#StrataData

The vindication of Big data.

How Hadoop is used in Santander UK to defend privacy.


WHO
HOW
WHEN
?
WHAT
WHY
WHERE

Why GDPR?

Stronger rules on data protection from May 2018
Evolution of data protection for the digital age

Customer data like money is an asset, valued, protected and owned by them

Improved Trust
Metadata investment driving efficiencies and empowering self service analytics

About us

Lidia Crespo – *Technical Business Manager*
Santander UK

https://www.linkedin.com/in/lidia-crespo-parra-12859145/

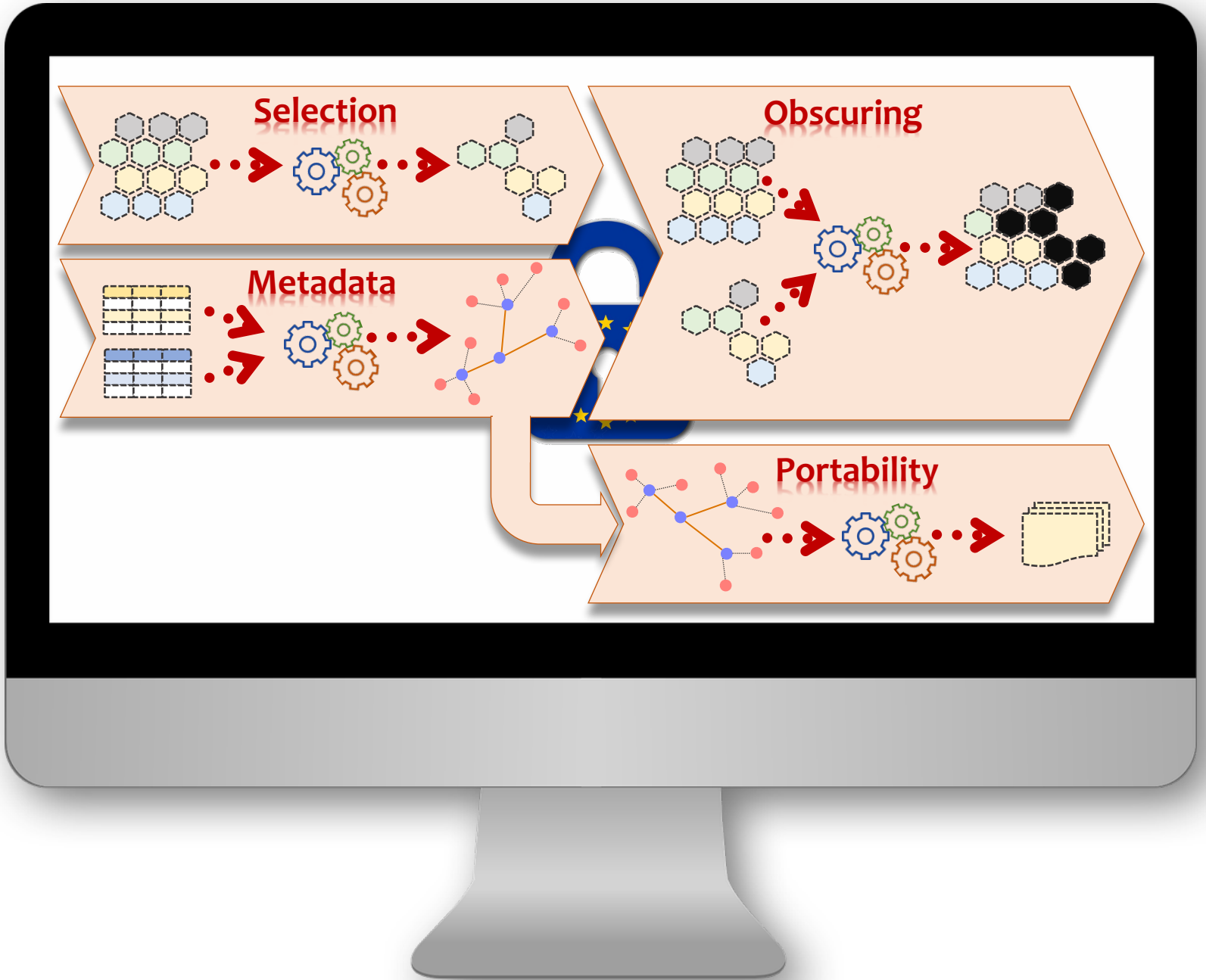
Mauricio Lins – *Big Data Architect*
Everis Consultancy UK

https://www.linkedin.com/in/mauricio-lins

Focus Areas

☆ New Technical Processes

- Multiple systems with customer data
- Selection + Right To Be Forgotten
- Data obscuring / deletion
- Feedback file to confirm all deletions
- Portability for the customer



Challenges

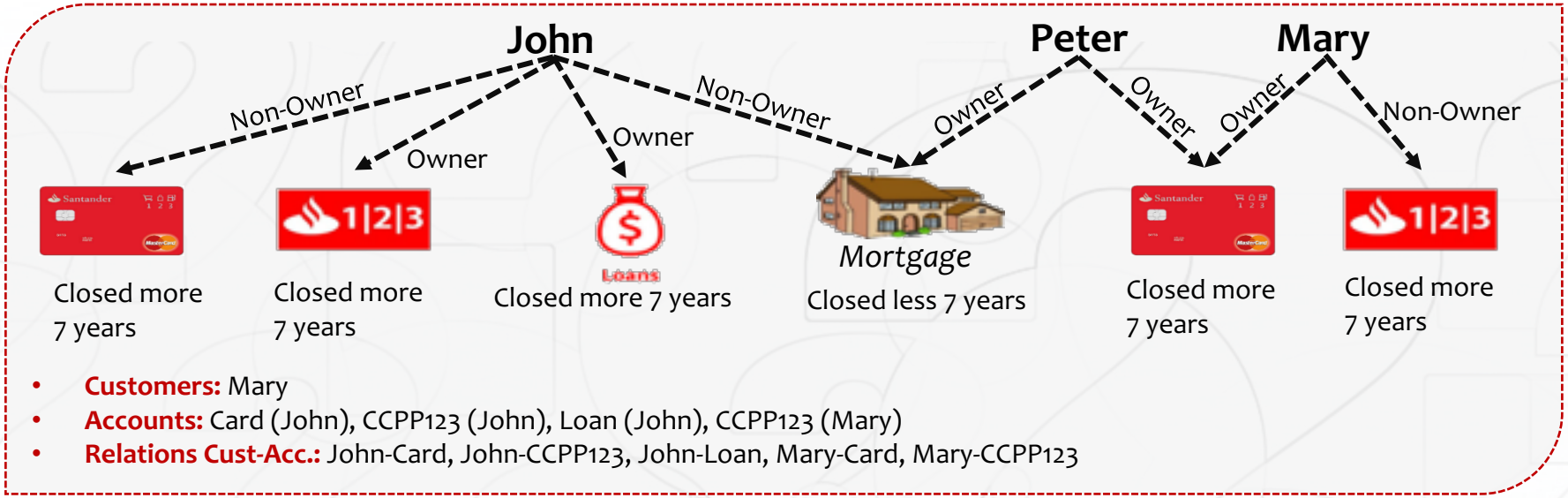
Scenarios to Select Data

Multiple Heterogeneous Systems

Data Status Synchronization

Statics Systems Become More Flexible

Scenarios to Select Data



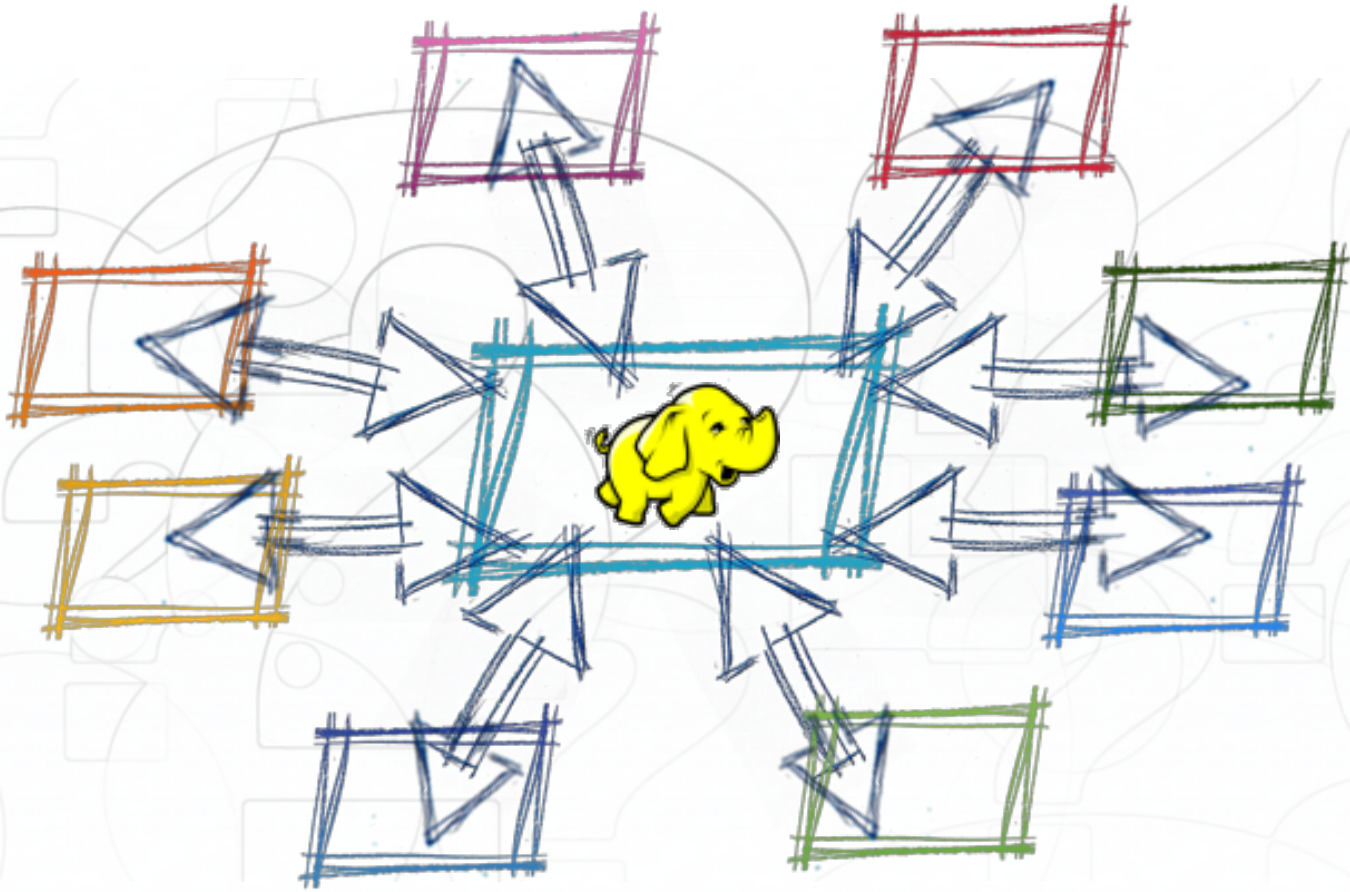
☆ Understand variety of customer relationships, joint accounts, business participants

☆ Comprehensive data retention needs: Active complaints, collections, remediation, regulatory landscape…

☆ Implement a robust exception process that allows flexibility to cover changing conditions in an agile way, with minimum impact in other parts of the process

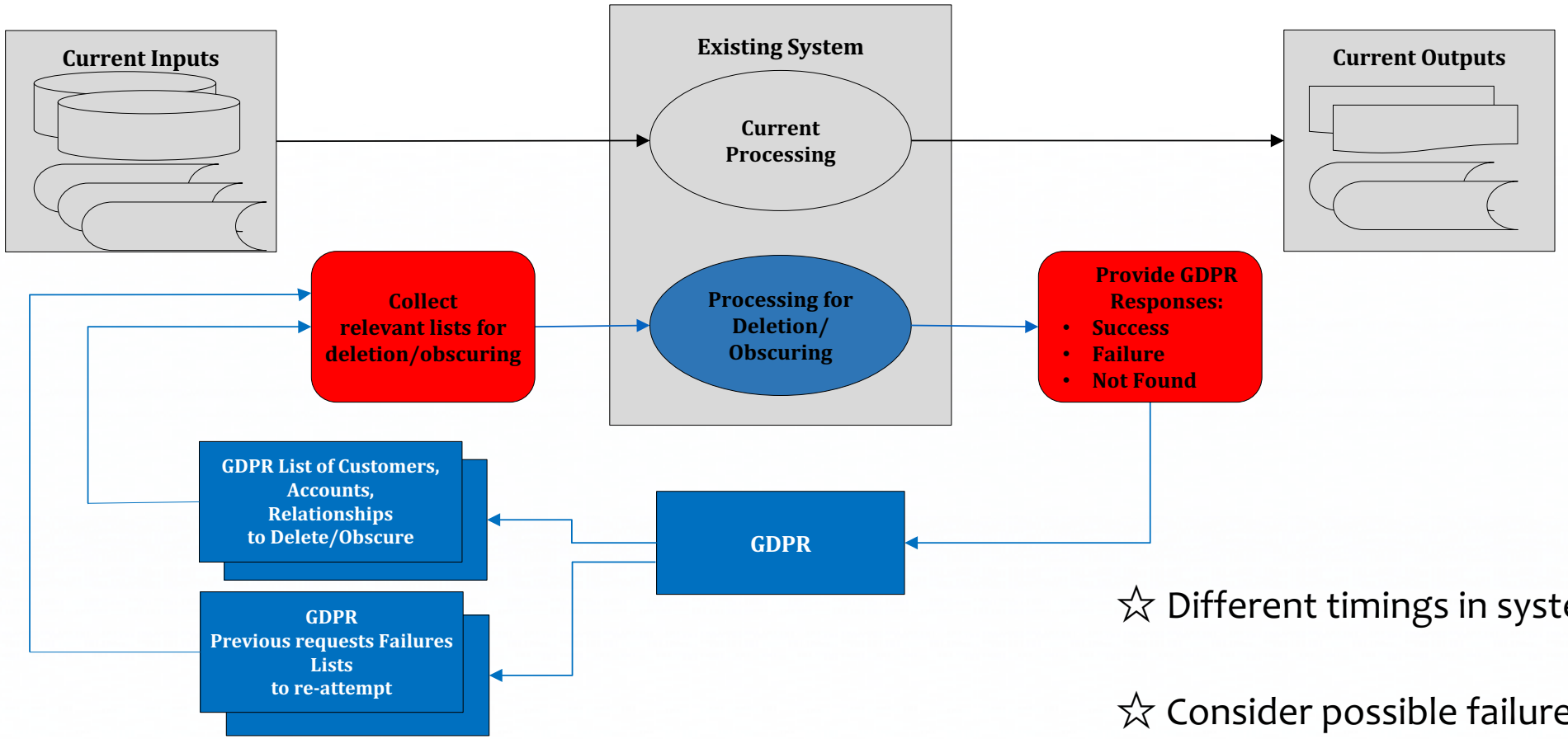
Multiple heterogeneous systems

- ☆ Complex ecosystem of IT systems and end user computing sources documented in the inventory – consolidation and completion of inventory
- ☆ Multiple systems with customer data
- ☆ Data flow processes - Downstream vs original operating systems



Data Status Synchronization

Coordination key to maintain consistency across systems



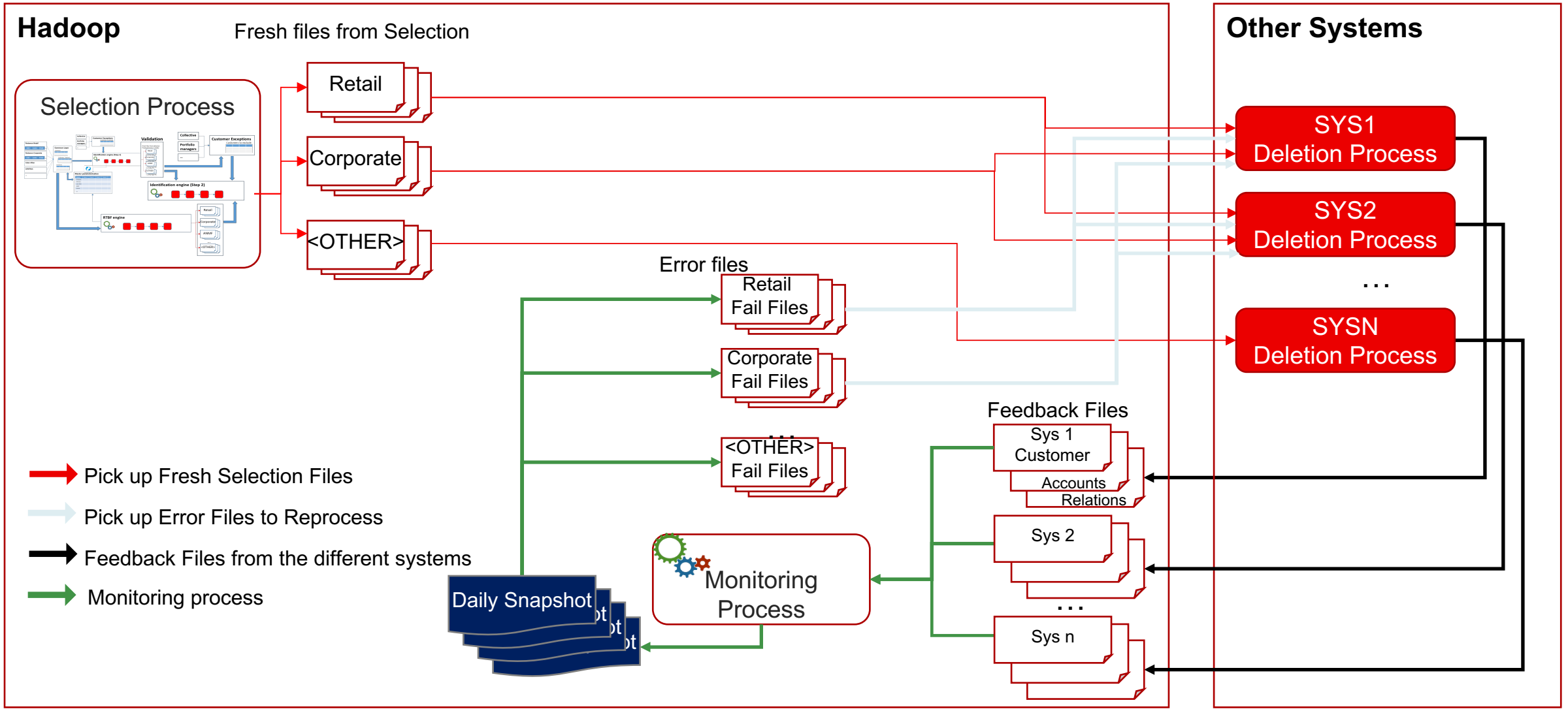
☆ Different timings in system processes

☆ Consider possible failures and allow reattempt

☆ Complex dependencies in downstream systems

Data Status Synchronization

Coordination key to maintain consistency across systems



Volume, variety and velocity of change

Systems become more agile, faster speed of change

Unstructured data with limited documentation

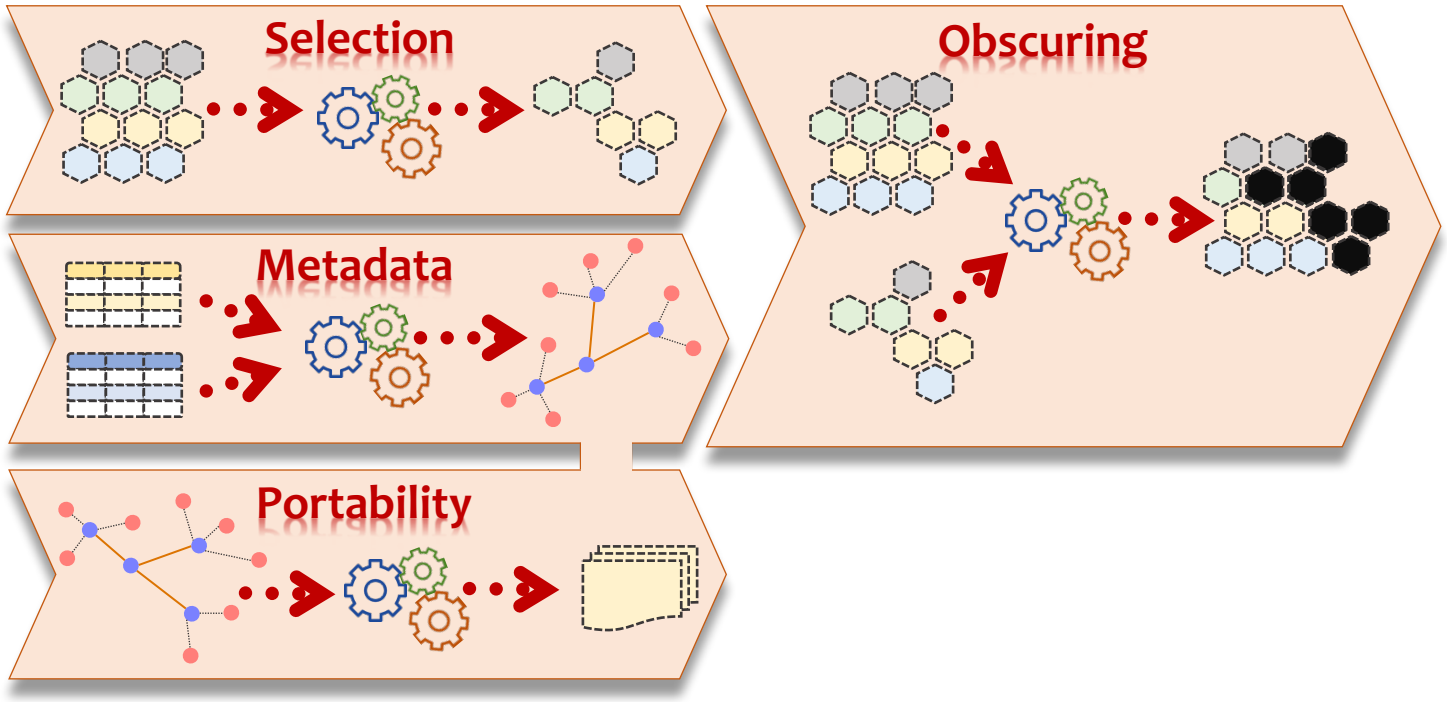
Discovery SW for detection of personal metadata, reducing resources and flagging possible manual errors

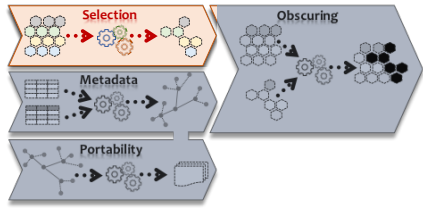
Importance of metadata to adapt to the changing landscape

Big Data Processes Implementation

GDPR Processes

Implementation Steps





How to Select

More Flexible approach…

DATA LAKE AS A SOURCE

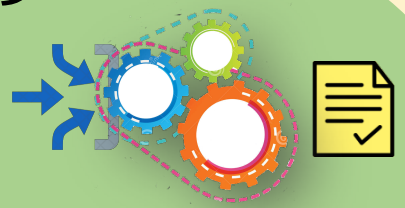
1

Applying the defined rules the eligible data is selected from the Data Lake

2

The business users validate the outputs, ensuring consistency, and create the exceptions list for special cases

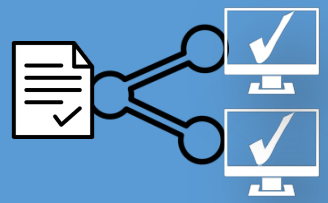
3



With the validation results and RTBF requests the Selection process is rerun to generate the final version

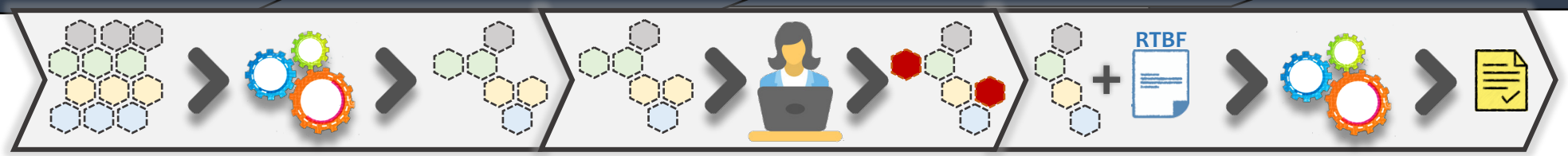
DATA LAKE AS A DRIVER

4



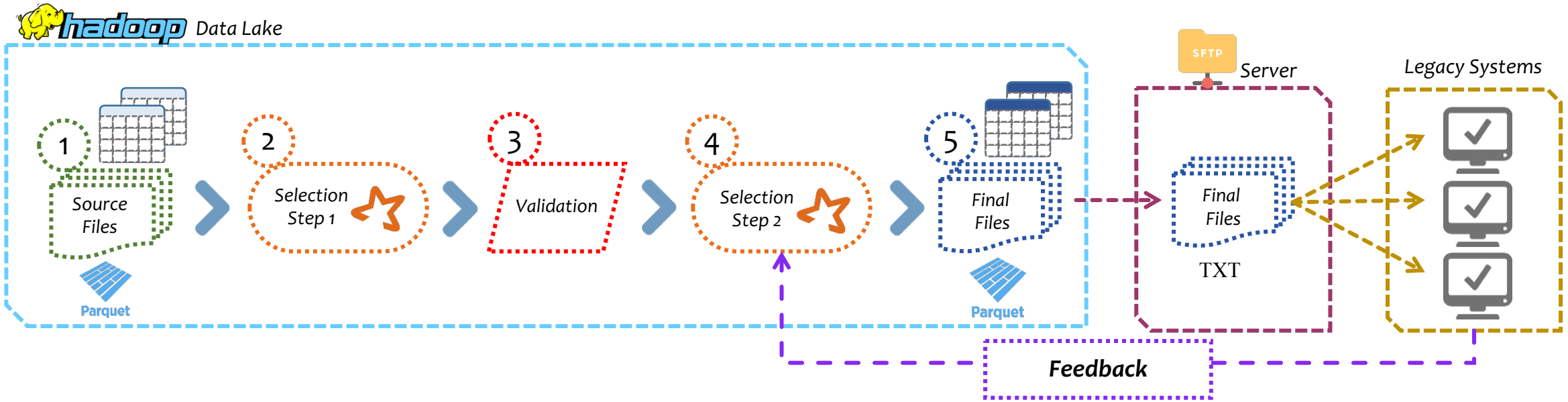
The final files are shared with all systems in scope to be obscured/deleted

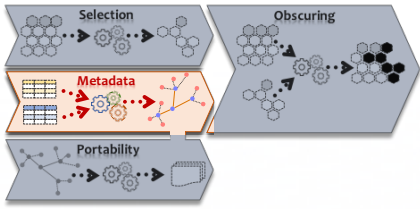
SELECTION STEP 1 | VALIDATION STEP | SELECTION STEP 2 | OUTPUTS



How to Select

In a technical view with the Legacy System's Feedback





Metadata

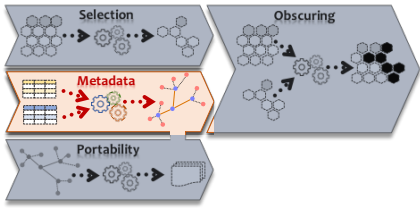
Importance of the Metadata & Detection of Personal Data

☆ What is PII?

- Scope clarity
- Logical (e.g. address) vs physical (tables/fields)
- Unstructured data

☆ Navigation

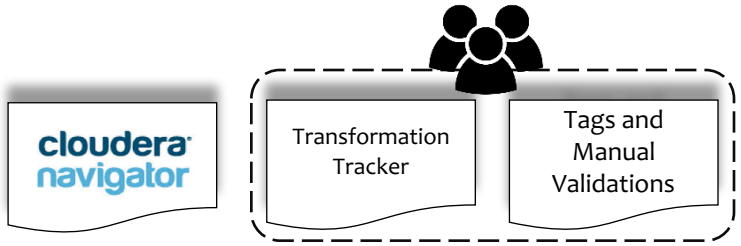
- Extended graph discovery to connect processes in separate products
- Built from the dictionary through discovery tools and traceability



Metadata

Importance of the Metadata & Detection of Personal Data

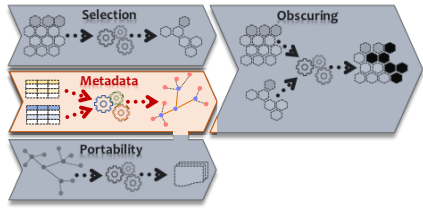
1. Periodic export of Cloudera Navigator Data To Hive Tables, containing the data lake representation



Managed by Users:

2. GDPR tags included on the transformation tracker

3. Discovery tools + Manual validation / correction on missing information

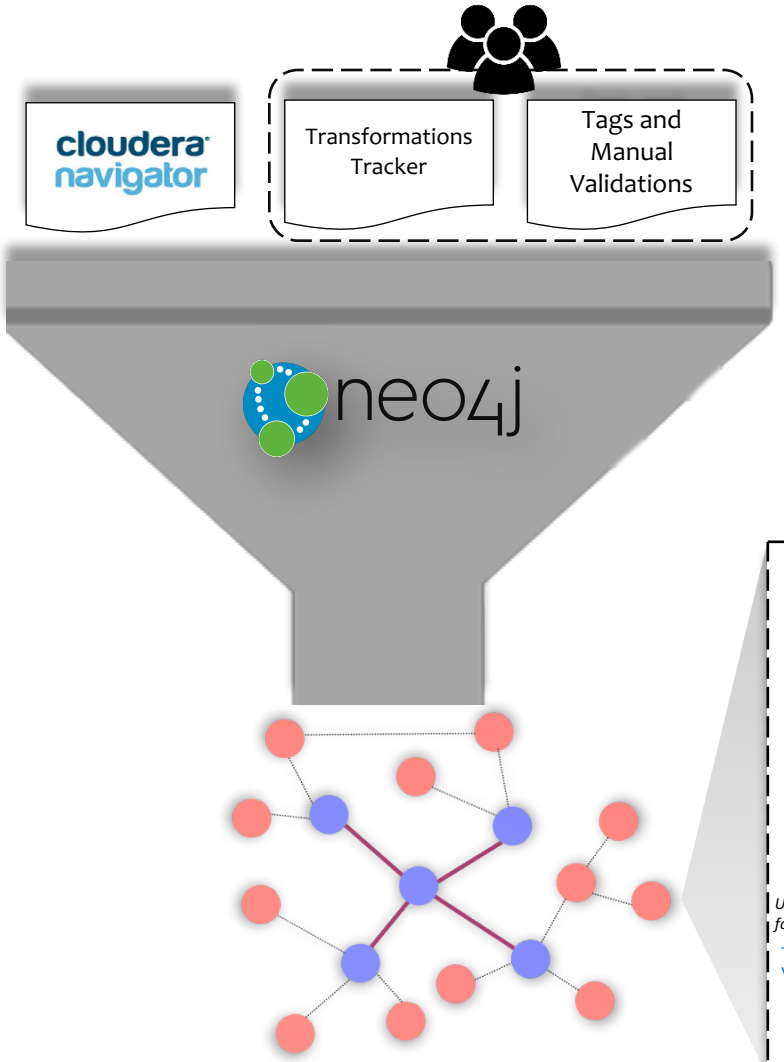


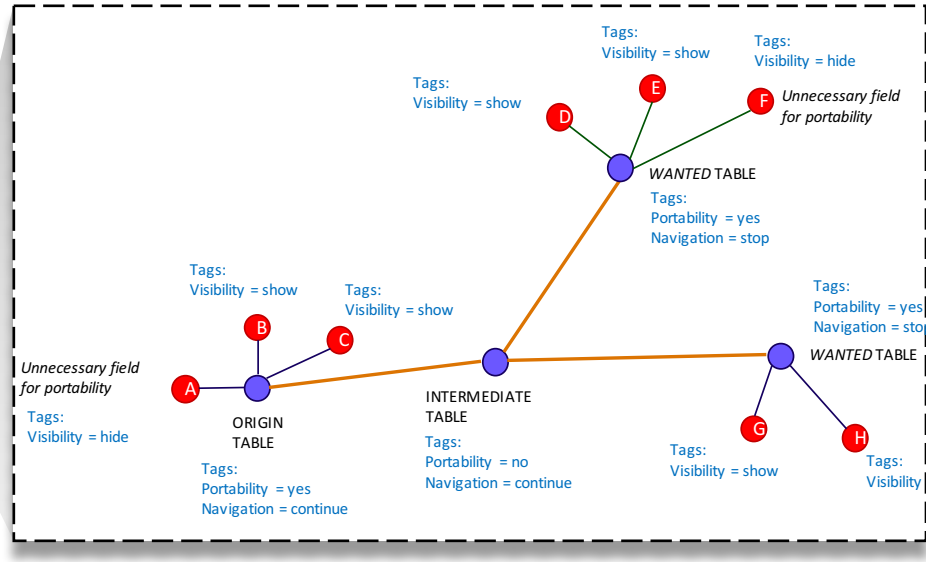
Metadata

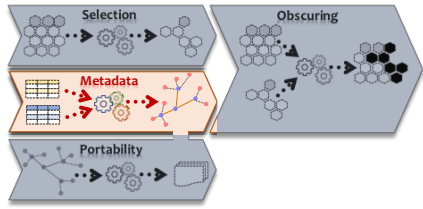
Importance of the Metadata & Detection of Personal Data

4 Metadata updates the Graph that represents the Data Lake with the new tags

5 The Graph Navigation Process obtains the model from neo4j and the records to search for from the selection files and navigates through the graph model to identify all records associated, existing in the metadata







Metadata

With More Details

Data Sources

Periodic export of Navigator Data To Hive Tables, containing the data lake representation

cloudera navigator

GDPR tags included on the transformation tracker

CDO

Transformation Tracker

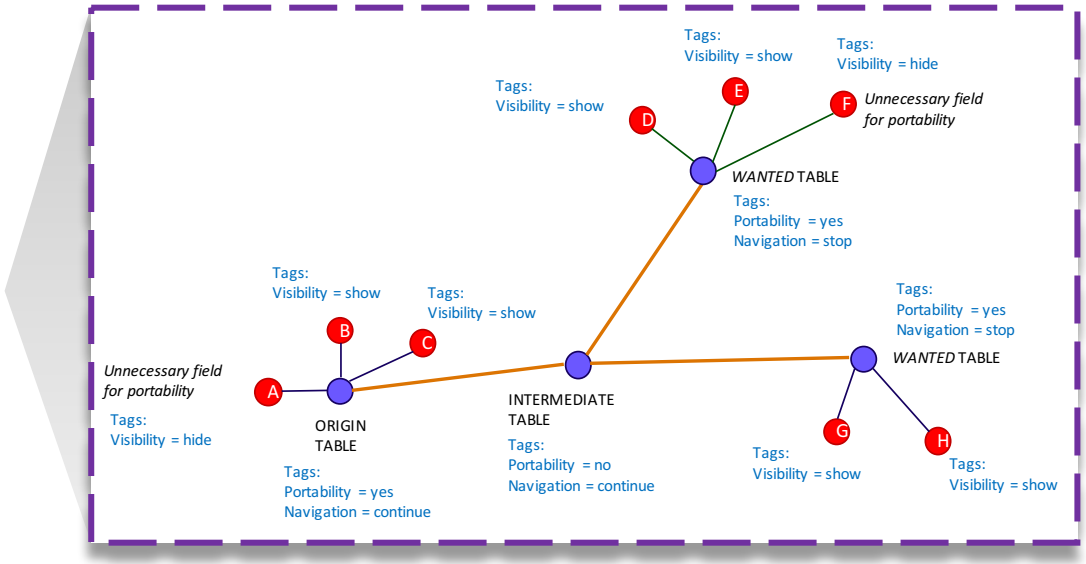
Source Tracker

Data Lake Metadata Update

Java

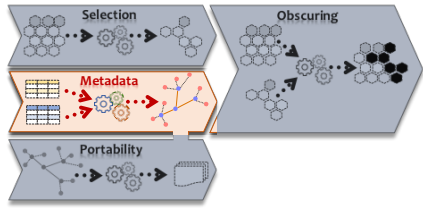
neo4j

Metadata updates the Graph that represents the Data Lake with the new tags



GDPR Tags:

- "Portability"
- "Visibility"
- "Navigation"



Metadata

With More Details

Data Sources

Periodic export of Navigator Data To Hive Tables, containing the data lake representation

cloudera navigator

CDO

GDPR tags included on the transformation tracker

Transformation Tracker

Source Tracker

Data Lake Metadata Update

Java

neo4j

Metadata updates the Graph that represents the Data Lake with the new tags

Java

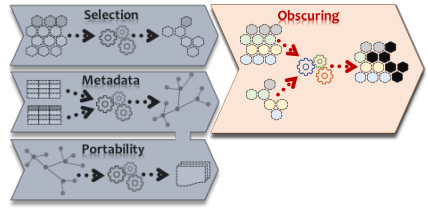
Metadata Output

JSON

JSON Sample:

{"db":"XYZ", "table":"tbABC",
"level":[…],
"fields":[…],
"joins":[…],…}

GDPR Obscuring Big Data Implementation



Data Obscuring

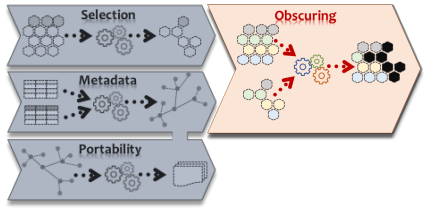
Obscuring vs Delete

Data Deletion

- ✓ Minimizes the volume of data stored, improving performance in many processes
- ✓ Cascade deletions to maintain Data Integrity across applications

Data Obscuring

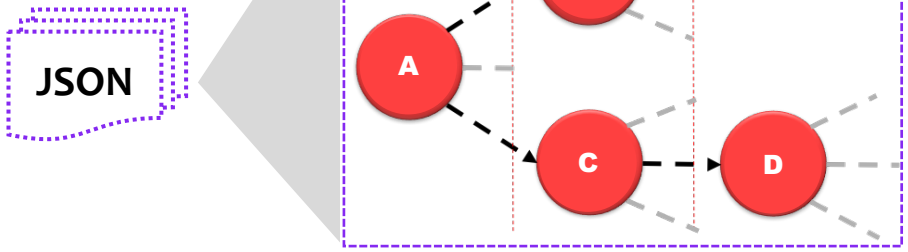
- ✓ Avoids possible impact on referential integrity
- ✓ Phased approach
  - x Incremental addition of systems to the scope
  - x Configurable addition of datasets and fields
- ✓ Certain non personal data can be used for improved customer service



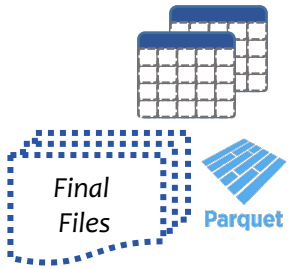
Pre Obscuring Process - Discovery Phase

The 1st step (Preparation)

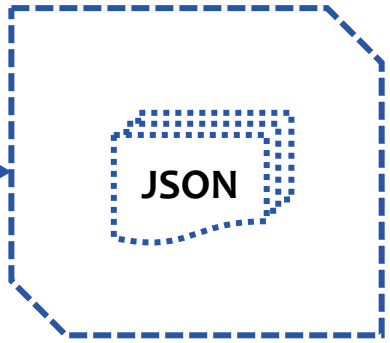
Metadata Output



Selection Output



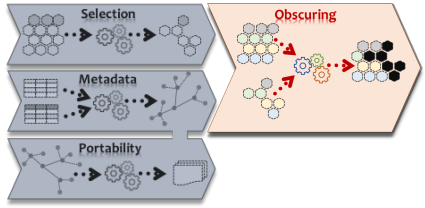
The process uses the Metadata to know which tables are in the scope and also uses the Selection results files to identify the ID's on the Metadata tables



JSON Sample:

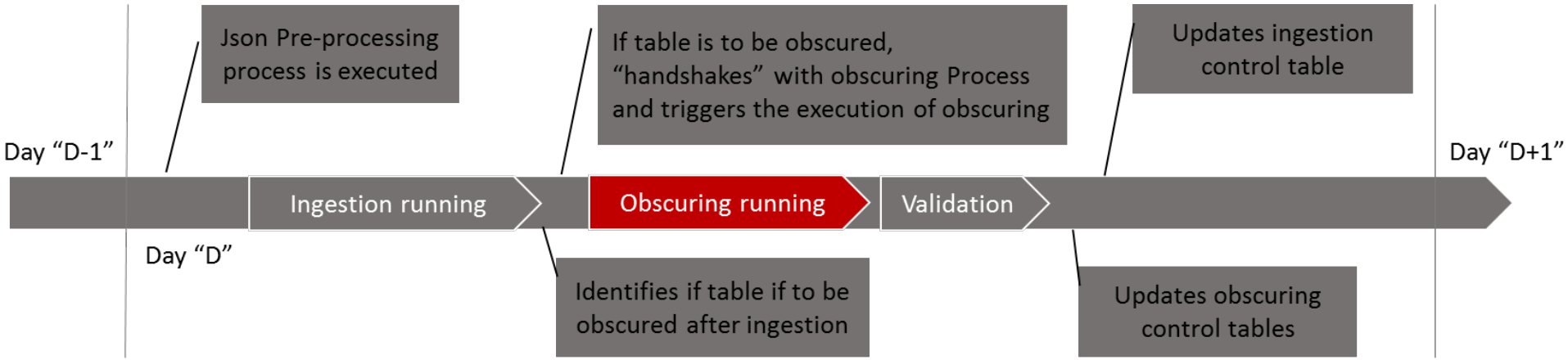
{"db":"XYZ", "table":"tbABC",
"keys":[…],
"fields":[…],
"actions":[…],…}

hadoop *Data Lake*



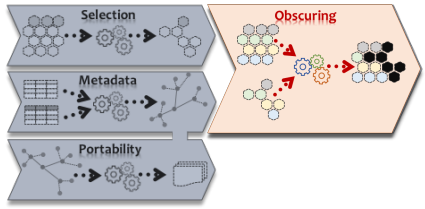
Obscuring Process - Strategy

Overview



Obscuring Process

The Data Flow



Used to store status and metrics for each table and JSON file

hadoop Data Lake

Trigger Process

This process is responsible to control and organize the executions, creating a queue of execution and also defining the resources needed for each table

Obscuring ./spark-submit

The obscuring process loads the table informed and the found information for this table in the JSON

Validation ./spark-submit

Here occurs the validation among the generated results and the JSON source. If the results are consistent the data is replaced

Discovery Process

Metadata Output

JSON

Level 1 Level 2 Level 3

A B C D

Selection Output

Final Files

Parquet

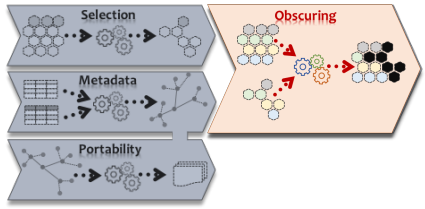
The process uses the Metadata to know which tables are in the scope and also uses the Selection results files to identify the ID's on the Metadata tables

Discovery Output

JSON

Data Lake Table

Parquet



Obscuring Process

Zoom In – Obscuring vs Deletion

hadoop *Data Lake*

Obscuring

{"db":"XYZ",
"table":"**SYSTEM-TABLE-A**",
"keys":[…],
"fields":[**"A1","A2","A7"**],
"action":"**OBS**",…}

Deletion

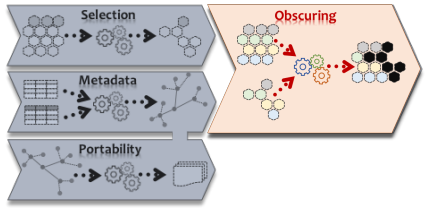
{"db":"XYZ",
"table":"**SYSTEM-TABLE-A**",
"keys":[…],
"fields":[**"A1","A2","A7"**],
"action":"**DEL**",…}

SYSTEM-TABLE-A

Temp copy

Parquet

Validation Process

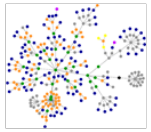


Obscuring Process – Trigger Phase

Details About Resources Allocation

Metadata Preparation

Data Dictionary



- Description of all data to consider on the data lake
- List of fields with sensitive information
- Fields are tagged with actions to perform in each row / field

JSON file with Information to delete / obscure:

- Database name
- Table name
- Field name
- Field value
- action

Obscuring Resource Manager Process

1. Identifies potential obscuring processes to run, depending on:
   - table size
   - partitioned / non-partitioned
   - Frequency of usage
2. Resource sizing. For the selected table(s) to delete, get:
   1. Number of simultaneous deletion processes
   2. Number of nodes to use
   3. Memory per process
3. Launches obscuring execution process(es) for the defined tables

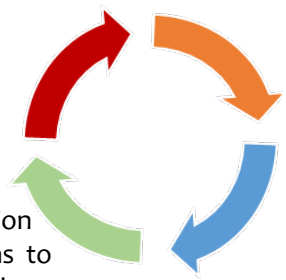
Real time performance indicators from the cluster

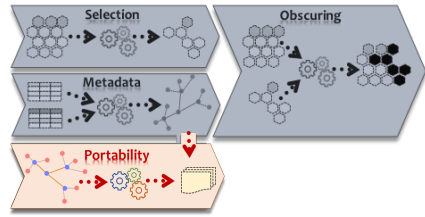
Load average indicators from the cluster

Table bucketing information
Group tables per size, existence of partitions, usage, etc

Obscuring execution

1. For a group of partitions in the table
2. It applies the obscuring /deletion logic
3. The new data is stored as parquet exactly the same original format data
4. The Validation process runs to check the data consistency and then apply the obscure / delete





Portability

Portability Process Execution

Portability Request

Portability Request File

Metadata Process

Data Sources

Periodic export of Navigator Data To Hive Tables, containing the data lake representation

cloudera navigator

CDO

GDPR tags includedon the transformation tracker

Transformation Tracker

Source Tracker

Java

Data Lake Metadata Update

neo4j

Metadata updates the Graph that represents the Data Lake with the new tags

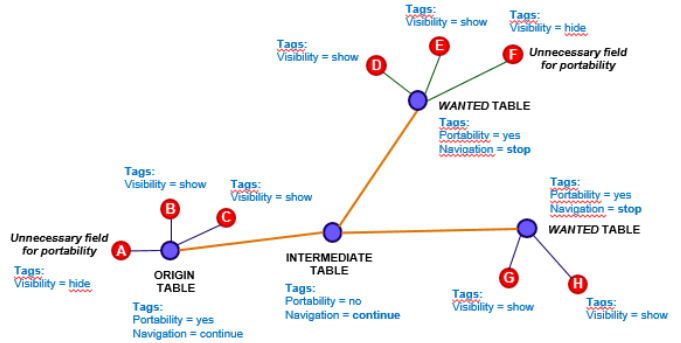
Java

Metadata Output

JSON

JSON Sample:

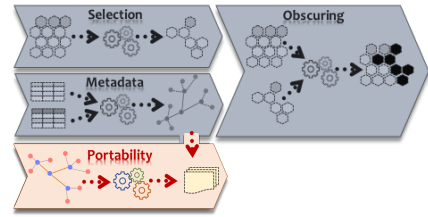
{"db":"XYZ", "table":"tbABC",
"level":[…],
"fields":[…],
"joins":[…],…}



The Graph Navigation Process obtains the model and the customers list from portability requests and navigates through the graph model to identify all records associated and key fields, based on the metadata.

Portability: Export Digital File

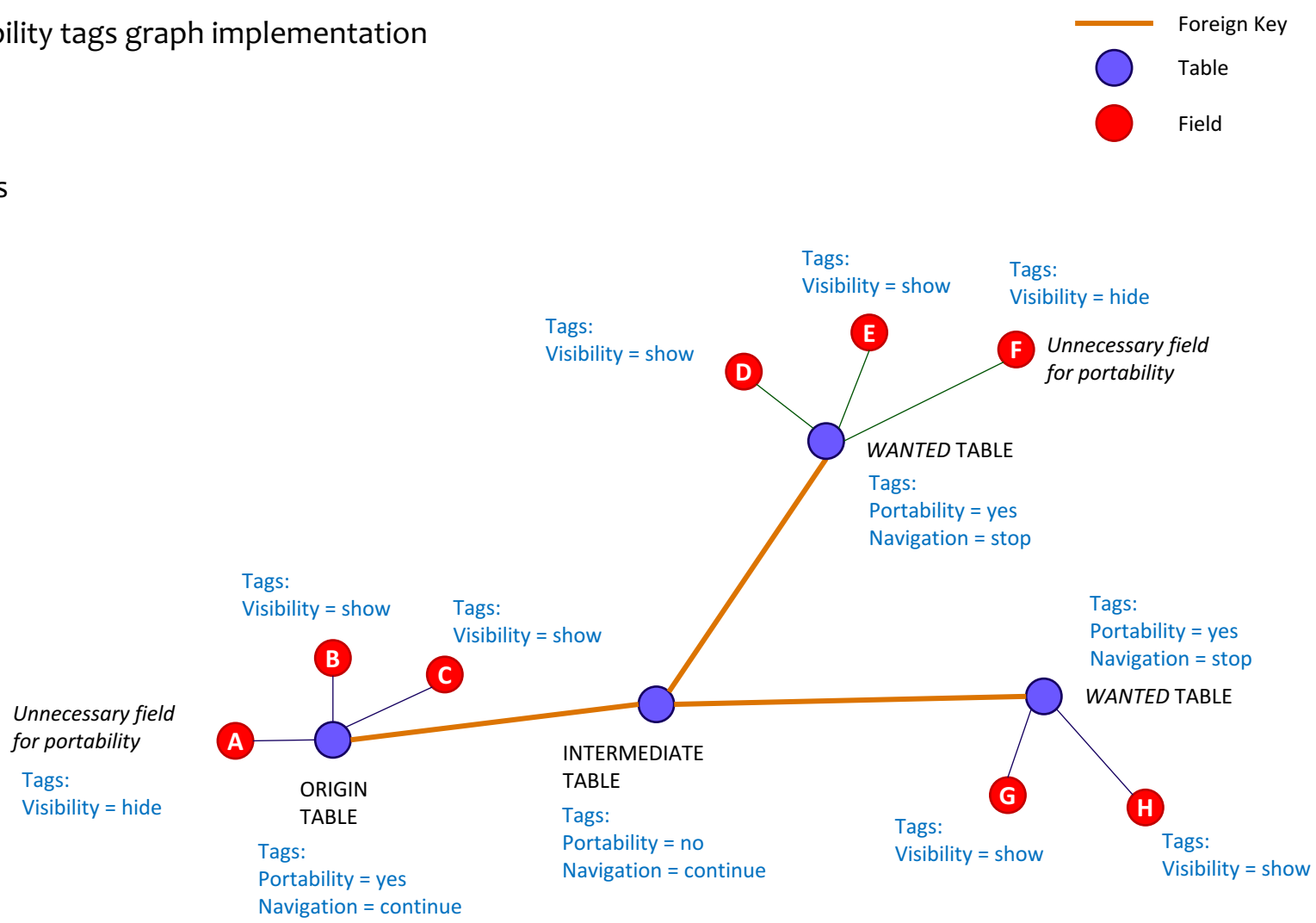
JSON files



Portability

Example with navigation, portability and visibility tags graph implementation

Digital file must compile data from
fields B and C from the driver table
and fields D, E, G and H from two other tables



Evolution

Future steps

☆ Benefit from the new process to automate anonymization

☆ Extend profiling to accelerate analysis phase in streaming processes

☆ Kudu for streaming analytics with more efficient data obscuring

☆ Machine learning to maintain the data inventory and complement subject matter expert tags

☆ More efficient history archival extending the use of slow changing dimensions and cold storage

Santander

Thank you