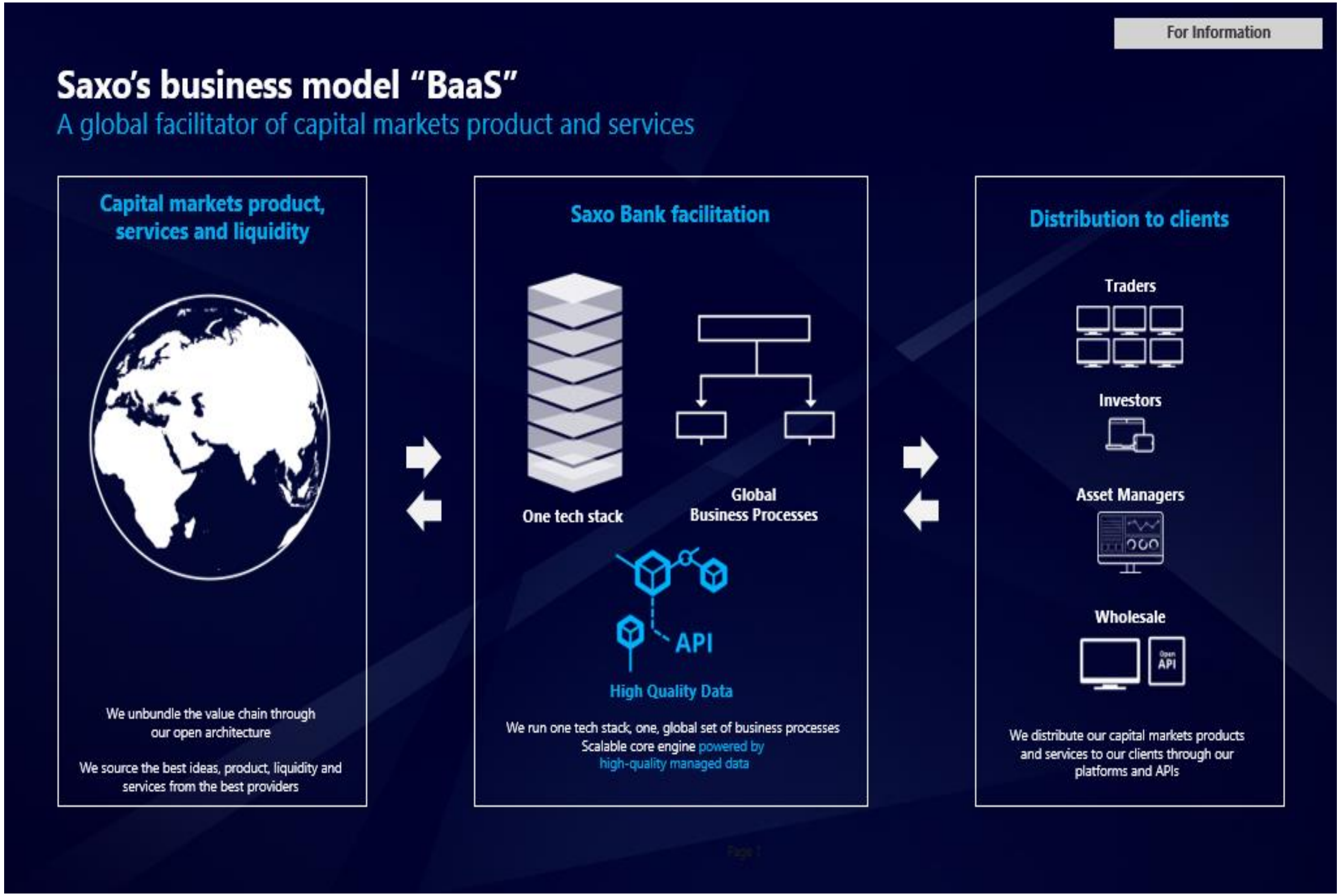
SAXO BANK

Data Workbench

Sheetal Pratik - Saxo Bank

August 2020



For Information

Saxo's business model "BaaS"

A global facilitator of capital markets product and services

Capital markets product, services and liquidity

We unbundle the value chain through our open architecture

We source the best ideas, product, liquidity and services from the best providers

Saxo Bank facilitation

High Quality Data

We run one tech stack, one, global set of business processes
Scalable core engine powered by high-quality managed data

Distribution to clients

We distribute our capital markets products and services to our clients through our platforms and APIs

About Saxo

We are leading fintech and regtech specialists, connecting traders, investors and partners to more than 35,000 instruments – across all asset classes – from a single account.

What we do

We build digital platforms to facilitate multi-asset market access and provide clients of all sizes with professional-grade tools, industry-leading prices and best-in-class service.

DATA GOVERNANCE FOR A DIGITAL NATIVE

- A data driven organization need to have multi-level Data Governance. Most of the tools are designed to fix the fact e.g. before a data warehouse load. What is needed is to ensure data integrity at the origin to prevent the "butterfly effect" in the downstream systems.

- The article "How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh", clearly emphasizes on how data platform with a centralized architecture can lead to failures by being bottleneck at certain point and have impact to stability. Also with ownership of data at the domain level, it becomes a failed attempt to manage the data dictionary centrally or duplicate the effort of creating and maintaining such data assets.

- Considering this, it is imperative that the solution has to be more futuristic and a straight implementation of any of the COTS products for Data Cataloguing might not be the right answer to Saxo's Data Governance implementation.

- The preferred strategy for tooling is to fix forward rather than attempting to fix the past by using some kind of crawler and using ML to extract the metadata from various data sources.

VISION

For Domain Teams

Who need visibility on the availability, meaning, usage, ownership and quality of data

The **Data Workbench** *(Owner's pride Neighbor's envy)*

Is a one-stop data shop

That provides transparency of Saxo's data ecosystem

Unlike our current state which is becoming increasingly complex as we grow

Our product will help Saxo to improve time to market and unlock new insights.

The **Data Workbench** is designed to be part of the new data architecture. It consists of two main components a ***Data Catalogue*** and a ***Data Quality Solution*.**

1. The Data Catalogue captures and exposes metadata. This provides transparency into the meaning and ownership of our data. The Data Catalogue is built on ***DataHub*** a data catalogue open-sourced by LinkedIn. LinkedIn is very supportive and are working closely with us helping with the adoption of the tool.

2. The Data Quality Solution is built on the open source solution **Great Expectations** supported by SuperConductive. Great Expectations is a declarative, flexible, and extensible data quality solution. It allows teams to define data quality rules and actively monitor the quality of their data.

MOTIVATION FOR THE SOLUTION

- **Federated Data Governance model** is an industry trend where the enterprise governance team facilitates the monitoring and management of the quality of enterprise critical data, with assistance from the business unit.

- LinkedIn's journey of its shift of approach from initial version of Data Governance solution called ***WhereHows to DataHub*** , is a typical example of the paradigm shift from *"a central metadata repository"* solution to a more **decentralized architecture** that puts domains before anything else to support the possibility of self-service data platform.

- We realized that a practical way of implementation would be to **stay lean and agile and iteratively work** with data domains while establishing the Data Governance framework and thus create a platform that is self-serviced, scalable and more relevant to stakeholders.

- We had a discussion with LinkedIn to understand their journey, learnings and lessons learnt that motivated them to evolve from **WhereHows to Datahub**. We acknowledged that, Saxo Bank is on a similar journey and we can fast forward the implementation by adopting **Datahub** open sources that best relates to the ecosystem of Saxo Bank.

- The LinkedIn datahub team has been extremely **responsive**.

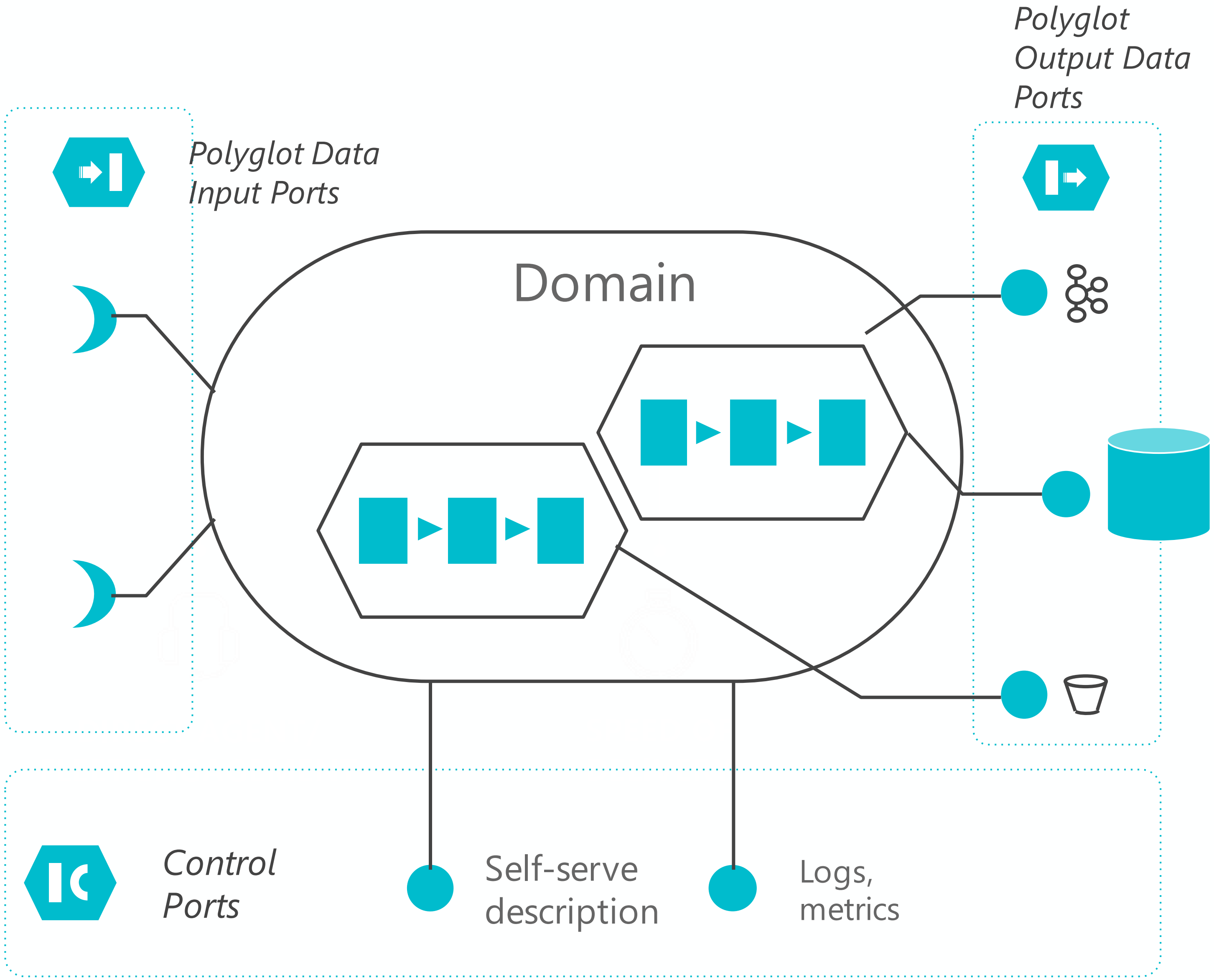
- Other digital natives have also recognized that the **incumbent solutions are not fit** for the modern age and have built their specific solutions.

PERSONAS: GOALS AND PAIN POINTS

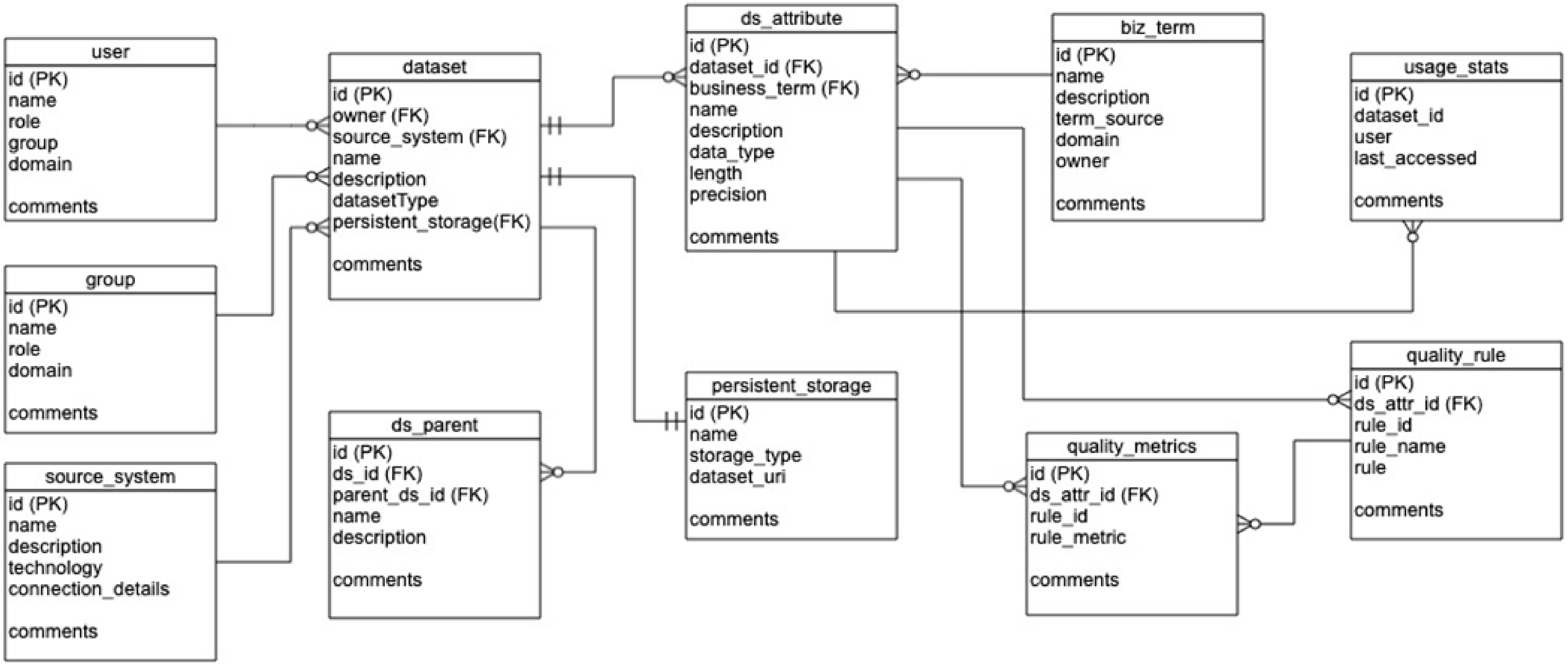
	Data Asset Owner	Data Steward	Data Governance Committee Member	Data Scientist	Data Consumer (Reporting)
Goal	*To be responsible and accountable for data*	*To define and document data standards*	*To get an overview of whole data in the org*	*To solve business problems based on data*	*To find data, its owner of data and anything else that helps compliance*
Pain Points	*Lack of clarity on ownership of data*	*New role in Saxo Bank, so yet to own full responsibility*	*Lack of clarity on what to mandate at what level (federated or data domain level)*	*Not sure if the data can be trusted for making the right decisions*	*Data missing/ incomplete* *See who can explain data elements*

DATA MESH APPROACH - PRODUCT THINKING

- Discoverable
- Addressable
- Self-describing
- Trustworthy
- Interoperable



TARGET METADATA MODEL



user
- id (PK)
- name
- role
- group
- domain
- comments

group
- id (PK)
- name
- role
- domain
- comments

source_system
- id (PK)
- name
- description
- technology
- connection_details
- comments

dataset
- id (PK)
- owner (FK)
- source_system (FK)
- name
- description
- datasetType
- persistent_storage(FK)
- comments

ds_parent
- id (PK)
- ds_id (FK)
- parent_ds_id (FK)
- name
- description
- comments

ds_attribute
- id (PK)
- dataset_id (FK)
- business_term (FK)
- name
- description
- data_type
- length
- precision
- comments

persistent_storage
- id (PK)
- name
- storage_type
- dataset_uri
- comments

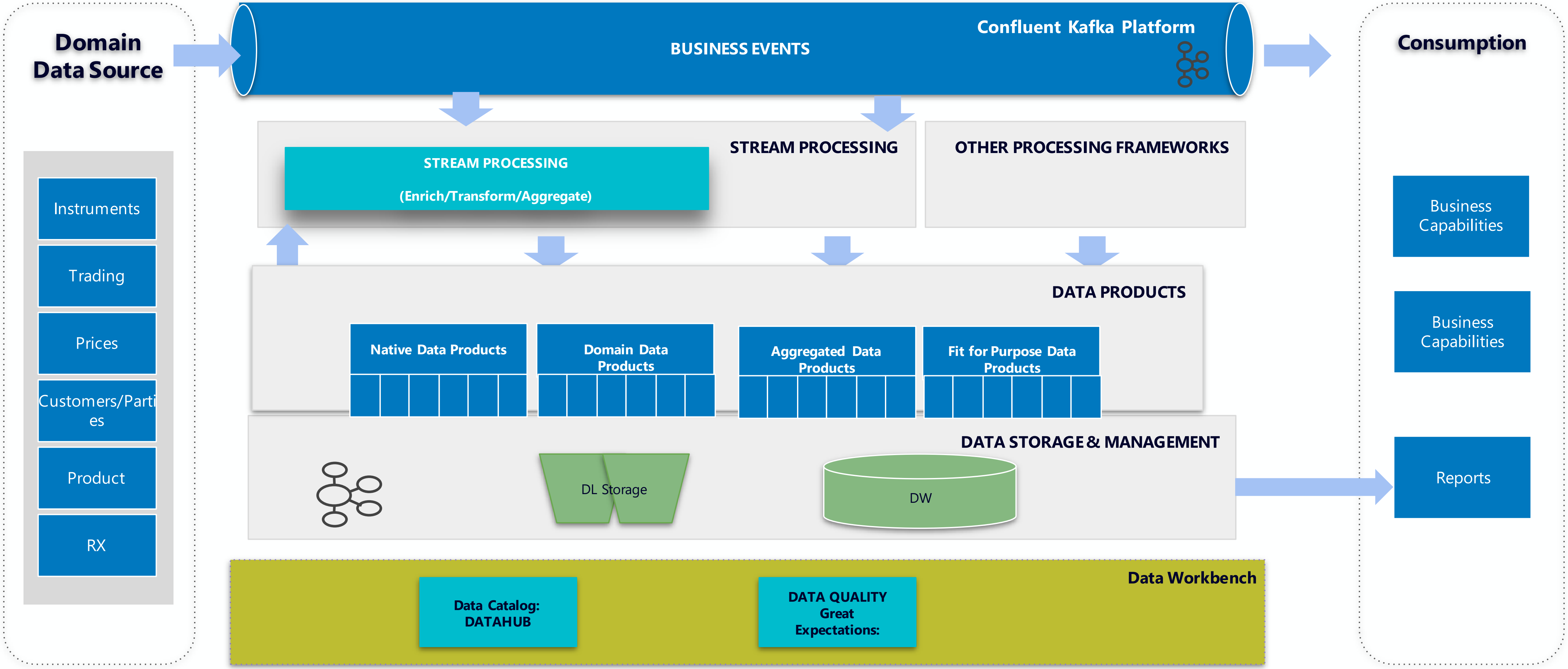
biz_term
- id (PK)
- name
- description
- term_source
- domain
- owner
- comments

quality_metrics
- id (PK)
- ds_attr_id (FK)
- rule_id
- rule_metric
- comments

usage_stats
- id (PK)
- dataset_id
- user
- last_accessed
- comments

quality_rule
- id (PK)
- ds_attr_id (FK)
- rule_id
- rule_name
- rule
- comments

DATA PLATFORM - HIGH LEVEL OVERVIEW



TOOLS EVALUATION PROCESS

Evaluation Criteria Definition

- SAXO features list
- SAXO initial evaluation params
- TW extended feature list

Shortlisted tools

- Includes initial & shortlisted list
- Data Catalog
- Data Quality

Evaluation Process

- Product documentation
- Software: Local installations
- Vendor questionnaire

SAXO BANK

DATA CATALOG TOOL EVALUATION

Prioritized Feature List

Metadata Search

- Full Text search on dataset name, attributes and tags
- Extensible search model

Metadata Export

- Export API

Architecture

- Cloud-native (Scalable & High Availability)
- Configurable
- Extensible

Security

- Authentication / LDAP
- Authorization / RBAC

Metadata UI

- Web-based UI to show metadata, governance attributes, tags and lineage
- Ability to edit and enrich attributes

Data Lineage

- Dataset lineage with upstream and downstream provenance
- Integration with data processing/orchestration tools

Alignment with Data Mesh

- Data as a Product
- Distributed Domain Driven Architecture
- Self-service platform

Metadata Ingestion

- Push-based REST API
- Pull-based adapters for Snowflake and CRM dynamics
- Extensibility

Data Stewardship

- Support for metadata enrichment and tagging
- Ability to flag a dataset

Total Cost of Ownership

- Licensing Cost
- Customization / Development cost

Support

- Release cycle
- Community support
- Commercial Support
- Documentation

Metadata Modelling

- Metadata entity for datasets, its users and attributes
- Business glossary & documentation
- Extensibility

Data Quality Integration

- Shows related Quality Attributes in UI
- Extensible to integrate with any DQ tool

Deprioritized

- Metadata Versioning
- Data Virtualization
- ML/AI capabilities

TOOLS LANDSCAPE

Data Catalog
Collibra
Informatica EDC
Alatian
Data.World
Azure Data Catalog - Prev2
Zeenea
Apache Atlas
Linkedin DataHub
Amundsen
Marquez

🟪	Commercial
🟦	Open Source
🟩	In House

TOOLS LANDSCAPE

Data Catalog
Collibra
~~Informatica EDC~~
~~Alatian~~
~~Data.World~~
~~Azure Data Catalog - Prev2~~
Zeenea
~~Apache Atlas~~
Linkedin DataHub
Amundsen
Marquez

🟪	Commercial
🟦	Open Source
🟩	In House

DATA CATALOG TOOL EVALUATION

SAXO BANK

Deep-dive analysis of the capabilities of shortlisted tools purely as per the teams understaning in Saxo's context.

	Datahub	Marquez	Amundsen	Collibra	Zeenea
Metadata Search *	🟩	🟧	🟩	🟨	🟩
Metadata UI Editable	🟨	🟨	🟨	🟩	🟩
Metadata Ingestion *	🟨	🟨	🟨	🟨	🟨
Metadata Modelling *	🟨	🟨	🟨	🟩	🟩
Data Lineage *	🟩	🟩	🟧	🟨	🟨
Metadata Export *	🟩	🟩	🟩	🟩	🟩
Data Stewardship	🟨	🟨	🟨	🟩	🟩
Data Quality Integration	🟨	🟨	🟧	🟩	🟧
Architecture *	🟩	🟨	🟨	🟨	🟧 Only supports AWS
Security *	🟨	🟧	🟨	🟩	🟩
Alignment with Data Mesh *	🟩	🟩	🟩	🟧	🟧
Support	🟨	🟨	🟧	🟩	🟩
Total Cost of Ownership *	🟩	🟨	🟨	🟨	🟨

🟦	Commercial
🟪	Open Source

🟩	Completely suitable
🟨	Partially suitable
🟧	No/Minimal suitability

***Disclaimer**:* *Based on Saxo evaluation criteria and interpretation of product capabilities*

- **Push Based approach that supports Event Driven architecture.** The solution built on the principle of **self-service** and producers know their data better and they can provide the rich metadata so that it helps in discover the data-assets and encourages consumption.
- Detailed Evaluation was carried out from Saxo perspective.
- Possibility of evolution
- Extensibility with the open source and evolve the tool as per needs. Right fit from feature perspective in terms of Data Governance maturity of the organization.
- Leverage from larger community needs and also influence internal process when needed.
- Reputation of LinkedIn, their success in Kafka and datahub scaling internally to LinkedIn volume
- Promising Roadmap

DATASET ONBOARDING - RESPONSIBILITIES

PRODUCERS

Metadata

- Dataset/Data Product Metadata
  - Ownership Information
  - Reader Information
  - Topic Configuration Details
  - Dataset Structure (AVRO Schema)
  - Business Term mapping
  - Source Dataset Definition (Optional)
- Quality Rules

Data Engineering

- Domain Transformations
- (Kafka Stream)

CONSUMERS

Metadata

- Consumer Details
- Usage Details
- Target Dataset details

Kafka PLATFORM-Lean Team

Engineering Capabilities

- Supporting New Domains
- Metadata Integration
- DQ Integration

Thank You

Q&A?