LF AI & DATA

The EGERIA Metadata Show

DataOps

What is it and Egeria's role

Mandy Chessell CBE FREng
Egeria Open Source Project Lead

Today's Agenda

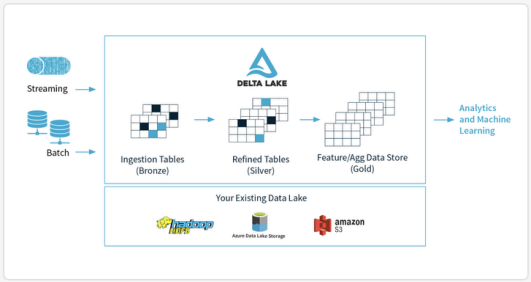
- New DataOps Technologies
- Value of Egeria to DevOps
- Information Management Practices
- Egeria's role in DataOps

Where is the interest coming from?

Delta Lake

Delta Lake - Reliable Data Lakes at Scale

Delta Lake is an open-source storage layer that brings ACID transactions to Apache Spark™ and big data workloads. (84 kB)



DataKitchen

The Complete Enterprise DataOps Platform

Simplify complex toolchains, environments, and teams. Automate your end-to-end data workflows so cross-functional teams can quickly innovate, test, and deliver error-free, on-demand insight.

Dec 10th, 2020 (9 kB)

LakeFS

Atomic Versioned Data Lake - LakeFS

lakeFS is an open-source tool that transforms your object storage to Git-like repositories. Start managing data the way you manage your code.

dvc.org

Use Cases

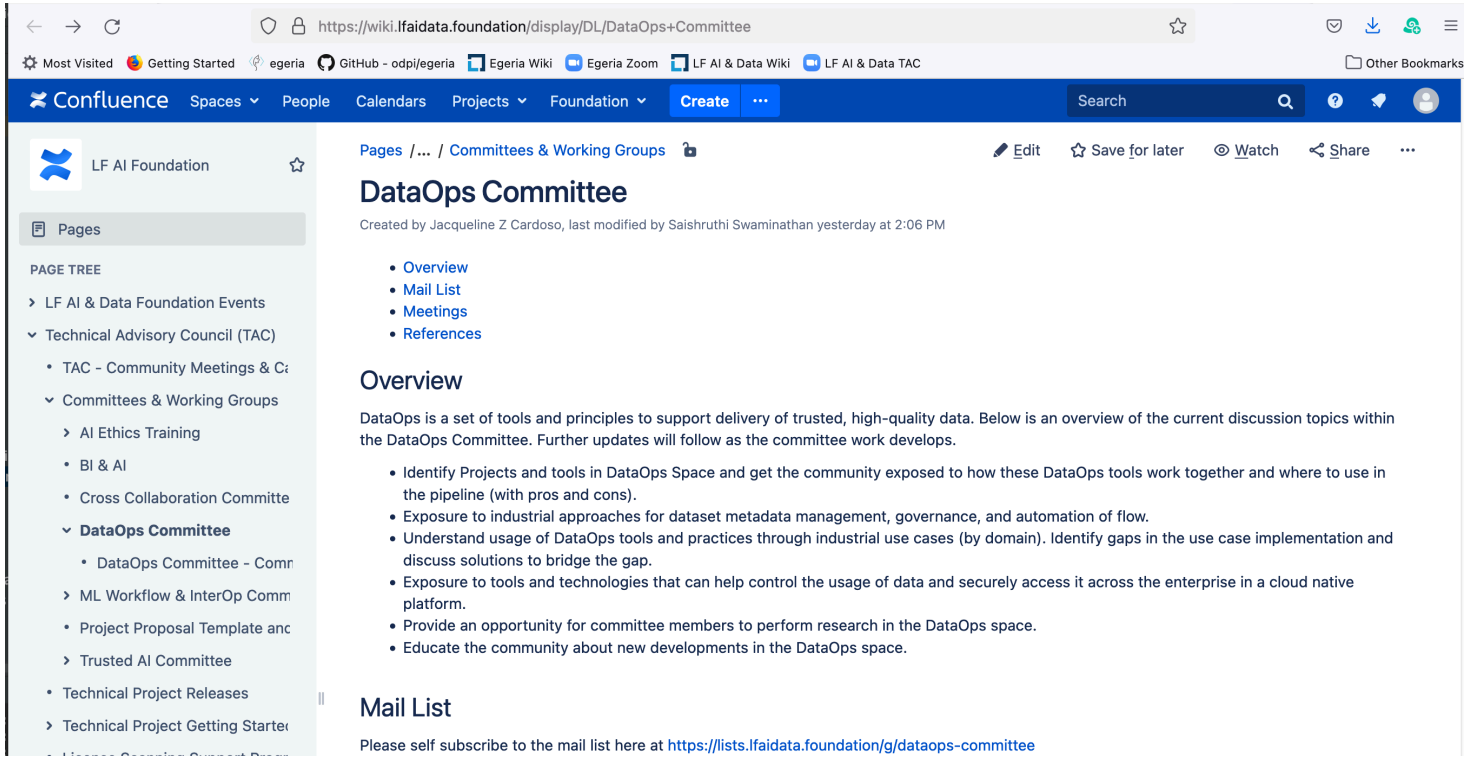
Open-source version control system for Data Science and Machine Learning projects. Git-like experience to organize your data, models, and experiments.

quiltdata.com

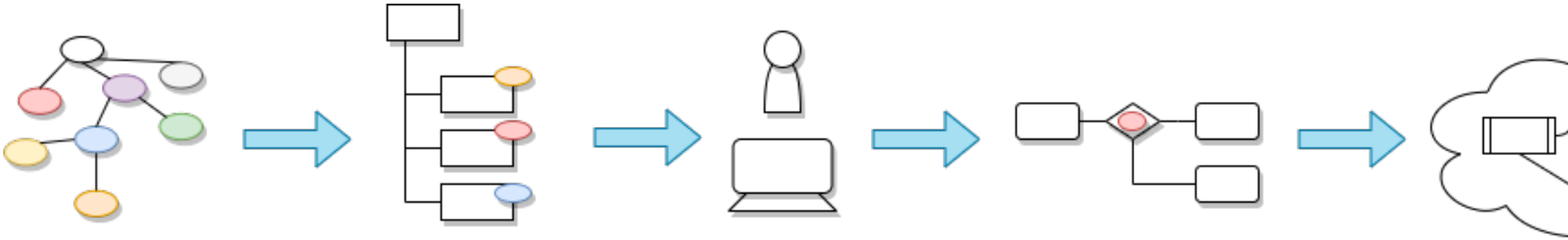
Quilt Data

Quilt is a versioned data hub for AWS. Quilt integrates files into datasets that your whole company can discover, understand, and trust. Quilt is instant infrastructure to bring drugs, targets, and therapies to market faster. (16 kB)

New DataOps committee



Tool chain example

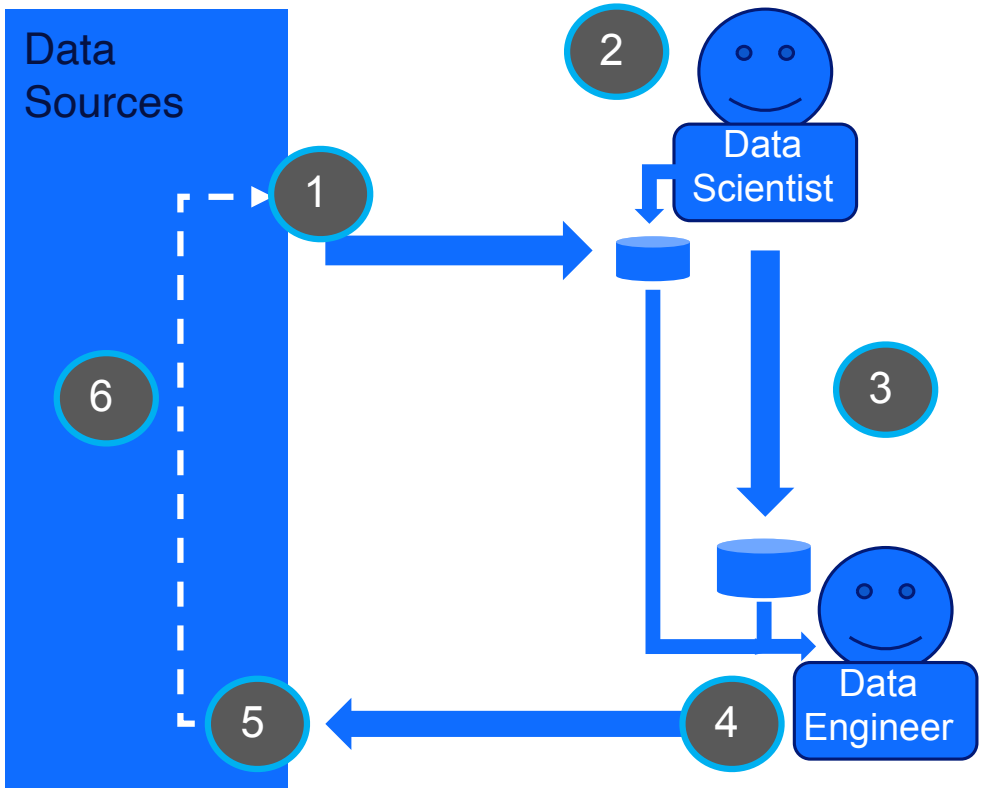


Conclusions from DevOps discussions

- Embedding knowledge about data into artefacts consumed by developer tools, it is possible to instrument the dev ops pipeline to ensure services processing sensitive information are deployed into appropriate environments.
- Increased automation enables the testing to be more focused on points of variability.

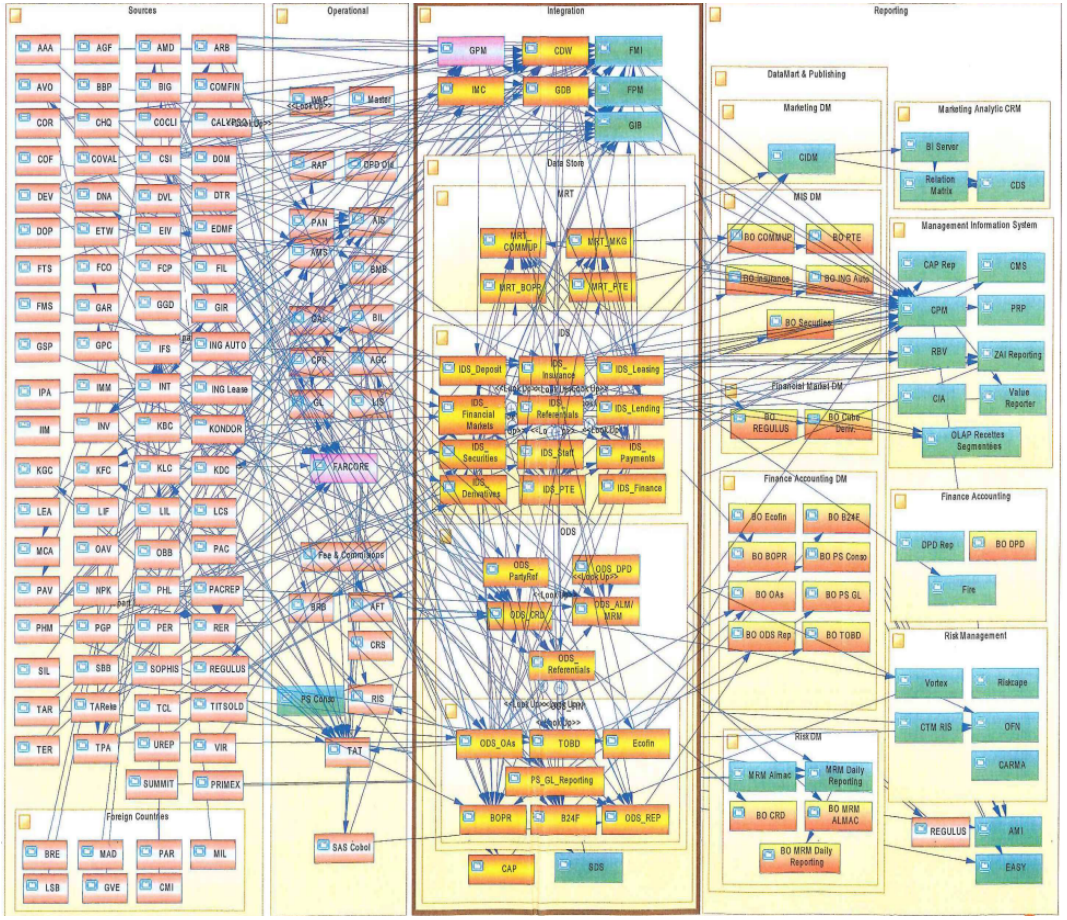
Data flows around analytics

1. Supply data
2. Refactor
3. Supply analytics model and test data
4. Embed analytics model in software service
5. Deploy
6. Results feedback to data scientist



Remember this picture …

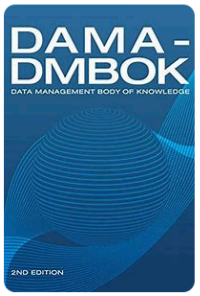
- How can you be sure which data set to use?
- How can you make the integrations more efficient and less error prone?
- If a link or a system goes down, how long do you have before data corruption occurs?
- If an invalid value was added to one of the systems yesterday, how far has the contamination spread?



Not a new science

- The principles and practices behind this are well understood.
- Organizations are inhibited from implementing them because they require the following behaviours:
  - Systematic
  - Purposeful
  - Collaboration with strangers
- Egeria seeks to lower the barrier on all three of these inhibitors:
  - Pervasive
  - Integrated
  - Broad in scope

DAMA-DMBOK: Data Management Body of Knowledge: 2nd Edition

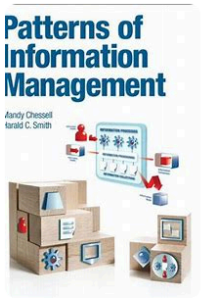


Goodreads 4.2/5

Amazon 4.7/5

The Data Management Body of Knowledge presents a comprehensive view of the challenges, complexities, and value of effective data management. Today's organizations recognize that managing data is central to their success. They ... +

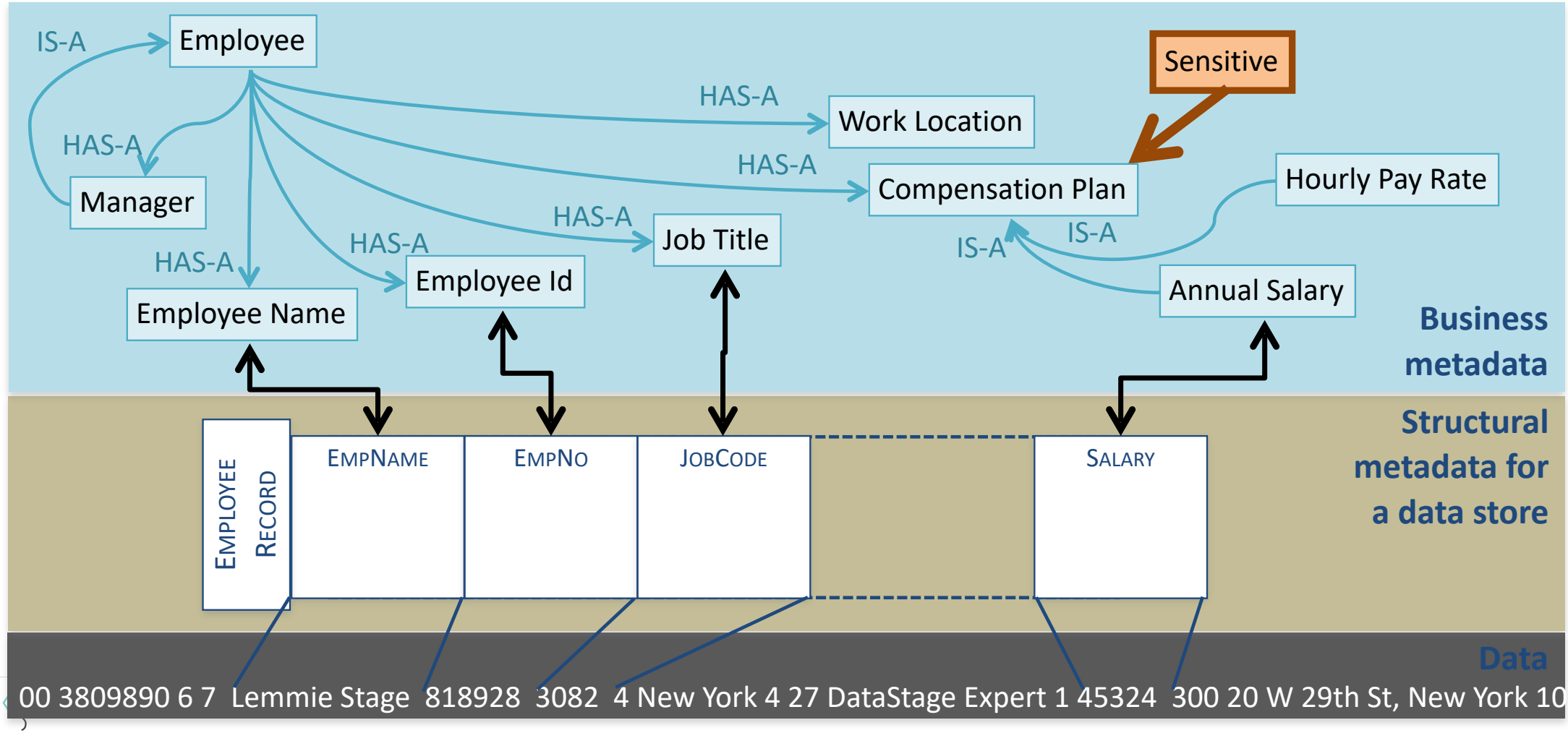
Patterns of Information Management



Look inside

Use Best Practice Patterns to Understand and Architect Manageable, Efficient Information Supply Chains That Help You Leverage All Your Data and Knowledge In the era of "Big Data," information pervades every aspect of the organization. Therefore, architecting and managing it is a multi-disciplinary task. Now, two pioneering IBM® arc... +

What is your view on the quality of this data?



Matching up records

account: a.steiff@mail.com
name: Alistaire Steiff
address: 4 Button Cottages, Tiletown

Order: 486801
Customer: Mr A Steif
Deliver to: 4 Button Cottages, Tiletown

Account: 0243-8005-5333-0055
Customer: Mr Alistair Steiff
Deliver to: 4 Button Cottages, Tiletown

John Smith … verses Bartholomew Virgil Fencepost

… plus date of birth and address

… 5/3/1979 verses 3/5/1979

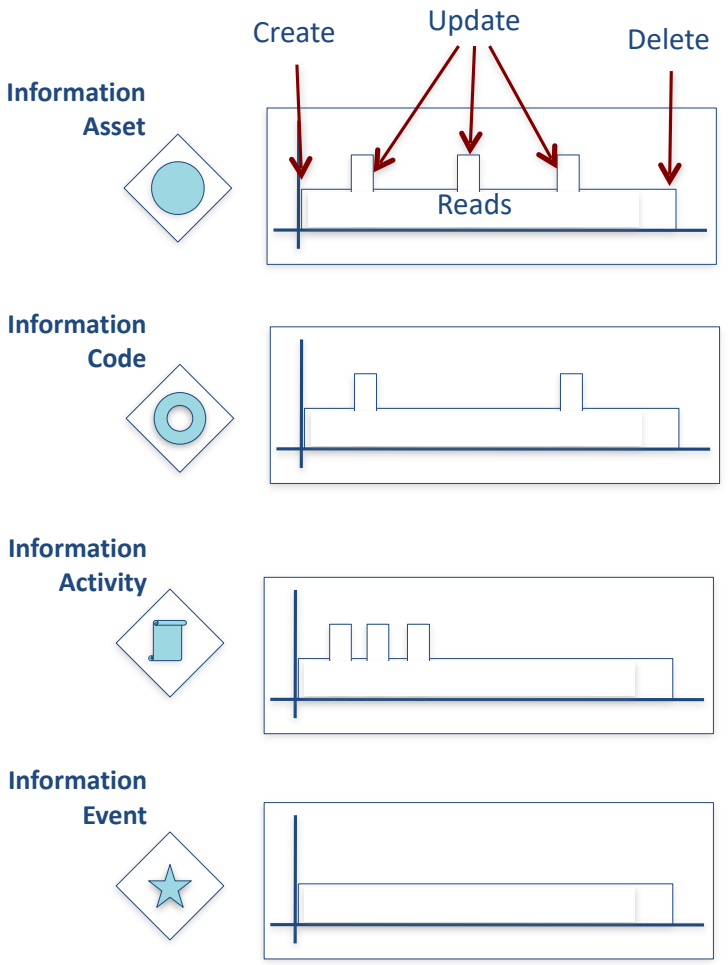
Frequency of values

Adding more data

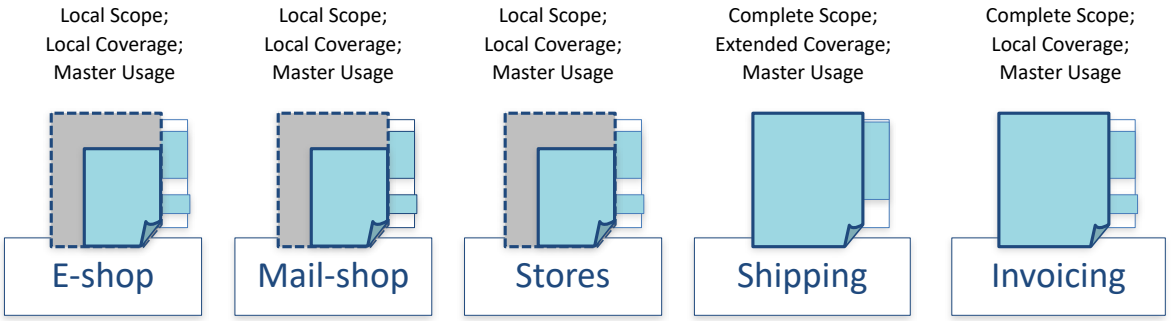
Standardising formats

Lifecycles of information

- Information Asset
  - Slowly changing.
  - Highly duplicated.
  - Multiple formats.
- Information Code
  - Set based information
  - Sets change frequently; individual codes rarely change.
  - Many representations; widely distributed and mapped.
- Information Activity
  - Rapid change during active phase then little to no change.
  - Limited distribution while active.
- Information Event
  - No change once created.
  - Distributed as required.

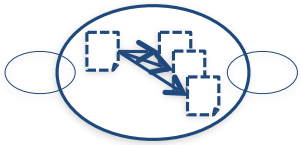


Example Product Information Collections

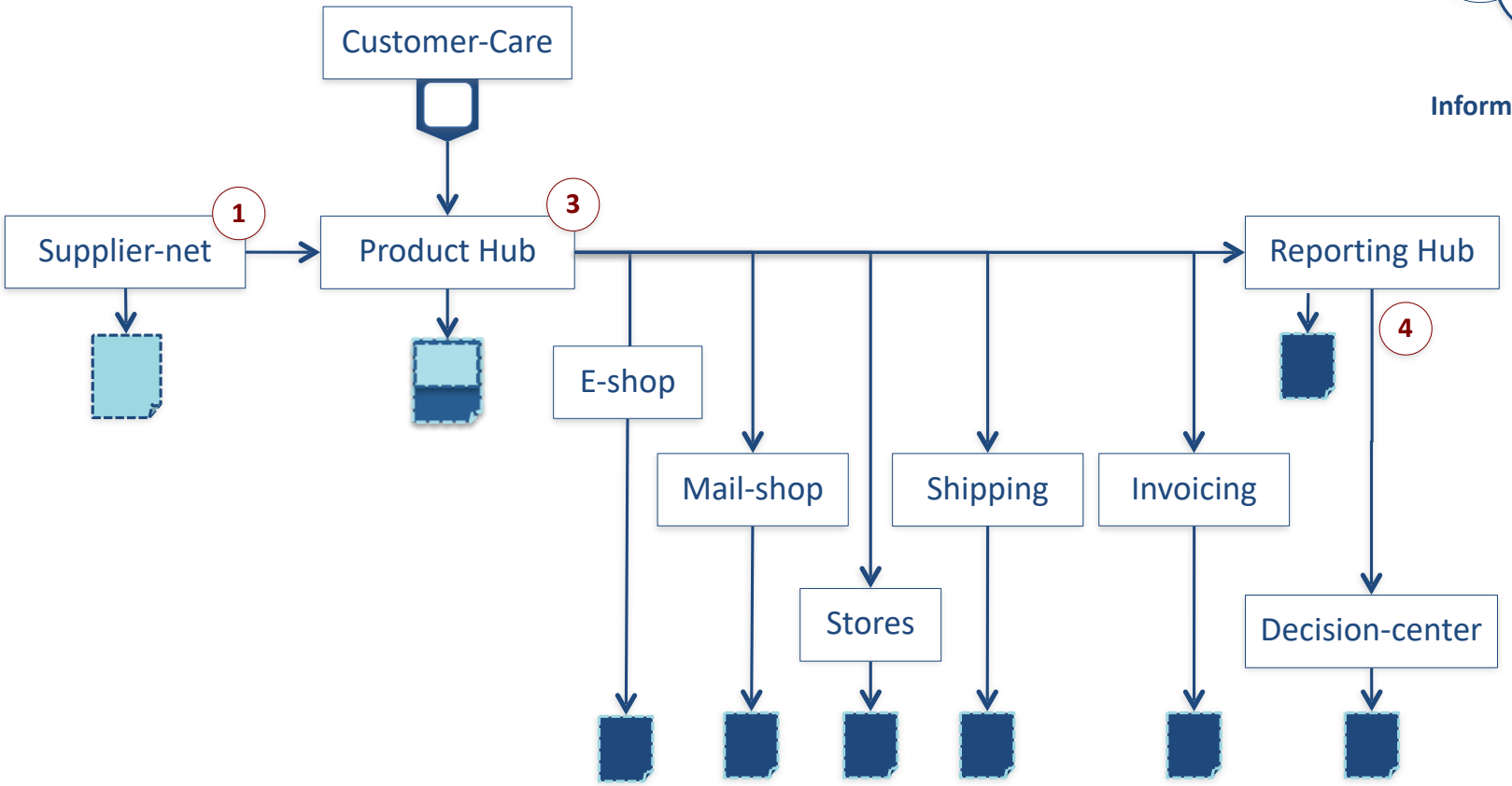


- Most application have local information collections with master usage for their principle information needs.
- Most applications have local scope. The exceptions are where there is a centralized function where complete scope is required.
- Coverage is rarely complete.

Example Product Information Supply chain

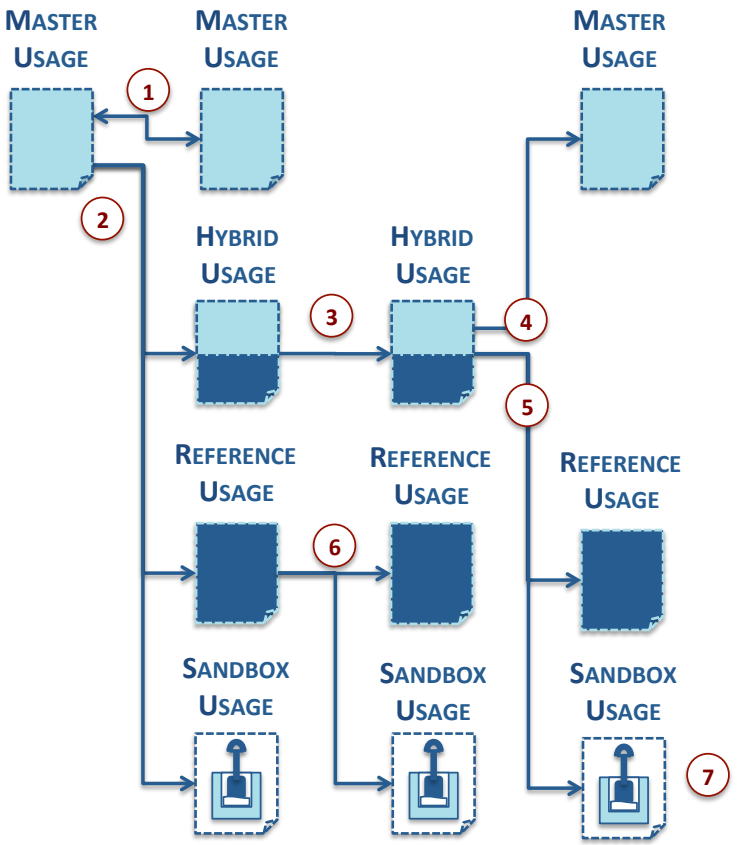


Cascading Information Supply Chain



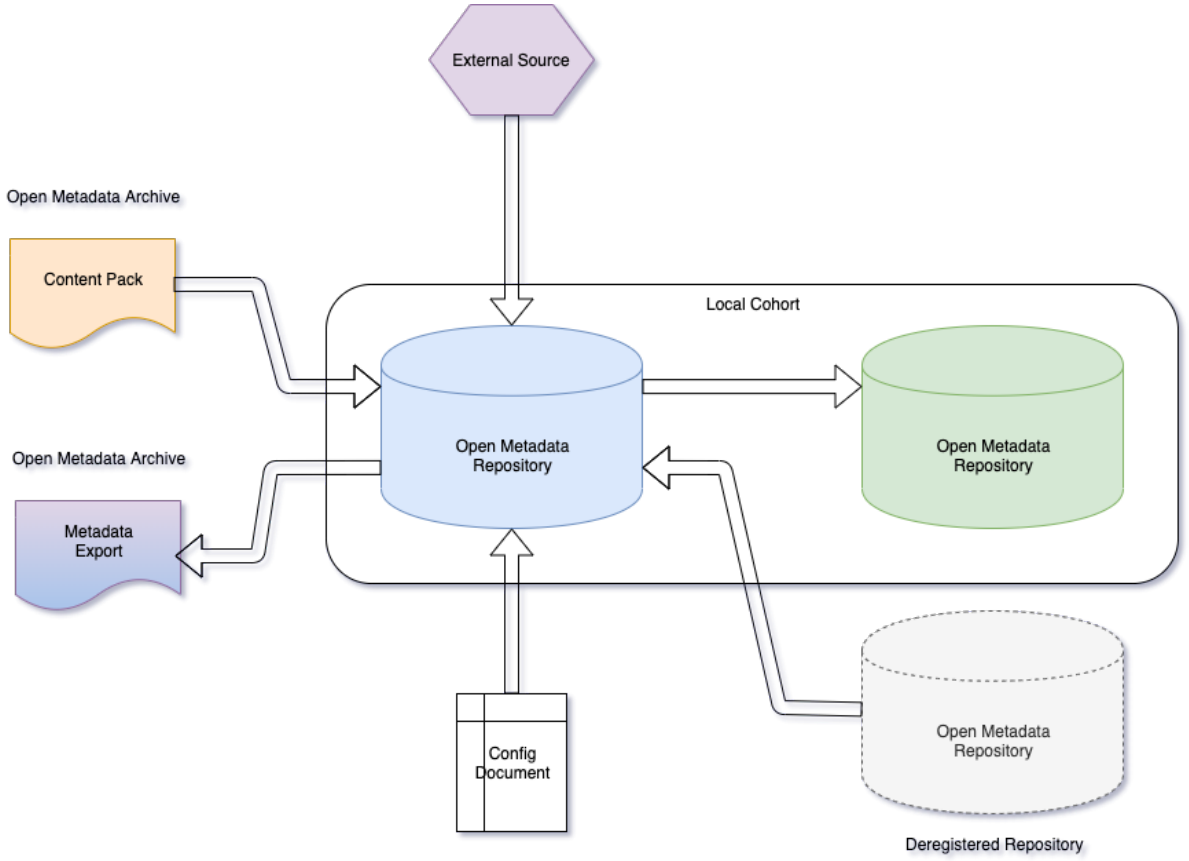
Example of an information supply chain

- As data is copied from its authoritative source, it becomes read-only

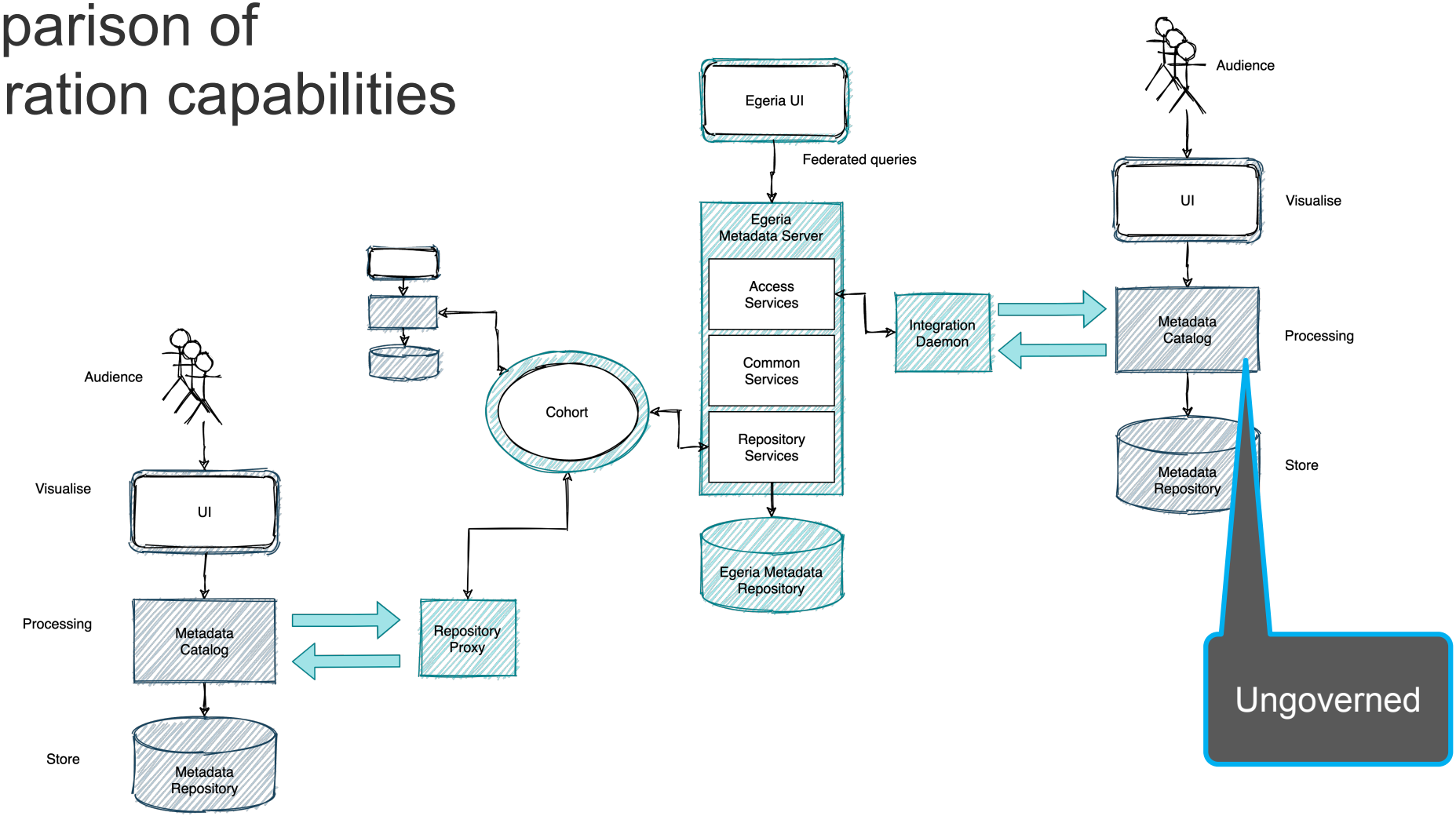


Egeria's metadata provenance

- Where are the update rules enforced?



Comparison of integration capabilities



Role of Egeria in DataOps

- Provision of context for the DataOps pipelines
- Capture of lineage

Open forum

THANK YOU!

LinkedIn: http://www.linkedin.com/pub/mandy-chessell/22/897/a49

Slack: https://lfaifoundation.slack.com/archives/C01F40J2XA8

Email: egeria-technical-discuss@lists.lfaidata.foundation