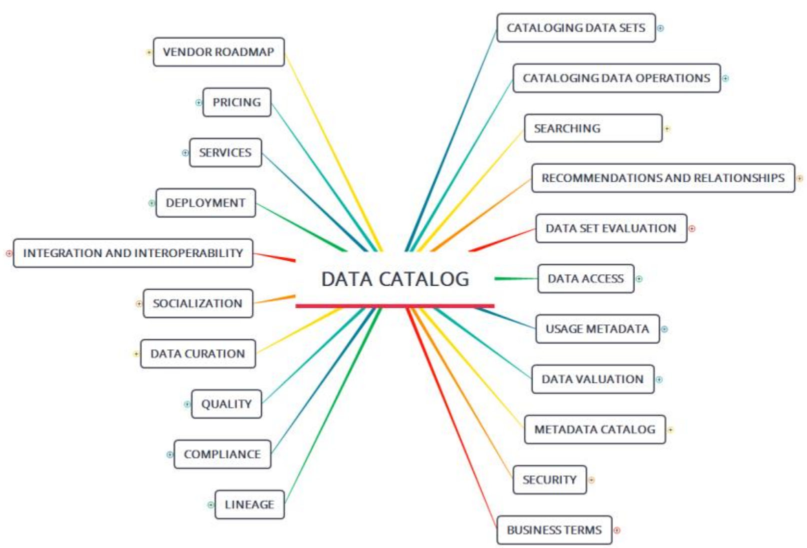
Example: Data Catalog Integration via Rest API



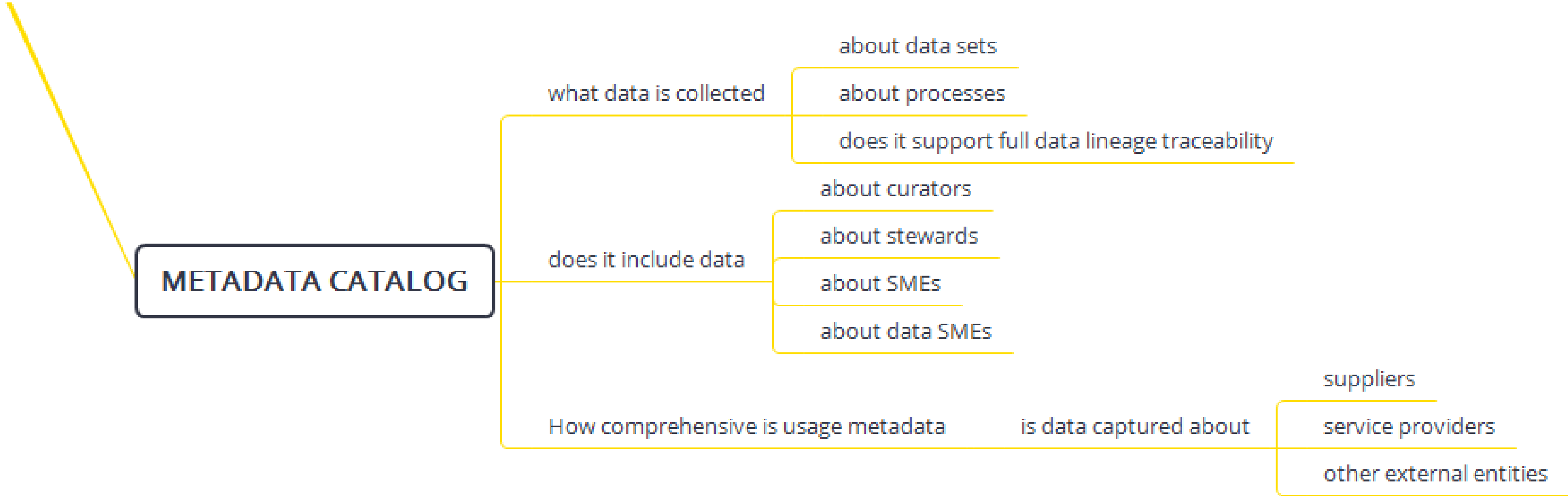
Übersicht: Data Valuation

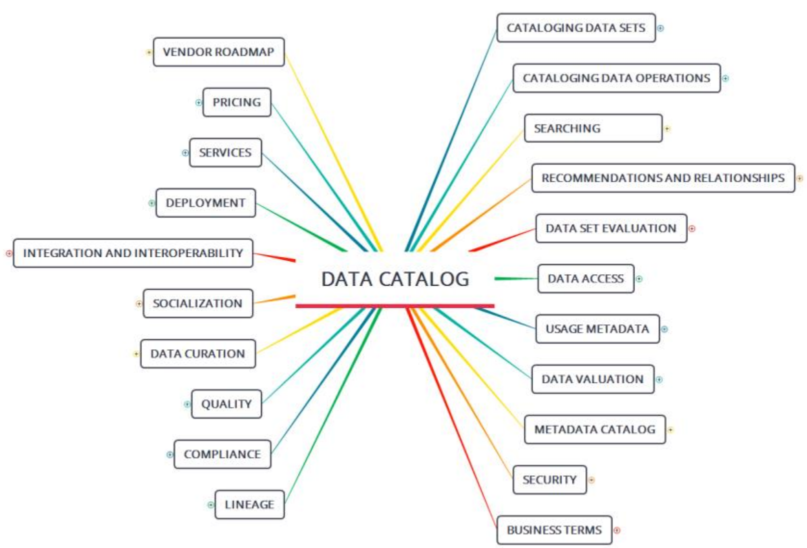
DATA VALUATION

- Quantify value of data assets
  - knowledge about frequency of use
  - knowledge about analytic use cases
  - does the data set include specific data valuation features
- No exact quantifiable value
  - usage information contribute to value estimation



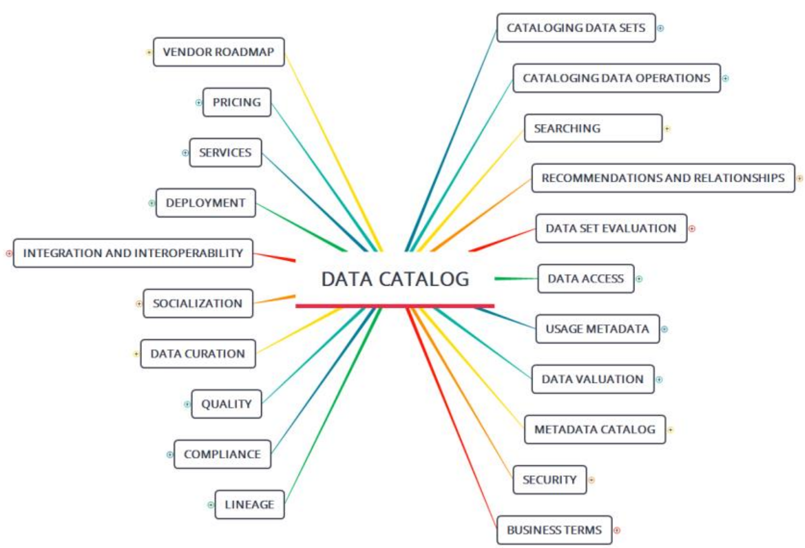
Übersicht: Metadata Catalog



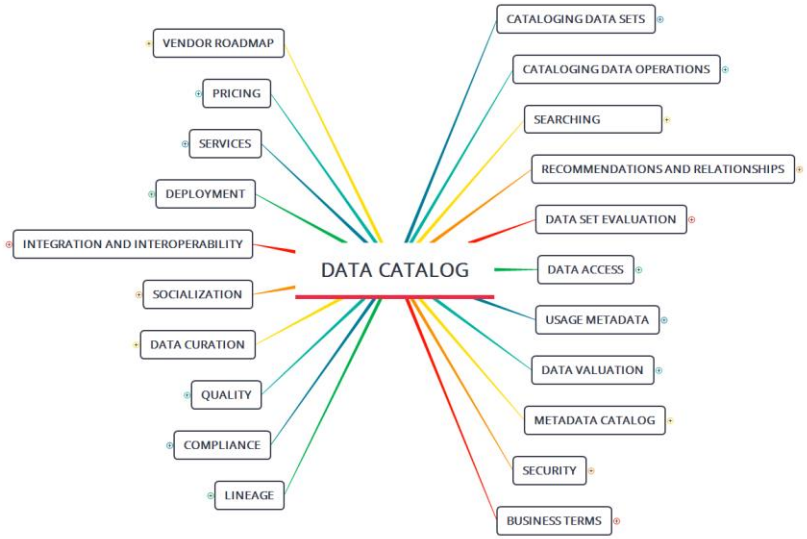


Übersicht: Security

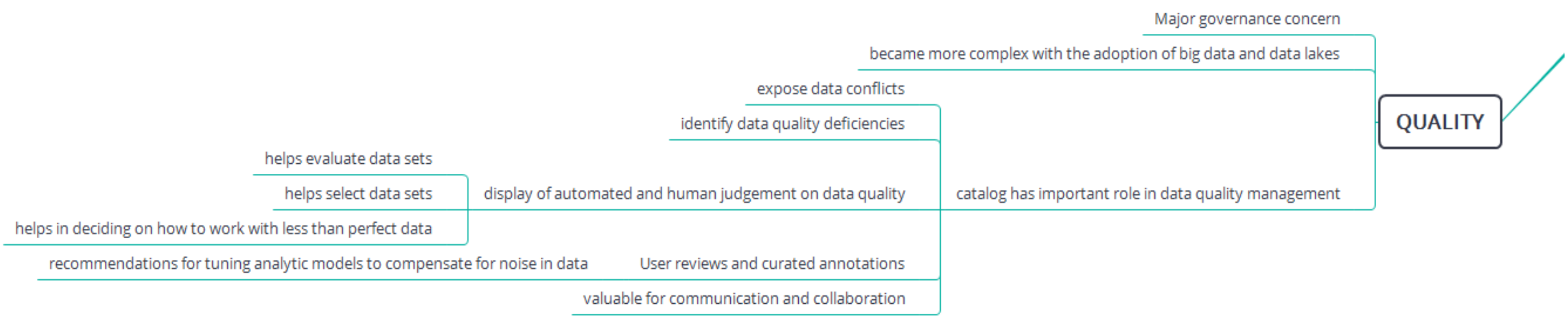
- **SECURITY**
  - Ability to work with existing security infrastructure
  - ability to work with existing processes for user authentication and authorization
  - user security should at a minimum distinguish
    - administrative users, such as curators
    - analytic users
  - level at which security constraints can be imposed
    - data set level
    - record or row level
    - column or field level
    - based on values of specific field

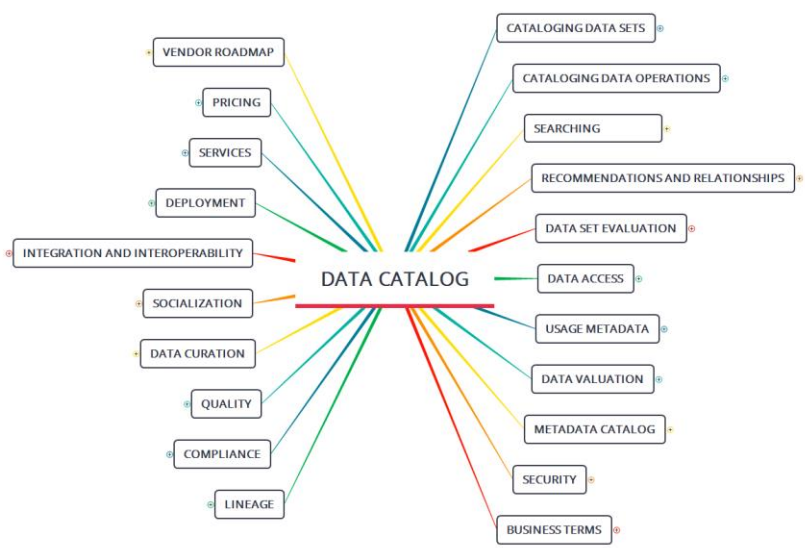


Übersicht: Business Terms

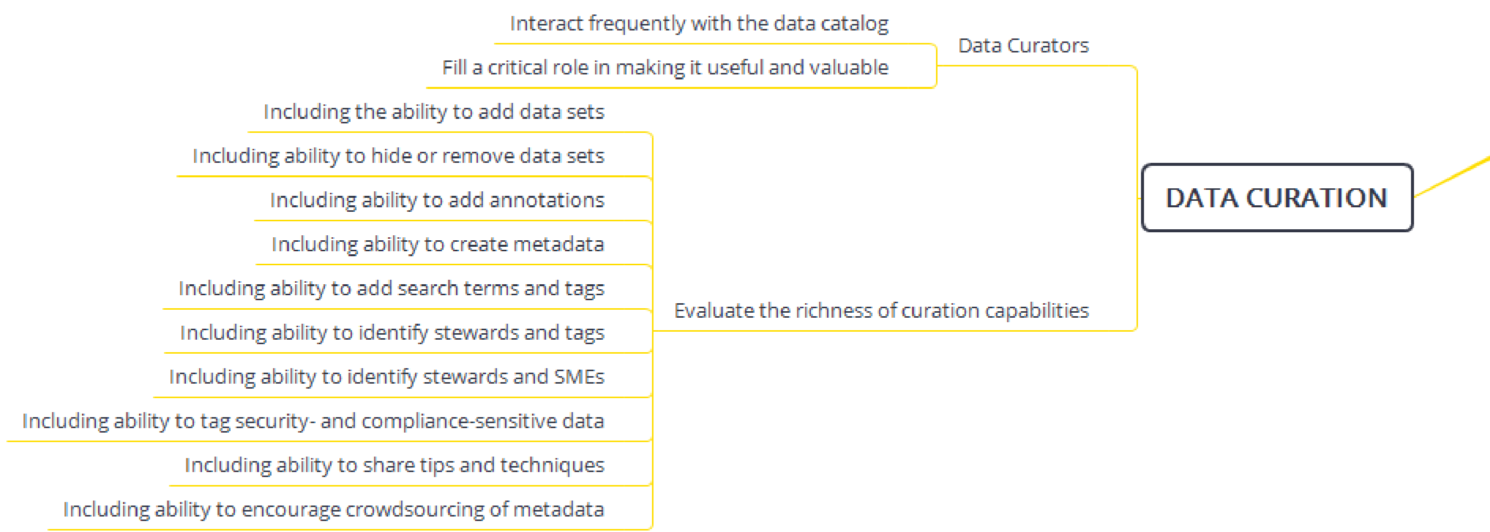


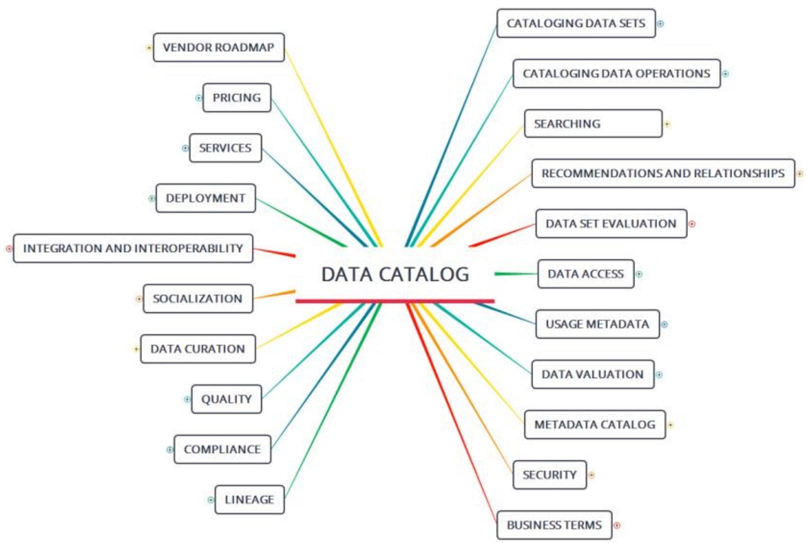
Übersicht: Quality



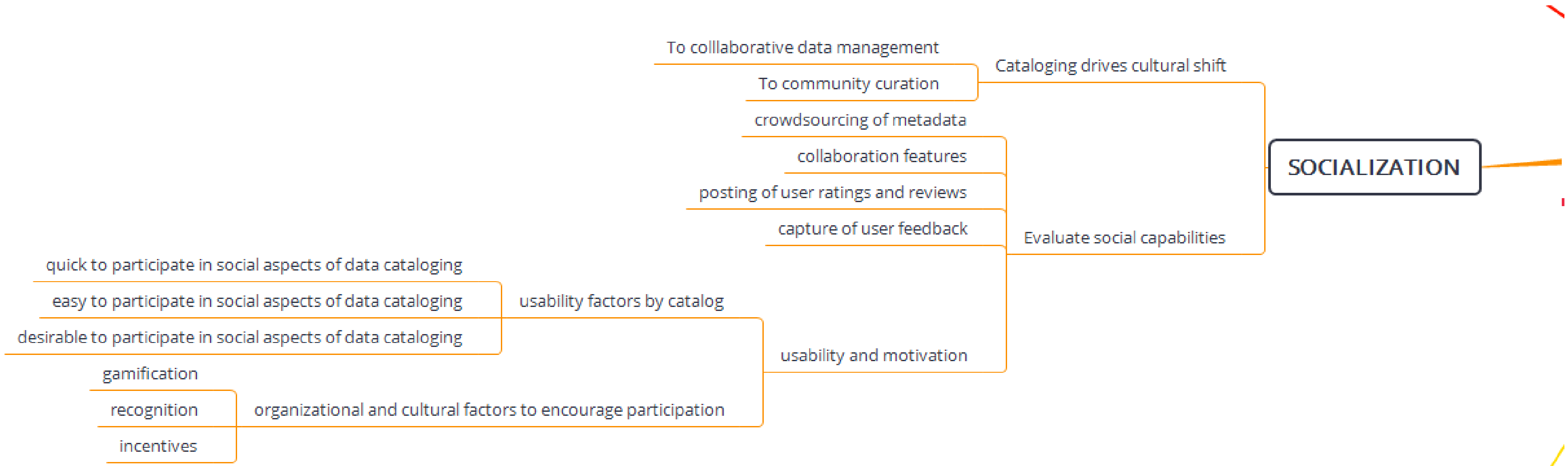


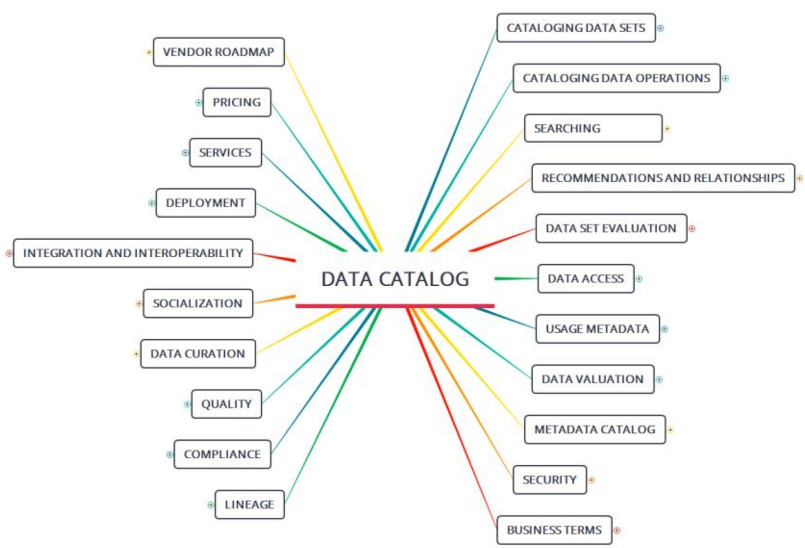
Übersicht: Data Curation



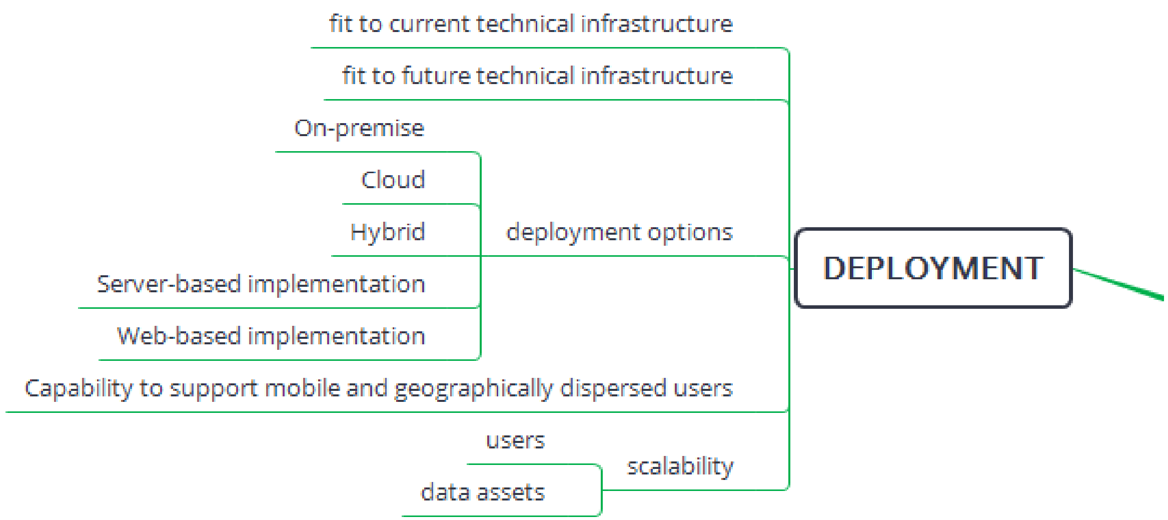


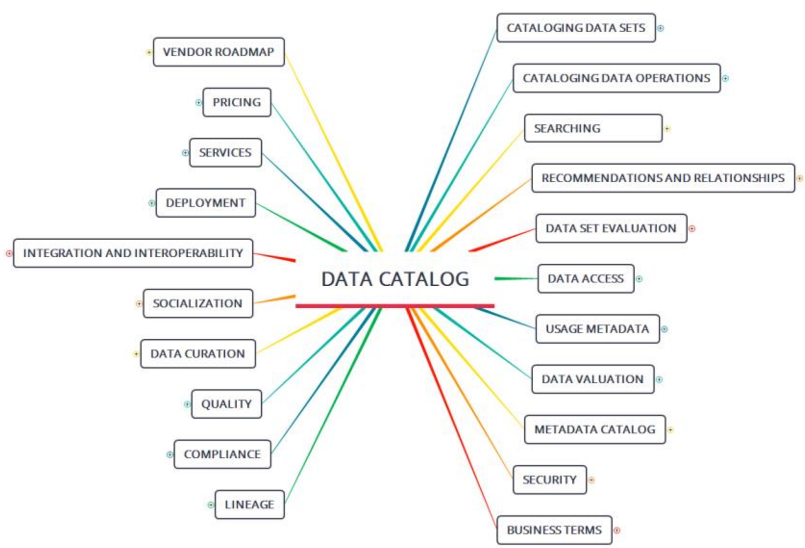
Übersicht: Socialization



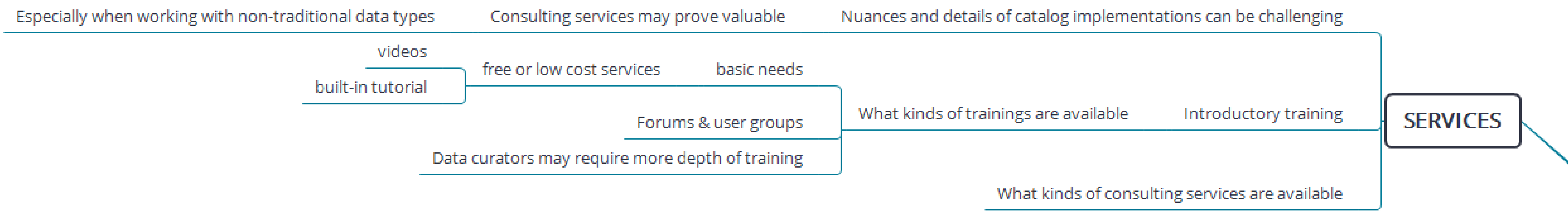


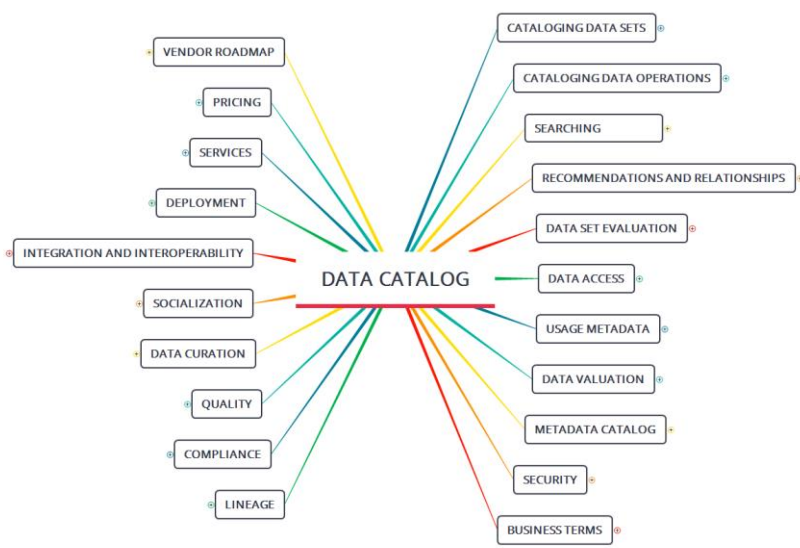
Übersicht: Deployment



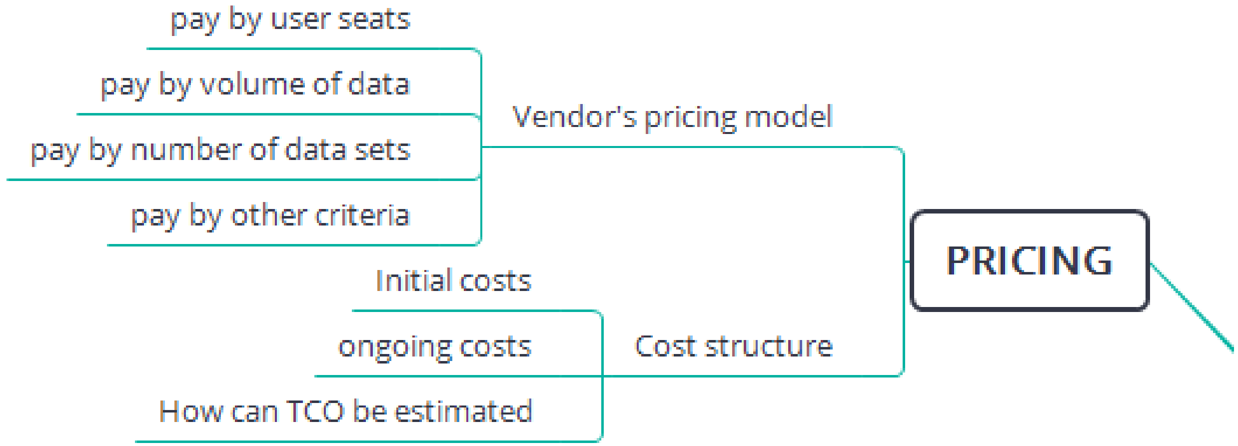


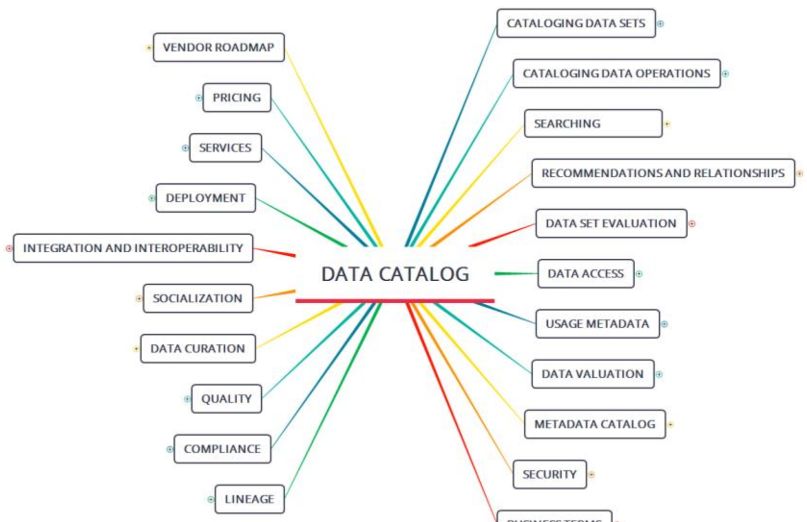
Übersicht: Services



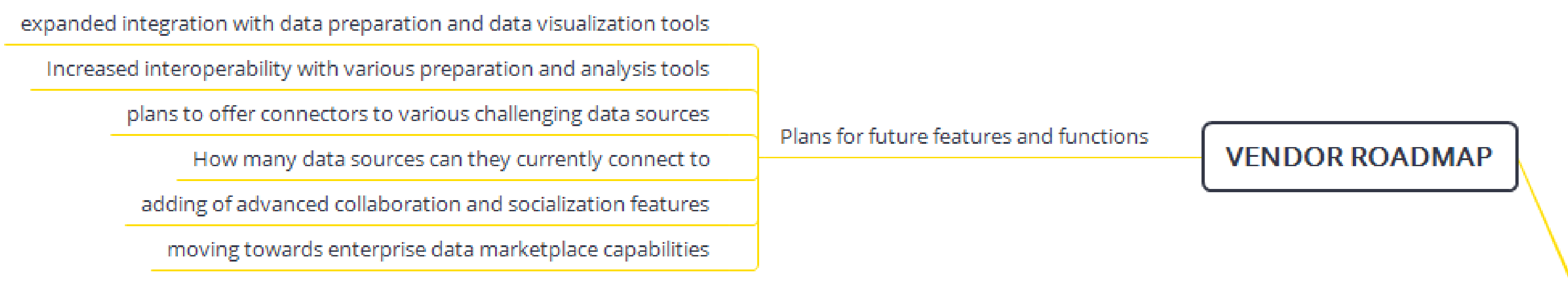


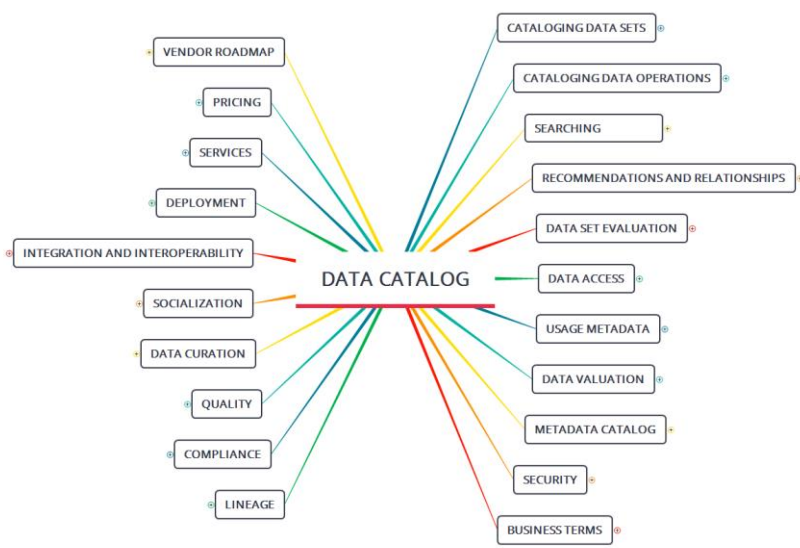
Übersicht: Pricing



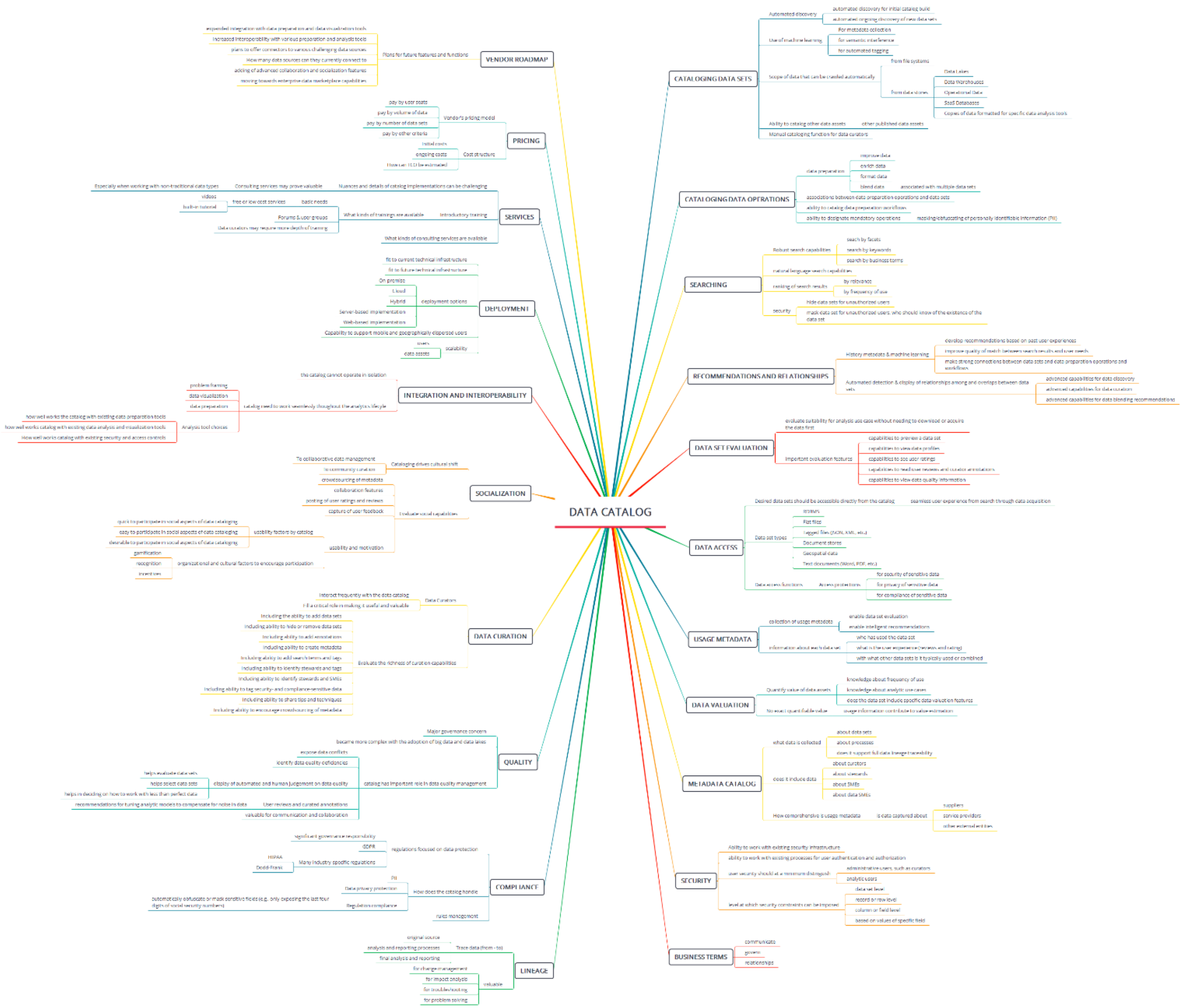


Übersicht: Vendor Roadmap





Übersicht



DATA CATALOG
VENDOR ROADMAP
PRICING
SERVICES
DEPLOYMENT
INTEGRATION AND INTEROPERABILITY
SOCIALIZATION
DATA CURATION
QUALITY
COMPLIANCE
LINEAGE
CATALOGING DATA SETS
CATALOGING DATA OPERATIONS
SEARCHING
RECOMMENDATIONS AND RELATIONSHIPS
DATA SET EVALUATION
DATA ACCESS
USAGE METADATA
DATA VALUATION
METADATA CATALOG
SECURITY
BUSINESS TERMS

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
   - Atlas Präsentation
   - Data Selfservice
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
   - Atlas Präsentation
   - Data Selfservice
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

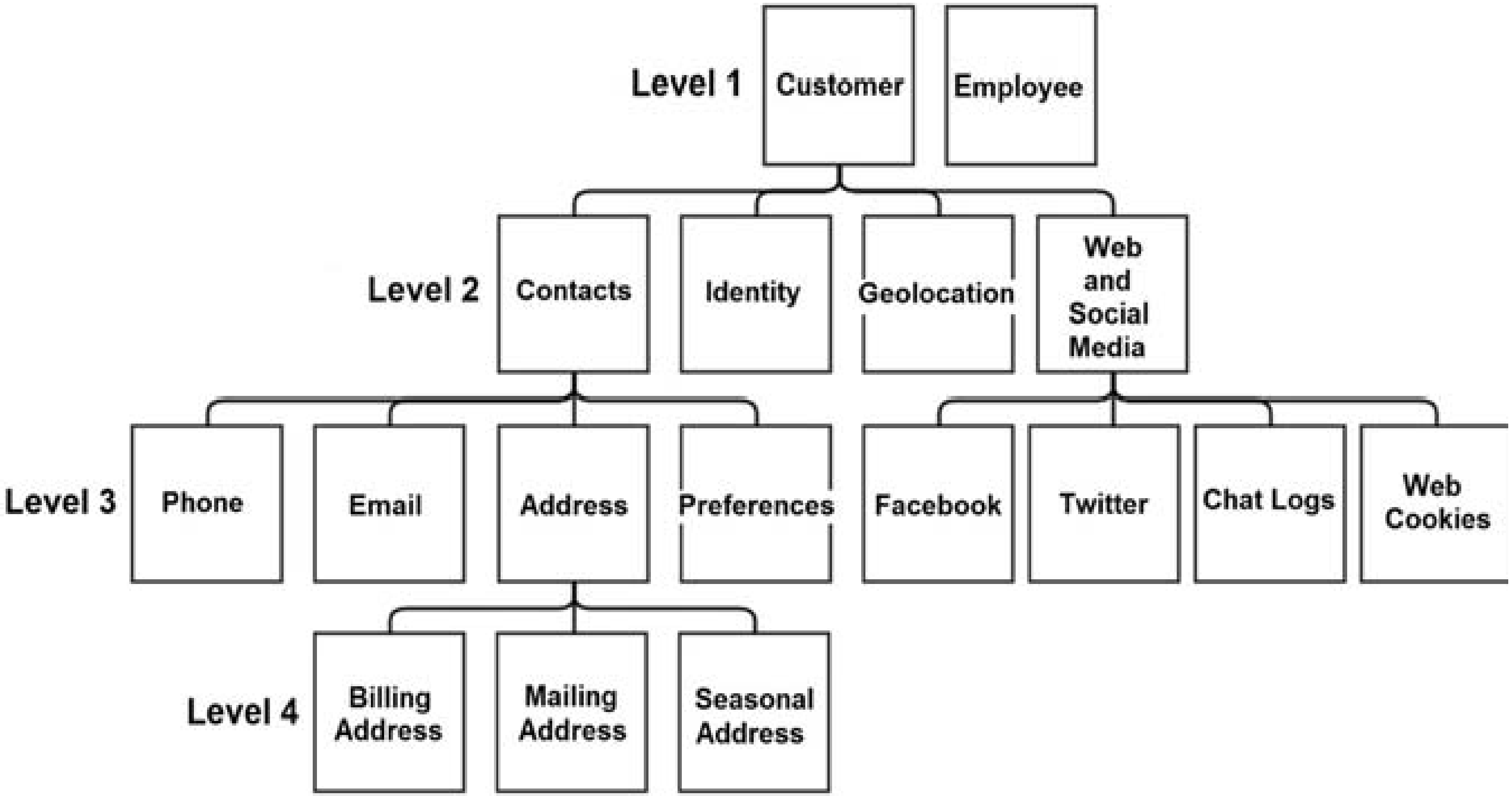
End-to-end approach to operationalize data governance for data sovereignty

1. Develop Policies, Standards, Processes & Controls
2. Create Data Taxonomy
3. Confirm Data Owners
4. Identify Critical Datasets & Critical Data Elements
5. Establish Data Collection Standards
6. Define Acceptable Use Standards
7. Implement Data Security
8. Conduct Data Protection Impact Assessments
9. Conduct Vendor Risk Assessments
10. Improve Data Quality
11. Stitch Data Lineage
12. Govern Analytical Models
13. Manage End User Computing
14. Govern the Lifecycle of Information
15. Set up Data Sharing Agreements
16. Enforce Compliance with Controls

End-to-end approach to operationalize data governance for data sovereignty

1. Develop Policies, Standards, Processes & Controls
2. Create Data Taxonomy
3. Confirm Data Owners
4. Identify Critical Datasets & Critical Data Elements
5. Establish Data Collection Standards
6. Define Acceptable Use Standards
7. Implement Data Security
8. Conduct Data Protection Impact Assessments
9. Conduct Vendor Risk Assessments
10. Improve Data Quality
11. Stitch Data Lineage
12. Govern Analytical Models
13. Manage End User Computing
14. Govern the Lifecycle of Information
15. Set up Data Sharing Agreements
16. Enforce Compliance with Controls

Sample hierarchy of data categories to support acceptable use



End-to-end approach to operationalize data governance for data sovereignty

1. Develop Policies, Standards, Processes & Controls
2. Create Data Taxonomy
3. Confirm Data Owners
4. Identify Critical Datasets & Critical Data Elements
5. Establish Data Collection Standards
6. Define Acceptable Use Standards
7. Implement Data Security
8. Conduct Data Protection Impact Assessments
9. Conduct Vendor Risk Assessments
10. Improve Data Quality
11. Stitch Data Lineage
12. Govern Analytical Models
13. Manage End User Computing
14. Govern the Lifecycle of Information
15. Set up Data Sharing Agreements
16. Enforce Compliance with Controls

Regulation sample	Description
European Union General Data Protection Regulation, Article 87	Processing of the national identification number— "Member States may further determine the specific conditions for the processing of a national identification number or any other identifier of general application. In that case the national identification number or any other identifier of general application shall be used only under appropriate safeguards for the rights and freedoms of the data subject pursuant to this Regulation."
European Union General Data Protection Regulation, Article 88(1)	Processing in the context of employment— "Member States may, by law or by collective agreements, provide for more specific rules to ensure the protection of the rights and freedoms in respect of the processing of employees' personal data in the employment context, in particular for the purposes of the recruitment, the performance of the contract of employment, including discharge of obligations laid down by law or by collective agreements, management, planning and organization of work, equality and diversity in the workplace, health and safety at work, protection of employer's or customer's property, and for the purposes of the exercise and enjoyment, on an individual or collective basis, of rights and benefits related to employment, and for the purpose of the termination of the employment relationship."
Singapore Personal Data Protection Act, Section 24	Protection of personal data— "An organization shall protect personal data in its possession or under its control by making reasonable security arrangements to prevent unauthorized access, collection, use, disclosure, copying, modification, disposal, or similar risks."
Regulation sample	**Description**
South Korea Personal Information Protection Act, Article 24(1)	Limitation to processing unique identifier— "The personal information processor shall not, except the cases stated in the following subparagraphs, process the identifier assigned so as to identify an individual in accordance with laws and regulations, as stated by the Presidential Decree (hereinafter referred to as the 'Unique Identifier'). . . ."

Template for critical data set or critical data element	
Attribute	**Description**
Name	Name of business term that is classified as a critical data set or CDE
Level 1 data category	Name of level 1 data category inthe information hierarchy (e.g., Employee)
Level 2 data category	Name of level 2 data category in the information hierarchy, if applicable
Level 3 data category	Name of level 3 data category in the information hierarchy, if applicable
Level 4 data category	Name of level 4 data category in the information hierarchy, if applicable
Definition	Meaning of the business term within the business context
Definition source	Source that provides the definition for the business term
Acronym	Shorthand formed from initial letters of term (e.g., DOB for Date of Birth)
Calculation	Algorithm used to produce a business term
Authoritative data source	Approved repository for a given business term, data element, or attribute
Synonyms	Relation between business terms with the same definition (e.g., vendor and supplier)
Is a type of/has types	Relation to show that a business term is a type of another
Related terms	Relation to link business terms that are not synonyms or types
Reference data	Defined set of values for a business term (e.g., U.S. state codes)
Information security classification	Classification for information security purposes (e.g., public, internal, confidential, highly confidential, secret, restricted)
Controls	Controls that govern the business term (e.g., the Do Not Collect control governs the business term “Race”)
Regulations	Regulations that govern the business term (e.g., CASL governs email address)
Data consumers	Business terms, reports, models, systems, and other artifacts that consume the CDE
Data producers	Business terms, models, systems, columns, fields, and other artifacts that produce the CDE
Data quality rules	Rules to define data quality classified by data quality dimension
Data owner	Name of individual who is ultimately accountable for the data
Data owner title	Title of individual who is ultimately accountable for the data
Data steward	Name of individual who works with data on a day-to-day basis
Data steward title	Title of individual who works with data on a day-to-day basis

CDE template for Patient Account Number	
Attribute	**Description**
Name	Patient Account Number
Level 1 data category	Patient
Level 2 data category	Identity
Level 3 data category	Account Number
Level 4 data category	Not applicable
Definition	This number is unique and is assigned to only one patient. It can be used to obtain electronic medical record data as well as patient payment history and medical claims data.
Authoritative data source	Customer Master Data Management Hub
Reports that consume this data	U.S. Department of Health and Human Services quarterly report, U.S. Health Insurance Portability and Accountability Act (HIPAA) reports, federal and state communicable disease reports
Information security classification	Protected
Data consumers	WebMD®, primary operational processing system, payment processing system, customer database
Data producers	Customer database
Related terms	Pt_Acct_Num
Data owner	Jane Wells
Data owner title	Vice President, Medical Policy
Data steward	Jack Smith
Data steward title	Medical Policy Data Manager

Critical data set template for Web Cookies	
Data set	**Description**
Name	Web Cookies
Definition	A web cookie is a small piece of data that a website asks the browser to store on the computer or mobile device. The cookie allows the website to “remember” a user’s actions or preferences over time.[1]
Information security classification	Internal: When cookie IDs are not combined with any other customer information. Confidential: When cookie IDs are combined with other customer information.
Controls	Obtain user consent before placing web cookies on user devices.
Regulation	Article 5(3) of the European Union ePrivacy Directive requires prior informed consent for storage of or access to information stored on a user’s terminal equipment. In other words, websites must ask users whether they agree to most cookies and similar technologies, such as web beacons and flash cookies, before the site starts to use them.[2]

End-to-end approach to operationalize data governance for data sovereignty

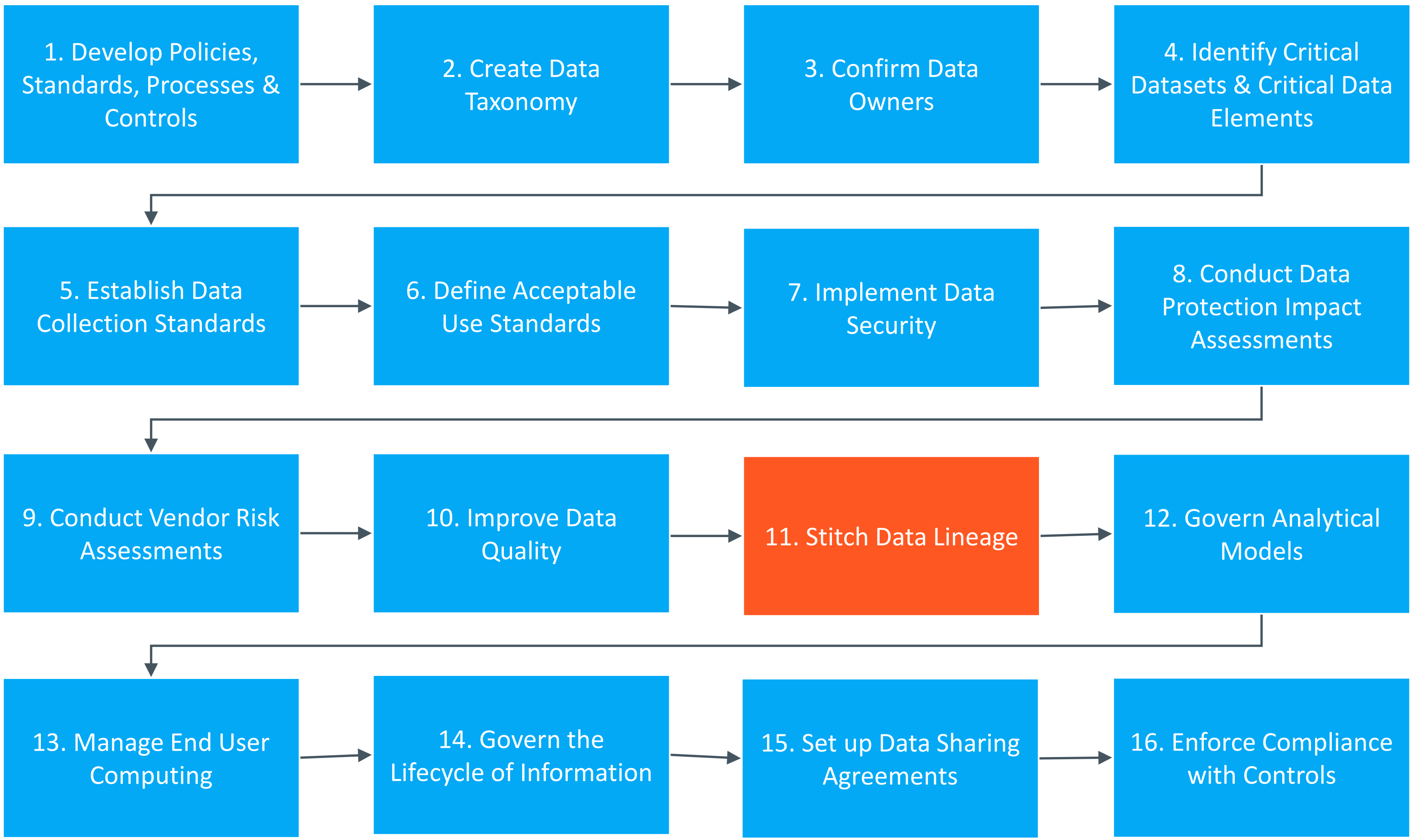
1. Develop Policies, Standards, Processes & Controls
2. Create Data Taxonomy
3. Confirm Data Owners
4. Identify Critical Datasets & Critical Data Elements
5. Establish Data Collection Standards
6. Define Acceptable Use Standards
7. Implement Data Security
8. Conduct Data Protection Impact Assessments
9. Conduct Vendor Risk Assessments
10. Improve Data Quality
11. Stitch Data Lineage
12. Govern Analytical Models
13. Manage End User Computing
14. Govern the Lifecycle of Information
15. Set up Data Sharing Agreements
16. Enforce Compliance with Controls

Regulation sample	Description
Canada's Anti-Spam Law, Provision 10(4)	Express consent, sections 6 to 8— "In addition to the requirements set out in subsections (1) and (3), if the computer program that is to be installed performs one or more of the functions described in subsection (5), the person who seeks express consent must, when requesting consent, clearly and prominently, and separately and apart from the license agreement. . . ."
European Union General Data Protection Regulation, Article 35(1)	Data protection impact assessment— "Where a type of processing in particular using new technologies, and taking into account the nature, scope, context, and purposes of the processing, is likely to result in a high risk to the rights and freedoms of natural persons. . . ."
Irish Data Protection Acts of 1998 and 2003, Amendment of Section 2 (Biometrics in the Workplace), Section 2C	Security measures for personal data— ". . .may have regard to the state of technological development and the cost of implementing the measures, and (b) shall ensure that the measures provide a level of security appropriate. . . ."

Table 8.1: Data protection impact assessment template	
Impacted systems or processes	List specific systems, databases, applications, or processes that are impacted (e.g., web portal for customer complaints, customer ordering).
Types of data	Identify data sets or data attributes that will be collected or used by the systems, databases, applications, or processes.
Data classification	Explain whether the data is private, protected, or for general use.
How system will use data	Explain the proposed usage of data within the system.
Is data necessary for function (Yes/No)?	Is the data required for proper functioning of the system or process?
Justification of necessity	Provide the business or technical justification for collecting and usage of data.
Data sovereignty laws	Identify any data sovereignty or other privacy laws that may impact this situation.
Data risks	Explain any data risks associated with this system or process (e.g., data breach).
Risk mitigation plan	Describe, at a high-level, a mitigation plan for planned risks (e.g., use sensitive flags, encrypt or mask).

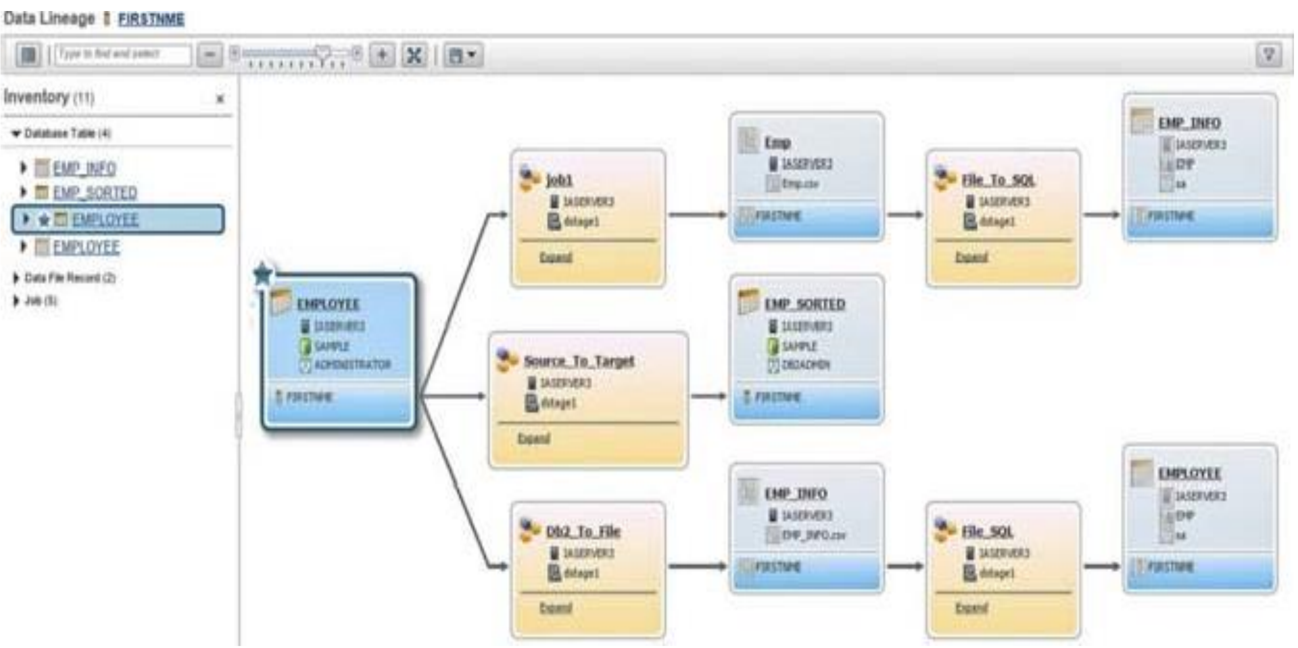
Table 8.2: Sample data protection impact assessment			
Name of impacted system	Modeling tool 'XYZ'	Data warehouse 'VRB'	Customer portal 'We See You'
How will system use data?	Physical model for customer and employee data	Store customer claim information	Expose customer data
Is data necessary for function?	Yes	Yes	Yes
Justification of necessity	'XYZ' will store all customer and employee data attributes by version	'VRB' data warehouse must capture all information sent by providers on claims	Regulations require that customers view and review data stored by company
Data subject risks identified	Private, protected, sensitive data will be captured, stored, and exposed	Private, protected, sensitive data will be captured, stored, and exposed	Private, protected, sensitive data will be exposed
Risk mitigation plan	When reviewing technical functionality, ensure the use of sensitive flags, encrypt or mask sensitive data where possible	During the functional and technical design phase, ensure that requirements are in place to use sensitive flags, encrypt data and mask as appropriate	During the functional design phase, ensure that requirements to restrict access and require user verification are included

End-to-end approach to operationalize data governance for data sovereignty



Regulation sample	Description
Hong Kong Personal Data Privacy Ordinance, Section 33	Prohibition against transfer of personal data except in specified circumstances— “A data user shall not transfer personal data to a place outside Hong Kong. . . .”
European Union General Data Protection Regulation, Article 17(1)	Right to erasure (“right to be forgotten”)— “The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies. . . .”
European Union General Data Protection Regulation, Article 22(1)	Automated individual decision-making, including profiling— “The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.”
Singapore Personal Data Protection Act, Act 59(1)	Preservation of secrecy— “Subject to subsection (5), every specified person shall preserve, and aid in the preservation of, secrecy with regard to — (a) any personal data an organization would be required or authorized to refuse to disclose if it were contained in personal data requested under section 21. . . .”
Australian Privacy Act of 1988, Principle 8.1	Cross border disclosure of personal information— “Before an APP entity discloses personal information about an individual to a person (the overseas recipient). . . .”
Laws of Malaysia Personal Data Protection Act 2010, Act 709, Section 44(1)	Record to be kept by data user— “A data user shall keep and maintain a record of any application, notice, request, or any other information relating to personal data that has been or is being processed by him.”

Figure 9.1: IBM InfoSphere Information Governance Catalog with detailed lineage



End-to-end approach to operationalize data governance for data sovereignty

1. Develop Policies, Standards, Processes & Controls
2. Create Data Taxonomy
3. Confirm Data Owners
4. Identify Critical Datasets & Critical Data Elements
5. Establish Data Collection Standards
6. Define Acceptable Use Standards
7. Implement Data Security
8. Conduct Data Protection Impact Assessments
9. Conduct Vendor Risk Assessments
10. Improve Data Quality
11. Stitch Data Lineage
12. Govern Analytical Models
13. Manage End User Computing
14. Govern the Lifecycle of Information
15. Set up Data Sharing Agreements
16. Enforce Compliance with Controls

Govern Analytical Modells I

Regulation sample	Description
France Data Protection Act, Article 8.7	Statistical processing— “. . .Statistical processing carried out by the National Institute for Statistics and Economic Studies or one of the ministerial statistical services, in compliance with Law No. 51-711 of 7 June 1951 on the obligation, coordination, and secrecy in statistics after consultation with the National Council for Statistical Information and under the conditions laid down in Article 25 of this Law. . . .”
France Data Protection Act, Article 10	Decisions based on automated processing— “No judicial decision involving an assessment of a person's conduct may be based on an automated processing of personal data intended to evaluate certain aspects of his personality. No other decision which has legal effects in relation to a person can be taken solely on the basis of an automated processing of data intended to define the profile of the person concerned or to assess certain aspects of his personality. . . .”
European Union General Data Protection Regulation, Article 21(1)	Right to object— “The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her which is based on point (e) or (f) of Article 6(1), including profiling based on those provisions. . . .”

Govern Analytical Models II

Table 11.1: Template for model metadata	
Attribute	**Description**
Model ID	Unique identifier for the model
Name	Name of the model
Description	Description of the model
Business purpose	Business use of the model
Methodology	Methodology used to develop the model (e.g., regression analysis, rules, logistic regression, random forest)
Application	Application used to develop the model (e.g., SAS®, R™, Hadoop®)
Level 1 model category	Name of the Level 1 category where the model can be best classified
Level 2 model category	Name of the Level 2 category where the model can be best classified, if applicable
Report and line item	Name of the line item in any report to which the model applies
Input variables	Names of the variables used as inputs into the models
Input models	IDs of models that are used as input into this model
Business rules governing inputs	Business rules that govern variable and model inputs into this model
Output variables	Names of the variables that are outputs from this model
Dependent models	IDs of models that depend on the outputs of this model
Business rules governing outputs	Business rules that govern the outputs of this model
Model creator	Name of individual who created the model
Model creator department	Department of individual who created the model
Model owner	Name of individual who owns the model
Model owner department	Department of individual who owns the model
Model creation date	Date that the model was created
Model deployment date	Date that the model was deployed
Model validation date	Date that the model was independently validated or will be validated
Model validation owner	Name of the individual who independently validated or will validate the model
Model validation department	Department of the individual who independently validated or will validate the model

End-to-end approach to operationalize data governance for data sovereignty

1. Develop Policies, Standards, Processes & Controls
2. Create Data Taxonomy
3. Confirm Data Owners
4. Identify Critical Datasets & Critical Data Elements
5. Establish Data Collection Standards
6. Define Acceptable Use Standards
7. Implement Data Security
8. Conduct Data Protection Impact Assessments
9. Conduct Vendor Risk Assessments
10. Improve Data Quality
11. Stitch Data Lineage
12. Govern Analytical Models
13. Manage End User Computing
14. Govern the Lifecycle of Information
15. Set up Data Sharing Agreements
16. Enforce Compliance with Controls

Regulation sample	Description
Russian Federal Law 152-FZ, Article 14(1)	Personal data subject's right to access his personal data— "The subject of personal data have the right to require the operator to clarify his personal data, their blocking or destruction if personal data are incomplete, outdated, inaccurate, or obtained illegally, are not necessary for the stated purpose of the processing, as well as to take measures prescribed by law for the protection of their rights."
European Union General Data Protection Regulation, Article 16	Right to rectification— ". . .the data subject shall have the right to have incomplete personal data completed. . . ."
European Union General Data Protection Regulation, Article 17(2)	Right to erasure ("right to be forgotten")— "Where the controller has made the personal data public and is obliged pursuant to paragraph 1 to erase the personal data, the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data."
European Union General Data Protection Regulation, Recital 85	Addressing personal data breach— ". . .Therefore, as soon as the controller becomes aware that a personal data breach has occurred, the controller should notify the personal data breach to the supervisory authority without undue delay and, where feasible, not later than 72 hours after having become aware of it, unless the controller is able to demonstrate, in accordance with the accountability principle, that the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons. . . ."

Continued

End-to-end approach to operationalize data governance for data sovereignty

1. Develop Policies, Standards, Processes & Controls
2. Create Data Taxonomy
3. Confirm Data Owners
4. Identify Critical Datasets & Critical Data Elements
5. Establish Data Collection Standards
6. Define Acceptable Use Standards
7. Implement Data Security
8. Conduct Data Protection Impact Assessments
9. Conduct Vendor Risk Assessments
10. Improve Data Quality
11. Stitch Data Lineage
12. Govern Analytical Models
13. Manage End User Computing
14. Govern the Lifecycle of Information
15. Set up Data Sharing Agreements
16. Enforce Compliance with Controls

Data Sharing Agreements I

Regulation sample	Description
Singapore Personal Data Protection Act, Act 26(1)	Transfer of personal data outside Singapore— "An organization shall not transfer any personal data to a country or territory outside Singapore except in accordance with requirements prescribed under this Act. . . ."
European Union General Data Protection Regulation, Article 44	General principle for transfers— "Any transfer of personal data which are undergoing processing or are intended for processing after transfer to a third country or to an international organization shall take place only if. . .the conditions laid down in this Chapter are complied with by the controller and processor, including for onward transfers of personal data from the third country or an international organization to another third country or to another international organization. . . ."
European Union General Data Protection Regulation, Article 50	International cooperation for the protection of personal data— ". . .(a) develop international cooperation mechanisms to facilitate the effective enforcement of legislation for the protection of personal data; (b) provide international mutual assistance in the enforcement of legislation for the protection of personal data, including through notification, complaint referral, investigative assistance and information exchange, subject to appropriate safeguards for the protection of personal data and other fundamental rights and freedoms;. . . ."
Australian Privacy Act of 1988, Principle 8	Cross-border disclosure of personal information— "Before an APP entity discloses personal information about an individual to a person (the overseas recipient): (a) who is not in Australia or an external Territory; and (b) who is not the entity or the individual; the entity must take such steps as are reasonable in the circumstances to ensure that the overseas recipient does not breach the Australian Privacy Principles (other than Australian Privacy Principle 1) in relation to the information."

Continued

Data Sharing Agreements II

Table 12.1: Sample template for data sharing agreement

Attribute	Description
Name	One-sentence description of the data sharing agreement
Type of data being shared	High-level description of the types of data included in the agreement (e.g., nature of personal data, purpose, and duration of processing)
Acceptable use	How the consuming system can use the data (i.e., approved guidelines for data once it has been persisted by the consum- ing system, who has access to it, what they can do with it)
Restricted additional sharing	Restrictions on how data may be shared with downstream consumers, used to derive new data, or propagated to other systems
Country of origin	Country in which the data originated
Country of destination	Country to which the data is being transferred
Consuming system/application	Name of the system/application that is consuming the data being shared
Consuming data owner	Name and title of the data owner who consumes the information (e.g., John Doe, CFO)
Producing system/application	Name of the system/application that is producing the data being shared
Producing data owner	Name and title of the data owner who produces the information (e.g., Jill Smith, CRO)
Critical data elements	Name of the fields in the report to which the data sharing agreement applies
Associated data feed(s)	Names of associated data feeds or web services that provide data or access to data supported by the specific parameters of this data sharing agreement
Define data quality responsibility	Consuming and producing parties' involvement in validating state of data and mitigation for materially altered data
Consuming system accountabilities	Accountabilities of the consuming system for ensuring data security (e.g., compliance with third parties, application of business rules)
Retention	How long the data in the data sharing agreement is allowed to be retained by the customer
Effective date	Effective date of the data sharing agreement

Continued

Data Sharing Agreements III

Table 12.2: Sample data sharing agreement	
Attribute	**Description**
Name	Customer Contact Details
Type of data being shared	Current customer personal contact details for marketing campaigns
Acceptable use	Data may only be used to contact customers who meet profile and demographic targets and who have opted to receive marketing information for legally approved marketing campaigns. Acceptable use of this critical data has been reviewed and approved by the legal, risk, and compliance offices. Use of data must adhere to approved acceptable uses, and comply with the identifiable legal basis for consent/use, processing, and transfer of per-sonal data (EU GDPR Article 44).
Restricted additional sharing	Data may not be propagated from the consuming system to any other downstream systems or shared in any other format outside this data sharing agreement as defined.
Country of origin	France
Country of destination	Belgium
Consuming system/application	Customer Digital Platform
Consuming data owner	Joe Bloggs – SVP – Marketing
Producing system/application	Customer Profile
Producing data owner	Jean Dupont – VP – Data Strategy
Critical data elements	First name, last name, customer ID, home address, personal email address, opt-in status, relationship type, marketing segment code
Associated data feed(s)	Cust_Profile – on-demand outbound web service
Define data quality responsibility	Data quality is the responsibility of the producer; as such, the Customer Digital Platform relies on Customer Profile for the completeness and accuracy of data. All critical data elements should meet established quality thresholds.
Consuming system accountabilities	Consuming system is accountable for securing data in accordance with the company's information security policies. Consumer must comply with any legislative data protection/privacy/sovereignty regulations applicable in the destination country.
Retention	Data must be replaced with each new feed and can only be used within two hours of feed.
Effective date	2017-15-01

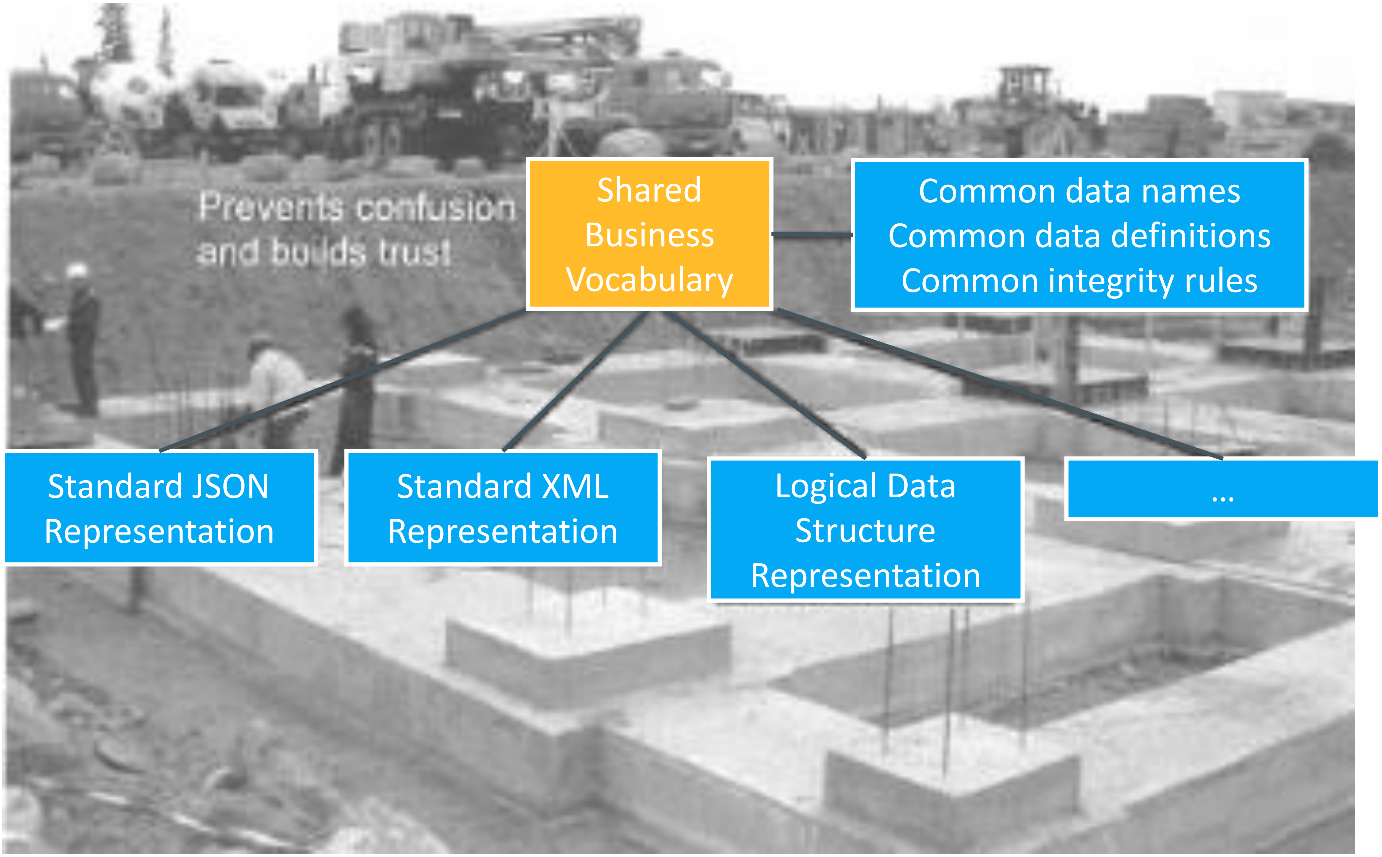
Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   – Data Sovereignty
   – Business Glossar
   – Alation Präsentation – Business Glossar anlegen
   – SVML-Präsentation
   – Data Catalog und Data Lake
   – Atlas Präsentation
   – Data Selfservice
   – IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

Agenda:

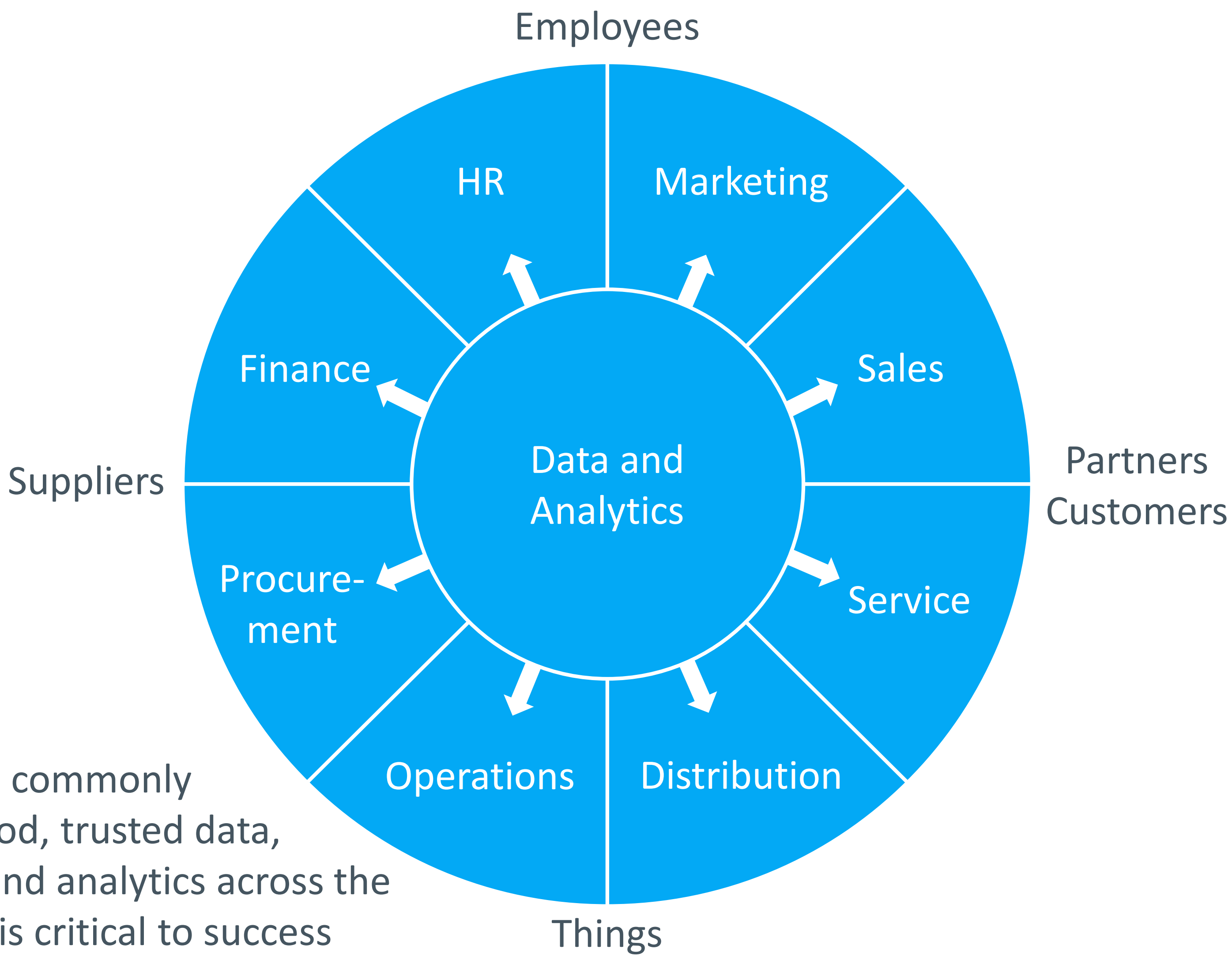
1. Data standardization using a shared business vocabulary (SBV)
2. The role of a SBV in MDM, RDM, SOA, DW and data virtualization
3. Planning for a business glossary
4. Organizing data definitions in a business glossary
5. Business involvement in SBV creation
6. Business glossary evolution to information governance catalogs

The Foundation for Smart Data Management, Governance And Creating Value Is A Common Vocabulary and Lineage



Acts as the foundation for sharing data across systems irrespective of whether those systems are on-premises or in the cloud – it is fundamental to getting rid of complexity

Why A Common Vocabulary? – Data and Analytics Have Moved to the Centre of the Enterprise for Use Everywhere



Access to commonly understood, trusted data, insights and analytics across the business is critical to success

Data Standardisation - The Role Of Shared Business Vocabulary

- A shared business vocabulary acts as a base for sharing data across applications and processes <u>irrespective of whether those systems are on-premises or in the cloud</u>
- Common metadata is built *incrementally* by identifying and mapping disparate data to common definitions
- It involves incrementally defining a set of *enterprise wide*
  - *Common* data names
  - *Common* data definitions
  - *Common* business integrity rules
  - *Common* reference data (e.g. code set valid values….)

  for at least
  - Master data
  - Transactional data
  - Metrics

  used in your business and then implementing this across all necessary infrastrucure to ensure consistency when sharing data across applications, processes and portals

Assign A Common Vocabulary To Ensure Common Understanding

Methodologies used to produce trusted data in a data lake

Used for structured data

Early definition of a common vocabulary for the data being produced

Used for semi-structured or unstructured data

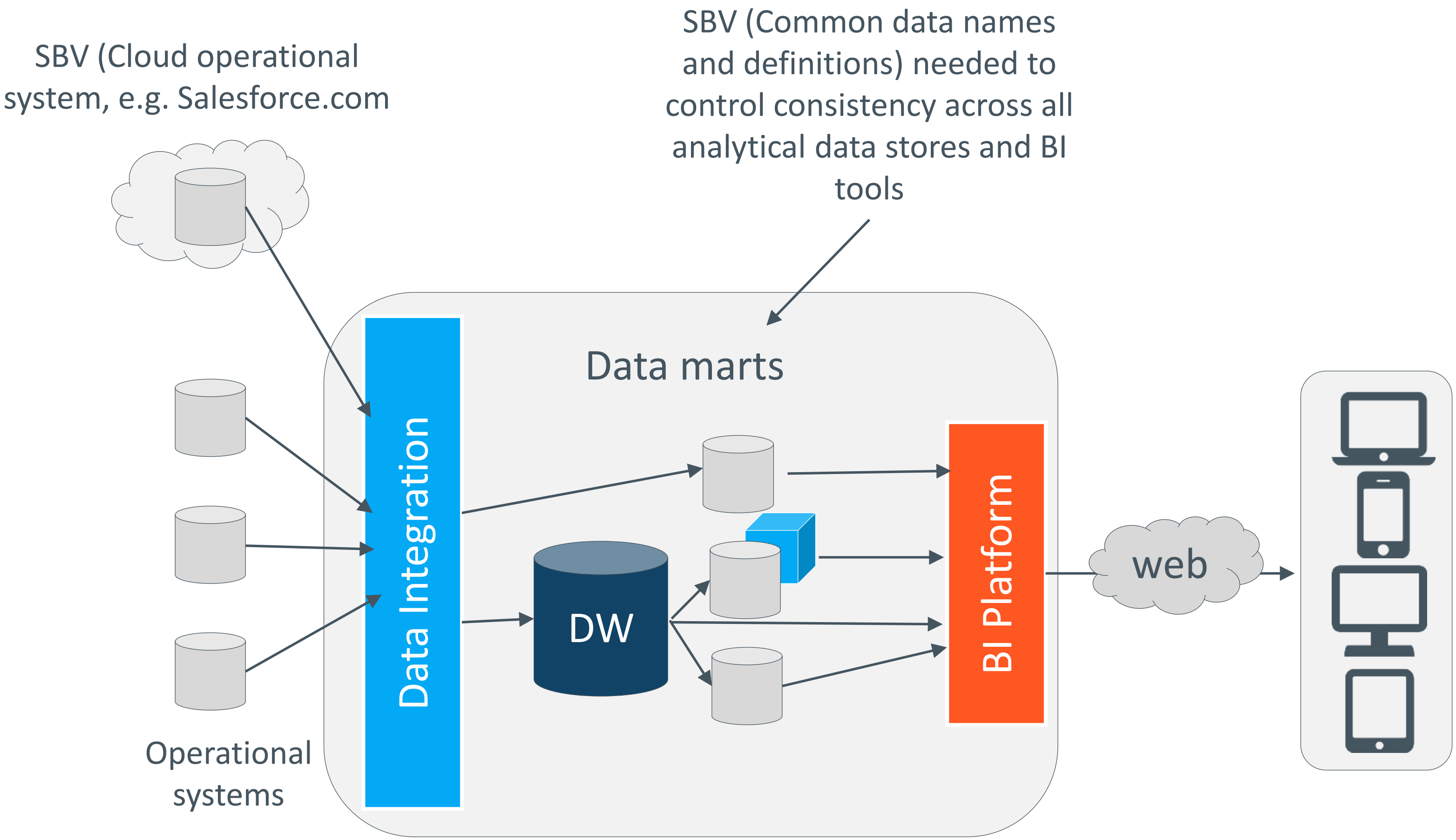
1. Collect data (including high volume and velocity ingest)
2. Crawl, auto profile, tag, classify and catalog data
3. Prepare data at scale
4. Analyse data
5. Produce new data and/or insights
6. Map insights to shared business vocabulary
7. Assign a schema using shared business vocabulary terms
8. Publish

Late definition of a common vocabulary for the produced data

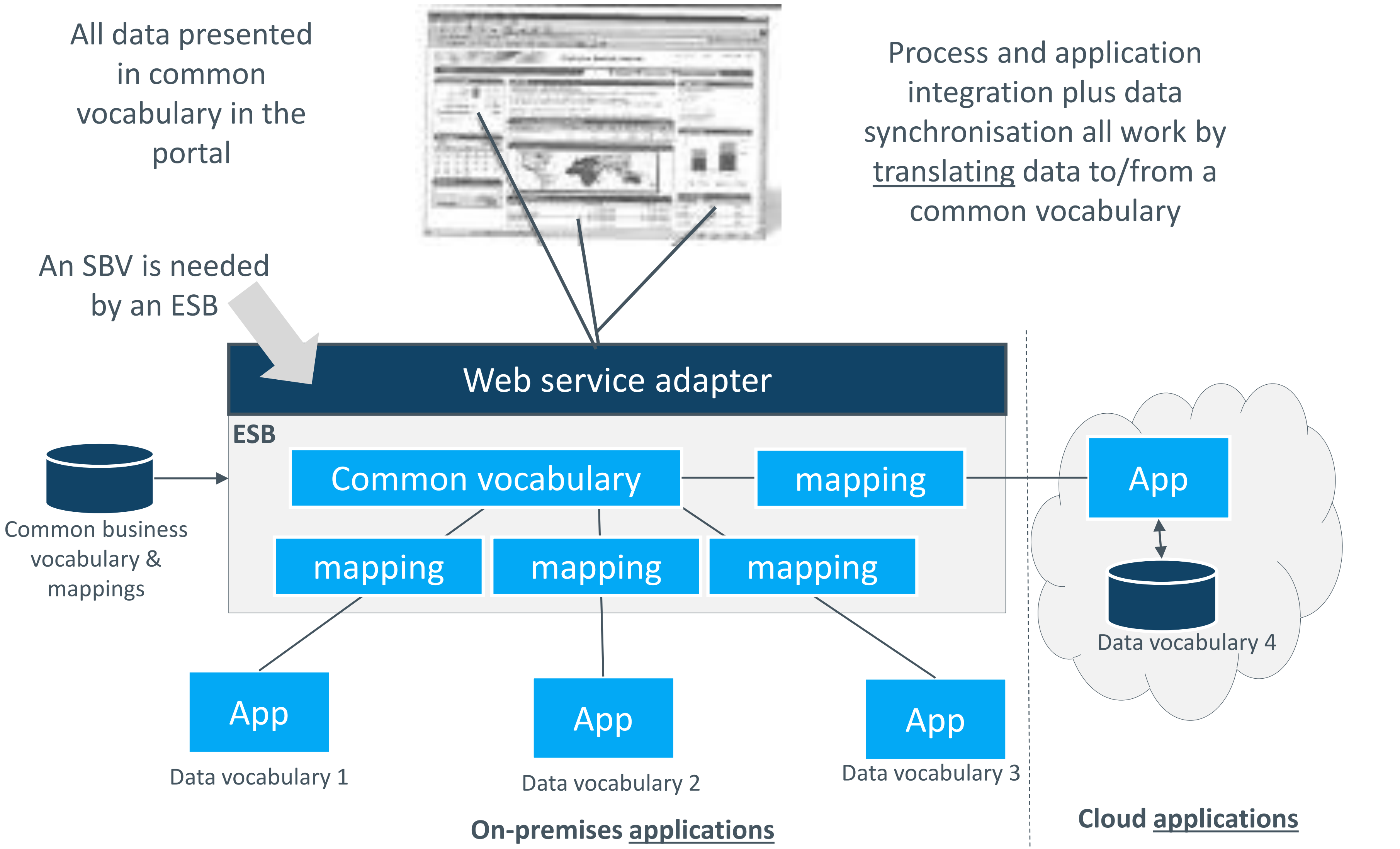
Agenda:

1. Data standardisation using a shared business vocabulary (SBV)
2. The role of a SBV in MDM, RDM, SOA, DW and data virtualisation
3. Planning for a business glossary
4. Organising data definitions in a business glossary
5. Business involvement in SBV creation
6. Business glossary evolution to information governance catalogs

Why is it so important? An SBV is Critical to keeping a BI Environment Consistent and is a best Practice



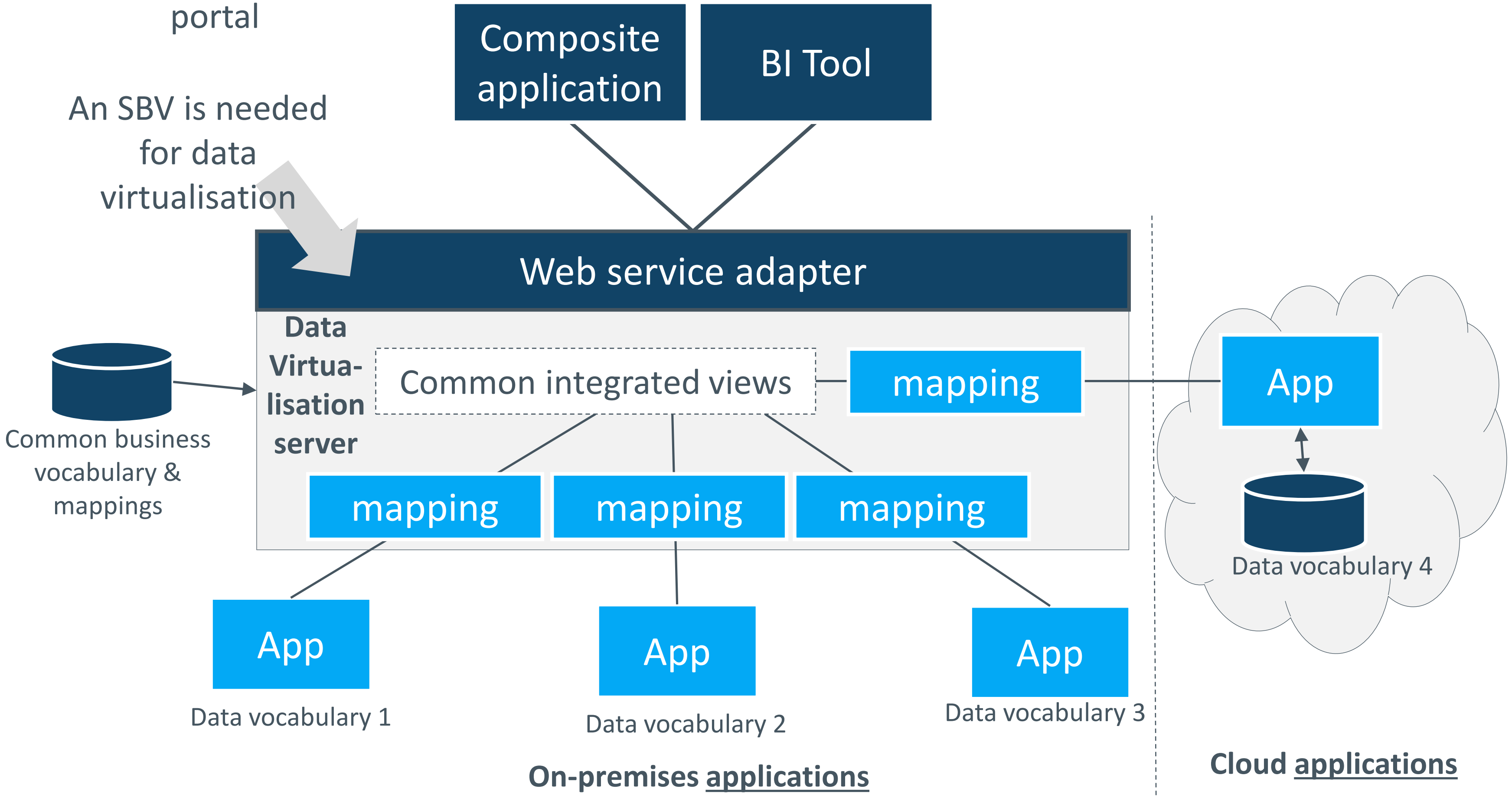
An SBV is ALSO needed for Business Processes AND Application Integration



Data Virtualisation works by giving Applications Integrated Virtual View of disparate Data defined using a SBV

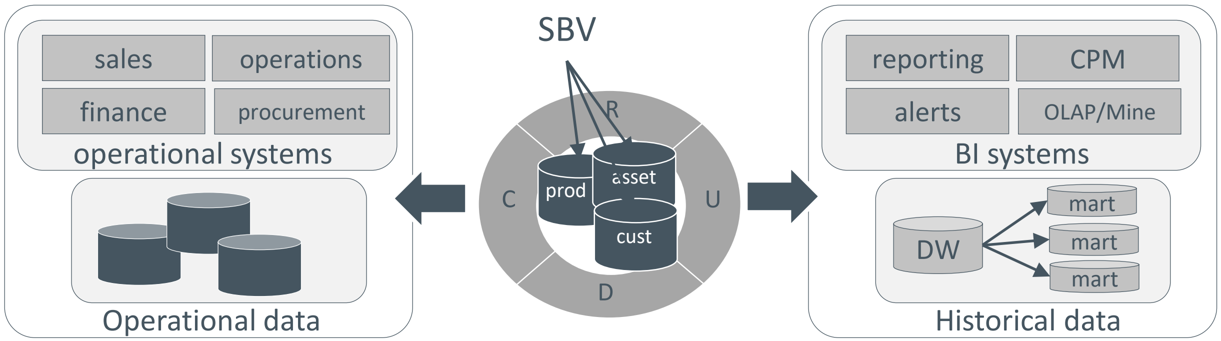
All data presented in common vocabulary in the portal

An SBV is needed for data virtualisation



Data Governance – The Shared Business Vocabulary (SBV) is also used to define commonly understood Master Data

- MDM is the end-to-end management of master data including
  - Defining master data using an SBV for use across the enterprise
  - Locating, cleansing, matching and mapping data to an SBV model
  - Persisting data in master data stores with common master data services to maintain master data
  - Supplying master data to on-premises, cloud operational and BI systems to ensure consistency and synchronisation
  - Metadata and metamodel changes have to be recorded in MDM applications to reflect product hierarchy changes, org. unit changes and customer detail changes



Master data &
Master data services

Agenda:

1. Data standardisation using a shared business vocabulary (SBV)
2. The role of a SBV in MDM, RDM, SOA, DW and data virtualisation
3. Planning for a business glossary
4. Organising data definitions in a business glossary
5. Business involvement in SBV creation
6. Business glossary evolution to information governance catalogs

Planning For A Business Glossary

- Identify any existing vocabularies to import into the glossary
- Determine categories and definitions needed
- Which categories contain which definitions?
- Determine top-level categories and subcategories (taxonomy)
- Determine custom attribute fields to be defined in the glossary to capture additional types of information
- Determine what roles to define around the glossary
  - E.g. Steward, Editor, Approver, Publisher…
- Define communities around data
- Define common approval workflows to govern data definitions
- Set up versioning to signal draft versus production definitions
- Define reports on SBV usage, ownership, development, etc.

Agenda:

1. Data standardisation using a shared business vocabulary (SBV)
2. The role of a SBV in MDM, RDM, SOA, DW and data virtualisation
3. Planning for a business glossary
4. Organising data definitions in a business glossary
5. Business involvement in SBV creation
6. Business glossary evolution to information governance catalogs

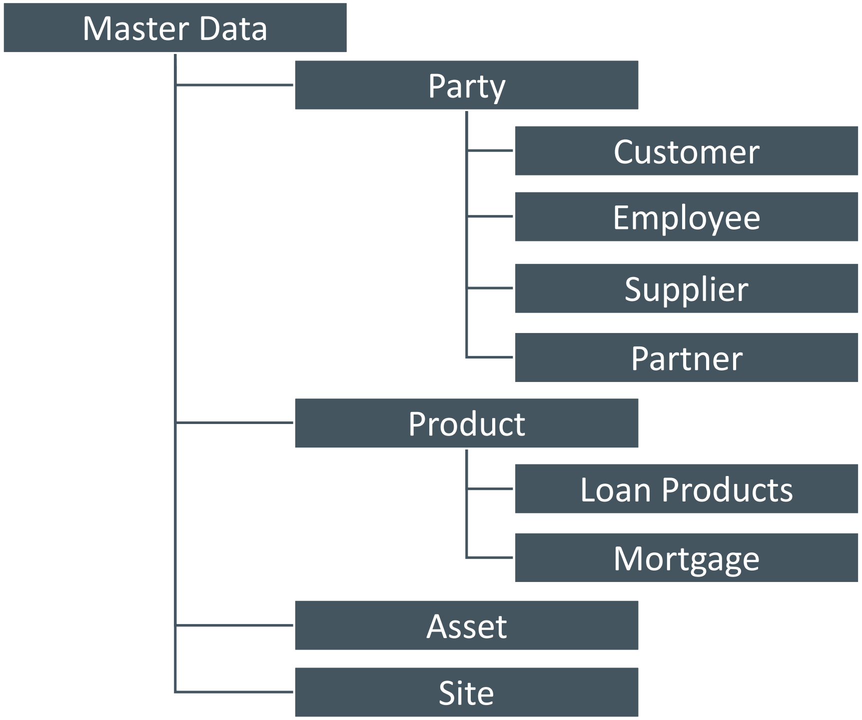
Category Structure Options

By department	• Sales, Marketing, Service, Finance, HR • …
By Line of Business	• E.g. Insurance – Property, casualty, motor, marine • Aviation, Professional Indemnity…
By application system	• ERP system, CRM system, SCM system • DW system…
By Business Data	• Master data – customer, product, asset… • Transaction data – orders, shipments, payment • Metrics, e.g. KPIs, KRIs

- Recommendation is by data because data cuts horizontally across the enterprise and is independent of location, line of business, organisational structure and applications

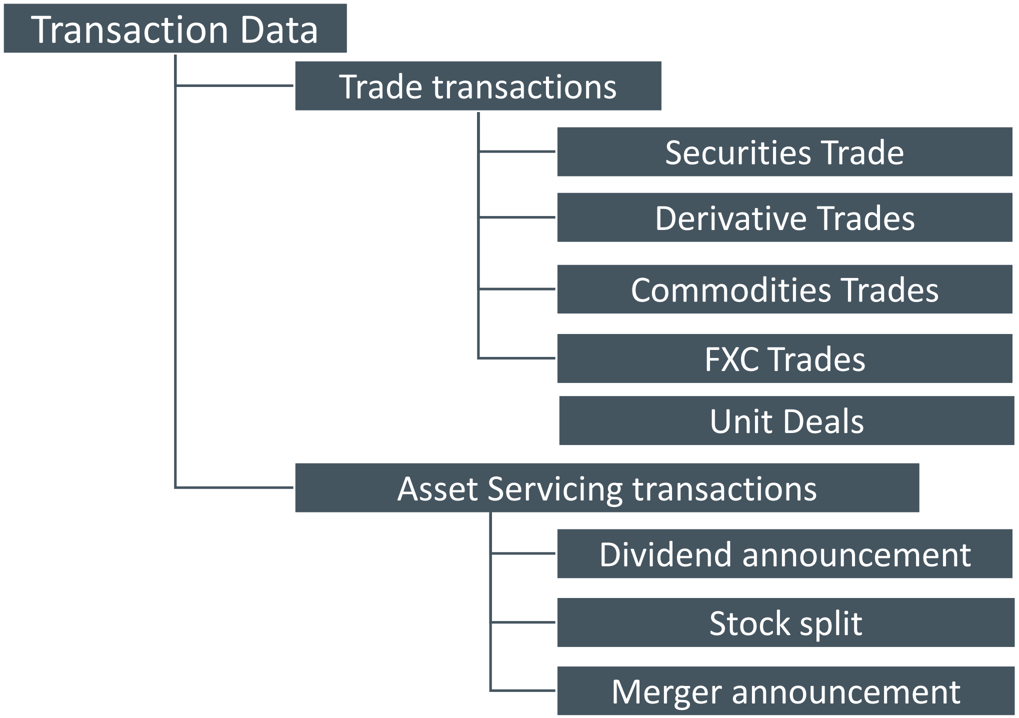
Categories And Data Definitions

- Categories
  - Used to group SBV data definitions to make them easy to find
  - Keeps the business glossary tidy and easy to navigate
- Categories can contain sub-categories and so introduce hierarchies into the glossary



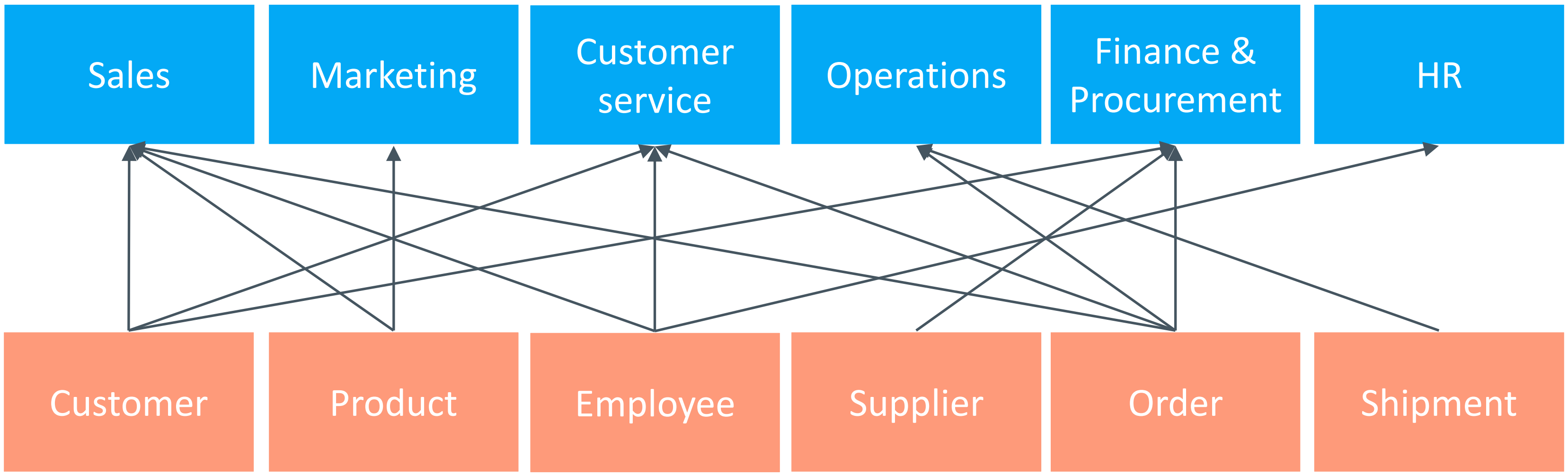
Categories And Data Definitions

- Categories
  - Used to group SBV data definitions to make them easy to find
  - Keeps the business glossary tidy and easy to navigate
- Categories can contain sub-categories and so introduce hierarchies into the glossary



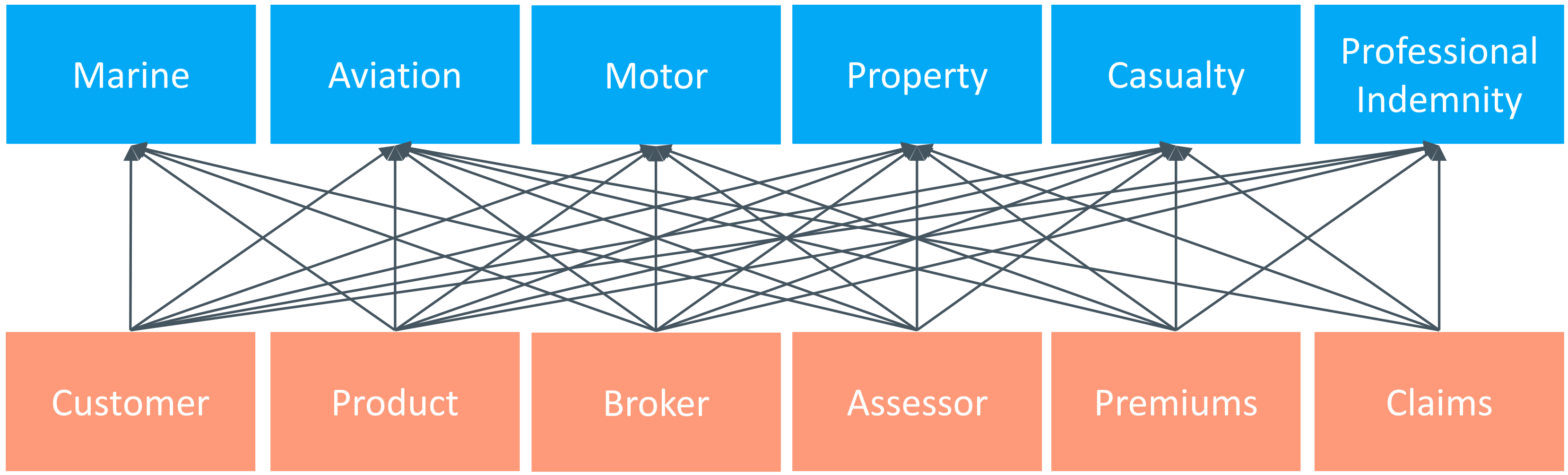
Organizing a Business Glossary by Data Provides a Foundation for other views to be created on top

See where in the business data is created and used



we can do the same to show LoB Views of the Glossary

e.g. Insurance



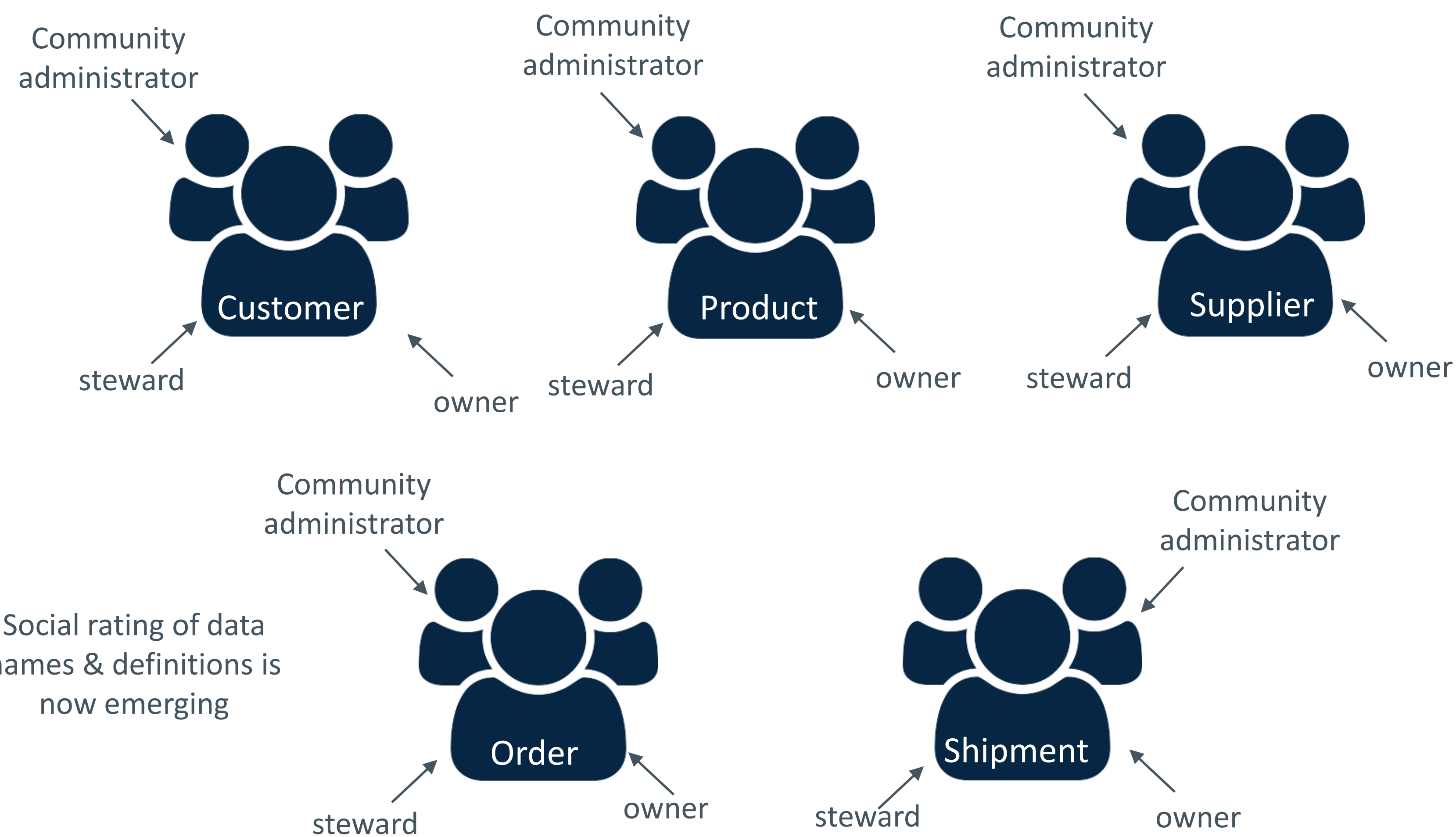
Business Glossary
- Key Roles And Responsibilities

- Data **owners**
  - Responsible for the data they own irrespective of where that data resides in the organization
- Data definition **editors**
  - Business users authorized to create new data definitions in the business glossary for approval by a data governance control board
  - They also maintain existing data definitions
- Data definition **approvers**
  - Business users who are **members of a data governance control** board authorized to approve the creation of new data definitions and changes to existing ones
  - Can publish a term changing its status from draft to production
- Data **stewards**
  - Responsible for monitoring and maintaining the quality of the data defined in the SBV

Agenda:

1. Data standardization using a shared business vocabulary (SBV)
2. The role of a SBV in MDM, RDM, SOA, DW and data virtualization
3. Planning for a business glossary
4. Organizing data definitions in a business glossary
5. Business involvement in SBV creation
6. Business glossary evolution to information governance catalogs

Several Communities are Likely to be Involved in defining common Data Names and Data Definitions



Social rating of data names & definitions is now emerging

Communities may include a mix of IT and Business users

Governance Needs To Be Applied To Data Definitions – e.g. Approval Process For Data Naming & Data Definition

- Only authorized business users can issue
  - Requests for new data items
  - Requests for decommission data items
  - Requests to change data item definitions

- All changes flow to a data governance council for approval

Business Glossary Governance - Benefits Of An Approval Process

- The approval process means each data item can have a lifecycle and be versioned
- Status of a data definition can be viewed so that the user knows what data item are considered:
  - Candidate
  - Accepted
  - Standard
  - Decommissioned
- A formal record is kept to make everything auditable
  - Who requested the change
  - Who approved it
  - When it was approved
  - etc.

Techniques For Achieving Business Participation To Populate A Business Glossary

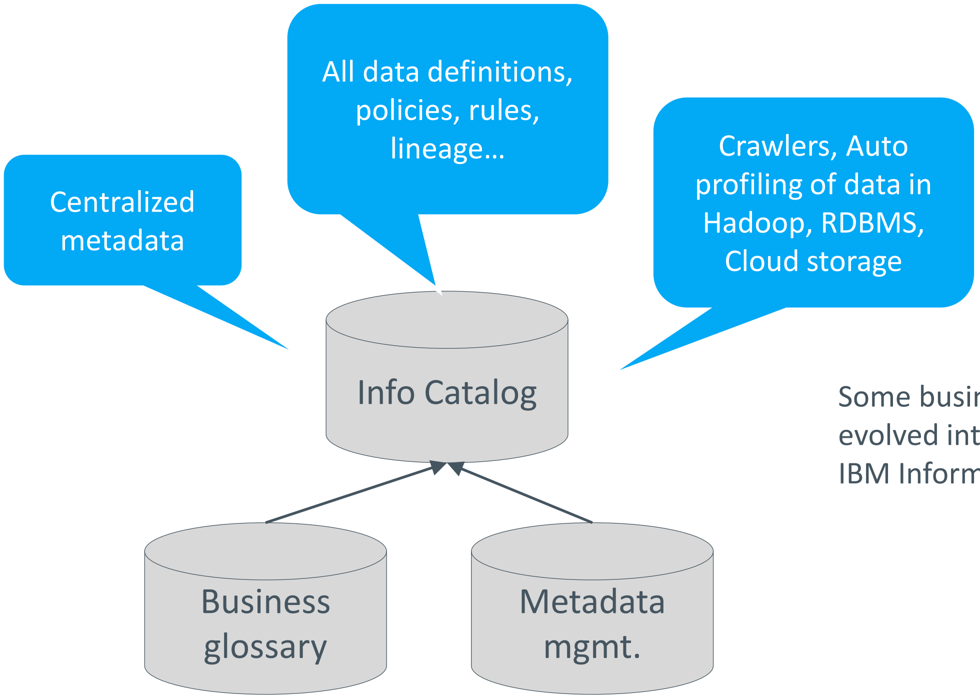
- Executive sponsorship and communication
- Broad communication of business case
  - Need to articulate why does this need to be done
- Breaking down the data into manageable categories
  - Master data categories
  - Reference data categories
  - Transaction data type categories
  - Metrics data KPIs and KRIs
- Multiple communities each associated with specific data

Customer

Agenda:

1. Data standardization using a shared business vocabulary (SBV)
2. The role of a SBV in MDM, RDM, SOA, DW and data virtualization
3. Planning for a business glossary
4. Organizing data definitions in a business glossary
5. Business involvement in SBV creation
6. Business glossary evolution to information governance catalogs

Evolution – Business Glossaries have and are Evolving into Information Catalogs



Some business glossary products have evolved into information catalogs e.g. IBM Information Governance Catalog

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
   - Atlas Präsentation
   - Data Selfservice
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

SARACUS CONSULTING

Demo: Alation –
Business Glossar anlegen

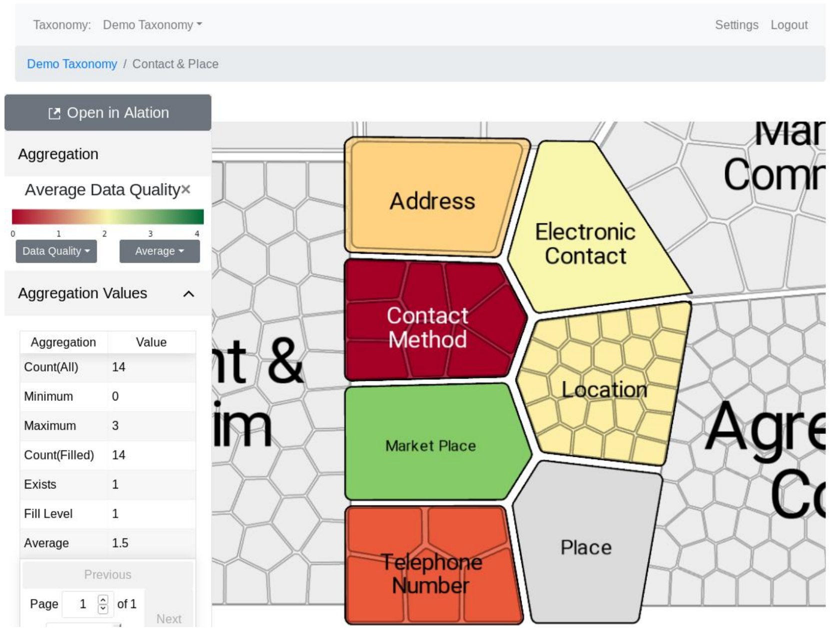
Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
   - Atlas Präsentation
   - Data Selfservice
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

Motivation: Taxonomy view

Broad Metadata Sources
- Logical
- Technical
- Operational
- Usage

Business Context
- Glossary
- Policies
- Process
- Usecases
- Models



Visualize and analyze metadata in a business scope

Self Service Analytics
[Data Analysts, Data Scientists]
- Google for enterprise data assets
- Which data is used in my model / usecase?
- Visualize Lineage

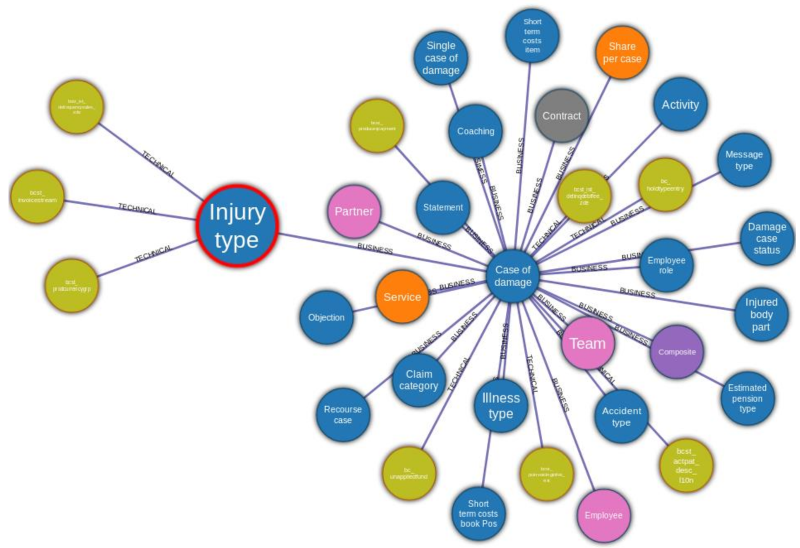
Data Governance
[Data Stewards]
- Associate Business glossary to technical objects
- Verify business to technical lineage
- Track key data elements compliance

Data Asset Management
[Architects, Developers]
- Analyze Lineage
- View transformation Logic
- Data asset and BI usage

Motivation: Graph view

Broad Metadata Sources
- Logical
- Technical
- Operational
- Usage

Business Context
- Glossary
- Policies
- Process
- Usecases
- Models

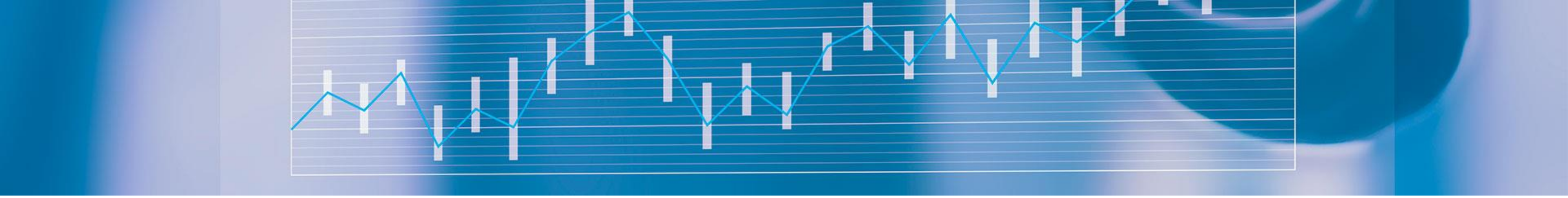


Find relations between business objects / technical data / users

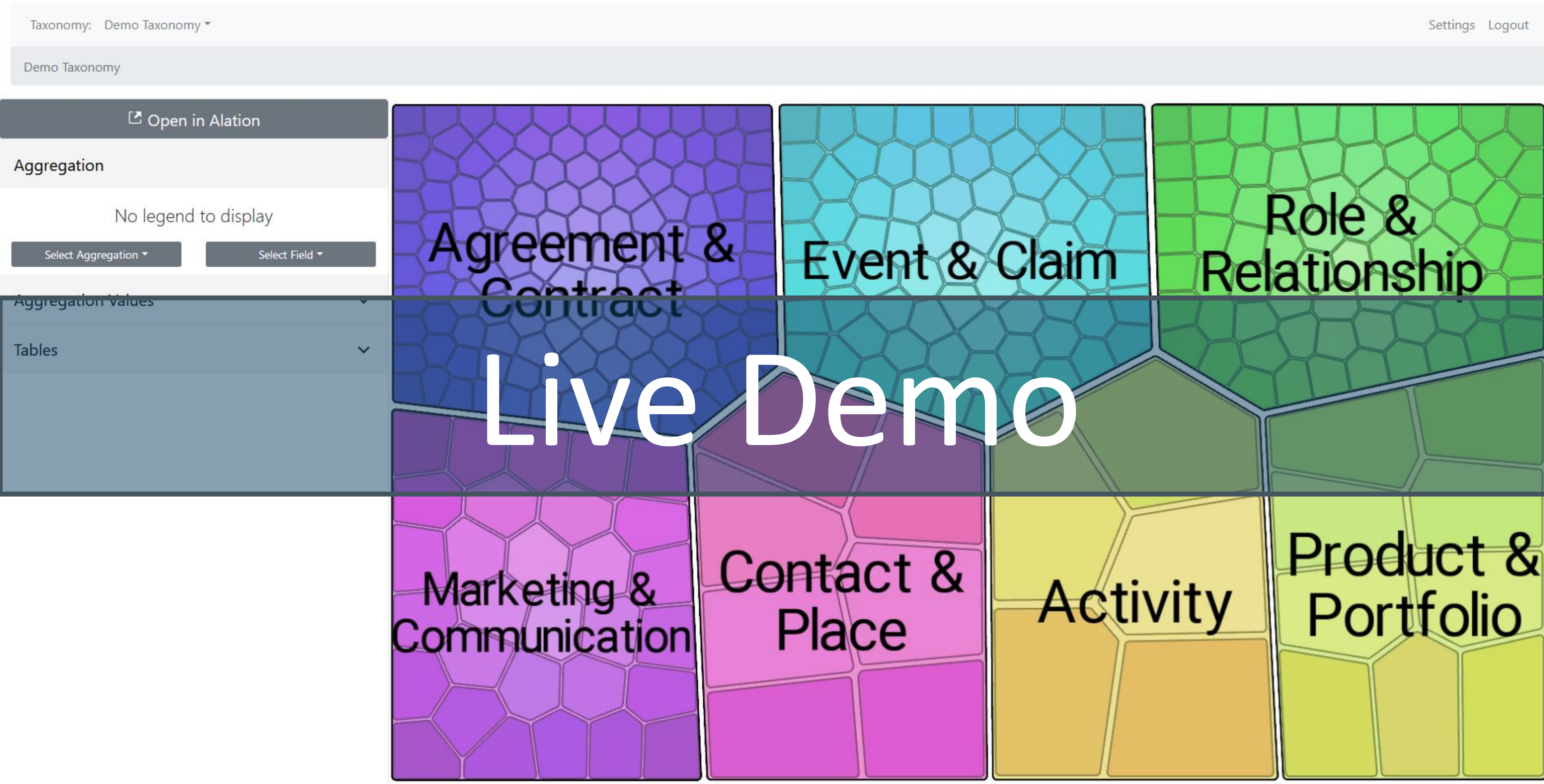
Self Service Analytics
[Data Analysts, Data Scientists]
- Google for enterprise data assets
- Which data is used in my model / usecase?
- Visualize Lineage

Data Governance
[Data Stewards]
- Associate Business glossary to technical objects
- Verify business to technical lineage
- Track key data elements compliance

Data Asset Management
[Architects, Developers]
- Analyze Lineage
- View transformation Logic
- Data asset and BI usage



Live Demo



Taxonomy: Demo Taxonomy
Settings Logout
Demo Taxonomy
Open in Alation
Aggregation
No legend to display
Select Aggregation
Select Field
Aggregation Values
Tables
Agreement & Contract
Event & Claim
Role & Relationship
Marketing & Communication
Contact & Place
Activity
Product & Portfolio

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
   - Atlas Präsentation
   - Data Selfservice
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

Agenda

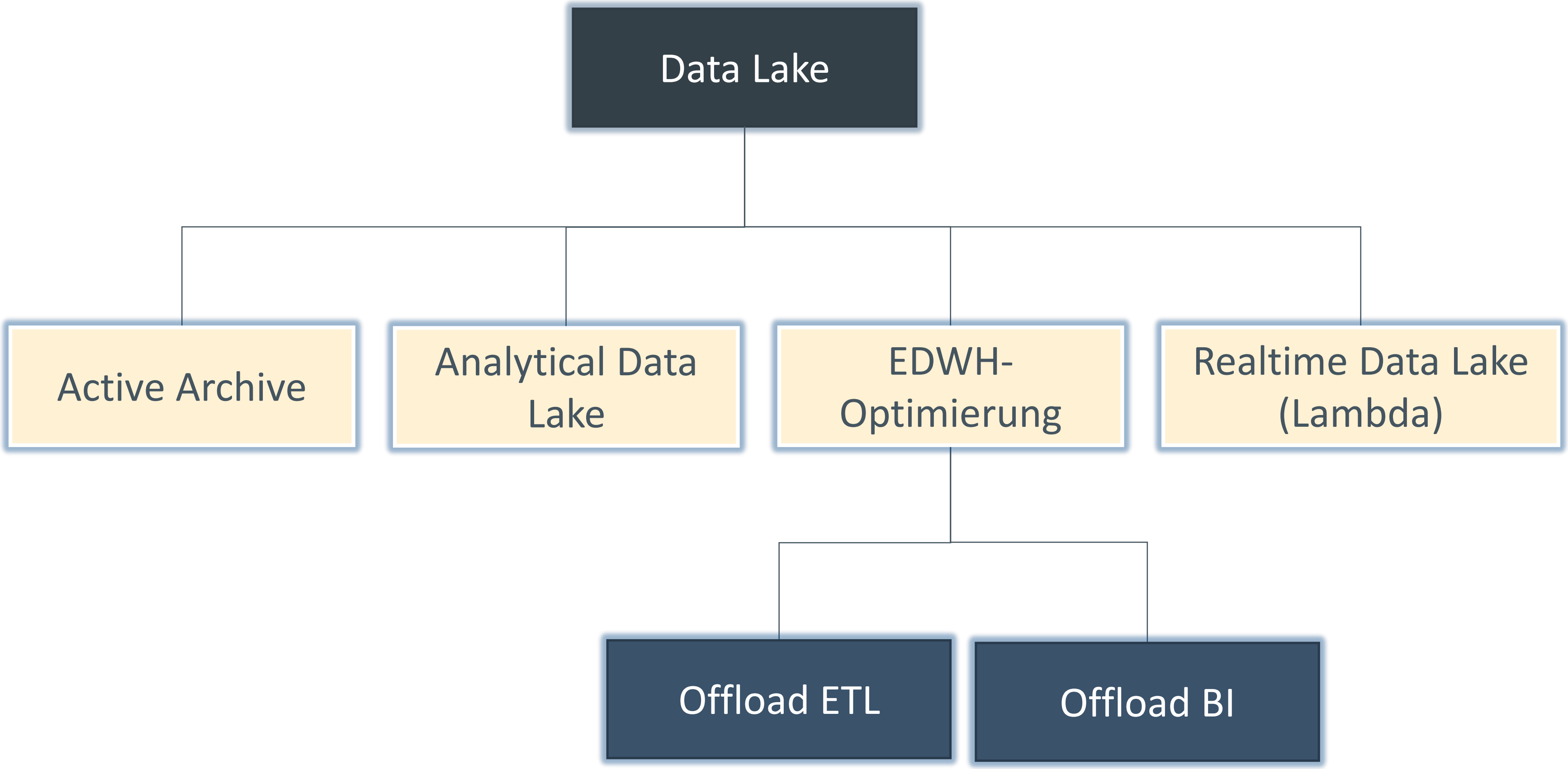
1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   – Data Sovereignty
   – Business Glossar
   – Alation Präsentation – Business Glossar anlegen
   – SVML-Präsentation
   – Data Catalog und Data Lake
     - Grundlagen Data Lake
     - Data Catalog in Data Lake
   – Atlas Präsentation
   – Data Selfservice
   – IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

What is a Data Lake

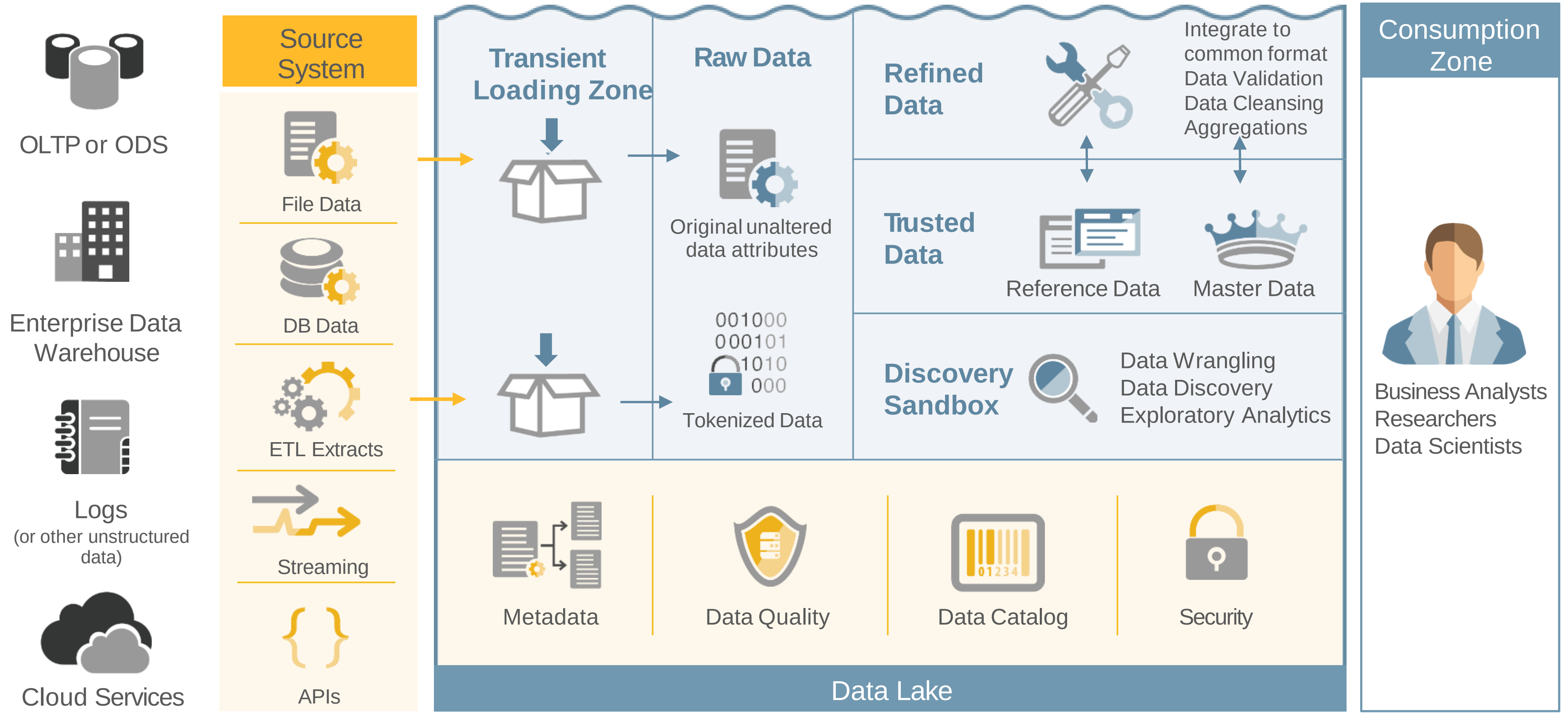
- Storage repository that can store large amounts of structured, semi-structured, and unstructured data
- Place to store every type of data in its native format with no fixed limits on account size or file
- High data quantity to increase analytic performance and native integration
- For better data consumption data is curated over different zones

Important Aspects of a Data Lake

Data Lake Types



Concept of a Data Lake



Tasks of the Different Data Lake Zones

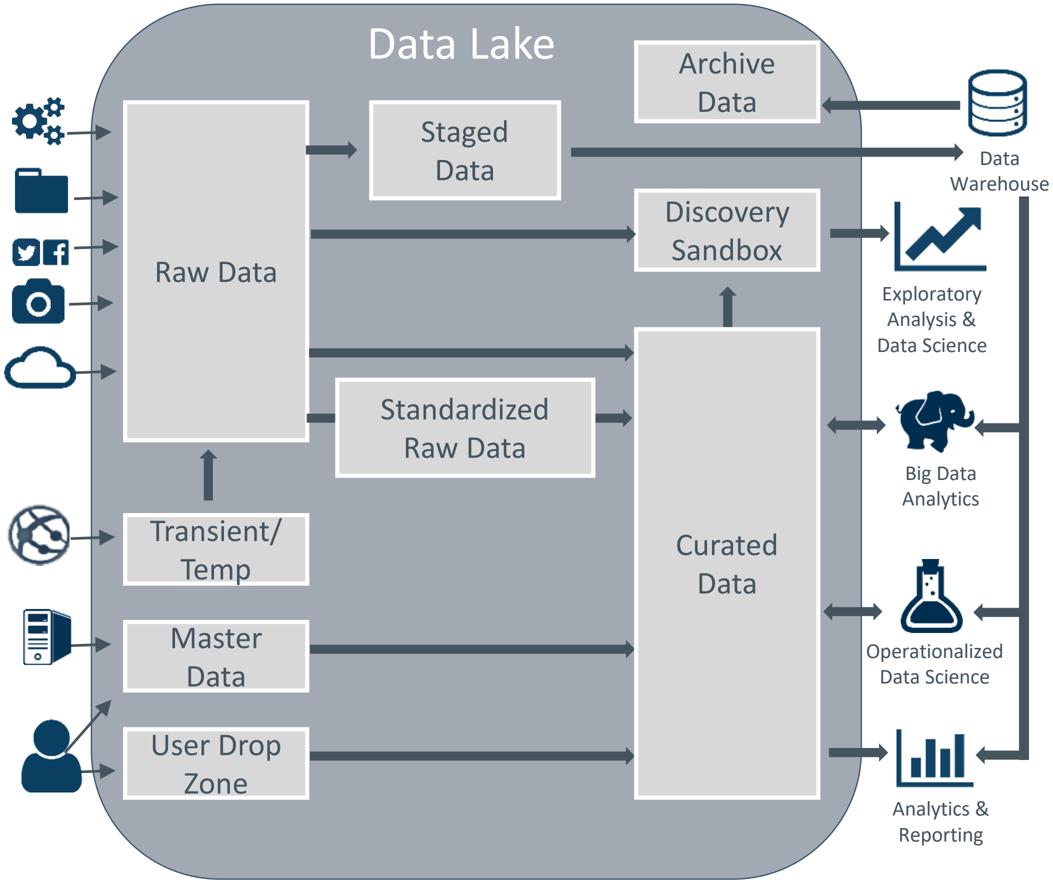
Raw Data Zone
- Exact copy of source data in native format (aka master dataset in the batch layer)
- Immutable to change
- History retained indefinitely
- Data access is highly limited to few people
- Everything downstream can be regenerated from raw

Transient/Temp Zone
- Selectively utilized
- Separation of „new data“ to ensure data consistency
- Transient low-latency data (aka speed layer)
- Data quality validations

Master Data Zone
- Reference data

User Drop Zone
- Manually generated data

Staged Data Zone
- Data staged for a specific purpose or application



Metadata | Security | Governance | Information Management

Standardized Raw Data
- Raw data which varies format or schema, such as JSON which is standardized into columns & rows (aka „semantic normalization“)
- File consolidations of data (i.e. to overcome performance issues with many small files)

Archive Data Zone
- Active archive of aged data, available for querying when needed

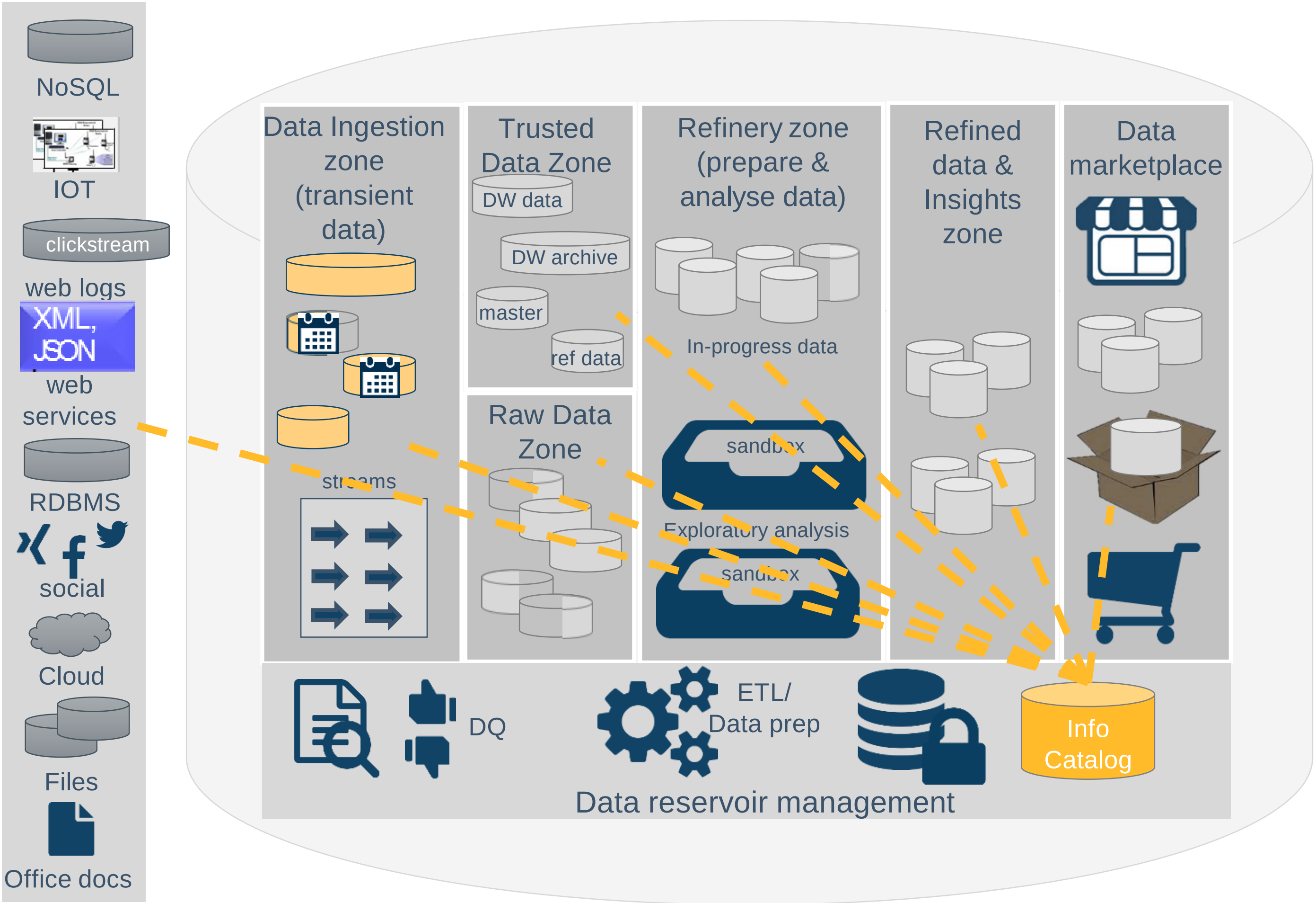
Discovery Sandbox
- Workspace for exploratory data science & analytics
- Valuable efforts are productionized to the curated data zone

Curated Data Zone
- Cleansed and transformed data, organized for optional data delivery (aka serving layer)
- Supports self-service
- Standard security, change management and governance

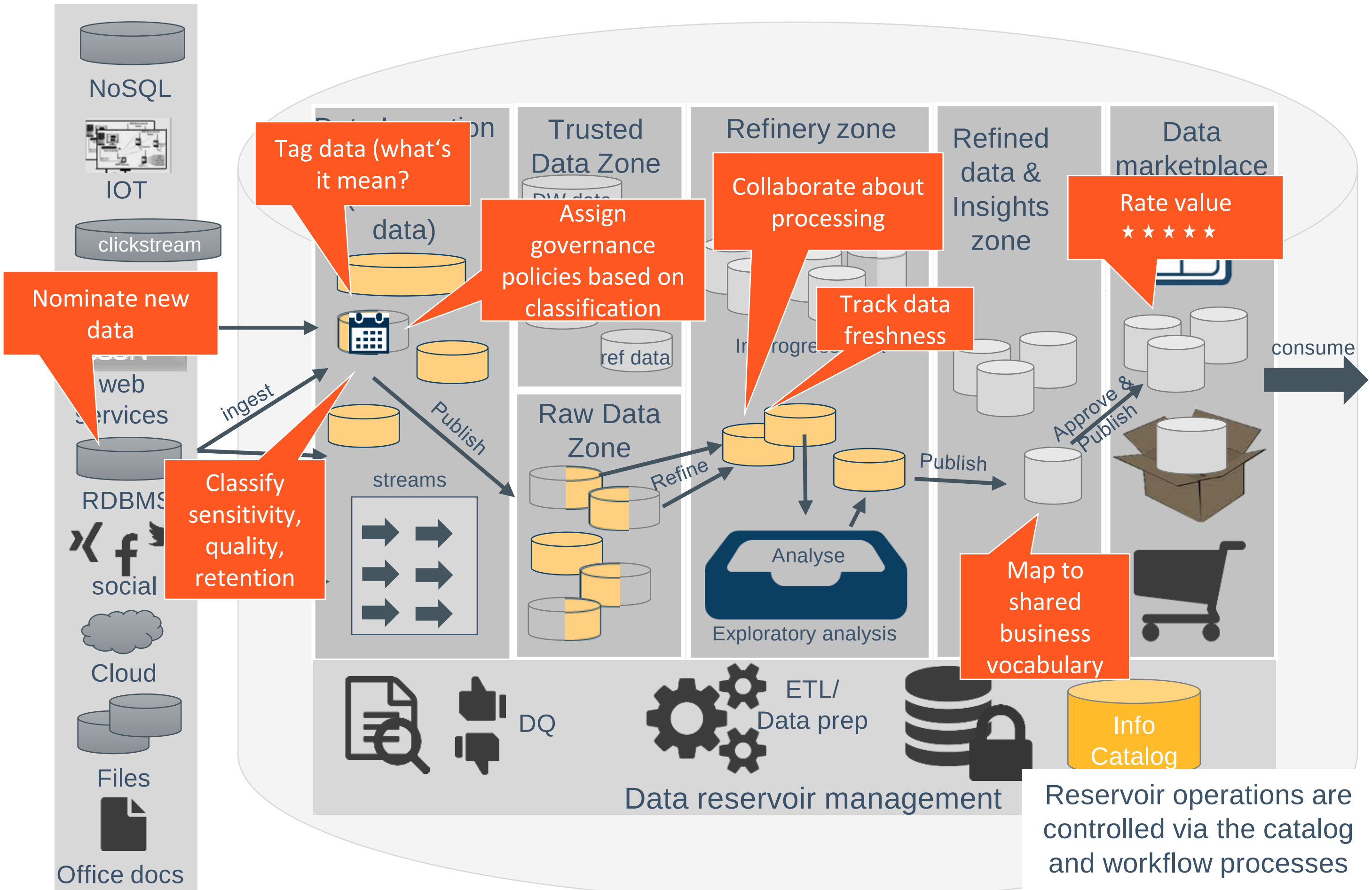
Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
     - Grundlagen Data Lake
     - Data Catalog in Data Lake
   - Atlas Präsentation
   - Data Selfservice
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

Organising Data In A Reservoir

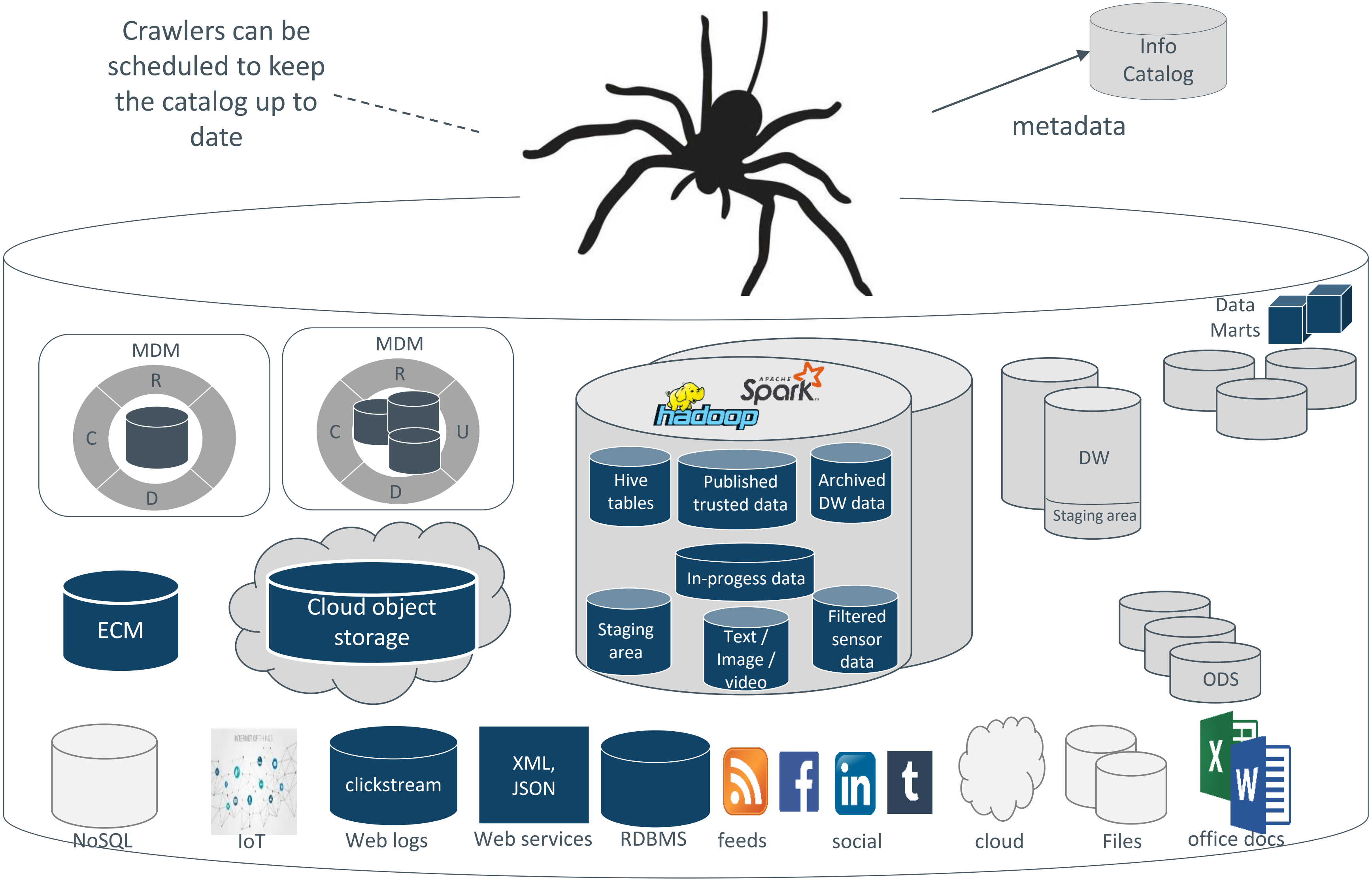


The Information Production Process Is A Production Line That Spans Data Lake Zones

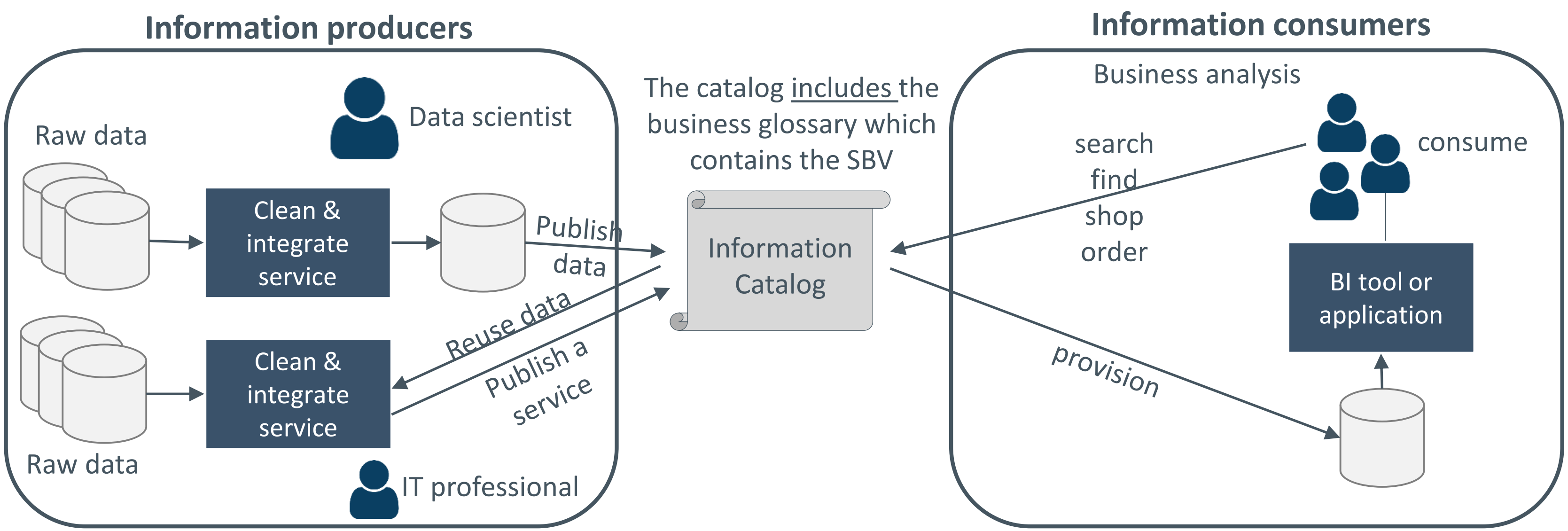


Reservoir operations are controlled via the catalog and workflow processes

Catalog Crawls the Data Lake

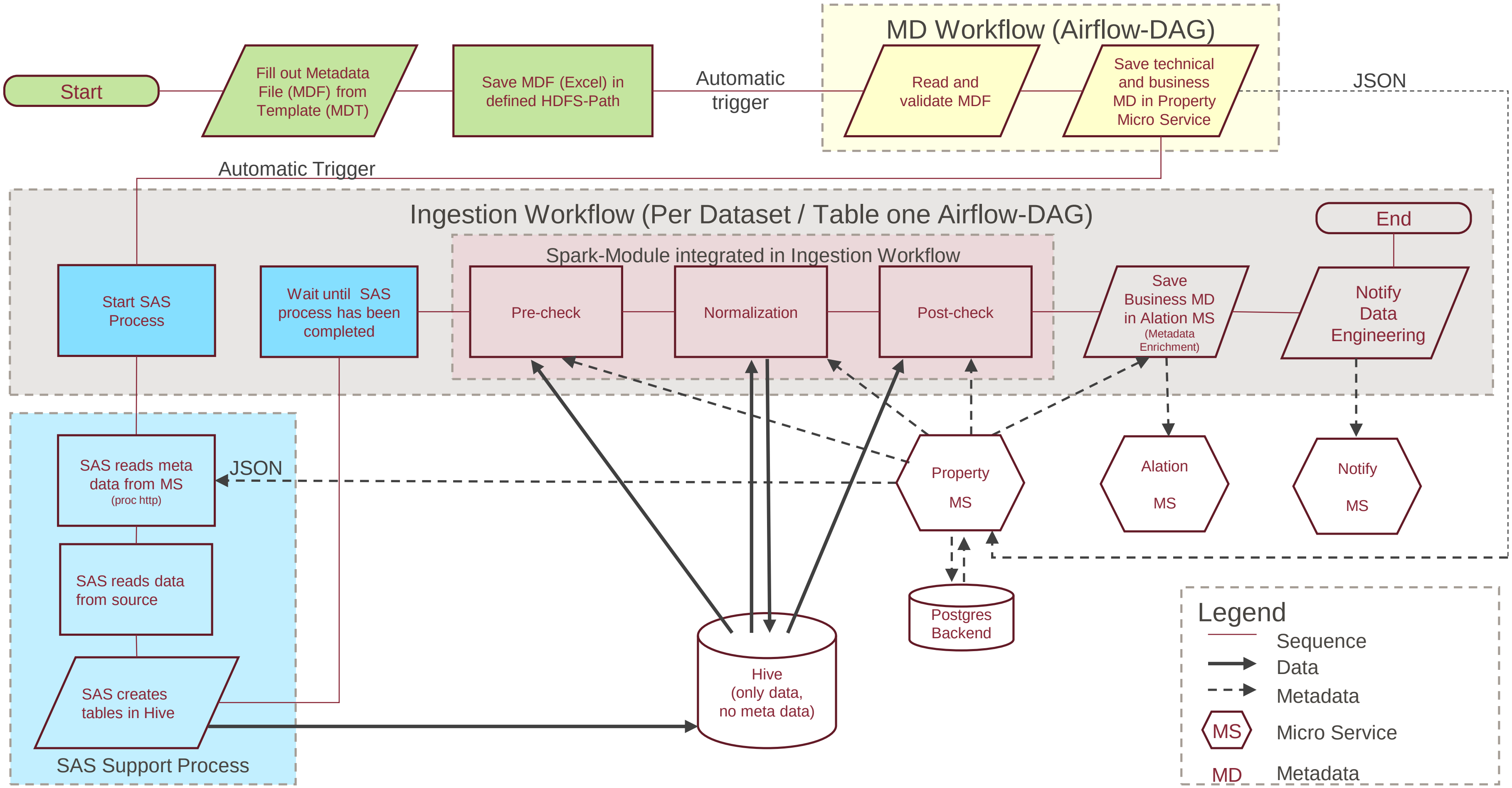


Why is a shared Business Vocabulary Relevant in a Data Lake?

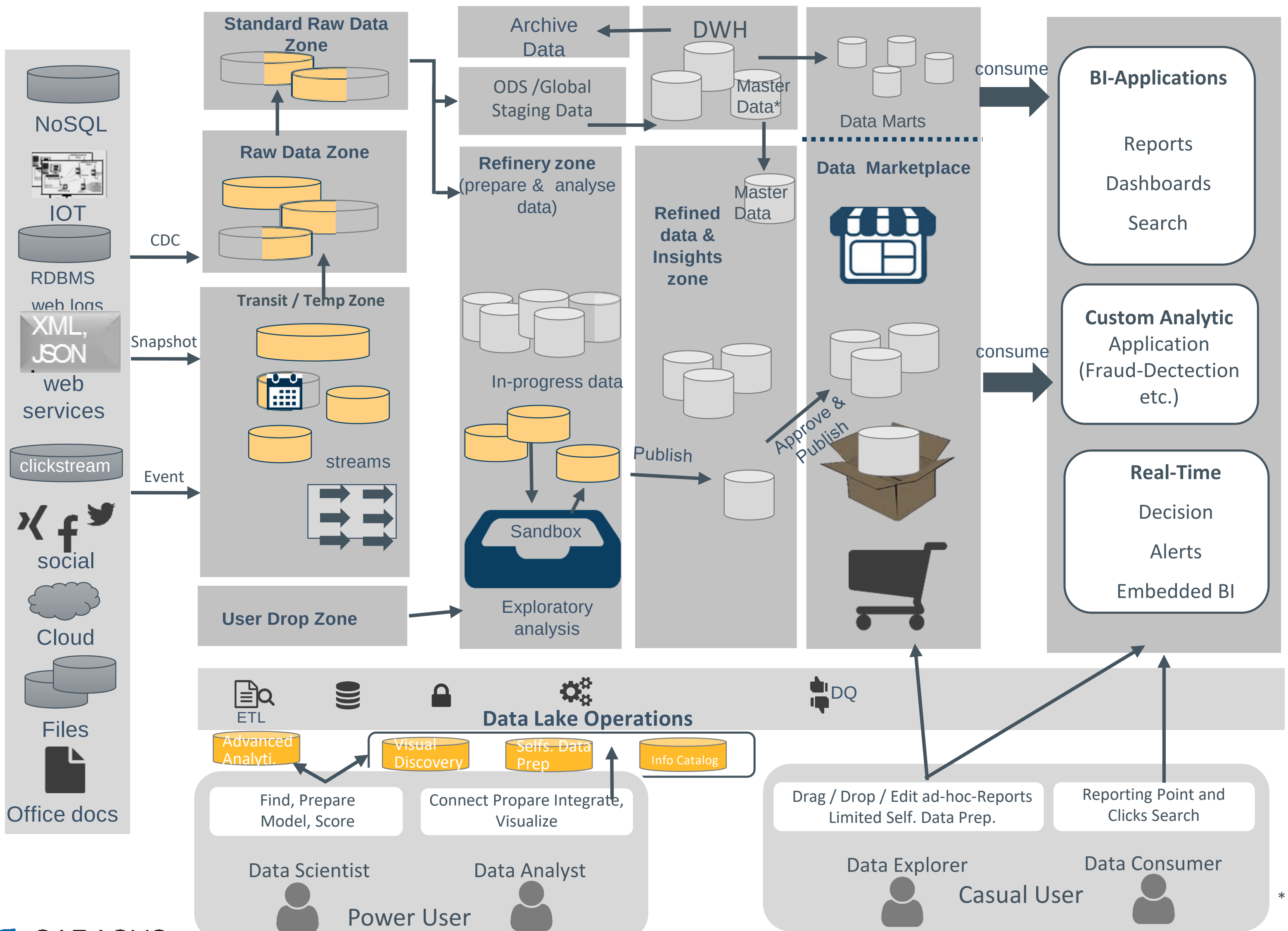


1. All trusted data published in the information catalog should be described (marked up) using SBV data names and definitions
2. All trusted data services published in the information catalog should produce data with SBV data names and data types when they execute
3. If totally new data is produced from a data lake, the SBV should be extended

Example: Data Pipeline and Data Catalog



Gesamtheitl. Analytische Plattform Informationsveredlung über Zonen



* Kein MDM-Tool

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   – Data Sovereignty
   – Business Glossar
   – Alation Präsentation – Business Glossar anlegen
   – SVML-Präsentation
   – Data Catalog und Data Lake
   – **Atlas Präsentation**
   – Data Selfservice
   – IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

SARACUS CONSULTING

Demo: Apache Atlas

Atlas Summary

Main features:

- De facto standard for Metadata Management within Hadoop Distributions
  - Will be the central Metadata Management tool in the merged Cloudera / Hortonworks Platform
- Automatic Lineage detection and creation for Hive, Sqoop, Storm, HBase and Kafka via Hooks and Bridges
- Open, very flexible and extensible Framework
- Tightly coupled with Apache Ranger (will be standard tool for Authorization within merged Cloudera / HW distro) for metadata-based authorization rules
- Tag based policies

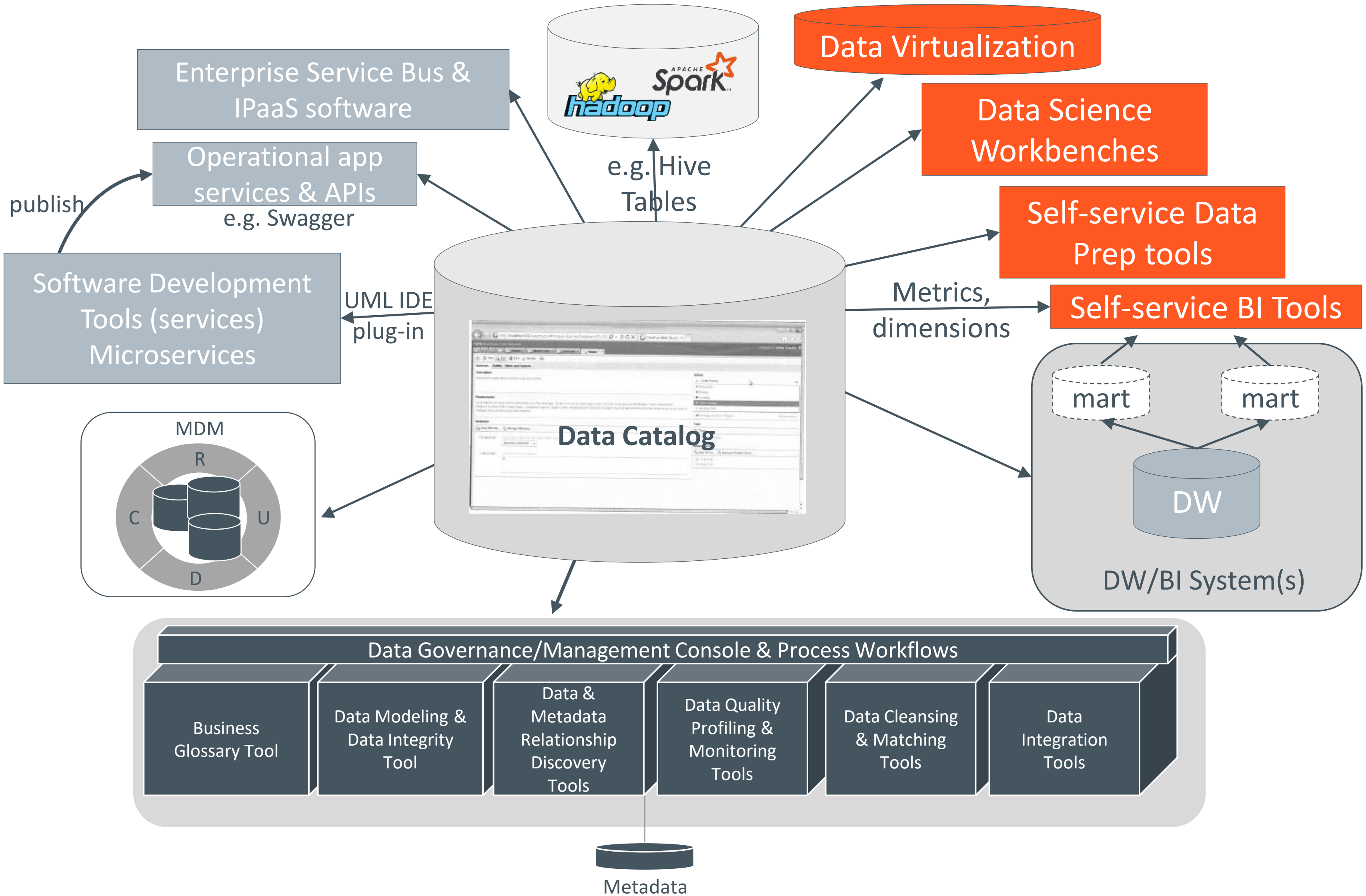
Technical aspects

- Metadata storage is based on HBase with Titan GraphDB, Kafka and Solr
- REST-API for Metadata Extraction and editing
- Java API
- Metadata CDC via Kafka (all changes to metadata are reported in a Kafka Topic)

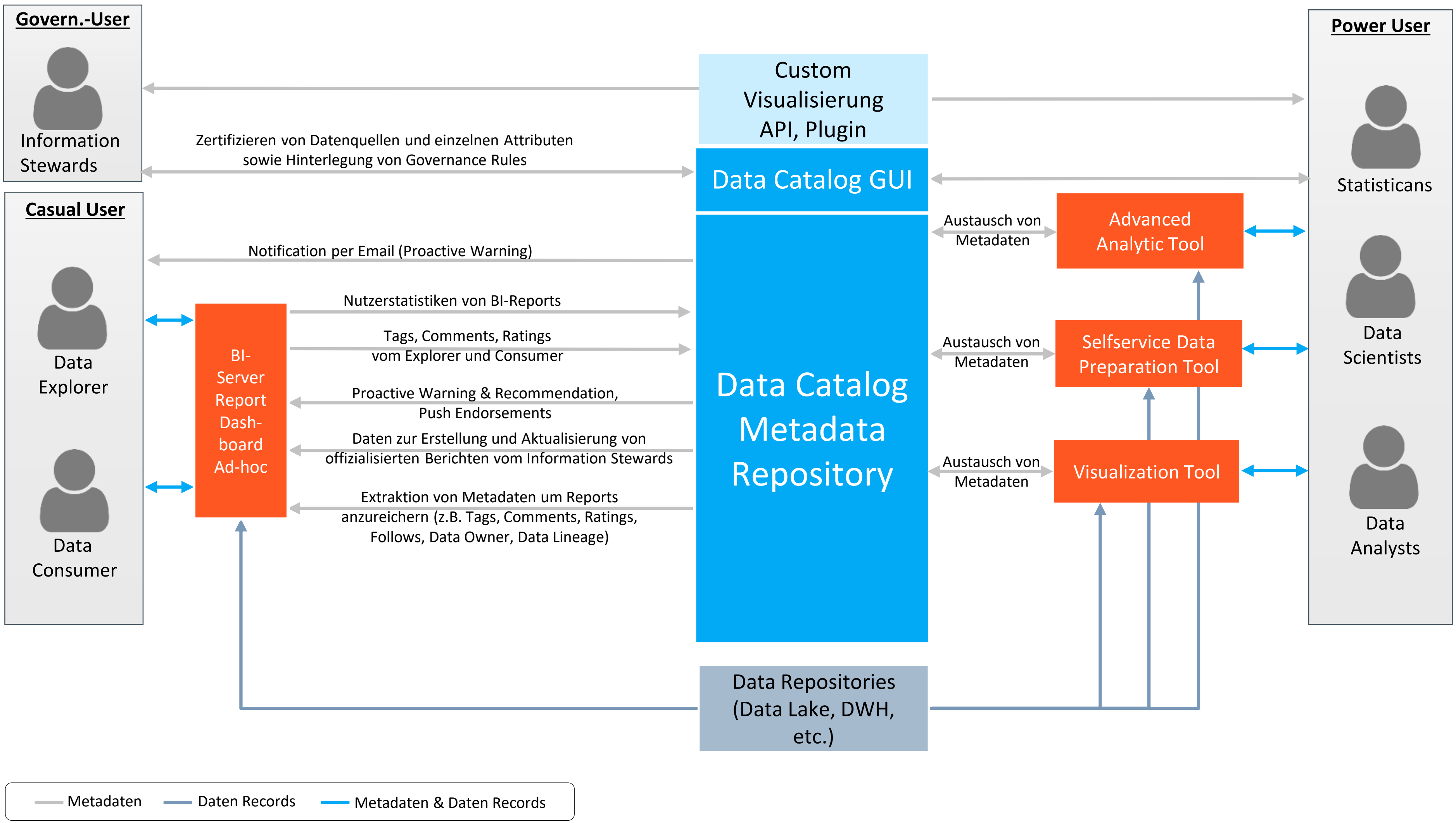
Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   – Data Sovereignty
   – Business Glossar
   – Alation Präsentation – Business Glossar anlegen
   – SVML-Präsentation
   – Data Catalog und Data Lake
   – Atlas Präsentation
   – Data Selfservice
   – IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

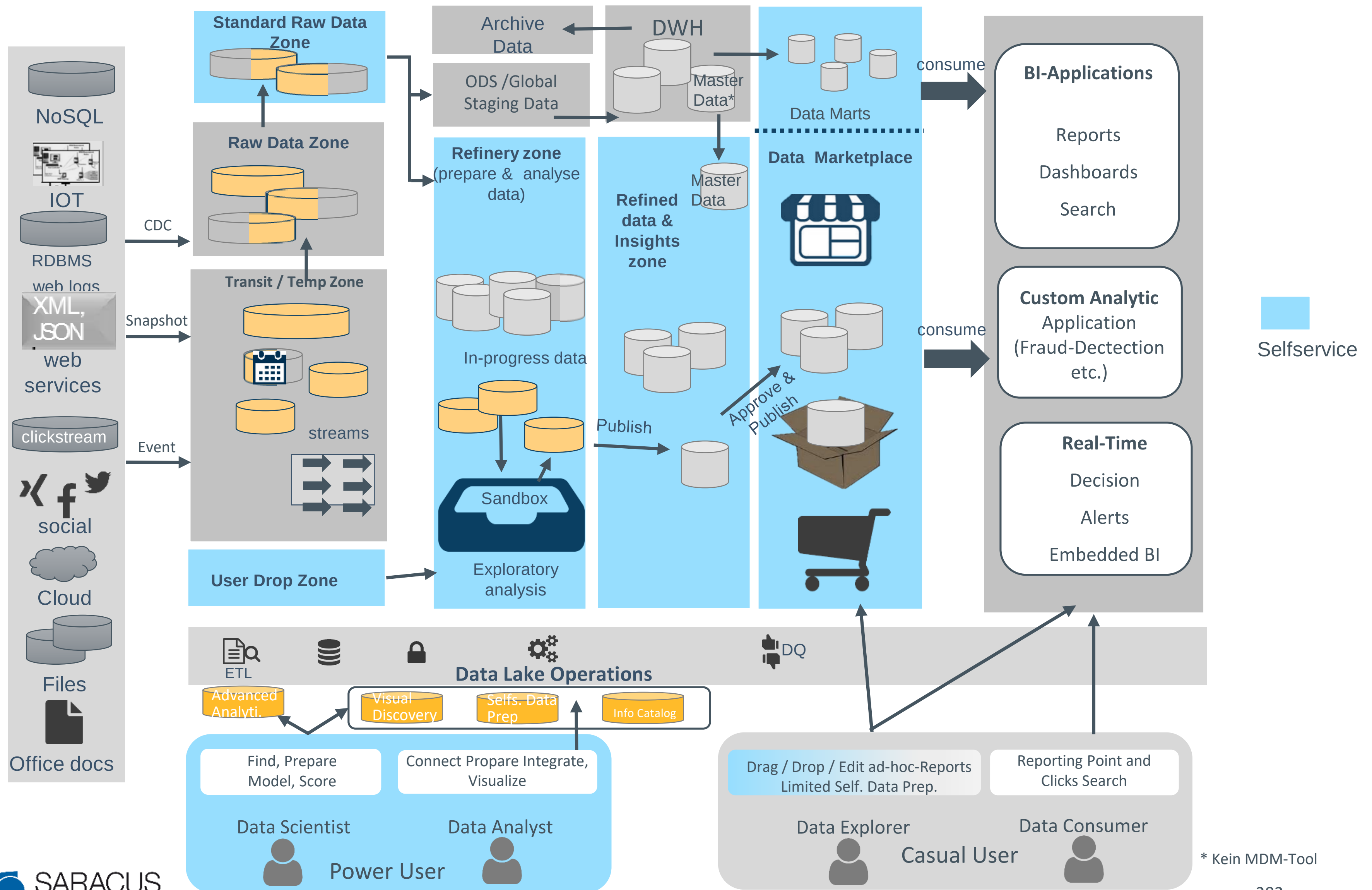
Selfservice Data Prep. & Analytics & Data Virtualization



Metadaten für Data Self Service



Selfservice in a Data Lake

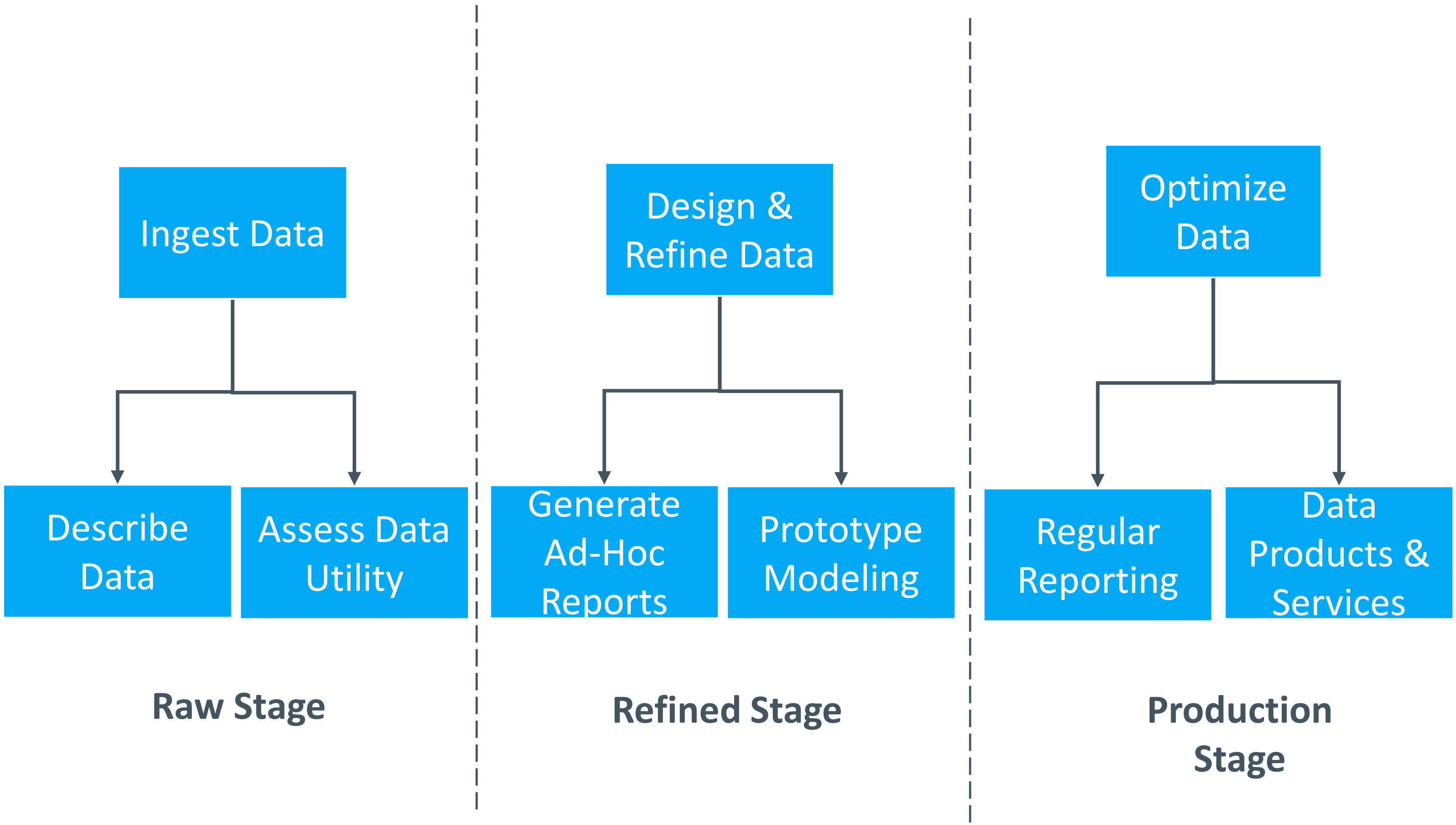


* Kein MDM-Tool

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
   - Atlas Präsentation
   - Data Selfservice
     - Selfservice Data Preparation
     - Selfservice Data Analytics
     - Data Virtualization
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

A holistic workflow framework for Selfservice Data Preparation



Raw Data Stage: The relative importance of each type of transformation and profiling

Raw Data Stage	Ingesting Data	**Structuring** Enriching Cleaning **Individual Value Based Profiling** Set Based Profiling
	Generating Generic Metadata	**Structuring** Enriching Cleaning **Individual Value Based Profiling** **Set Based Profiling**
	Generating Proprietary Metadata	Structuring **Enriching** **Cleaning** Individual Value Based Profiling **Set Based Profiling**

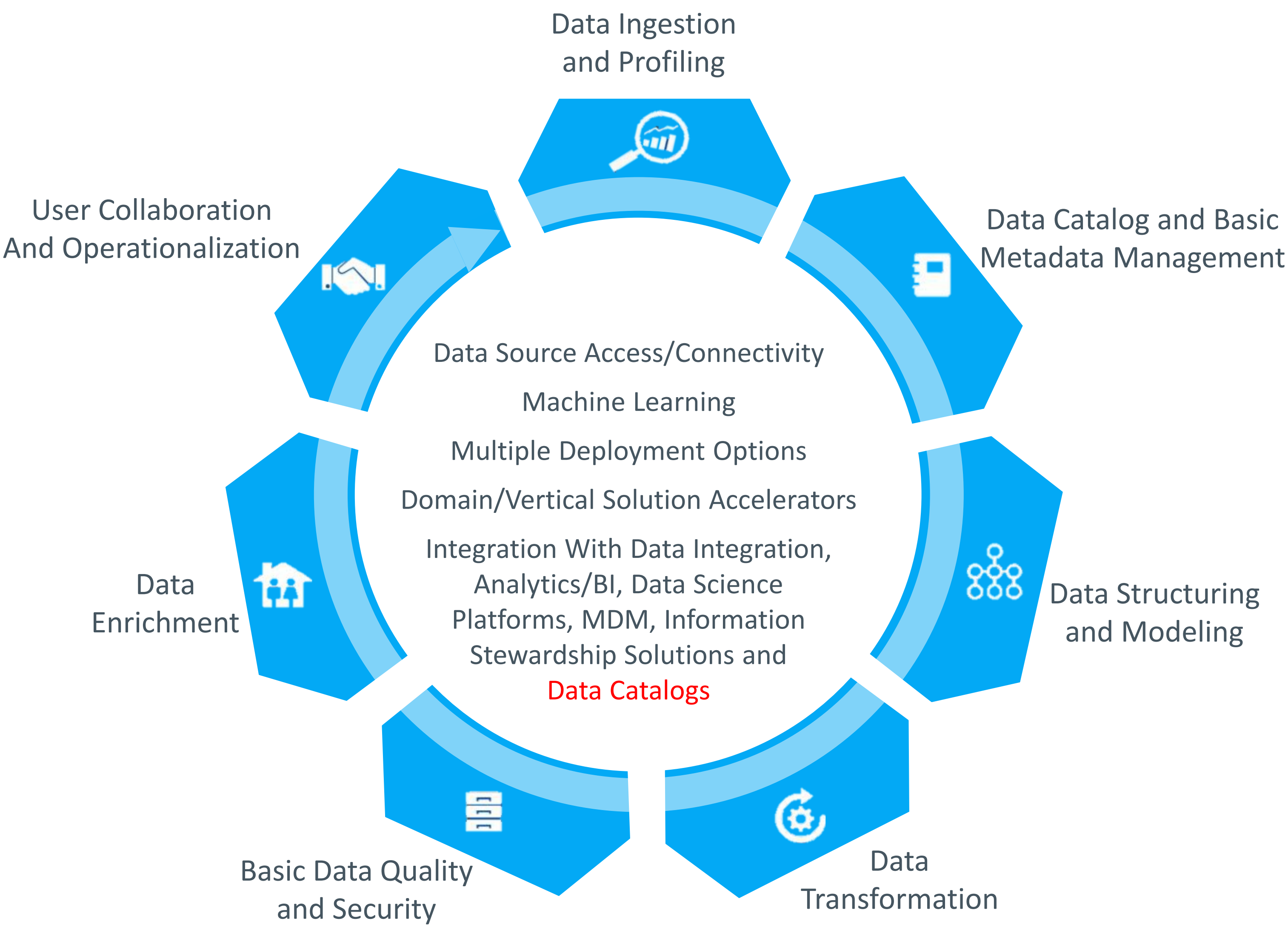
Refined Data Stage: The relative importance of each type of transformation and profiling

Refined Data Stage		
	Designing and Building Refined Data	**Structuring** **Enriching** **Cleaning** **Individual Value Based Profiling** **Set Based Profiling**
	Ad-Hoc Reporting	**Structuring** Enriching Cleaning Individual Value Based Profiling Set Based Profiling
	Exploratory Modeling and Forecasting	**Structuring** Enriching Cleaning Individual Value Based Profiling Set Based Profiling

Production Data Stage: The relative importance of each type of transformation and profiling

Production Data Stage		
	Designing and Building Optimized Data	Structuring Enriching Cleaning **Individual Value Based Profiling** **Set Based Profiling**
	Regular Reporting	**Structuring** Enriching Cleaning **Individual Value Based Profiling** **Set Based Profiling**
	Building Products and Services	**Structuring** Enriching Cleaning **Individual Value Based Profiling** **Set Based Profiling**

What is Self-Service Data Prep?
Key Capabilities of a Modern Data Preparation Tool



Self-Service Data Prep Landscape

talend

TRIFACTA

alteryx

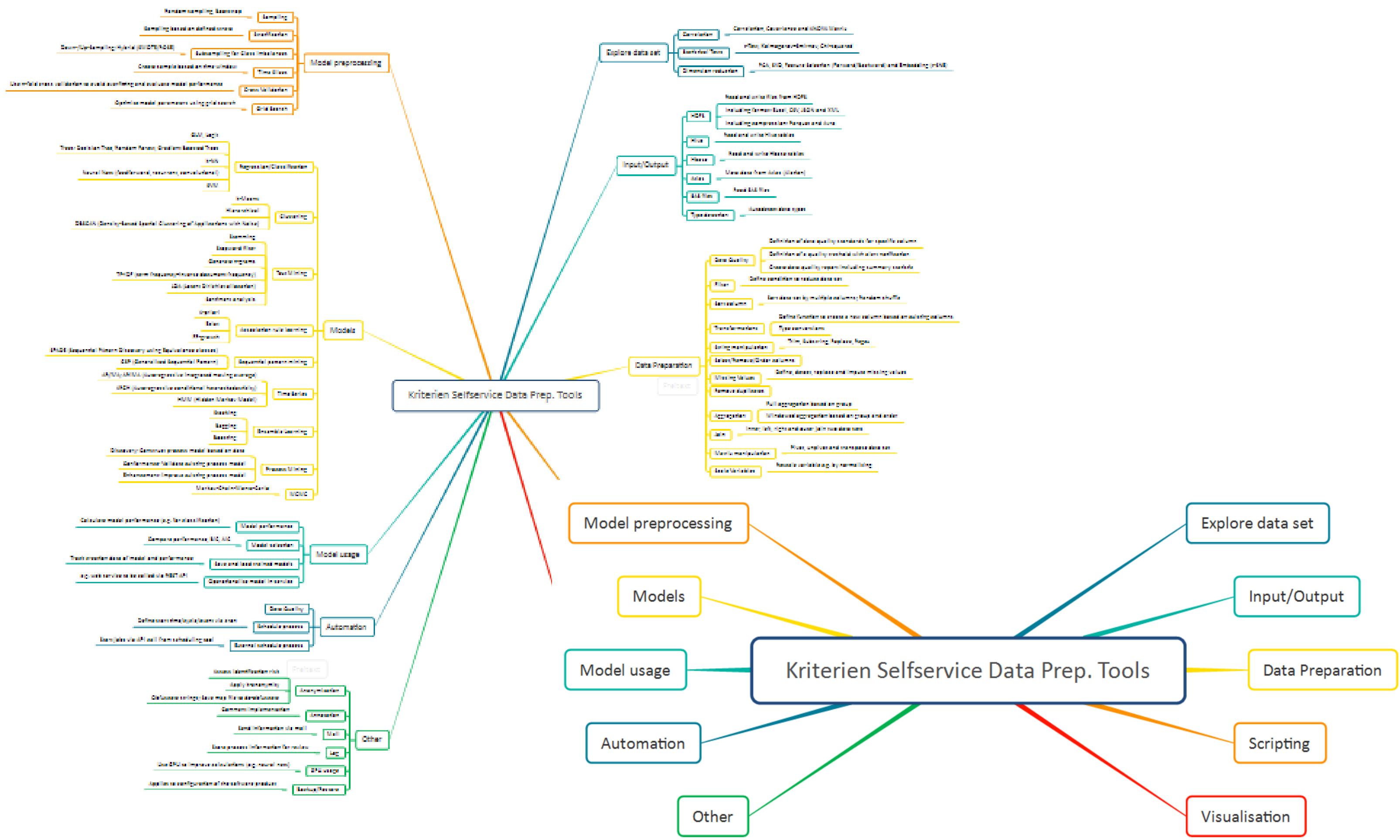
pentaho®

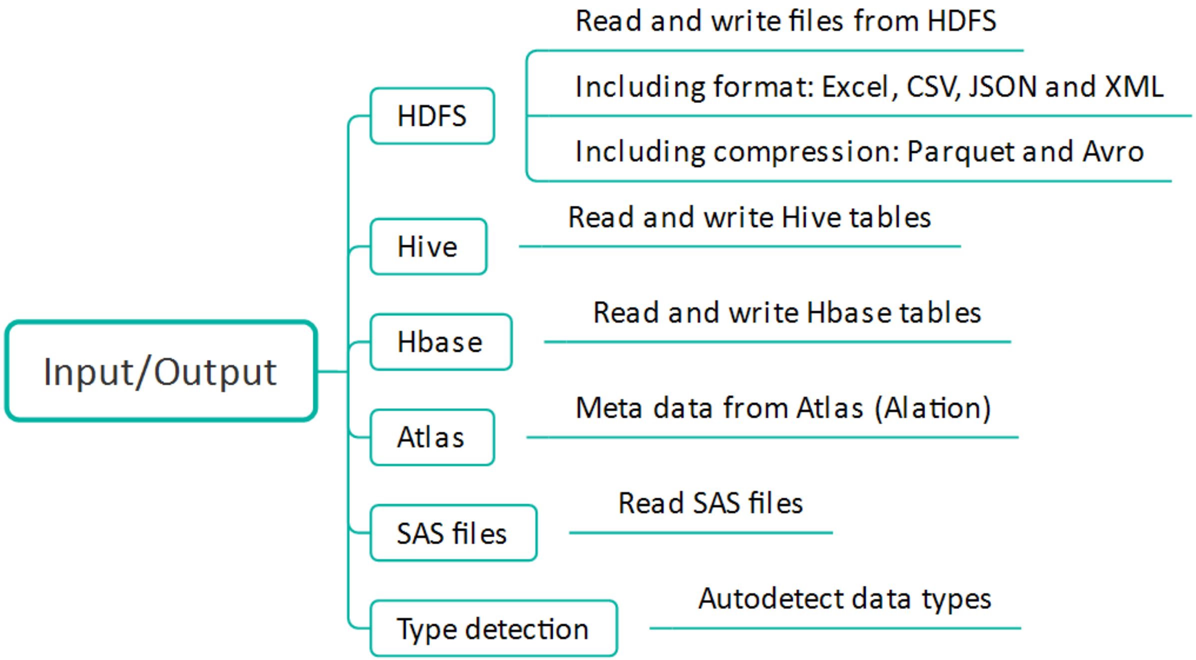
Paxata

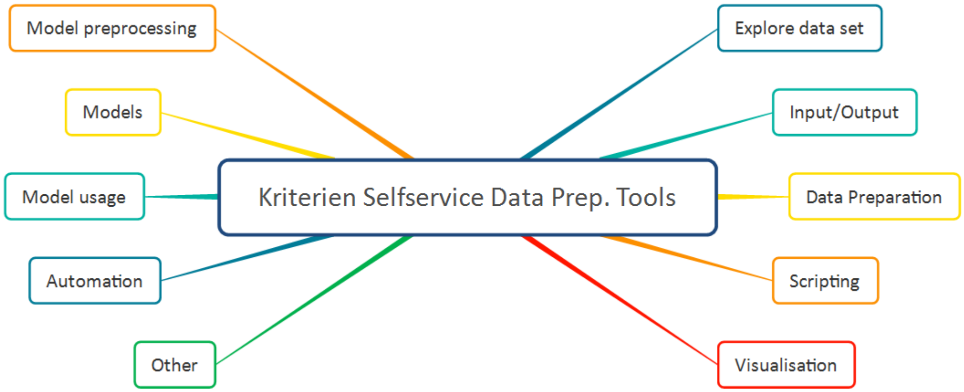
tamr

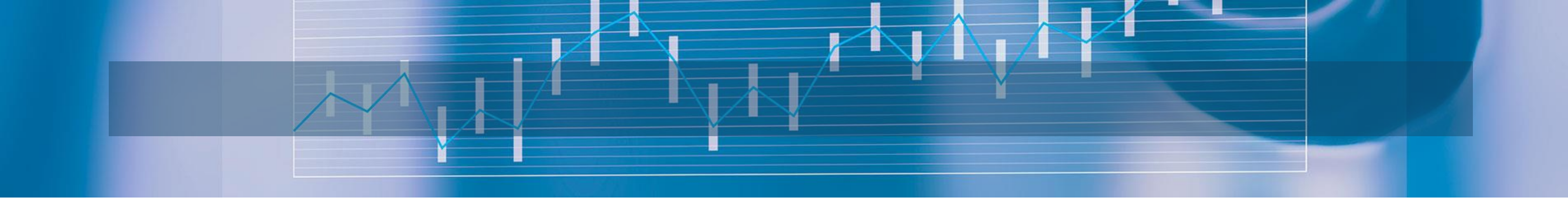
unifi

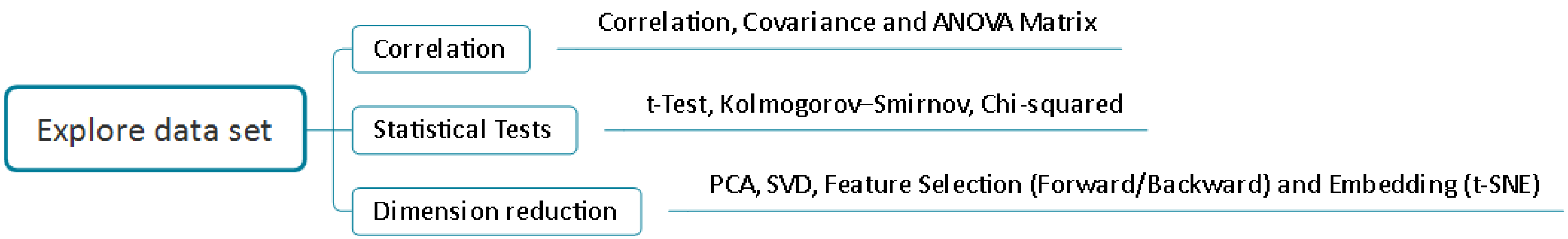
Kriterien für Toolevaluation



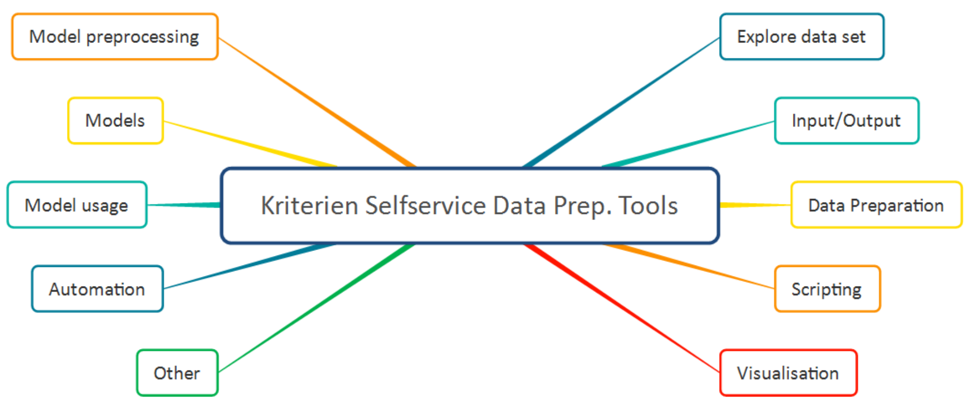

Input/Output
HDFS
Read and write files from HDFS
Including format: Excel, CSV, JSON and XML
Including compression: Parquet and Avro
Hive
Read and write Hive tables
Hbase
Read and write Hbase tables
Atlas
Meta data from Atlas (Alation)
SAS files
Read SAS files
Type detection
Autodetect data types


Kriterien Selfservice Data Prep. Tools
Model preprocessing
Models
Model usage
Automation
Other
Explore data set
Input/Output
Data Preparation
Scripting
Visualisation

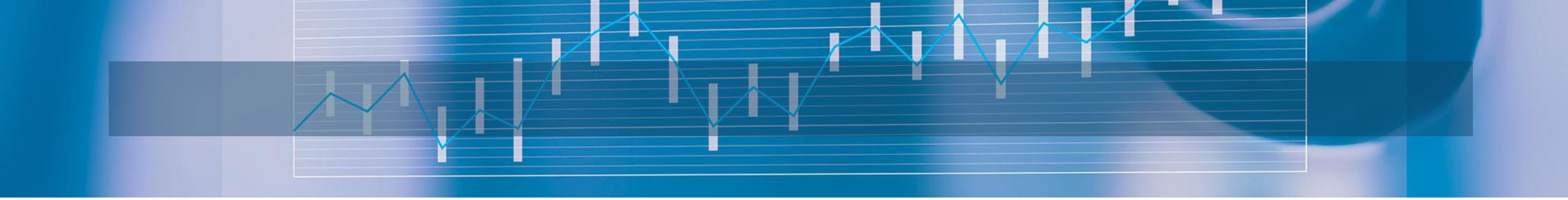




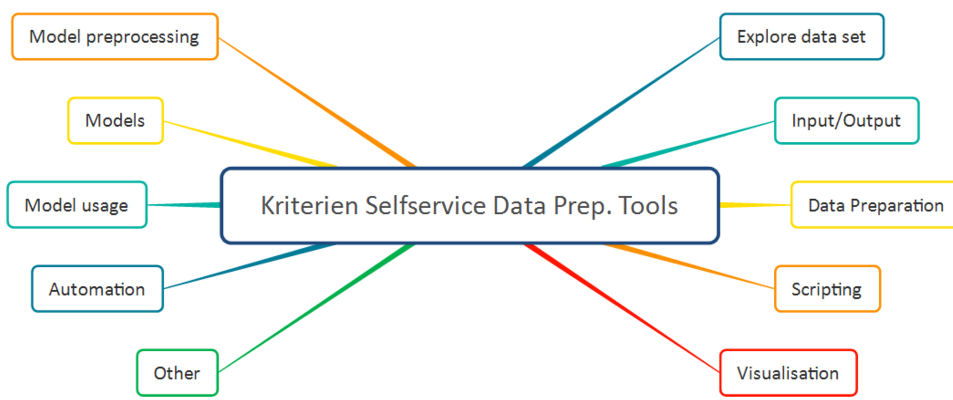
Explore data set
Correlation
Correlation, Covariance and ANOVA Matrix
Statistical Tests
t-Test, Kolmogorov–Smirnov, Chi-squared
Dimension reduction
PCA, SVD, Feature Selection (Forward/Backward) and Embedding (t-SNE)

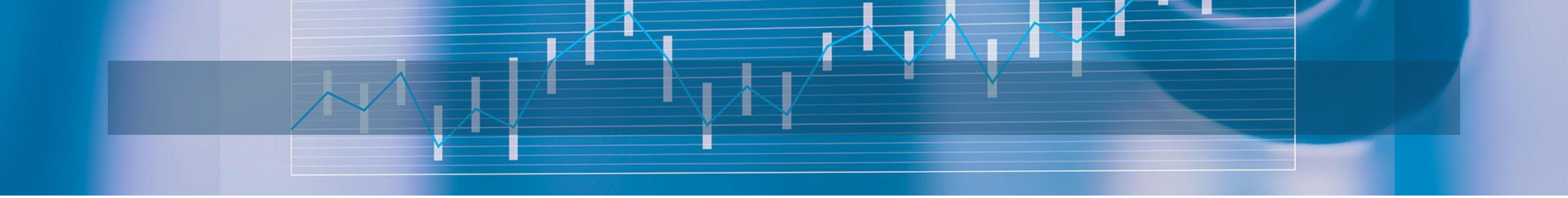


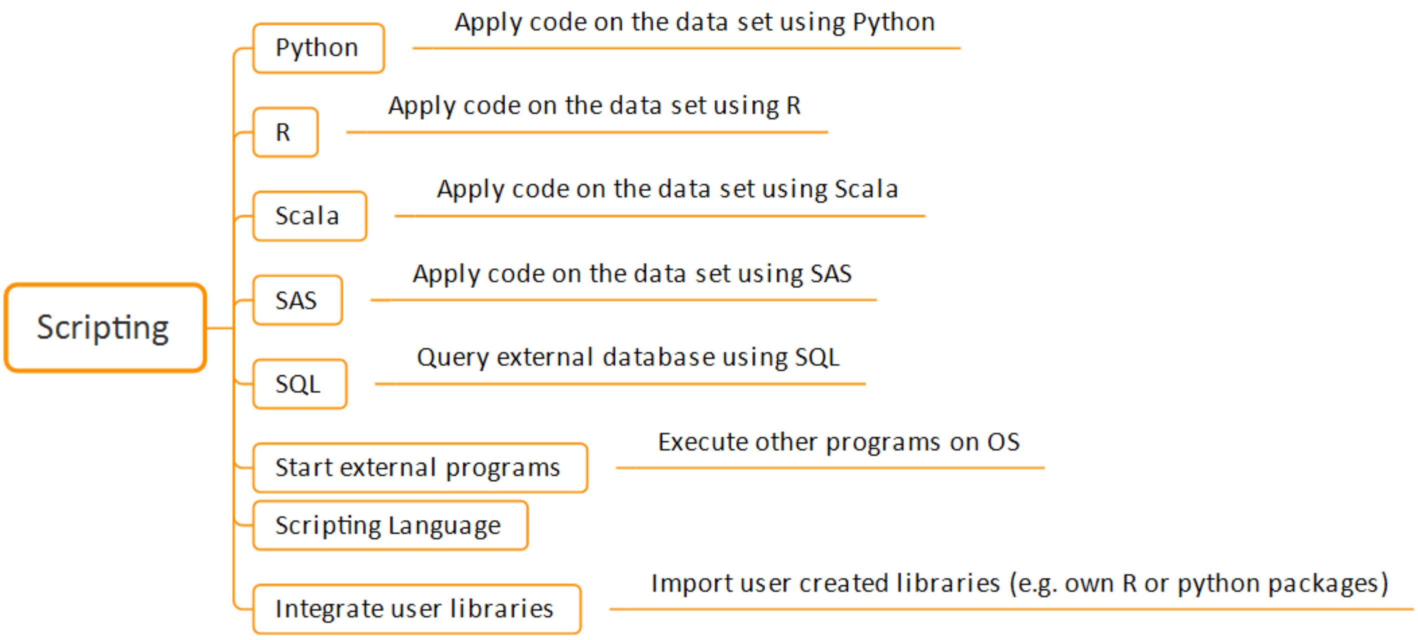
Kriterien Selfservice Data Prep. Tools
Model preprocessing
Models
Model usage
Automation
Other
Explore data set
Input/Output
Data Preparation
Scripting
Visualisation

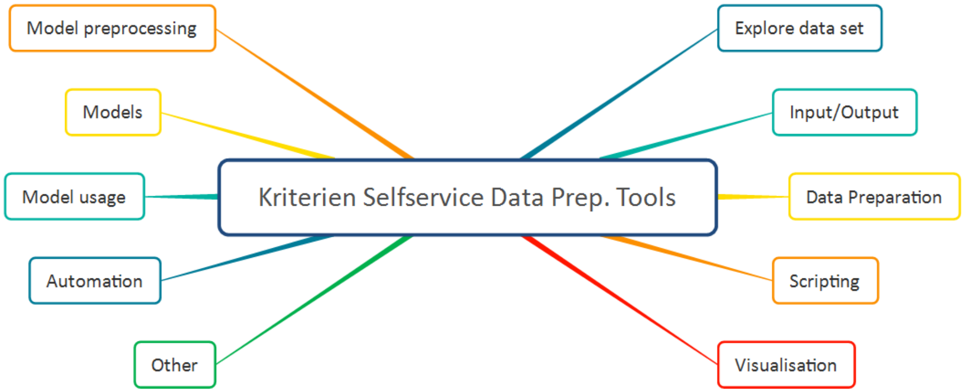


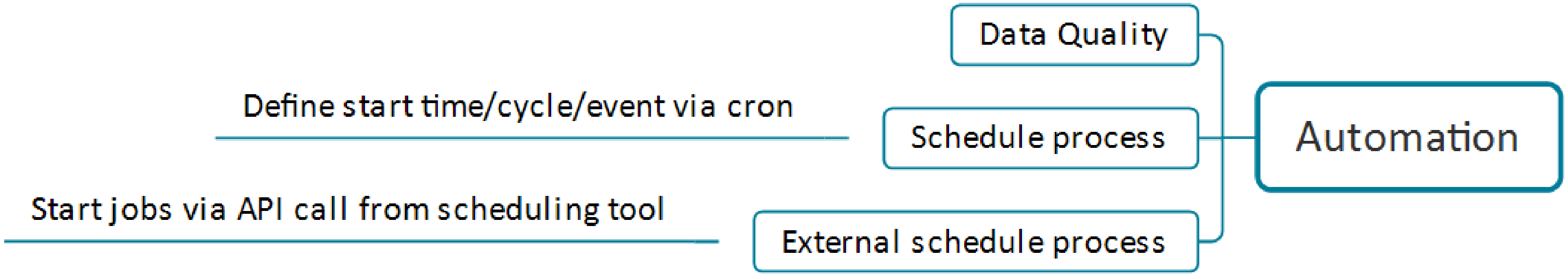

Data Preparation
Data Quality
Definition of data quality standards for specific column
Definition of a quality treshold with alert notification
Create data quality report including summary statistic
Filter
Define condition to reduce data set
Sort column
Sort data set by multiple columns; Random shuffle
Transformations
Define function to create a new column based on existing columns
Type conversions
String manipulation
Trim, Substring, Replace, Regex
Select/Remove/Order columns
Missing Values
Define, detect, replace and impute missing values
Remove duplicates
Aggregation
Full aggregation based on group
Windowed aggregation based on group and order
Join
inner, left, right and outer join two data sets
ıspose data set
nalizing

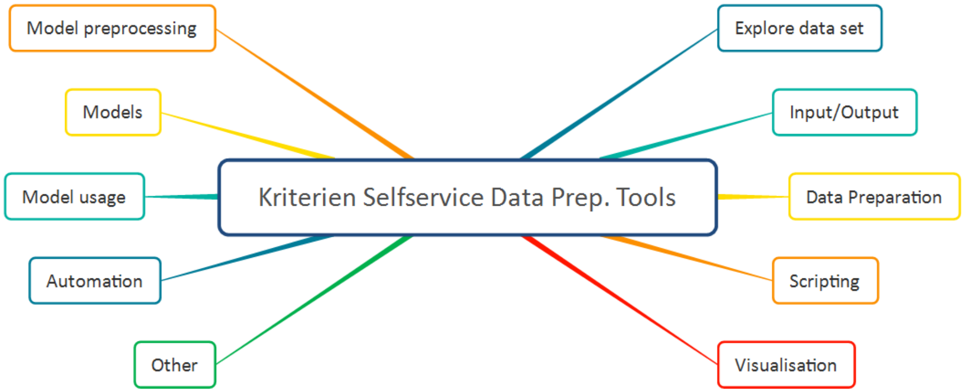

Kriterien Selfservice Data Prep. Tools
Model preprocessing
Models
Model usage
Automation
Other
Explore data set
Input/Output
Data Preparation
Scripting
Visualisation

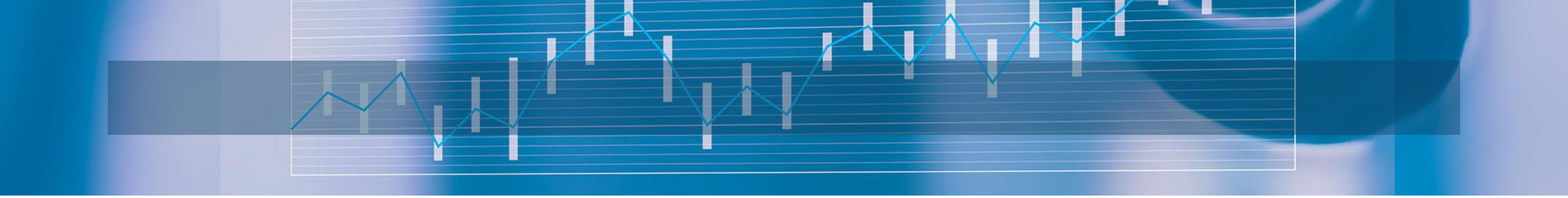


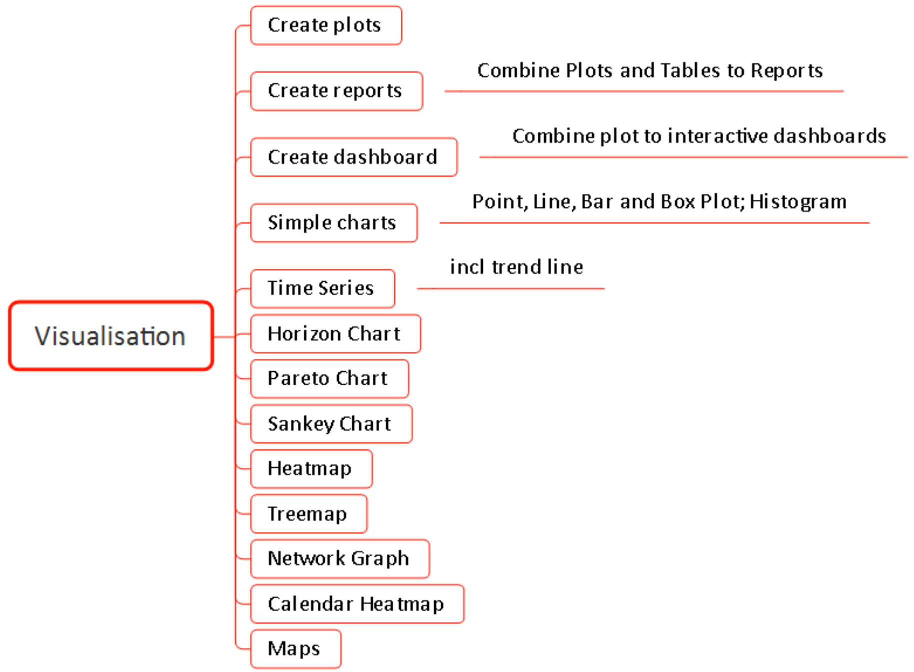

Scripting
Python
Apply code on the data set using Python
R
Apply code on the data set using R
Scala
Apply code on the data set using Scala
SAS
Apply code on the data set using SAS
SQL
Query external database using SQL
Start external programs
Execute other programs on OS
Scripting Language
Integrate user libraries
Import user created libraries (e.g. own R or python packages)


Kriterien Selfservice Data Prep. Tools
Model preprocessing
Models
Model usage
Automation
Other
Explore data set
Input/Output
Data Preparation
Scripting
Visualisation

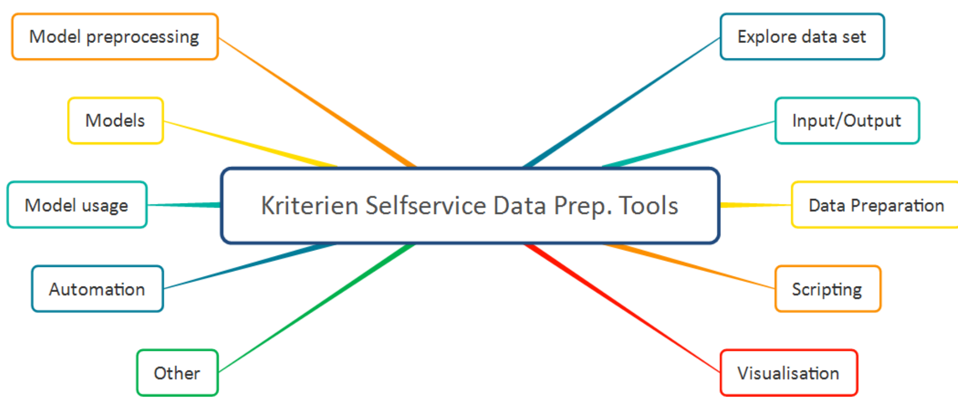

Data Quality
Define start time/cycle/event via cron
Schedule process
Automation
Start jobs via API call from scheduling tool
External schedule process


Model preprocessing
Explore data set
Models
Input/Output
Model usage
Kriterien Selfservice Data Prep. Tools
Data Preparation
Automation
Scripting
Other
Visualisation

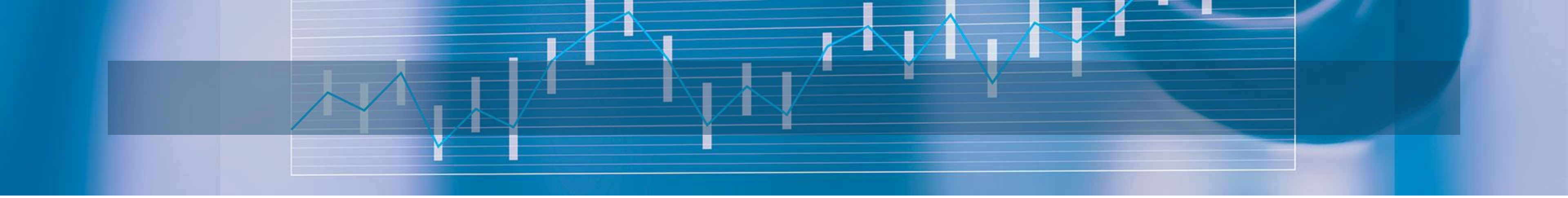


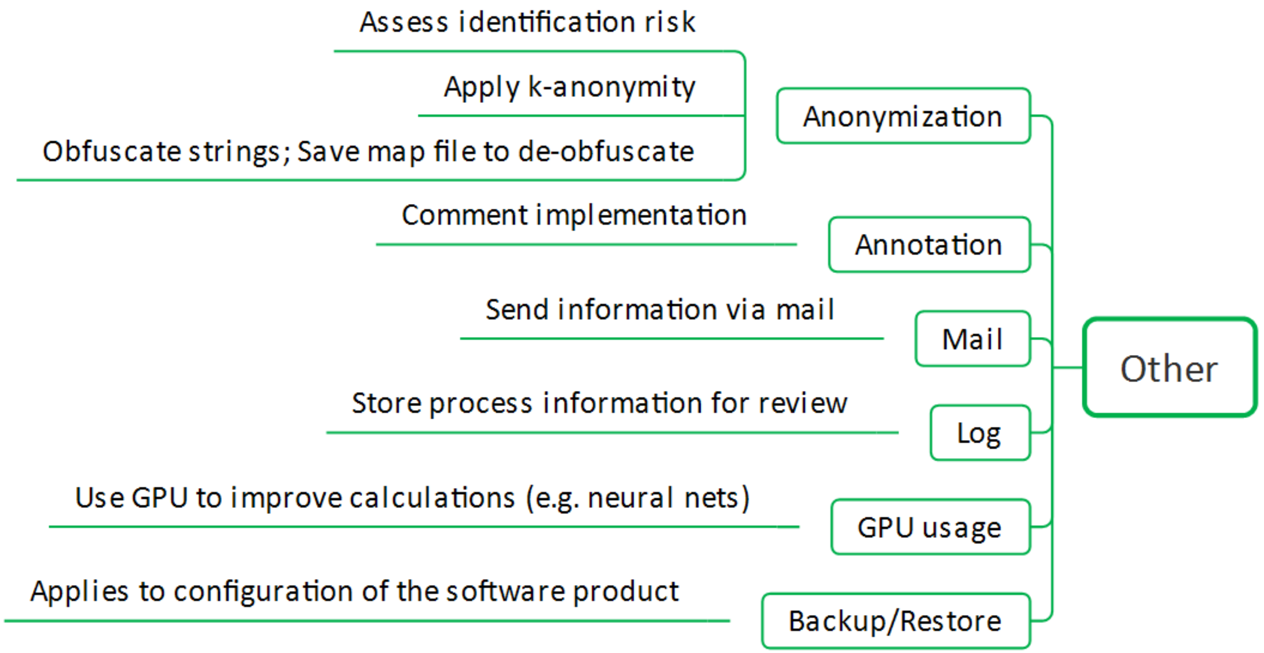


Visualisation
Create plots
Create reports
Combine Plots and Tables to Reports
Create dashboard
Combine plot to interactive dashboards
Simple charts
Point, Line, Bar and Box Plot; Histogram
Time Series
incl trend line
Horizon Chart
Pareto Chart
Sankey Chart
Heatmap
Treemap
Network Graph
Calendar Heatmap
Maps

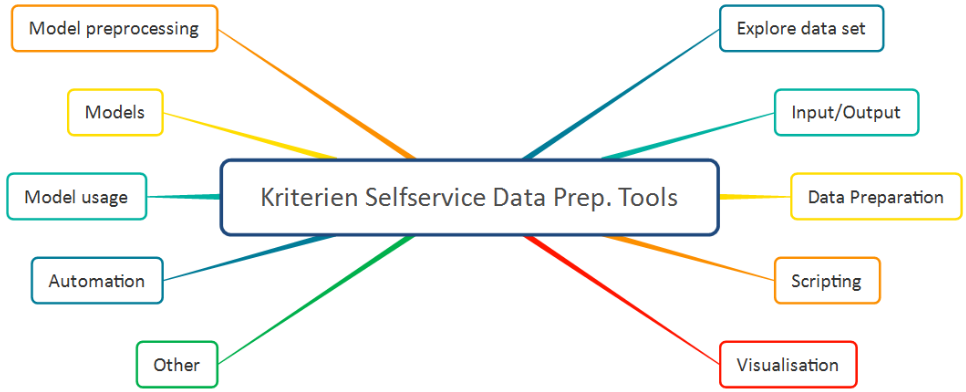


Kriterien Selfservice Data Prep. Tools
Model preprocessing
Models
Model usage
Automation
Other
Explore data set
Input/Output
Data Preparation
Scripting
Visualisation





Assess identification risk
Apply k-anonymity
Obfuscate strings; Save map file to de-obfuscate
Anonymization
Comment implementation
Annotation
Send information via mail
Mail
Store process information for review
Log
Use GPU to improve calculations (e.g. neural nets)
GPU usage
Applies to configuration of the software product
Backup/Restore
Other



Model preprocessing
Models
Model usage
Automation
Other
Kriterien Selfservice Data Prep. Tools
Explore data set
Input/Output
Data Preparation
Scripting
Visualisation

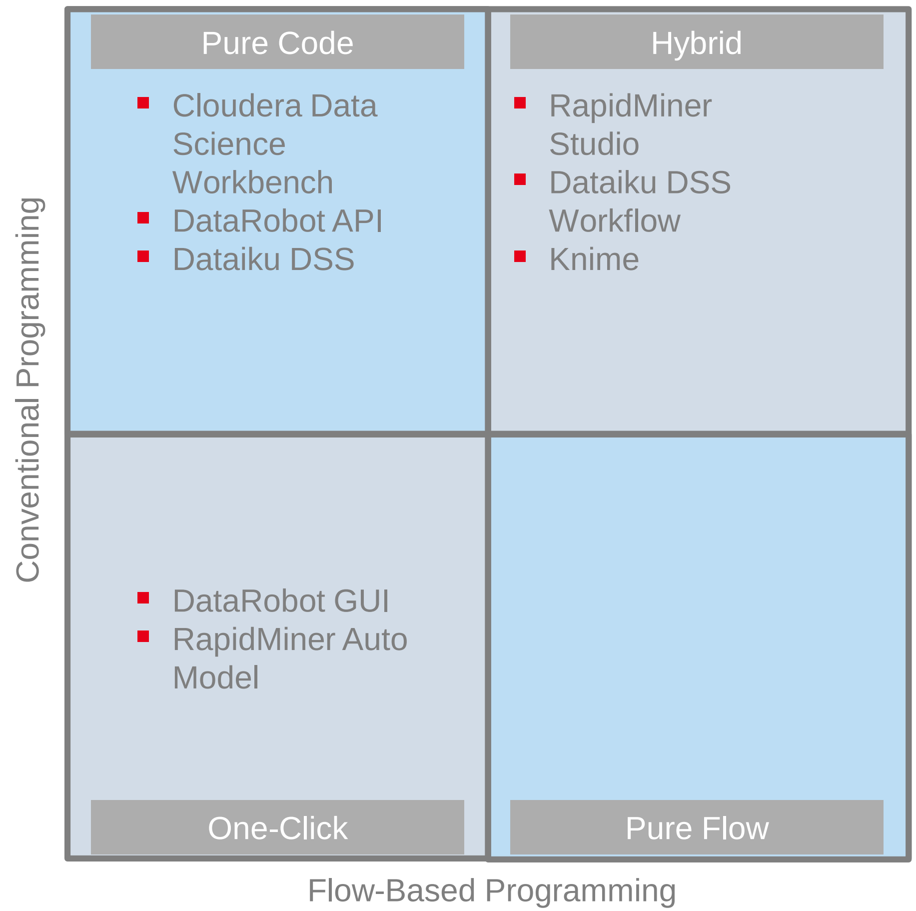
Agenda

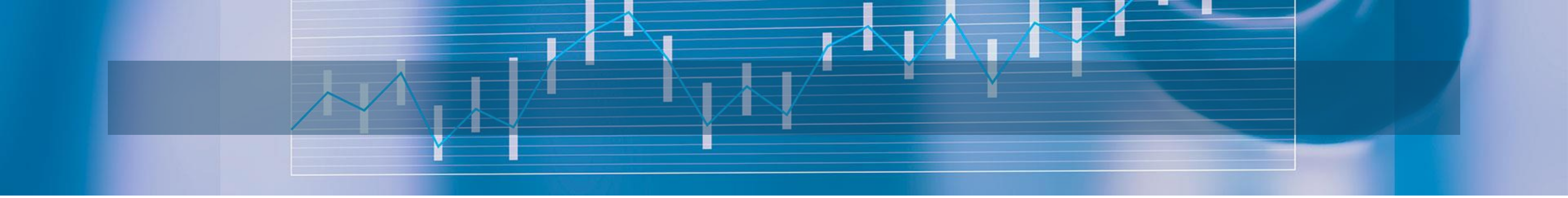
1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
   - Atlas Präsentation
   - Data Selfservice
     - Selfservice Data Preparation
     - Selfservice Data Analytics
     - Data Virtualization
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

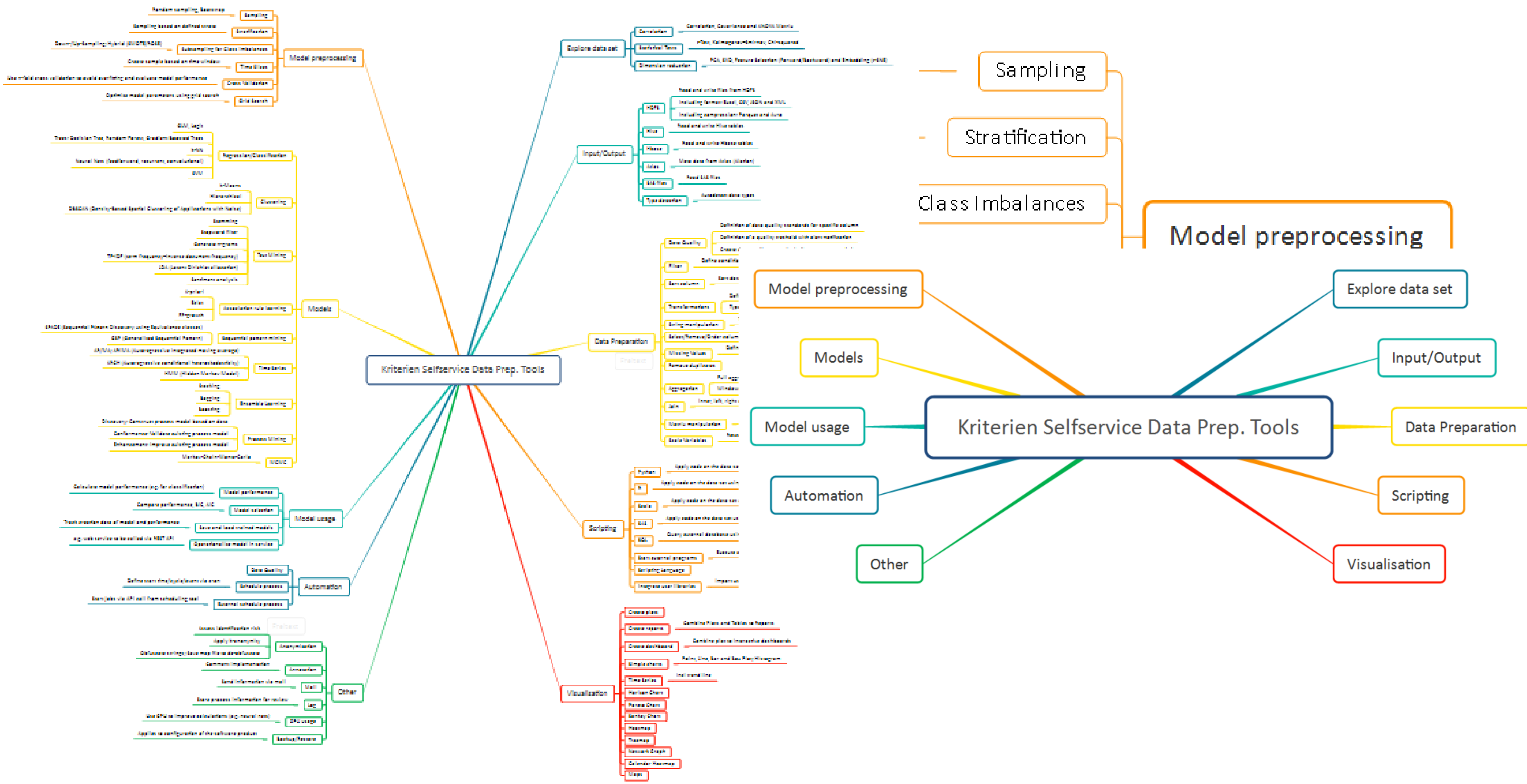
Self Service Data Analytics

- Self Service Analytics Tools können …
  - Citizen Data Scientists komplexere Datenanalysen zugänglich machen
  - Data Scientists bei der Organisation und Produktivsetzung von Modellen helfen
- Drei Gruppen von Tools
  - Automatisierungstools
  - Data Science Plattforms (kommerziell und GUI-geführt)
  - Code-first Plattformen
- Einführung einer Plattform erfordert gründliche Bedarfsanalyse und sorgfältige Evaluation

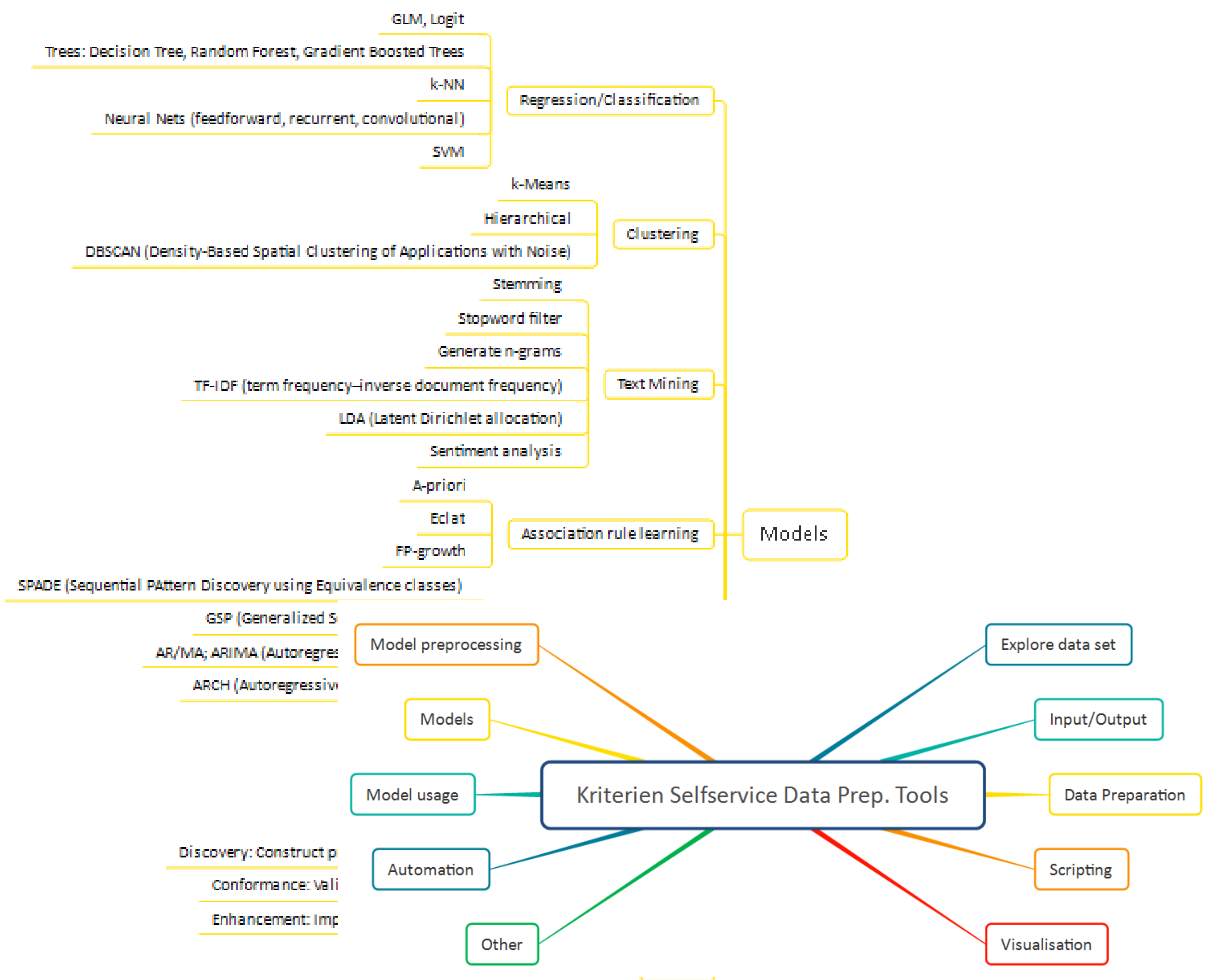
Kategorisierung von Machine Learning Frameworks

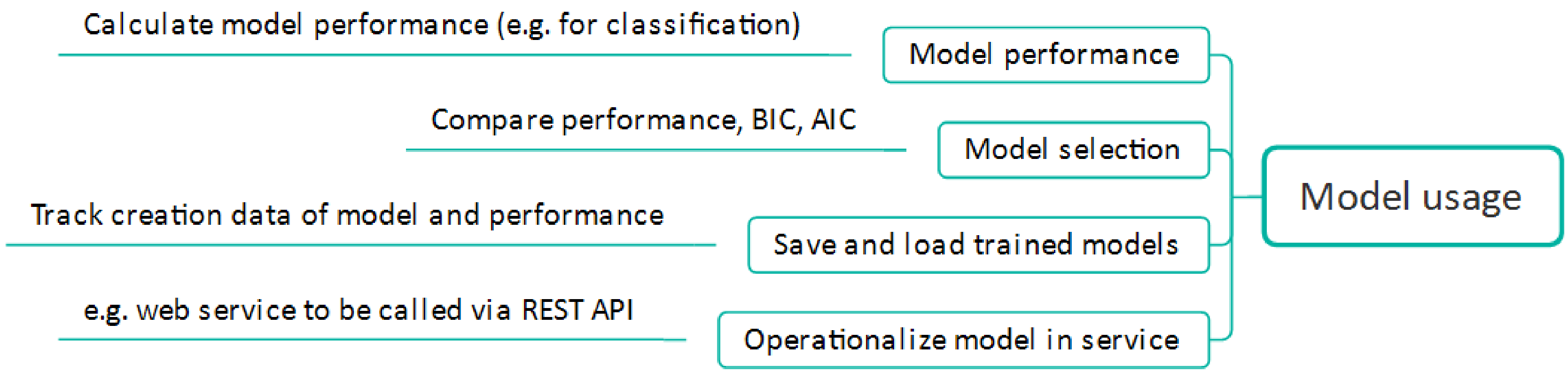


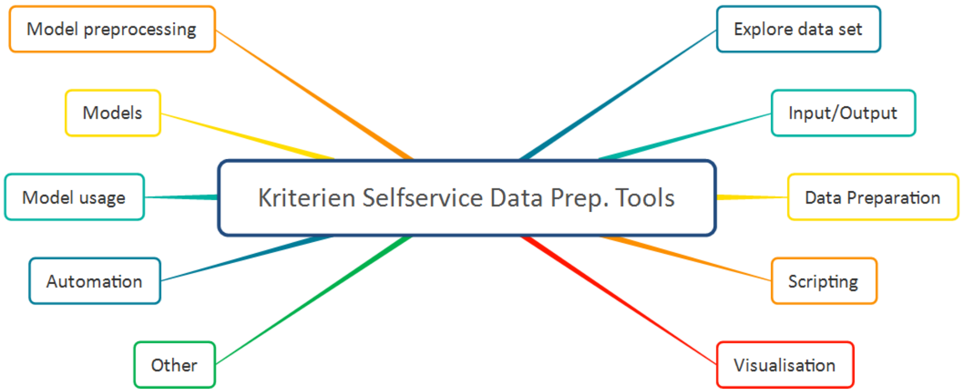




Sampling
Stratification
Class Imbalances
Model preprocessing
Model preprocessing
Models
Model usage
Automation
Other
Kriterien Selfservice Data Prep. Tools
Explore data set
Input/Output
Data Preparation
Scripting
Visualisation


GLM, Logit
Trees: Decision Tree, Random Forest, Gradient Boosted Trees
k-NN
Neural Nets (feedforward, recurrent, convolutional)
SVM
Regression/Classification
k-Means
Hierarchical
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
Clustering
Stemming
Stopword filter
Generate n-grams
TF-IDF (term frequency–inverse document frequency)
LDA (Latent Dirichlet allocation)
Sentiment analysis
Text Mining
A-priori
Eclat
FP-growth
Association rule learning
Models
SPADE (Sequential PAttern Discovery using Equivalence classes)
Model preprocessing
Explore data set
Models
Input/Output
Model usage
Kriterien Selfservice Data Prep. Tools
Data Preparation
Automation
Scripting
Other
Visualisation


Calculate model performance (e.g. for classification)
Model performance
Compare performance, BIC, AIC
Model selection
Track creation data of model and performance
Save and load trained models
e.g. web service to be called via REST API
Operationalize model in service
Model usage


Kriterien Selfservice Data Prep. Tools
Model preprocessing
Models
Model usage
Automation
Other
Explore data set
Input/Output
Data Preparation
Scripting
Visualisation

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   - Data Sovereignty
   - Business Glossar
   - Alation Präsentation – Business Glossar anlegen
   - SVML-Präsentation
   - Data Catalog und Data Lake
   - Atlas Präsentation
   - Data Selfservice
     - Selfservice Data Preparation
     - Selfservice Data Analytics
     - Data Virtualization
   - IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

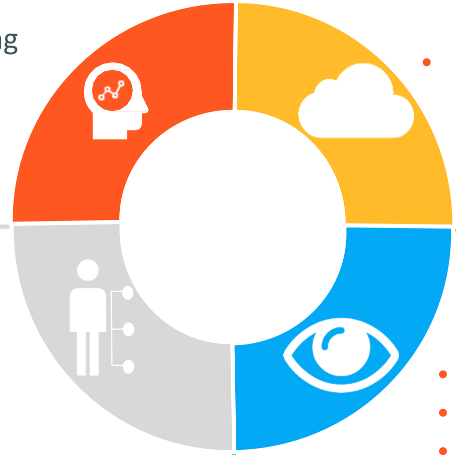
Data Virtualization Use Cases

Agile BI and Analyticals

- Logical Data Warehouse
- Virtual Data Marts
- Federation of Data Warehouses
- Operational BI/Analytics
- Hybrid DV-ETL
- Virtual Sandboxes & prototyping
- Self-Service BI and Reporting

Big Data Lake

- Logical Data Lake
- DV for Analytical Data Integration
- Data Warehouse Offloading
- Hadoop as an Analytical Sandbox
- Hadoop as an extra Data Warehouse
- Hadoop for ETL processing



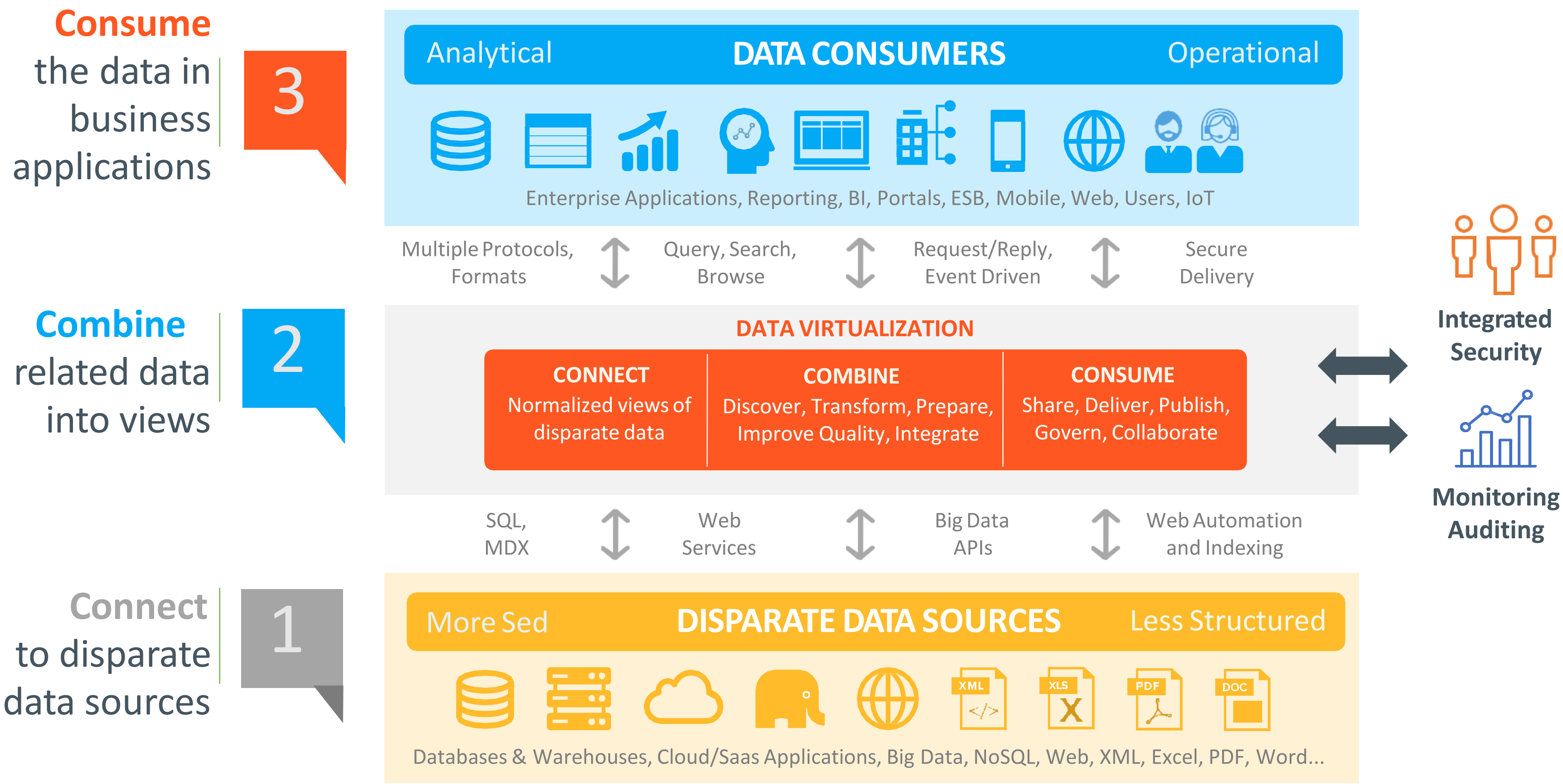
Digital Transformation

- Enterprise Data Layer
- Legacy Application Modernization
- Mirgration from Enterprise to Cloud
- Enterprise Data Marketplace
- Mergers & Acquisitions – Data Consolidation
- Agile Application Development (Mobile/Web/SOA)

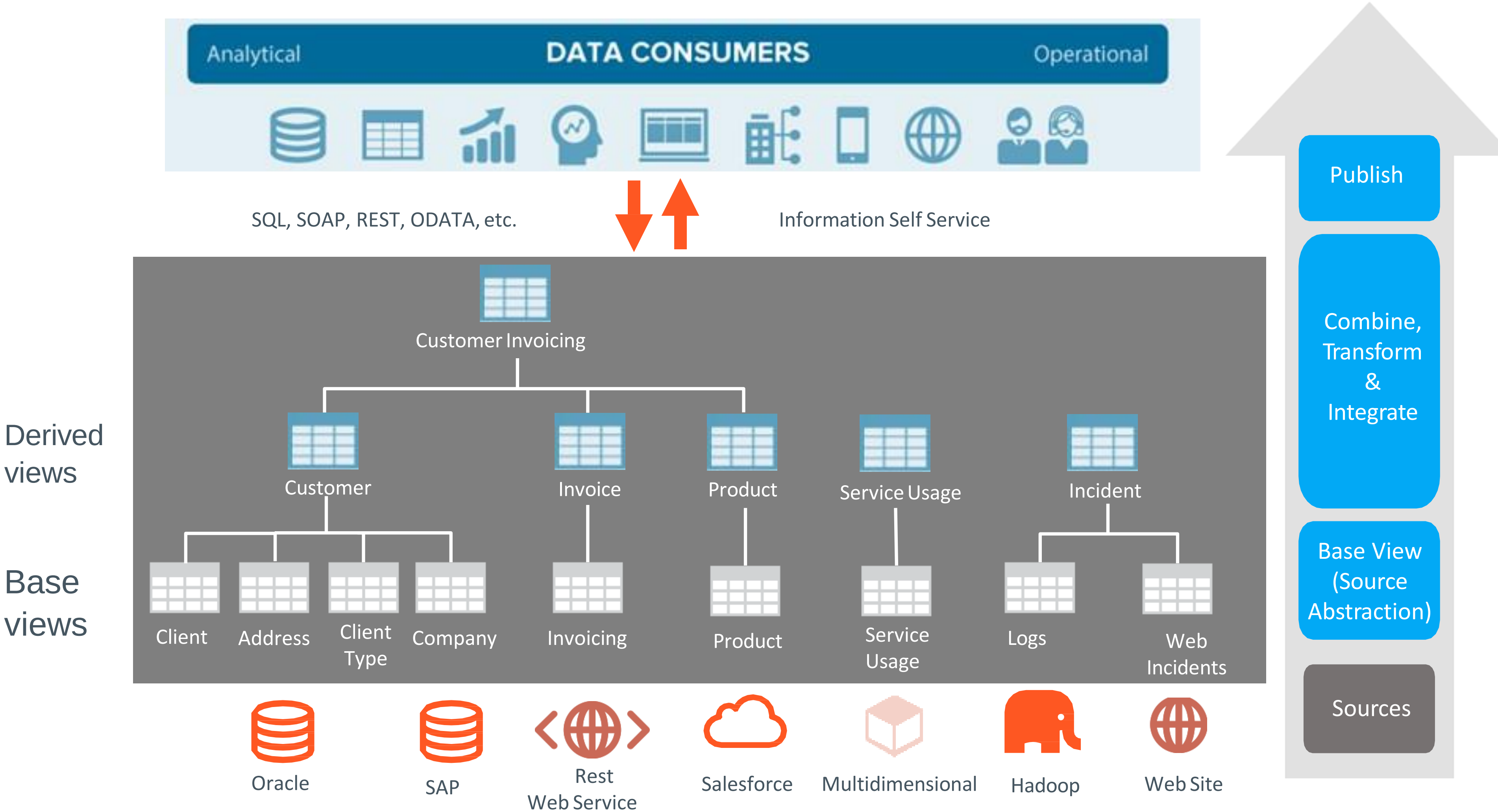
Single View Applications

- Single View of Customer
- Products/product Catalogs
- Vertical Specific (e.g. Well or Physician Data)

Data Virtualization - Data Abstraction Layer



Virtualization: Modeling



Data Virtualization Tool Vendor Landscape

Cambridge Semantics
- Data Virtuality
- Denodo
- Dremio
- eQ Technologic
- fraXses
- Gluent
- IBM
- Informatica
- Microsoft
- OpenLink Software
- Oracle
- Progress
- SAP
- SAS
- Stone Bond Technologies
- TIBCO Software

Denodo Stärken

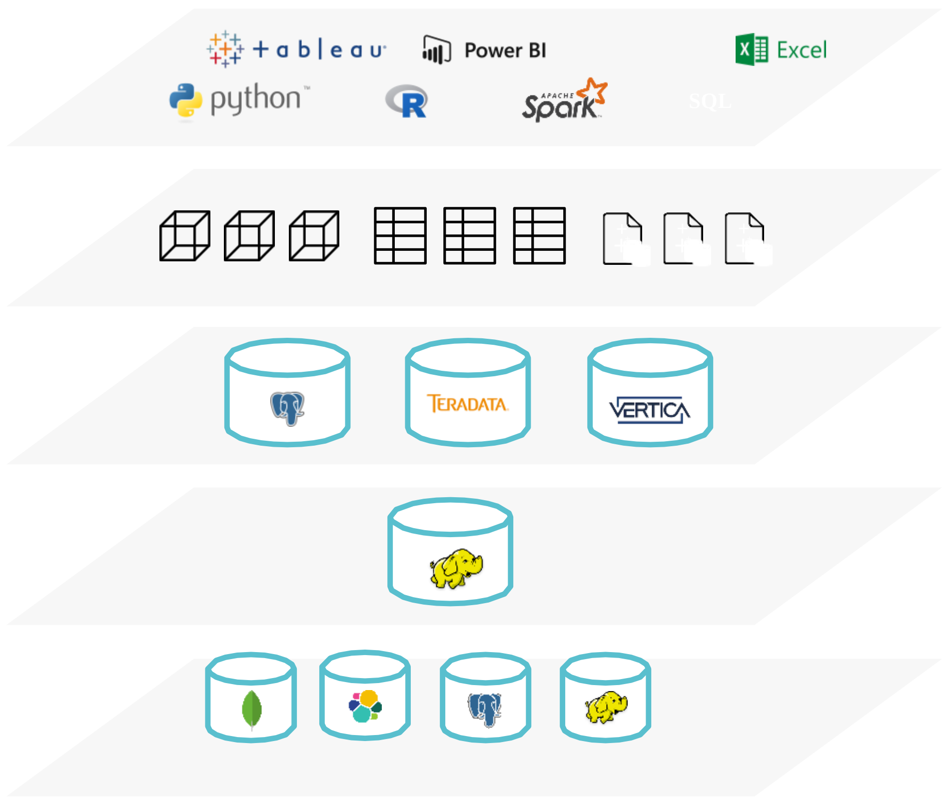
- Connectors zu vielen verschiedenen Datenhaltungssystemen
  - DB2, Oracle, SQLServer, SAP, Hadoop, NoSQL-DBs, AWS, SAS, …
- Feingranulare Security
  - Rollenbasiert, AD/LDAP-Integration, Custom Policies, Row / Column Masking Optionen
- Einfache Anbindung von verschiedenen Datenkonsumenten
  - Qlik, Talend, Cognos, JDBC Clients, …
- Verschiedene Data Governance Features
  - Lineage, Data Privacy / Maskierung, Datenmodelle der virtuellen Datasets, Auditing
- Versionierung der Objekte per Versionskontrolle (Git, Subversion, TFS,…) möglich
- Einfache Queries über heterogene Datenhaltung gut möglich

Denodo Herausforderungen

- Pushdown in Datenquelle wird nicht immer optimal durchgeführt
  - Pushdown für z.B. SAS nicht möglich
- Je komplexer die Queries desto eher wird kein Pushdown mehr durchgeführt
  - Optimizer trifft nicht immer die beste Entscheidung
- Benchmark mit Qlik als Datenkonsument zeigt bis zu 70% schlechtere Performance bei der Nutzung von Denodo im Vergleich zum direkten Datenbankzugriff

Alternative mit Self-Service Option: Dremio

Data is a massive engineering project today



Cubes, BI Extracts & Aggregation Tables

Data Warehouse

Data Staging

- Data sprawl
- Governance issues
- Slow to update

Dremio

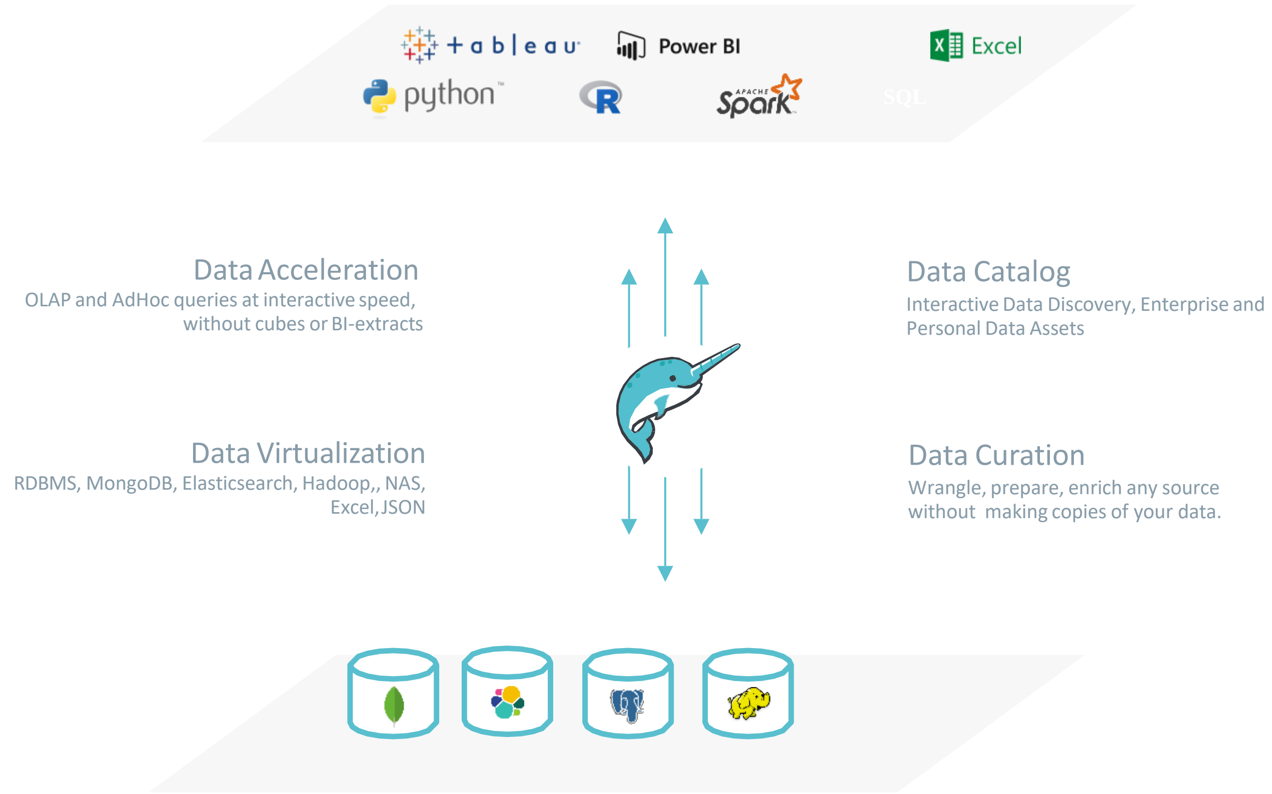


tableau
Power BI
Excel
python
Spark
Data Acceleration
OLAP and AdHoc queries at interactive speed, without cubes or BI-extracts
Data Catalog
Interactive Data Discovery, Enterprise and Personal Data Assets
Data Virtualization
RDBMS, MongoDB, Elasticsearch, Hadoop,, NAS, Excel,JSON
Data Curation
Wrangle, prepare, enrich any source without making copies of your data.

Dremio

- Skalierbar:
  - Durch Deployment auf Hadoop Infrastruktur kann Dremio auf mehrere 1000 Nodes linear skaliert werden
  - Dremio lässt sich nahtlos in bestehende Hadoop Systeme integrieren (Ressourcenmanagement über Yarn, Datenhaltung im HDFS, …)
- Data Reflections:
  - Hochperformantes spaltenbasiertes Dateiformat als Cache für schnelle ad hoc Abfragen
- Query Pushdown in Datenquellen
- Self-Service Data Preparation Komponente
- Feingranulare Security Optionen

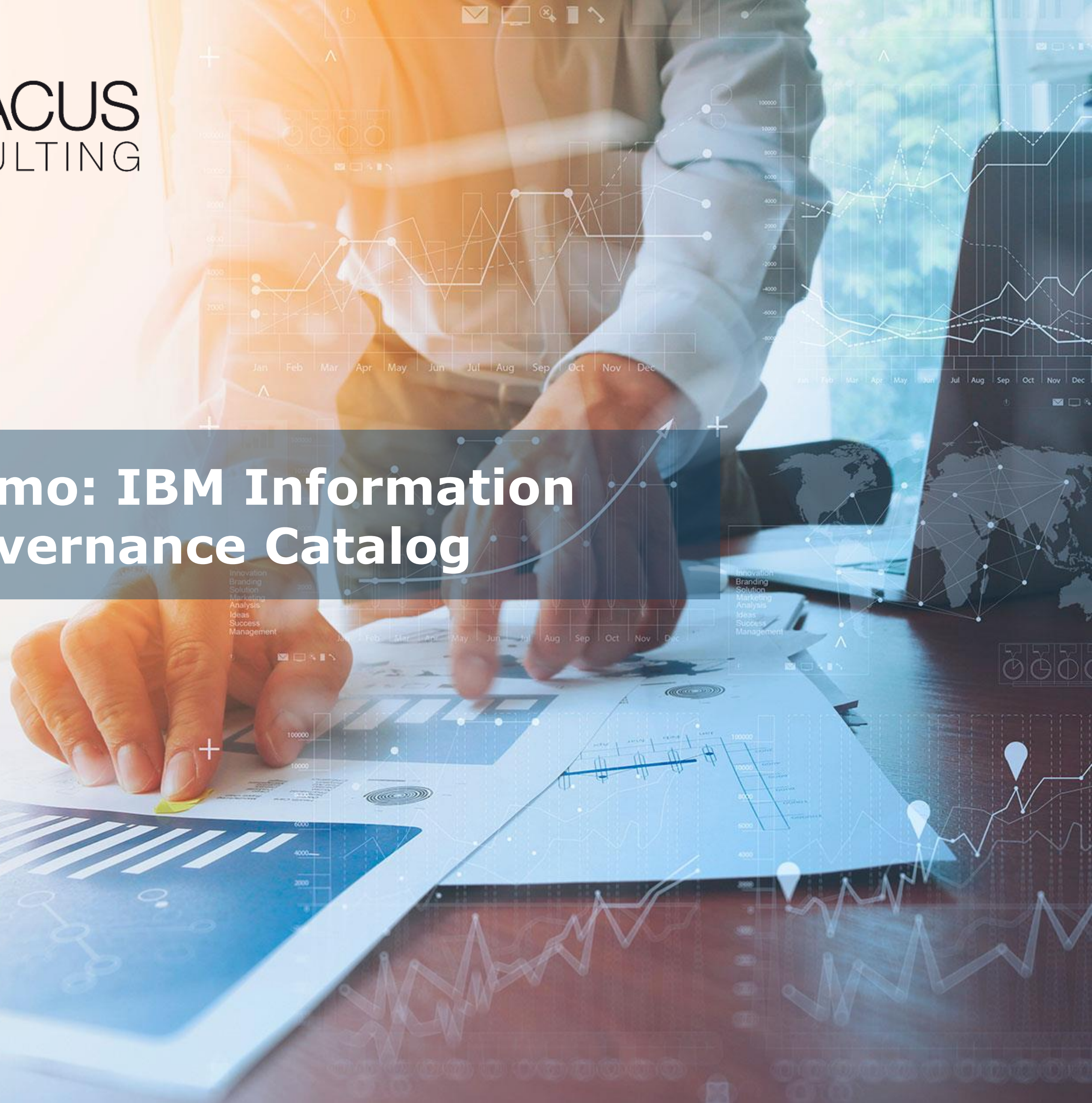
Dremio Herausforderungen

- Pushdown in Datenquelle nicht immer möglich
- Exotischere Datenbanken werden nicht immer nativ unterstützt
- Fokus bisher eher auf Open Source Komponenten
  - Dremio Enterprise Version bietet extra Enterprise Features

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
   – Data Sovereignty
   – Business Glossar
   – Alation Präsentation – Business Glossar anlegen
   – SVML-Präsentation
   – Data Catalog und Data Lake
   – Atlas Präsentation
   – Data Selfservice
   – IBM IGC Präsentation
5. Metadata Strategy und Data Catalog-Einführung

SARACUS CONSULTING



Demo: IBM Information Governance Catalog

IGC Summary

Features:

- Metadata Catalog coming with connectors to many different Database Systems
  - DB2, Hive, Oracle, SQLServer, Teradata, …
- Integration of various asset types:
  - Information assets (data resources, data models, applications, stored procedures, …)
  - Glossary assets (terms, categories, governance rules, …)
- Linage capabilities (e.g. integration of DataStage jobs)

Technical aspects

- Export/Import assets in CSV/XML/XMI file format
- REST-API for Metadata Extraction and editing

Agenda

1. Grundlagen Metadaten
2. Metadaten Management und Data Catalogs
3. Funktionalitäten von Data Catalogs
4. Data Catalogs: Ausgewählte Themen
5. **Metadata Strategy und Data Catalog-Einführung**

Agenda: Metadata Strategy & Implementation

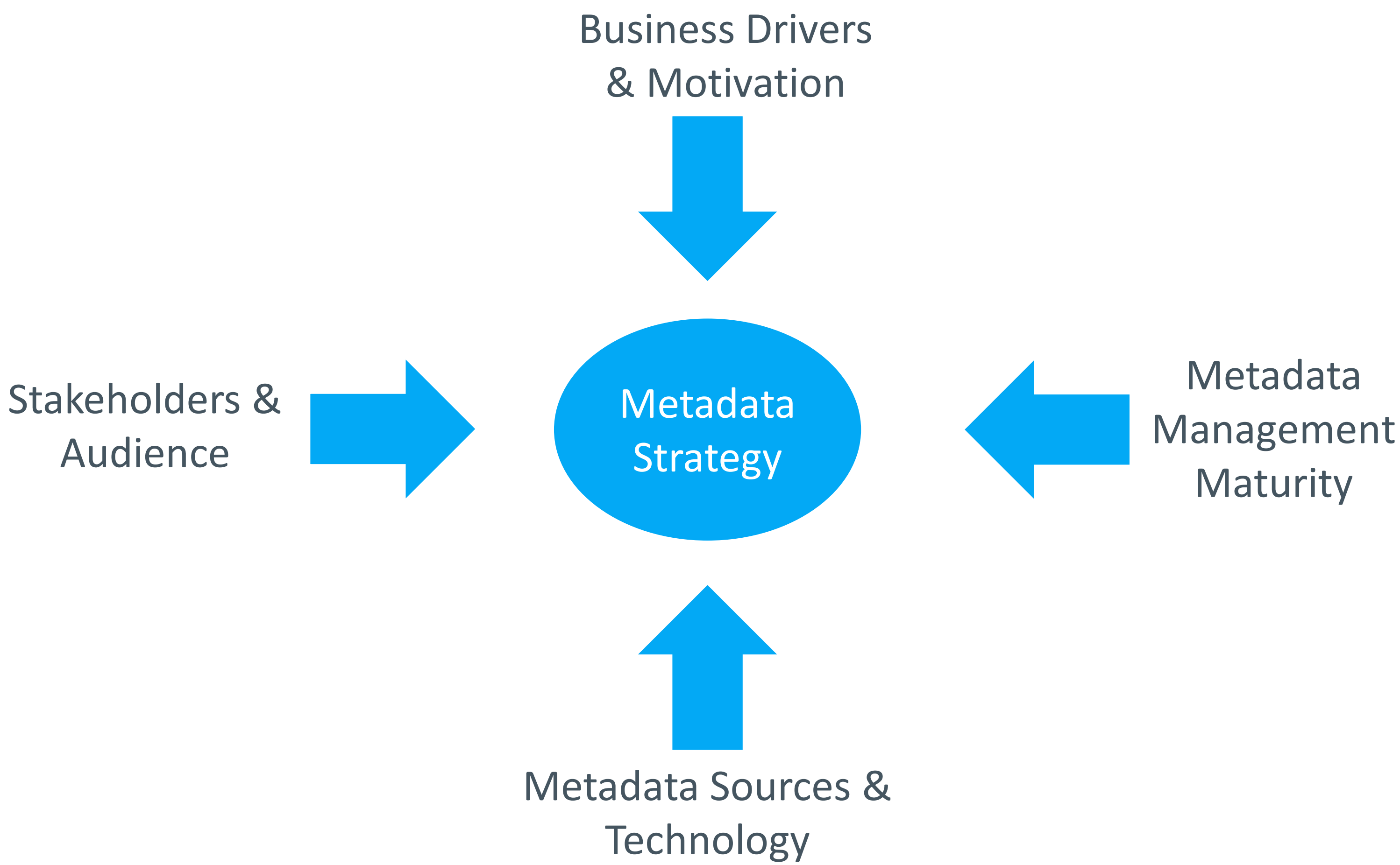
1. Creating a Metadata Strategy for your Organization
   - Business Drivers
   - Stakeholder Needs
   - Metadata Source Inventory
   - Technology Selection
   - Metadata Management Maturity Assessment
2. Metadata Implementation & Rollout
   - Identifying High-Priority Activities & “Quick Wins”
   - Defining an Actionable Roadmap

Key Components of Metadata Management

Metadata Strategy	Metadata Capture & Storage	Metadata Integration & Publication	Metadata Management & Governance
Alignment with business goals& strategy	Identification of all internal& external metadata sources	Identification of alltechnical metadata sources	Metadata roles & responsibilities defined
Identification of & feedback from key stakeholders	Population/import mechanism for all identified sources	Identification of key stakeholders & audiences (internal & external)	Metadata standards created
Identification of other Initiatives (Selfservice, Data Virtualization, Data Science Workbench, etc.)	Identification of existing metadata storage	Integration mechanism for key technologies (direct integration, export, etc.)	Metadata lifecycle management defined & implemented
Prioritization of key activities aligned with businessneeds & technical capabilities	Definition of enterprise metadata storage strategy	Publication mechanism for each audience	Metadata quality statistics defined & monitored
Prioritization of key data elements/subject areas		Feedback mechanism for each audience	Metadata integrated into operational activities & related data management projects
Communication Plan developed			

Metadata Strategy

- A successful metadata strategy requires input from multiple factors.



Aligning with Business Priorities

- Before you begin any metadata management initiative, it is important to determine the key business drivers & priorities.
- Some may be business-driven, and some may be IT-driven, for example:
  - Business Drivers
    - Better customer information for an upcoming marketing campaign
    - Data Lineage for financial audit
    - Sharing information with other organizations – R&D, Supply Chain, etc.
    - Data Governance support
  - IT Drivers
    - Impact analysis for application development
    - Reuse & efficiency through standards
    - Data lineage for data warehousing and integration
- There are likely a number of drivers – it is important to document & prioritize them.

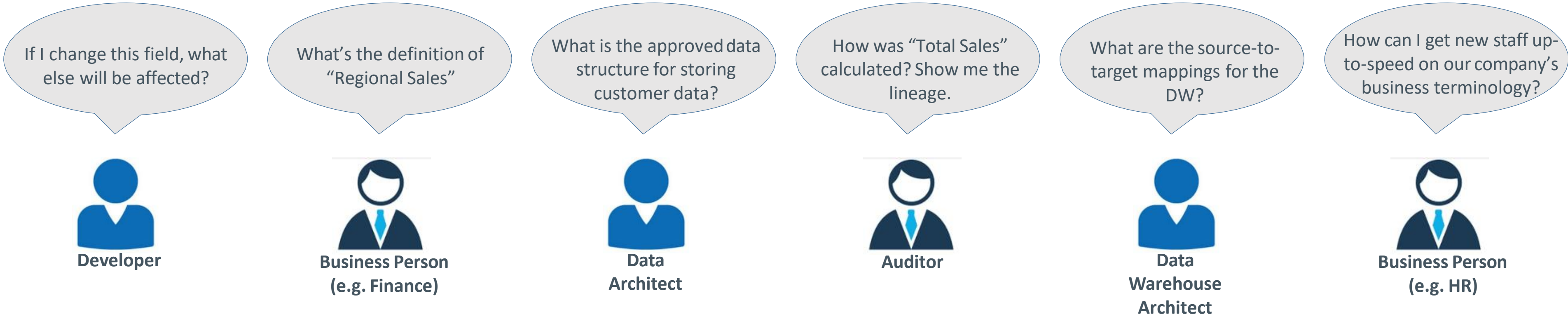
Internal and External Drivers

- There are both internal & external business drivers, and it is important to evaluate both.
  - Pay attention to external drivers: Your company might be perfecting the manufacture of horse shoes while the industry is building cars!

Internal Drivers	External Drivers
• Improve marketing campaigns with better customer info • Faster time-to-market for new applications • Increase efficiency & reduce costs	• Digital e-Commerce driving market • Increased Regulatory pressures • Community & social media-driven marketing

Who Uses Metadata?

- In addition to sharing metadata between tools and via export, many users across both IT & the business want to view the metadata through reports, portals, etc.



Stakeholder Analysis

- Stakeholders are key to the success or failure of your data program. Like data assets, they should be analyzed and managed.
- A number of tools and techniques exist to help manage stakeholders.
  - Stakeholder Map: Listing of key stakeholders with their roles, contact information, location, etc.
  - Interest/Influence matrix: Rank stakeholders by level of interest vs. amount of influence they hold.
  - Interest matrix: Identify key interest areas and map their importance to each stakeholders or stakeholder group.
  - Interview Schedule & Key Questions: Plan the interview schedule to respect stakeholders' time. Identify key questions ahead of the meeting.
  - Preferred Communication Styles: Identify the Styles of communication preferred by stakeholders & their communication styles (email, face to face meeting, coffee, introvert/extrovert, etc.)
  - Communication Plan: Develop a phased communication plan including feedback, reporting, metrics, etc.

Stakeholder Matrix

- Keeping track of “who’s who”: Create a simple stakeholder matrix outlining the key stakeholders, their roles, involvement, influence, impact, etc.

Stakeholder Matrix

Stakeholder Name / Group	Job Title/Role	Location	Involvement					Influence	Impacted	Phone	Email
			R	A	C	I		H / M / L	H / M / L		
EXECUTIVE REVIEW											
Mary Smith	CIO	Plano, TX	X			X		H	H	+1 (214) 555-1212	mary.smith@thisco.com
Robert Quantiles	CFO	New York, NY			X	X	e data	H	H	+1 (212) 555-1212	robert.quantiles@thisco.com
STEERING GROUP											
Stuart Ling	Director of Enterprise Architecture	San Francisco, CA	X	X			Core working group	H	H	+1 (415) 555-1212	stuart.ling@thisco.com
Ian Wordingham	Director of Data Strategy	London, UK	X	X			Core working group	H	H	+44 (020) 1234 1234	ian.wordingham@thisco.com
Melissa Smith	Strategic Consultant	Edinburgh, UK			X		Core working group	H	L	+44 131 123 1234	melissa.smith@thisco.com
DATA ARCHITECTURE											
Eric Wong	Data Architect	Plano, TX			X	X	Recommendations & input on data architecture	M	H	+1 (214) 555-1212	eric.wong@thisco.com
Wendy Collington	Data Architect	San Francisco, CA			X	X	Recommendations & input on data architecture	M	H	+1 (415) 555-1212	wendy.collington@thisco.com
Myles Stuart	DBA	Plano, TX				X	Historical input on legacy systems	L	M	+1 (214) 555-1212	myles.stuart@thisco.com
ETC - Other IT Groups listed											
FINANCE											
Lisa Winston	Director of Finance	New York, NY			X	X	Input into US finance needs for data	H	H	+1 (214) 555-1212	lisa.winston@thisco.com
Timothy Preston	EMEA Finance Lead	London, UK			X	X	Input into EMEA finance needs for data	H	H	+44 (020) 1234 1234	timothy.preston@thisco.com
Juan Morales	Latin America Finance Lead	Santiago, CL			X	X	Input into LATAM finance needs for data	H	H	+56 2 12345678	juan.morales@thisco.com
ETC - Other Business Groups listed											

RACI *:
R: Responsible
A: Accountable
C: Consulted
I: Informed

Stakeholder Interviews

- Prepare for the Interviews
  - Research as much about the business and stakeholders' goals as possible
  - Prepare targeted questions to prompt discussion
  - Group stakeholders for interviews or workshops
- During the Interviews
  - Use questions as a guide to encourage discussion. It shouldn't feel like a "quiz".
  - LISTEN. Playback information to ensure understanding.
  - Record the sessions, and take notes
  - Use the attendees' own language, avoid technical jargon
- After the Interviews
  - Summarize the findings and playback to stakeholders
    - Ensures Understanding
    - Helps Gain Buy-In and Show Progress
  - Group feedback into Business Drivers & Priorities

Stakeholder Feedback

- Determine key business issues & drivers through direct feedback.

There is limited ownership or enforcement of common practices and standards across the projects

We have 15 customer databases – with many duplications.

$12m has been spent on projects to clean up the data over the past 2-3 years

Where do I go to get the definition of "default banking standard"?

I didn't know we had any documented data standards

I just joined the company and don't understand all of the acronyms!

There was an error in reporting products by customer & region that was noticed by upper management.

Key subject matter experts are relied upon to review detailed data from various systems to ensure accuracy

I need a central, accurate view of all my customers worldwide.

I hear that our competitors are using the Semantic Web. Should we?

Issue Matrix

- An Issue Matrix lists:
  - Key Themes & Issues around metadata
  - Which teams are interested in each issue / theme
- Creates a “heat map” of priorities

Key Issues & Themes	Leadership	Sales	Finance	Marketing	Support	R&D	HR	Legal	Compliance
Improved Customer Information	X	X	X	X	X	X	X	X	X
No Cross-Domain Integration view (Sales, Marketing, Support, etc.)	X	X	X	X	X	X	X	X	X
Inconsistent Definitions of Key Business Terms	X	X	X	X	X			X	X
Inconsistent Summarization/Timing (e.g. Monthly view)	X	X	X						
External data integration needed				X	X				
Faster Time-to-Market for New Applications	X	X		X		X			
Lack of standards creating quality issues & rework					X	X			
Siloes of information slow development across teams	X	X		X	X	X	X		
Increase Efficiency & Reduce Costs									
System Redundancy	X	X	X	X	X	X			
Staff spend extra hours looking for information	X	X	X	X	X	X	X	X	X
Rework needed due to incorrect definitions			X		X	X			
Etc.									

Mapping Business Drivers to Metadata Management Capabilities

Business Drivers

External Drivers

- Digital Self Service
- Increasing Regulatory Pressures
- Online Community & Social Media
- Community Building

Internal Drivers

- Targeted Marketing
- Reducing Time to Market of Adv. Analytical Models
- 360 View of Customer
- Efficient ETL Processes

Stakeholder Challenges

1. **Lack of Business Alignment**
   - Data spend not aligned to Business Plans
   - Business users not involved with data
2. **Integrating Data**
   - Siloed systems
   - No common view of key information
3. **Data Quality Issues**
   - Bad customer info causing Brand damage
   - Completeness & Accuracy Needed
4. **Cost of Data Management**
   - Manual entry increases costs
   - System redundancy
   - No reuse or standards
5. **No Audit Trails**
   - No lineage of changes
   - Fines had been levied in past for lack of compliance
6. **Big Data Exploitation**
   - Exploiting Unstructured Data
   - Access to External & Social Data

Metadata Capability

- **Metadata Strategy** 1 2 3 4 5 6
- **Metadata Capture & Storage** 2 3 4 5 6
- **Metadata Integration & Publication** 1 2 3 4 5 6
- **Metadata Management & Governance** 2 3 4

Shows “Heat Map” of Priorities

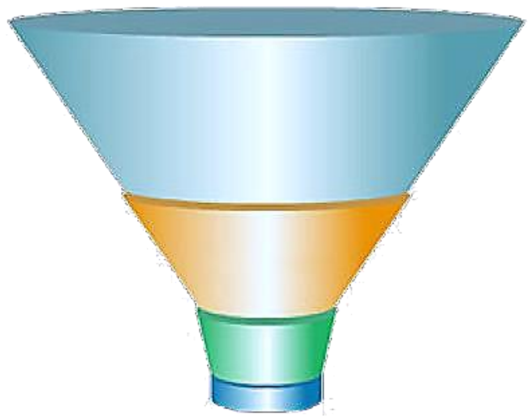
Identify High-Priority Data Elements

- It’s often not feasible to manage metadata for the entire organization, so it’s important to focus on the data that matters most.

Identify Key Business Driver

Launch of New Product – Marketing Campaign requires better customer information

Filter Data Elements Aligned with Business Driver



Targeted Project to Show Short-Term Results

Focus Efforts on Key Data

Customer **Product** **Partner** **Region** **Vendor**

Inventory Current Metadata Sources

- Inventory current metadata sources, both internal & external.

Metadata Sources	Internal	External
Relational Databases		
MySQL	X	
Oracle	X	
SQL Server	X	
Sybase	X	
Etc.		
BI Tools		
Tableau	X	
Qlik	X	
Etc.		
Open Data		
Data.gov – agricultural data		X
Etc.		

Inventory & Usage Mapping

- It’s also important to determine which teams are using these technologies to create a “heat map” of usage & priority.

Metadata Sources	Leadership	Sales	Finance	Marketing	Support	R&D	HR	Legal	Compliance
Relational Databases									
MySQL				X					
Oracle		X	X	X	X	X	X	X	X
SQL Server		X	X						
Sybase			X						
Etc.									
BI Tools									
Tableau		X			X	X	X	X	X
Qlik	X		X	X					
Etc.									
Open Data									
Data.gov – agricultural data	X			X		X			
Etc.									

Technology & Tool Selection - Interfaces

- Based on the inventory of metadata sources, evaluate what data standard & tools are necessary
  - Existing tools
  - New tools for purchase
- This evaluation will help determine whether:
  - a new tool is needed
  - existing tools suffice
- a combination of tools may work together

Data Sources	Tool A – In House	Tool B – In House	Tool C – Purchase	Tool D – Purchase
Relational Databases				
MySQL	X		X	X
Oracle	X			X
SQL Server	X			X
Sybase	X			X
BI Tools				
Tableau	X			X
Qlik	X		X	X
Open Data				
Data.gov – agricultural data		X		
ETL Tools				
Informatica				X

Technology & Tool Selection - Standards

- Be aware of industry standards that should be considered.

Data Sources	Tool A – In House	Tool B – In House	Tool C – Purchase	Tool D – Purchase
Relational Databases				
CWM	X			X
BI Tools				
CWM	X			X
Open Data				
Open Data Metadata Schema		X		
ETL Tools				
CWM	X			X

Technology & Tool Selection – Interfaces & Publication

- When devising a strategy for metadata integration & publication, first consider the audience for the metadata solution:
  - Technical users
    - Interfaces & Export to in-use tools & technologies (e.g. Data Modeling Tools, BI Tools, ETL Tools, etc.)
    - Intuitive visualization of Impact Analysis, Lineage, etc.
    - Publication of common standards
    - Leverage the Technology Inventory & Usage Mapping
  - Business Users
    - Intuitive interfaces for business terms, glossary information
    - Easy search
    - Integration with in-use tools & technologies (e.g. Self-Service BI)
    - Leverage the Stakeholder Matrix

Key Issues & Themes	Leadership	Sales	Finance	Marketing	Support	R&D	HR	Legal	Compliance
Relational Databases									
MySQL				x					
Oracle		x	x	x	x	x	x	x	x
SQL Server		x	x						
Sybase			x						
Etc.									
BI Tools									
Tableau		x			x	x	x	x	x
Qlik	x		x	x					
Etc.									
Open Data									
Data.gov – agricultural data	x			x		x			
Etc.									

Stakeholder Matrix

Stakeholder Name / Group	Job Title/Role	Location	Involvement				Role on Project	Influence	Impacted
			R	A	C	I		H / M / L	H / M / L
EXECUTIVE REVIEW									
Mary Smith	CIO	Plano, TX	X			X	Executive Sponsor	H	H
Robert Quantiles	CFO	New York, NY			X	X	Executive Champion for Finance data	H	H
STEERING GROUP									
Stuart Ling	Director of Enterprise Architecture	San Francisco, CA	X	X			Core working group	H	H
Ian Wordingham	Director of Data Strategy	London, UK	X	X			Core working group	H	H
Melissa Smith	Strategic Consultant	Edinburgh, UK			X		Core working group	H	L
DATA ARCHITECTURE									
Eric Wong	Data Architect	Plano, TX			X	X	Recommendations & input on data architecture	M	H
Wendy Collington	Data Architect	San Francisco, CA			X	X	Recommendations & input on data architecture	M	H
Myles Stuart	DBA	Plano, TX				X	Historical input on legacy systems	L	M
ETC - Other IT Groups listed									
FINANCE									
Lisa Winston	Director of Finance	New York, NY			X	X	Input into US finance needs for data	H	H
Timothy Preston	EMEA Finance Lead	London, UK			X	X	Input into EMEA finance needs for data	H	H
Juan Morales	Latin America Finance Lead	Santiago, CL			X	X	Input into LATAM finance needs for data	H	H
ETC - Other Business Groups listed									

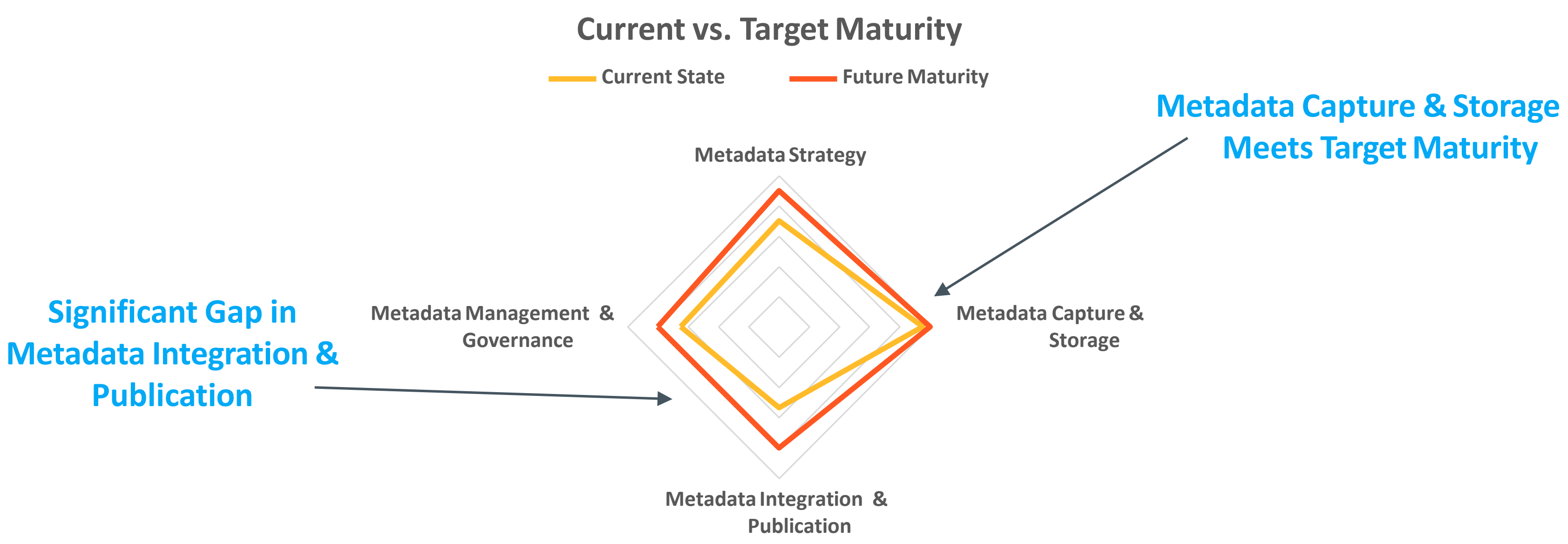
Metadata Management Maturity Assessment

- Ask a detailed set of questions for each metadata functional area.
- Compare with desired future state – you don't have to be a "5" in everything!

	Current State	Future State	Full Maturity
Metadata Strategy	3.5	4.5	5
We have a Metadata Strategy for the capturing, integrating, processing, delivery and presentation of data within our organization	5	5	5
Our Metadata Strategy is aligned to our Business Strategy.	3	5	5
We have executive and/or senior-level business support and sponsorship for our strategy	5	5	5
We have published a plan to achieve our Metadata Strategy that includes organization support, process, and IT.	3	4	5
We have policies, organizations, and budgets in place to support our Metadata Strategy and Plan.	3	4	5
Our Metadata Strategy is published and well understood across lines of business and technology groups.	2	4	5
Etc…			
Metadata Capture & Storage	4.8	5.0	5
We have a centralized metadata repository which stores metadata from all sources across the organization	4	5	5
We have metadata storage for individual tools and sources	5	5	5
Automated population is available for all of our metadata sources	5	5	5
We have a common metamodel for our metadata storage across sources	5	5	5
Etc…			
Metadata Integration & Publication	2.7	4.0	5
We use industry standards where available	4	4	5
We establish, publish, and maintain definitions of business terms in a centralized location visible across business and IT.	2	5	5
We link business and technical metadata to establish clear lines of communication and vocabulary between business and IT.	3	5	5
We can trace the data path (lineage) through events and processes to understand its origination and what happens to data as it flows through a system.	3	5	5
We use metadata to perform impact analysis (i.e. to understand the downstream effects of changes to a data element).	3	5	5
We publish intuitive reports or have an online portal for end users that are actively used.	1	5	5
Etc….			
Metadata Management & Governance	3.3	4	5
We have common metadata standards that are used and implemented	4	4	5
We have defined reuse and integration rules for rationalizing and integration metadata sources	4	4	5
We have roles clearly defined, communicated, and progress measured as part of employee reviews	2	4	5
There are defined metrics for metadata quality that are actively monitored for continual improvement.	3	4	5
Metadata management is integrated into key operational activities and related data management projects.			
Etc…			

Metadata Management Maturity Assessment

- A Radar Chart ("Spider Chart") can be a helpful way to visualize the relative strengths & weaknesses in various capability areas.



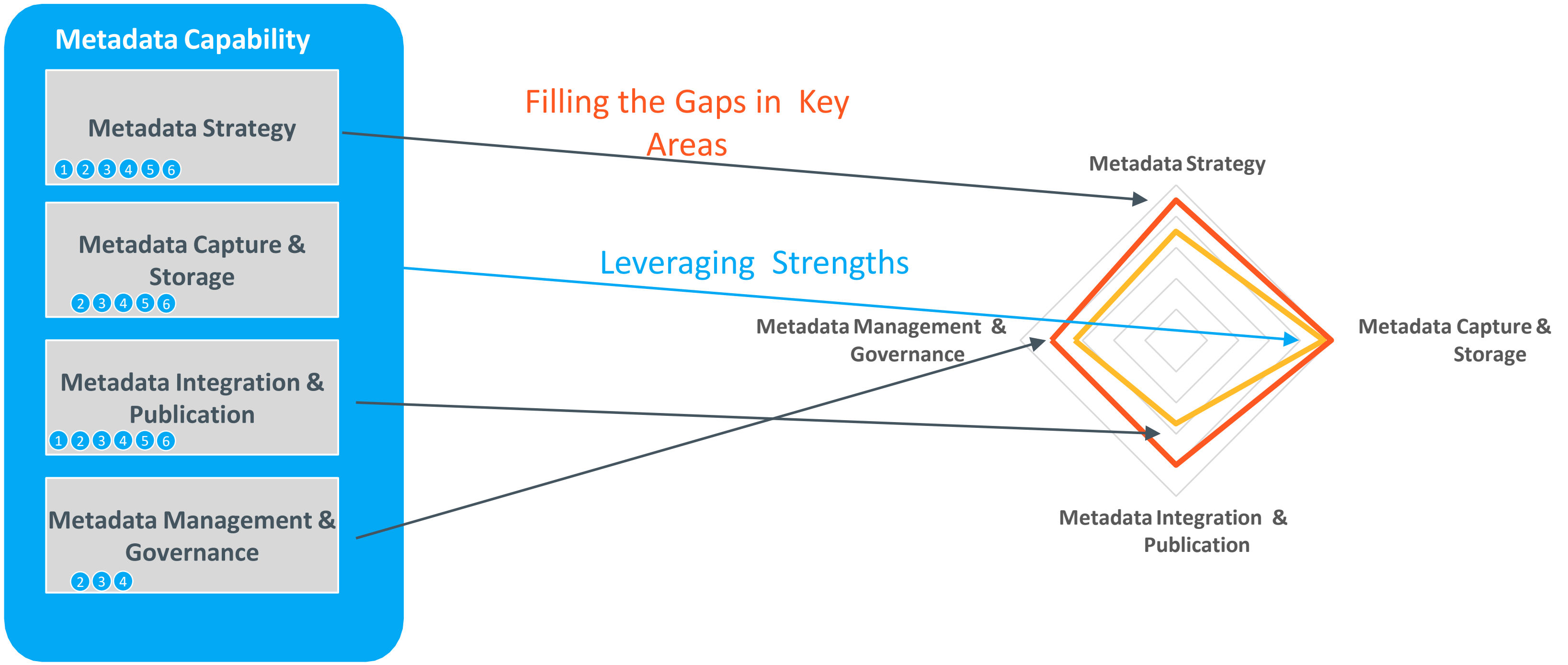
Priorities with Capability

- Aligning high-priority capabilities with current & target maturity helps with prioritization.

- Fixing what's broken
- Highlighting what's good

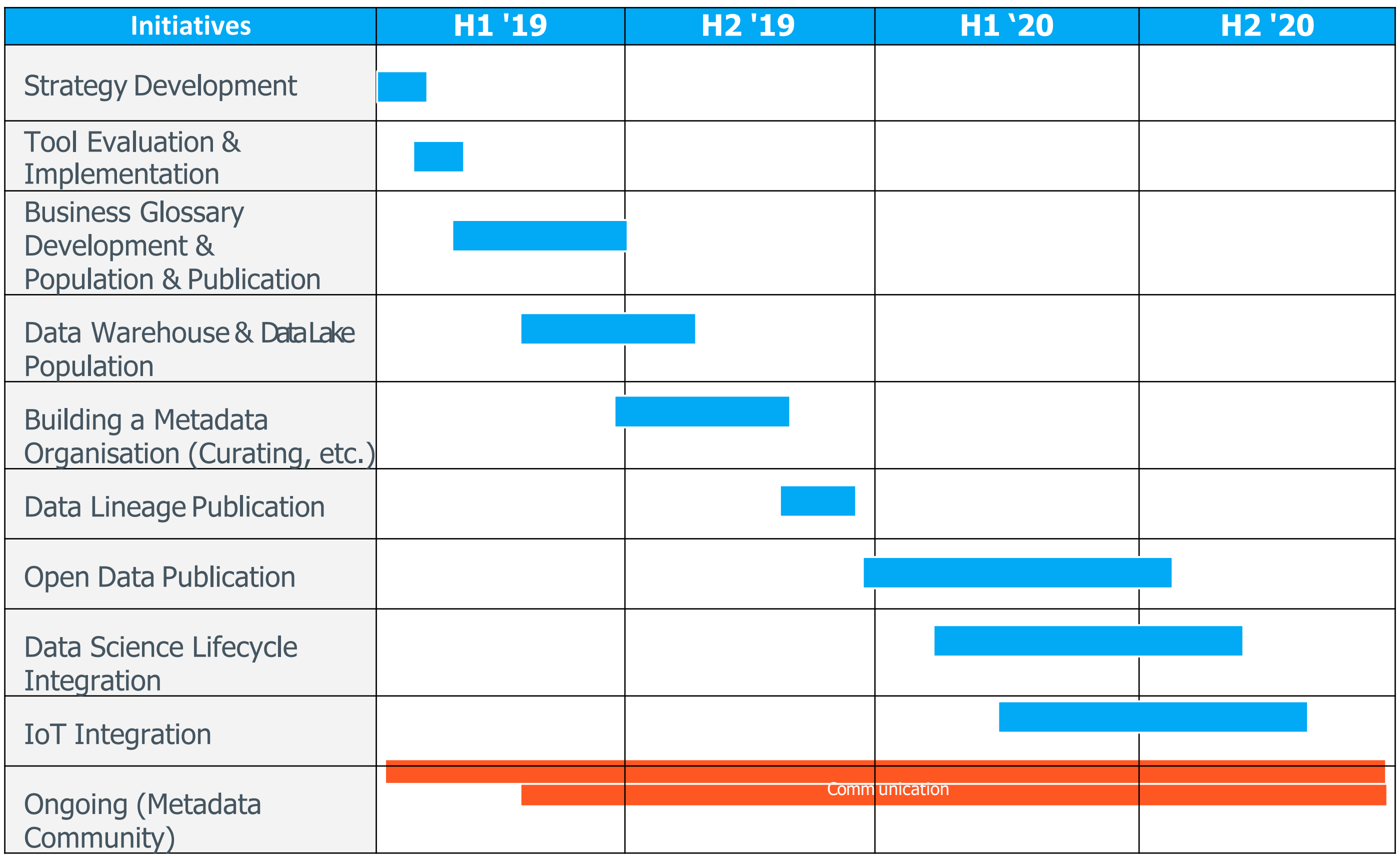
Priorities are:

- High priority & High maturity
- High priority & Low maturity



Defining an Actionable Roadmap

- Develop a detailed roadmap that is both actionable and realistic
  - Show quick-wins, while building to a longer-term goal
  - Balance Business Priorities with Data Management Maturity



Metadata Roles & Responsibilities

- It's important to establish formal roles & responsibilities for your metadata effort.
- Some may be part-time, and some full-time, but they should be clearly defined and communicated so that staff has understanding of and accountability for their roles.
  - Executive Sponsor/Champion: Understands & communicates the importance of metadata management across the organization.
  - Steering Group: As part of a metadata management effort, or part of a larger data governance effort, the steering group prioritizes & sets direction for key activities.
  - Data Stewards: Responsible for business definitions & rules for key data elements.
  - Metadata Repository Administrator: Manages the administration, population, and interfaces of a metadata repository.
  - Metadata Publicist: Establishes reports & publication methods to end users.
  - Metadata Consumers: Actively use metadata as part of their daily jobs, and are held accountable for using published standards.
    - Data Modelers
    - Developers
    - Business Users
    - Report Developers
    - Etc.

Monitoring Metadata Quality & Metrics

- Metadata is a key driver of data quality, and to support this, the metadata itself must be of high quality.
- In order to ensure that quality metadata is maintained, it must be actively managed and monitored. Dashboards & Reports can be used to monitor key quality indicators.
- Key metadata quality indicators include:
  - Completeness: e.g. Do definitions exist for all key data elements?
  - Accuracy: e.g. Are current definitions correct? Do data types accurately represent currently implemented standards?
  - Currency/ Timeliness: e.g. Are metadata definitions current or outdated?
  - Consistency: e.g. Are metadata standards defined, published & implemented consistently across the organization?
  - Accountability: e.g. Are data stewards or owners defined?
  - Integrity: e.g. Are linkages and relationships established between critical metadata items?
  - Privacy: e.g. Is any metadata subject to privacy restrictions?
  - Usability: e.g. Are people actually using this metadata?

Summary

- A successful metadata strategy considers both business and technology needs
  - Evaluate both internal & external business drivers
  - Interview Stakeholders to understand their requirements
  - Create a Metadata Source Inventory mapped to stakeholder usage
  - Implement Technology Selection based on documented technical & business requirements
- Metadata Implementation & Rollout
  - Identifying High-Priority Activities & “Quick Wins”
  - Align with current Metadata Management Maturity
  - Define an actionable Roadmap
  - Metadata management is an ongoing process. Define formal roles, and measure & monitor progress.
  - Communicate to stakeholders throughout the entire process

Ihre Fragen?

Vielen Dank für ihre Aufmerksamkeit