Enterprise Data Catalog 10.5 Overview

Product Management Team
Enterprise Data Catalog & Metadata Intelligence

Informatica

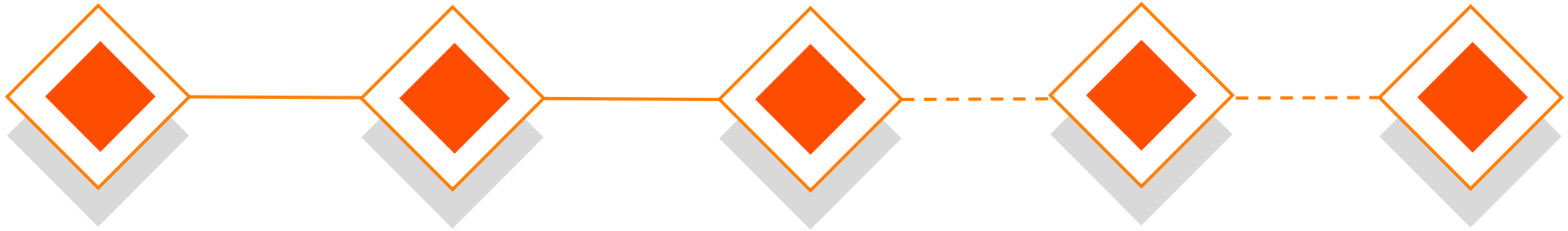
Safe Harbor

The information being provided herein is for informational purposes only. The development, release and timing of any Informatica product or functionality described herein remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made herein are based on information currently available, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

Release Plan

* Monthly CP releases on 10.4.0 and 10.4.1 trains



10.4.1
Jun 2020

10.4.1.1
Aug 2020

10.4.1.2
Oct 2020

10.4.1.3
Jan 2021

10.5
Mar 2021

Informatica

10.5 Release Themes

Enterprise Readiness

- **Security enhancements**
- **Non-Hadoop Platform**
- Profiling on Databricks
- Performance Enhancements
  - Relationship API
  - Lineage diagram rendering
  - Selective backup

User Experience

- Search result page enhancements
- BG Term bulk curation
- Data Flow Analytics (Preview)

CLAIRE

- Similarity Discovery enhancements

Scanners/Connectivity

Product integration of 15 Advanced Scanners

- Code: Oracle, SQL Server, Teradata, IBM DB2, Netezza, Sybase
- BI: SAS, Microsoft SSAS & SSRS,
- Legacy: Cobol, JCL
- ETL: Oracle Data Integrator, Talend DI, IBM DataStage, Microsoft SSIS

Standard Scanners

- Axon scanner enhancements (Add. Facets)
- S3 compatible filesystem (Scality)
- Enhanced Snowflake scanner
- SAP S/4 HANA (GA)

4

EDC Architecture Change in 10.5

Re-architecture to keep up with market trend

- Replacing

- With

- **Motivations**
  - HDP going EOL end of 2021
    - Customers can continue to use old stack on 10.4.1 till March'22.
  - Be relevant to the market trend
- **What to expect?**
  - No additional hardware or software prerequisites
  - No functional loss, similar or improved performance and scale
  - Seamless upgrade and content migration
  - Continued and improved full support on EDC and all deployed services
    - better support EDC customers on the longer term
    - faster turn around for OS support, security patches

Informatica

Upgrade process

Much like previous version upgrade process

- Upgrades supported from EDC 10.4 and 10.4.1 (latest CPs)
- Backup catalog content
- Clean embedded cluster hosts – (run infacmd.sh ihs cleanCluster)
- Upgrade the platform
- Deploy new ICS service (should be taken care by the installer)
- Restore catalog content
- Upgrade the content
- Re-index the content

FAQs

Q: Is there an automatic procedure to migrate the current internal cluster data to new mongoDB?
A: We will produce automated migration script. However, we also would have step-by-step guide for the deployment.

Q: Will nomad+mongodb will be shipped with EDC install and do like HIS for internal cluster today?
A: Yes, it will be one installation for overall new architecture deployment.

Q: Will it be bundled with embedded Mongo? Will it still support Hadoop pushdown when scanning customers Hadoop env?
A: we will bundle MongoDB and continuer to pushdown to source cluster

Q: How will this impact our AWS, Azure, & GCP Marketplace offerings?
A: New 10.5 marketplace listing will be published post the march release.

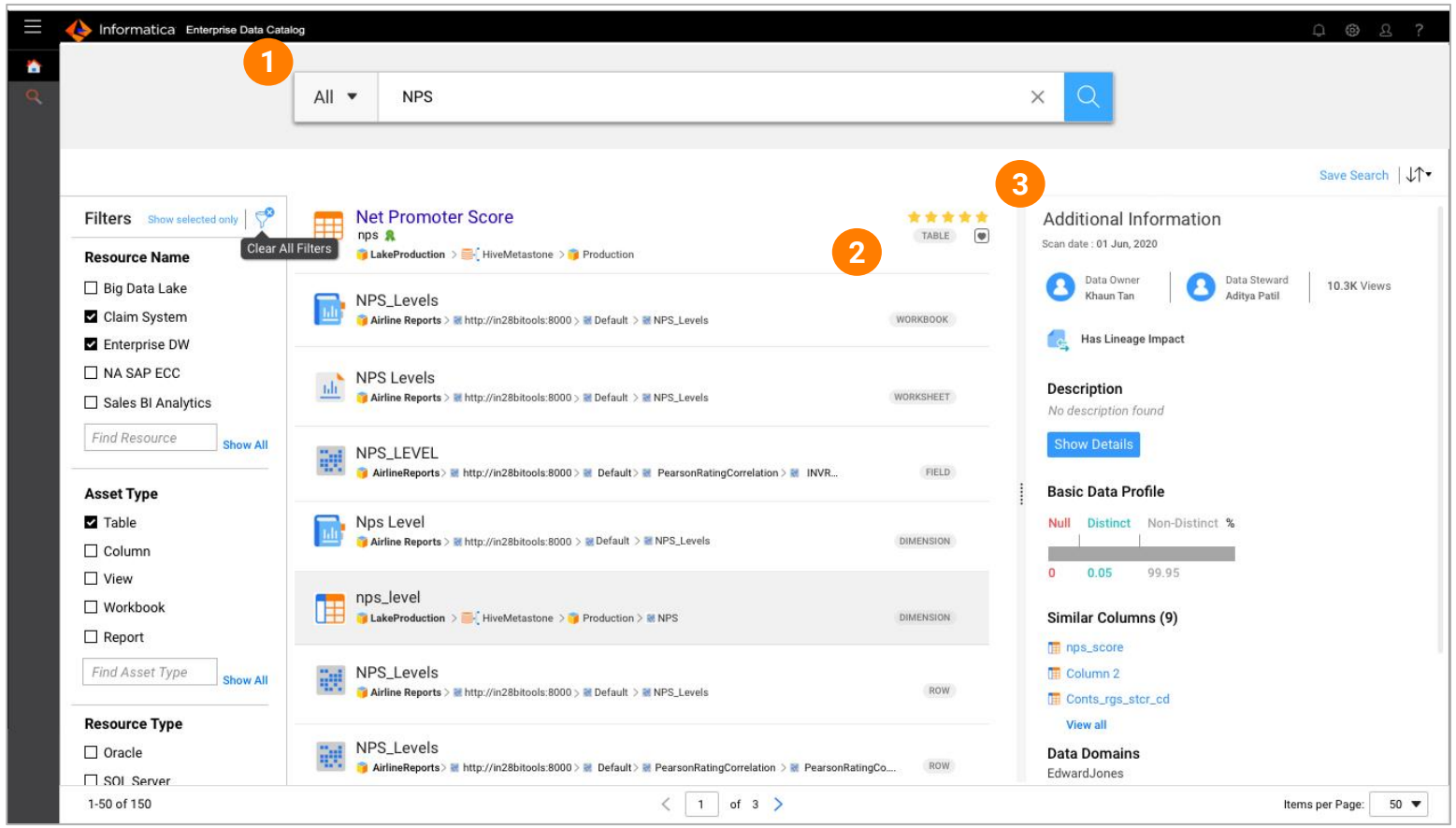
Q: Will there be Kerberos in Nomad cluster?
A: No kerberos, security will be handled through other mechanism, mTLS auth and encryption implemented.

Q: What happens to EDP for 10.5 is it also going to use Mongo?
A: EDP customers would just have to upgrade the EDC deployment. The EDP deployment will work as is.

More FAQs: https://knowledge.informatica.com/s/article/Enterprise-Data-Catalog?language=en_US

Search Result Page Enhancements

1. Redesigned Search Bar
   - Search Prefilter replaced Search Tab for better context
2. Simplified Search Result
   - Asset Type more visible in search result list
3. Added Asset Additional Information Pane
   - Enable user to quickly determine the most relevant to explore



Search Result Page Enhancements

Search Prefilter *replaced Search Tab for better context*

Asset Type more visible in search result list

Added clear search string icon

Additional Information Pane on hover. *Removed Show Related Search & Show/Hide Details*

Added Save Search

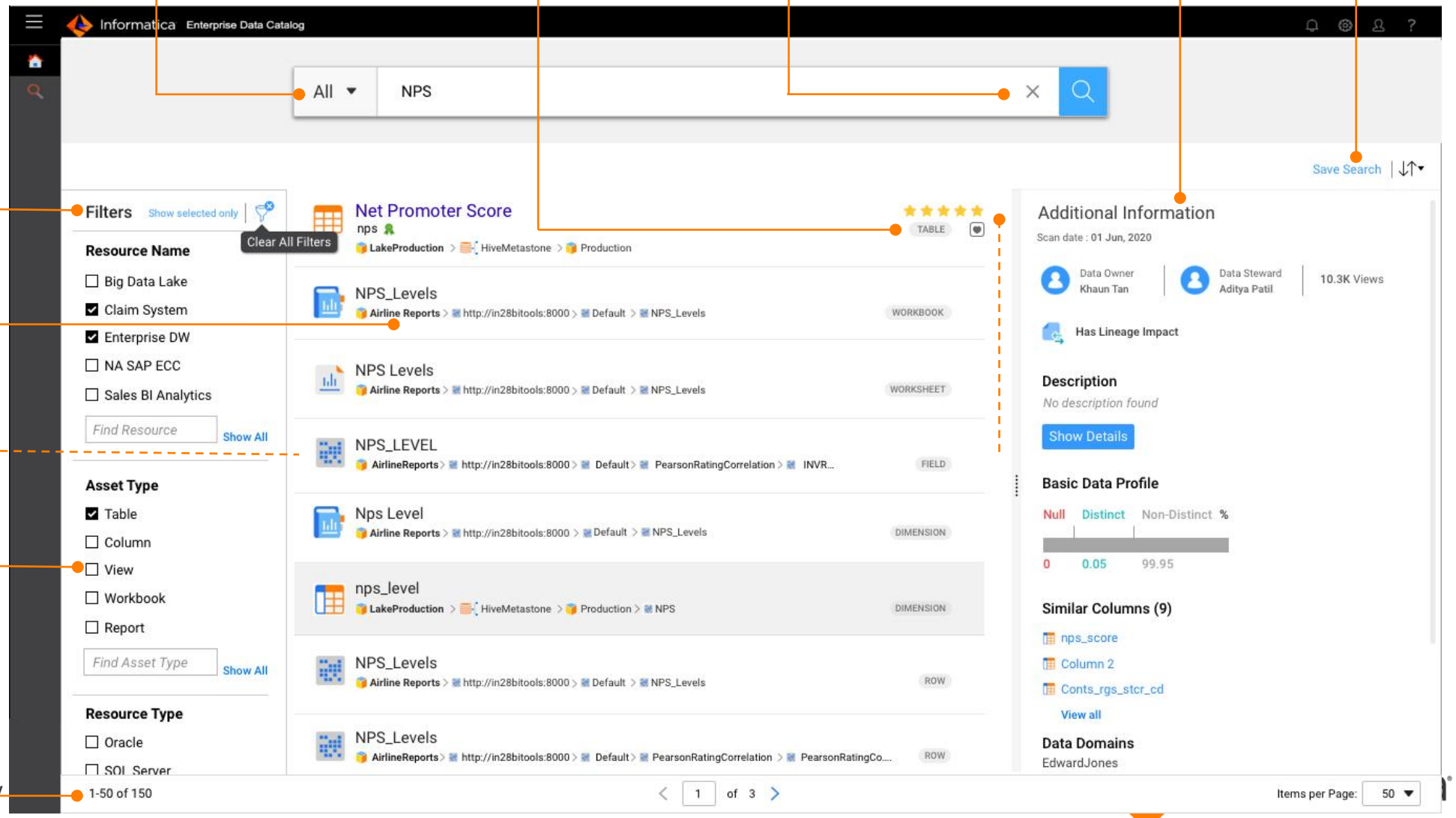
Filter pane simplified

Resource Name visible on breadcrumb

Clean-up Rating icon cluster

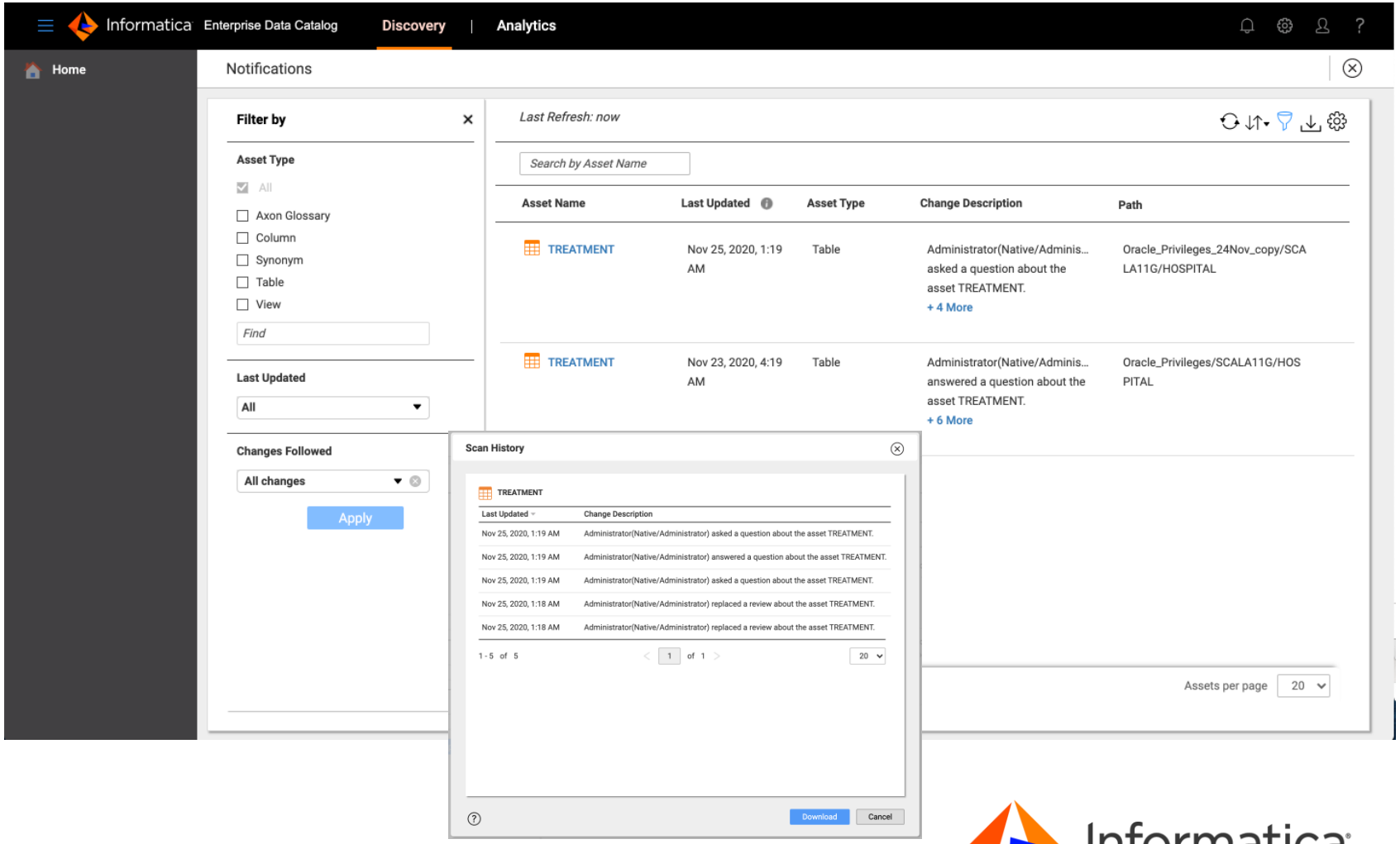
Combined identical Asset Types from different resource types

Combined assets # and pagination



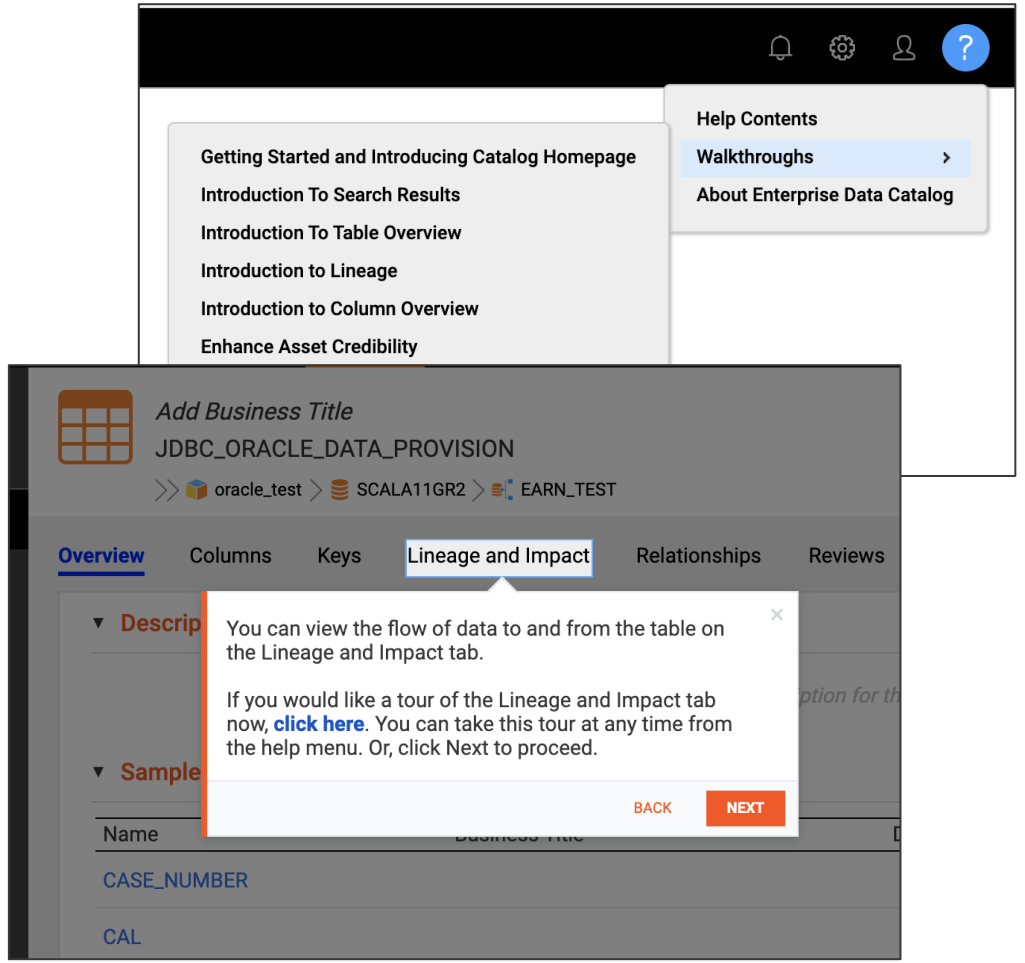
Notifications Page Enhancements

- 10.4.1 Change Summary enhancements extended to 10. 5 Notifications Center
- Key Enhancements:
  - Filter by:
    - Asset Name
    - Time Period
    - Type of change (Source, Enrichment, Collaboration)
  - Download as CSV



10.5 Walkthroughs Update

- On-demand, Context-based, in app product walkthroughs
- Expanded Walkthrough List
  - Catalog UI
    - Homepage
    - New Search Results *(New)*
    - Tables & Columns
    - Lineage
    - Asset Enrichment
    - Data Domain Overview *(New)*
    - Data Domain Curation *(New)*
    - Business Term Overview *(New)*
    - Resource Overview *(New)*
    - Application Configuration *(New)*
  - Catalog Admin
    - Catalog Admin Homepage *(New)*
    - Creating Resources *(New)*
    - Creating Custom Attributes *(New)*
    - Creating Data Domains *(New)*
    - Configuring Security and Permissions *(New)*



Discovery changes in EDC 10.5 (new functionalities)

Resource grouping for similarity computations

- Allow users to logically group resources for similarity computation
- Improves the performance and accuracy
- Goes through the scanner framework and lifecycle (create, edit, purge and delete)

Reduce false positives in similarity by smart feature choice based on Type

- Consider the data types for date/timestamp and pure numeric while similarity computation

Allow similarity computation based on features enabled

- Each resource group can be configured for the features against similarity is run

Support for s3 compatible storage (Scality)

- Support for data discovery from EDC and IDE

New Architecture
EDC 10.5

Informatica

New Tech Stack

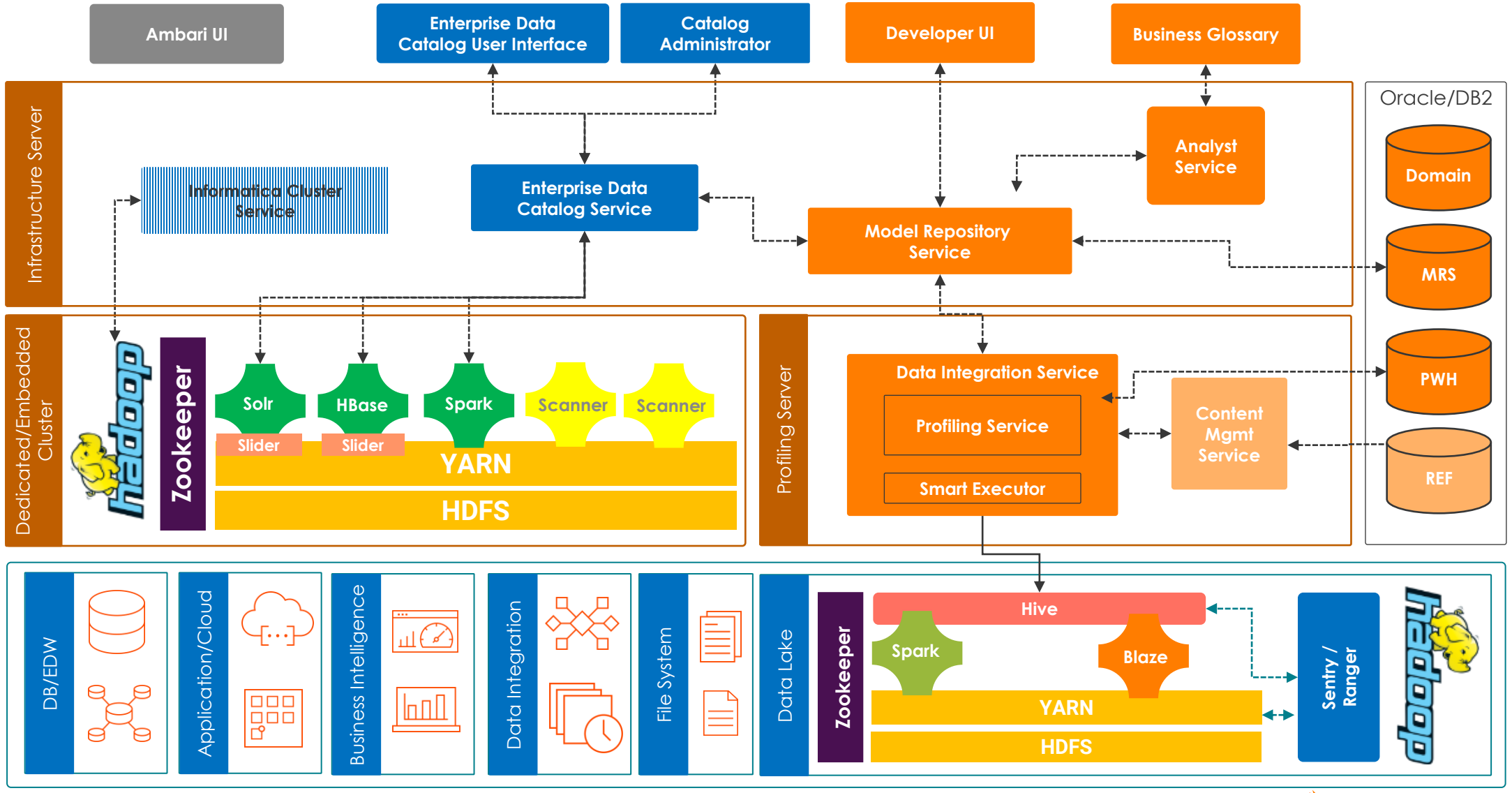
EDC/DPM 10.5 will use:

- For Storage:
  - MongoDB, MongoDB GridFS
  - PostgreSQL
  - SolR
  - Elastic search (DPM)
- For orchestration and security:
  - Nomad
  - Vault
- For compute:
  - Native Java (Scanners, ingestion service)
  - Spark (DPM)

Store/Engine	Before (10.4.1)	After (10.5.0)
Asset Store	HBase	MongoDB
Graph Store	JanusGraph	MongoDB
Index Store	SOLR, Elastic (DPM)	SOLR, Elastic (DPM)
Event Store (DAA)	Relational (MRS)	Relational (MRS)
Stage Store	HBase	MongoDB
Scan Content Store	HDFS	MongoDB GridFS
Similarity Store	HBase	Postgres
Config Store	Relational (MRS)	Relational (MRS)
Monitoring Store	Relational (MRS)	MongoDB
Compute	Native, Spark (YARN)	Native on Nomad, Spark Standalone (DPM)

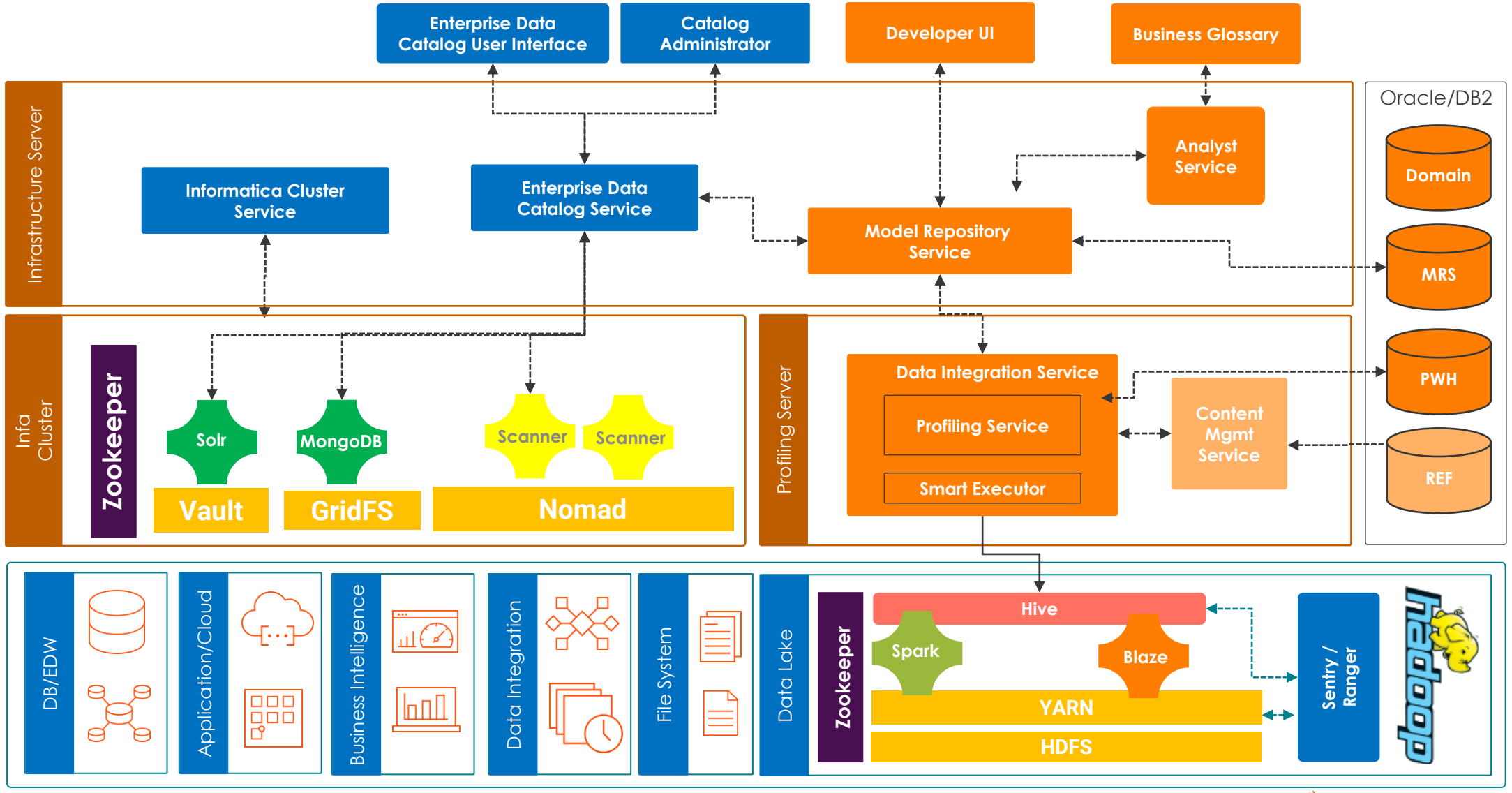
Informatica

EDC Architecture 10.4 – Services



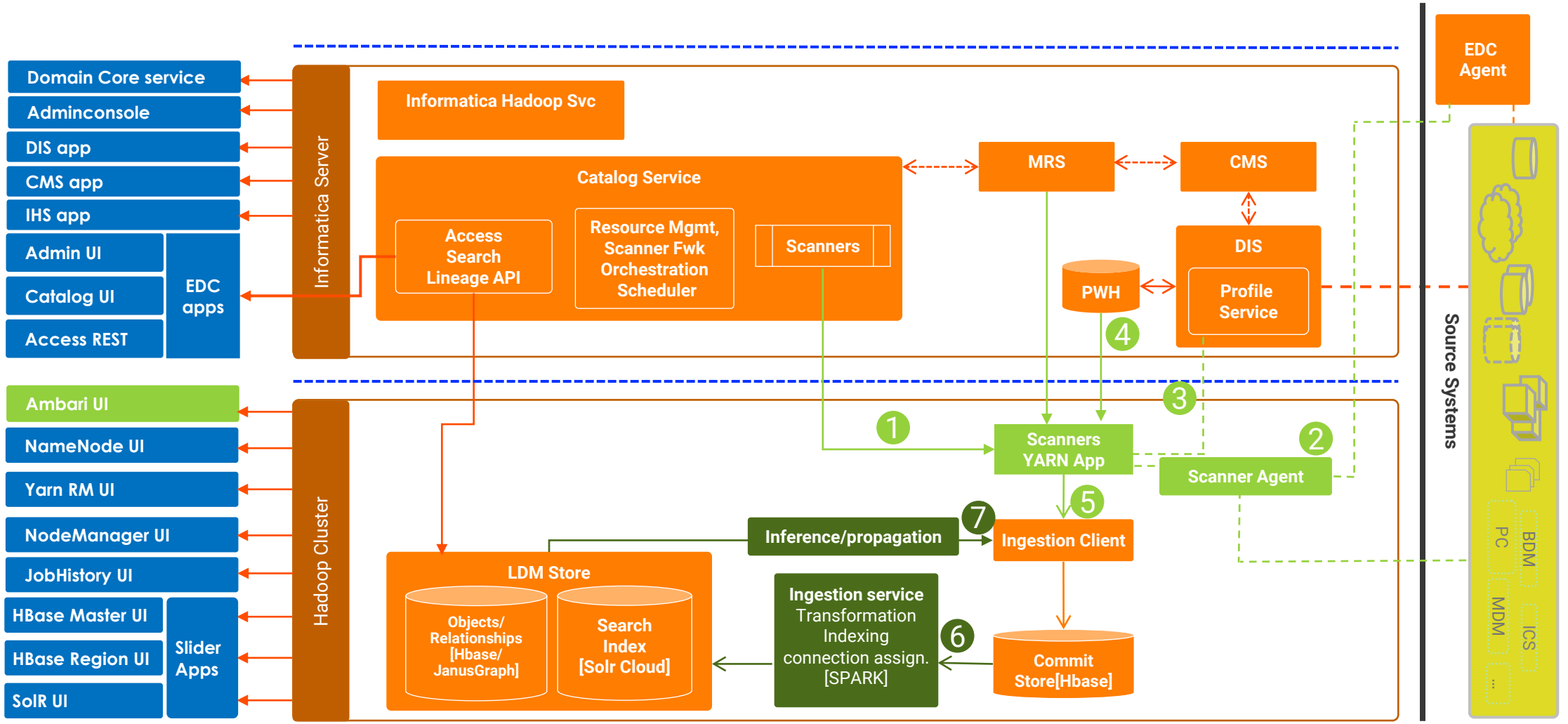
Informatica

EDC Architecture 10.5 – Services



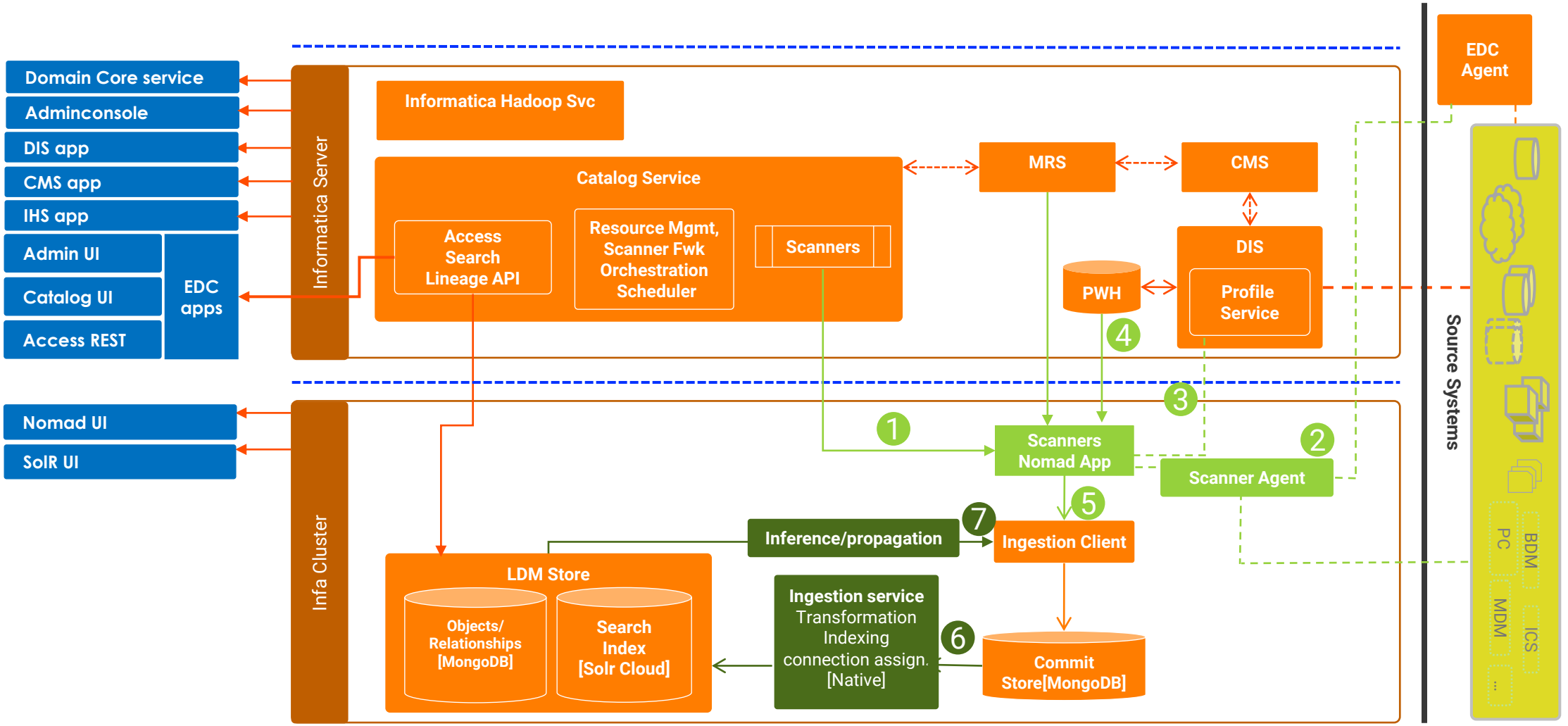
Informatica

EDC Internals 10.4 – Scanner process



Informatica

EDC Internals 10.5 – Scanner process



Informatica

ICS Enterprise readiness

- Security
  - Authentication enable for all services (mTLS), Vault to manage internal certificates
  - Encryption of data in transit with TLS
  - Encryption of data at rest with platform encryption mechanism (AES-256 in 10.5)
- Highly available (HA)
  - No more Single Point Of Failure (SPOF)
  - Each component is deployed in HA mode across the node in the cluster
  - Possibility to provide a PostgreSQL HA custom setup
- Disaster Recovery (DR) support
  - Hot backup support for regular replication to the DR site

EDC Sizing guideline (summary)

Unchanged requirements for 10.5.0

- Refer to Sizing and Performance Tuning Guide for sizing recommendations, parameter tuning and more.

		Infrastructure			Metadata Processing					Infa Cluster			
Env. Size	# of conc. (total) users	CPU	RAM	Disk	Metadata Resources	# of objects	CPU	RAM	Disk	# of nodes	CPU	RAM	Disk
Small	20 (200)	16	32 GB	200 GB	30-40	1 Million	16	32 GB	20 GB**	1	8	24 GB	120 GB***
Medium	50 (500)	24	32 GB	200 GB	200-400	20 Million	32	64 GB	100 GB**	3	24	72 GB	2 TB***
Large	100 (1000)	48	64 GB	300 GB	500-1000	50 Million	32	64 GB	500 GB**	6	48	144 GB	12 TB***

** 1 to 4 disks for profiling

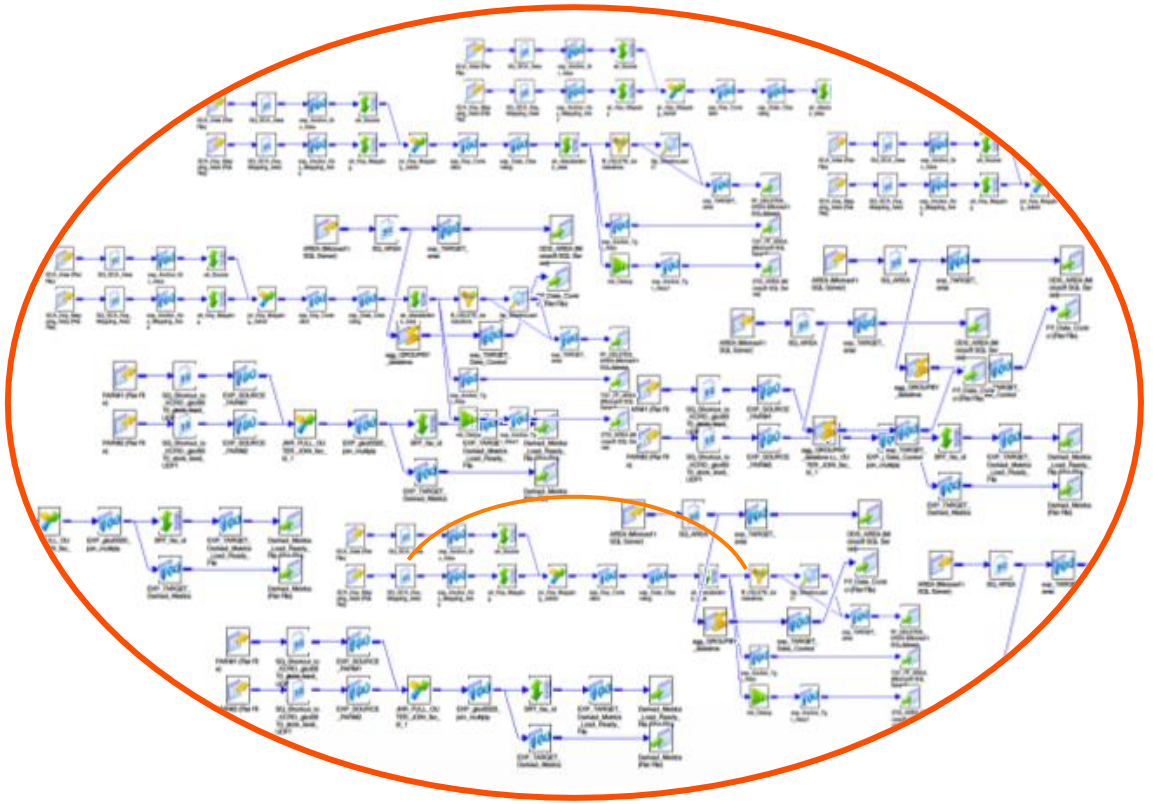
*** 4 to 6 disks recommended on cluster nodes

Data Flow Analytics

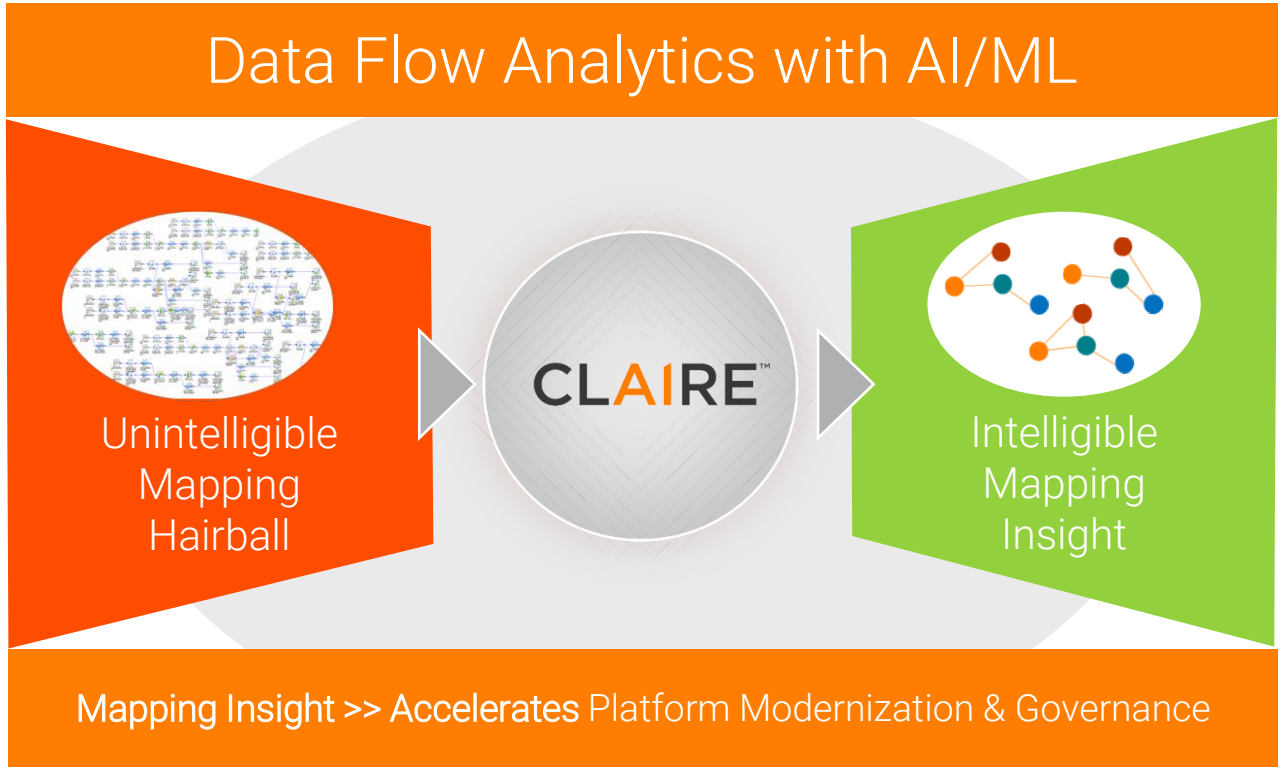
Data Mapping Complexity Inhibits Digital Modernization

- Data mapping is crux of data integration and management
- Making sense of the massive # of mappings is nearly impossible with
  - 10,000s of Mappings
  - Created by 25+ ETL Developers
  - Over 10+ years
- Lack of mapping insight
  - Inhibits digital modernization
  - Prevents govern sensitive data flows
  - Increases mapping cost and technical debt

Mapping Hairball Challenge



Data Flow Analytics Overview



Scale Mapping Insight with AI/ML Automation

- Convert mappings into graphs
- Label each node on the graphs with mapping properties depending on the use case
- Apply AI/ML algorithms on graphs

Mapping Discoveries and Insight

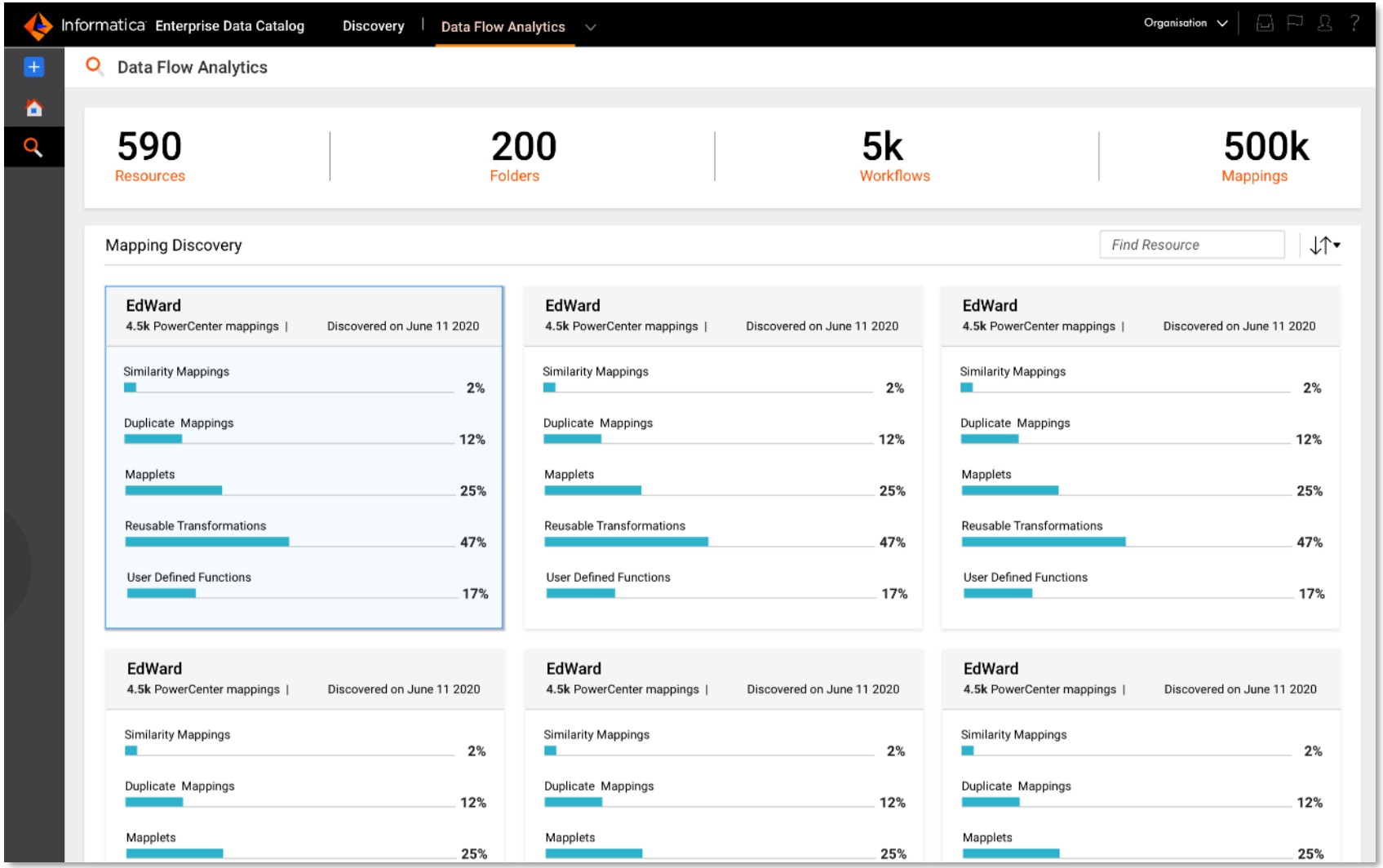
- Similar Mappings
- Duplicate Mappings
- Reusable Mapplets, Transformations, & User Functions for complex expressions

Measure & Improve Mapping Quality

- Reduce data errors, inconsistencies, and risks
- Optimize development efficiencies
- Reduce mapping operational cost

DFA - Dashboard

UX Mockup



Key metrics on mapping resources in EDC*

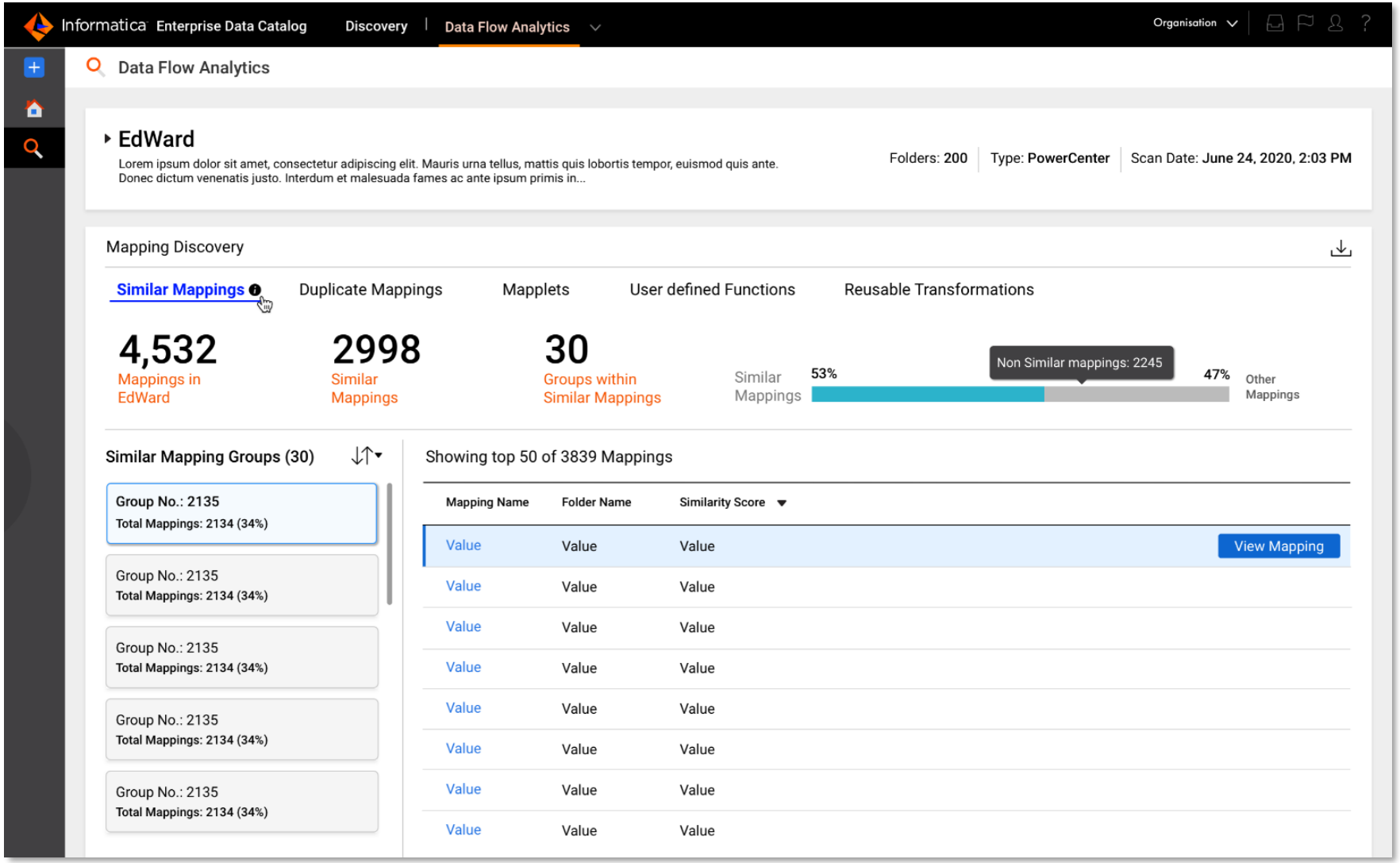
Mapping discovery metrics for each mapping resource

* = DFA (beta) supports PowerCenter mappings

Informatica

Mapping Cluster Discovery

UX Mockup



Mapping Resource

Mapping discovery metrics for the resource

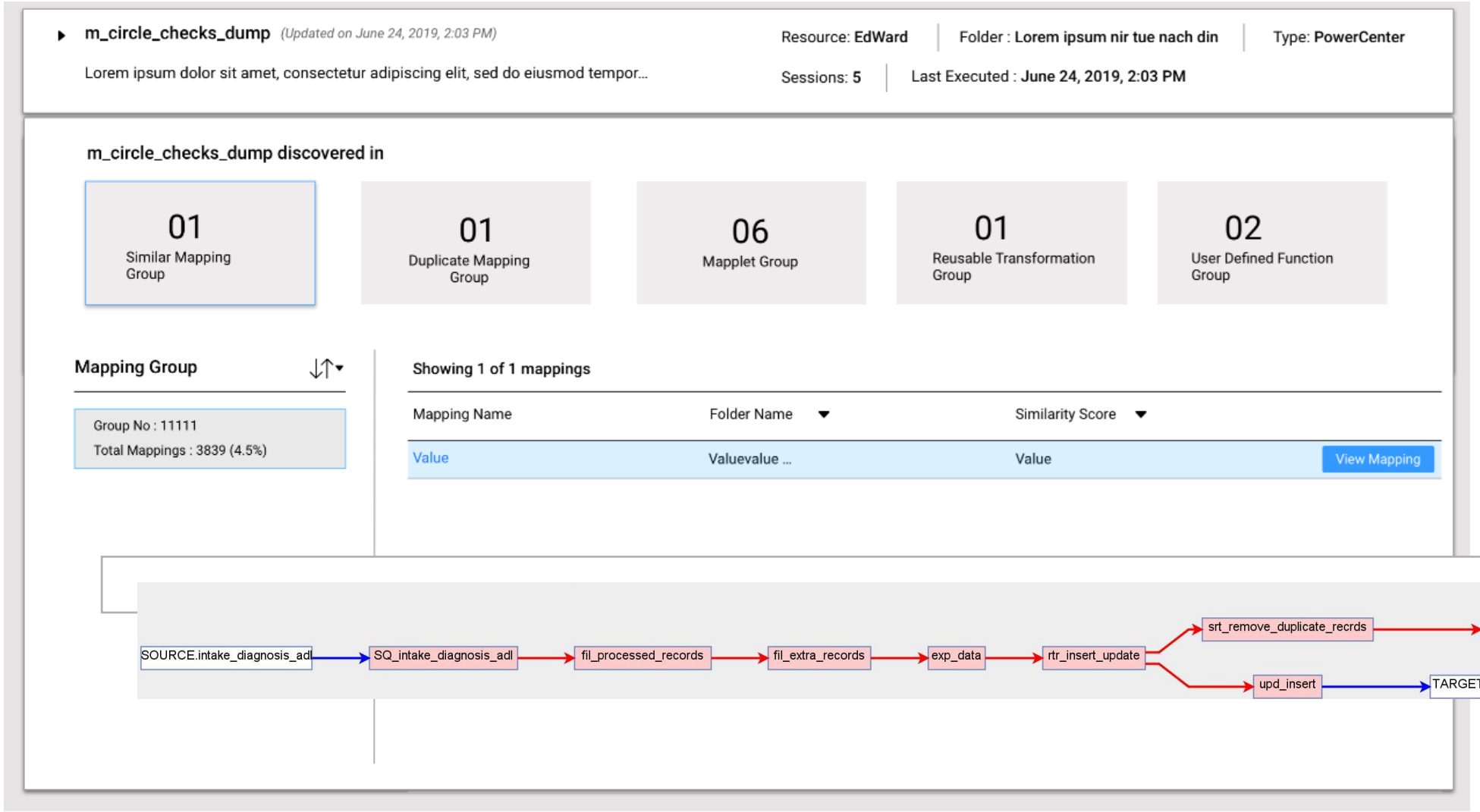
Mapping discovery group details

* = DFA (beta) supports PowerCenter mapping

Informatica

Mapping Details

UX Mockup



Mapping Discoveries to which a mapping belongs

Mapping Diagram

Informatica

Thank You

Devashish Sharma
dsharma@informatica.com
Louis-Noel Trapadoux
ltrapadoux@informatica.com