

# Enterprise Data Catalog 10.5 Overview

*Product Management Team  
Enterprise Data Catalog & Metadata Intelligence*

# 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



# 10.5 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)

# 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

# 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](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

The screenshot displays the Informatica Enterprise Data Catalog search results for 'NPS'. The search bar at the top (1) shows 'All' and 'NPS'. The results list (2) includes assets like 'NPS\_Levels' (TABLE), 'NPS Levels' (WORKSHEET), 'NPS\_LEVEL' (FIELD), 'Nps Level' (DIMENSION), 'nps\_level' (DIMENSION), and 'NPS\_Levels' (ROW). The 'Additional Information' pane (3) on the right shows details for the selected 'NPS\_Levels' asset, including scan date (01 Jun, 2020), data owner (Khaun Tan), data steward (Aditya Patil), 10.3K Views, and a basic data profile with 0 Null, 0.05 Distinct, and 99.95 Non-Distinct values.



# 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**

**Additional Information**  
Scan date : 01 Jun, 2020  
Data Owner: Khaun Tan | Data Steward: Aditya Patil | 10.3K Views  
Has Lineage Impact  
Description: No description found  
Show Details  
Basic Data Profile: Null (0), Distinct (0.05), Non-Distinct (99.95)  
Similar Columns (9): nps\_score, Column 2, Conts\_rgs\_stcr\_cd  
Data Domains: EdwardJones

1-50 of 150

1 of 3

Items per Page: 50

# 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

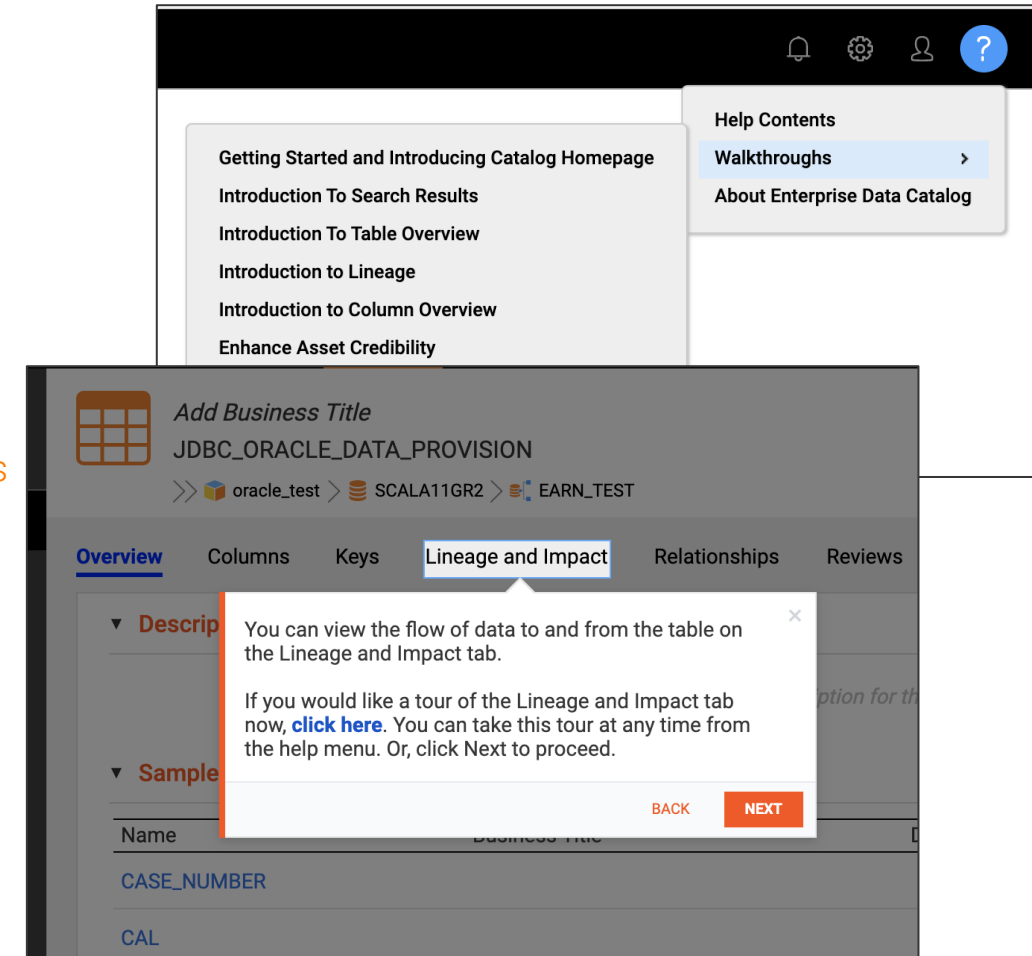
The screenshot shows the Informatica Enterprise Data Catalog interface. The main page is titled 'Notifications' and displays a table of notifications for the asset 'TREATMENT'. The table has columns for 'Asset Name', 'Last Updated', 'Asset Type', 'Change Description', and 'Path'. Two notifications are visible, both for 'TREATMENT' (Table type). The first notification is dated 'Nov 25, 2020, 1:19 AM' and describes an administrator asking a question. The second is dated 'Nov 23, 2020, 4:19 AM' and describes an administrator answering a question. A 'Scan History' modal window is open, showing a detailed log of changes for the asset 'TREATMENT'. The log includes columns for 'Last Updated' and 'Change Description', with five entries showing various actions like asking and answering questions and replacing reviews. The modal also includes a 'Download' button and a 'Cancel' button.

Asset Name	Last Updated	Asset Type	Change Description	Path
TREATMENT	Nov 25, 2020, 1:19 AM	Table	Administrator(Native/Adminis... asked a question about the asset TREATMENT.	Oracle_Privileges_24Nov_copy/SCA LA11G/HOSPITAL
TREATMENT	Nov 23, 2020, 4:19 AM	Table	Administrator(Native/Adminis... answered a question about the asset TREATMENT.	Oracle_Privileges/SCALA11G/HOS PITAL

Last Updated	Change Description
Nov 25, 2020, 1:19 AM	Administrator(Native/Administrator) asked a question about the asset TREATMENT.
Nov 25, 2020, 1:19 AM	Administrator(Native/Administrator) answered a question about the asset TREATMENT.
Nov 25, 2020, 1:19 AM	Administrator(Native/Administrator) asked a question about the asset TREATMENT.
Nov 25, 2020, 1:18 AM	Administrator(Native/Administrator) replaced a review about the asset TREATMENT.
Nov 25, 2020, 1:18 AM	Administrator(Native/Administrator) replaced a review about the asset TREATMENT.

# 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

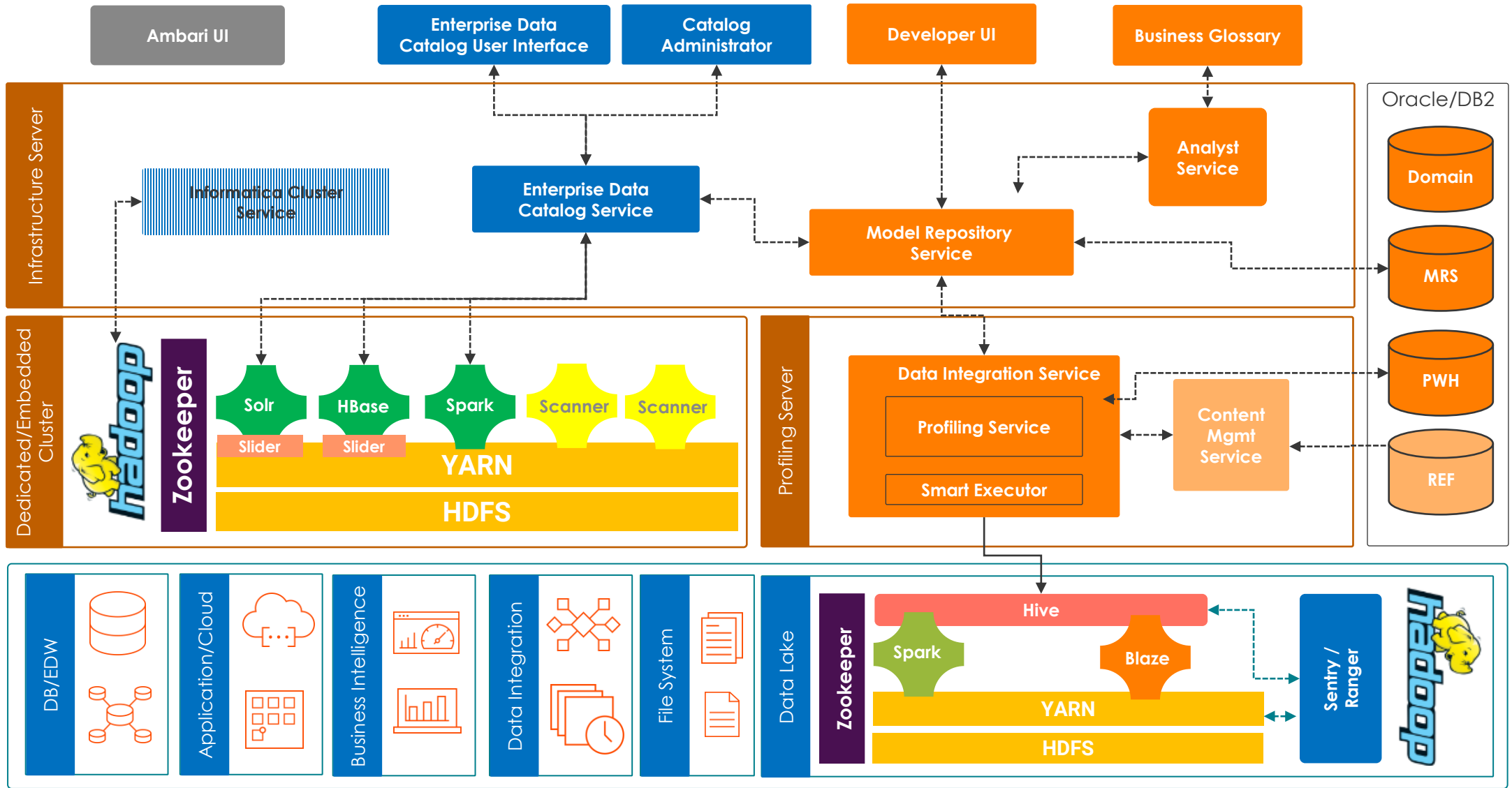
# New Tech Stack

## EDC/DPM 10.5 will use:

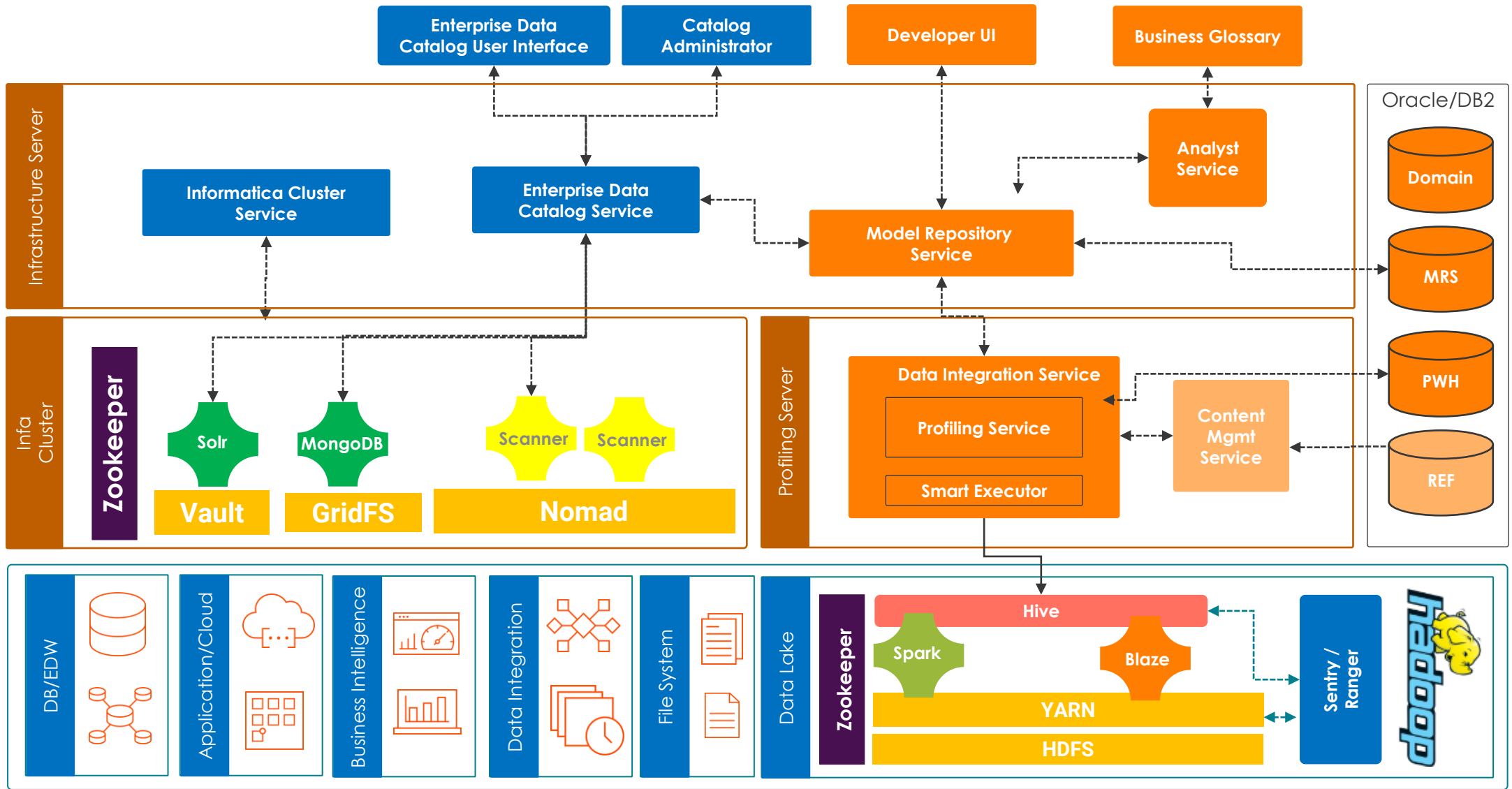
- For Storage:
  - MongoDB, MongoDB GridFS
  - PostgreSQL
  - SolR
  - Elastic search (DPM)
- For orchestration and security:
  - Nomad
  - Vault
- For compute:
  - Native Java (Scanners, ingestion service)
  - Spark (DPM)

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)

# EDC Architecture 10.4 – Services

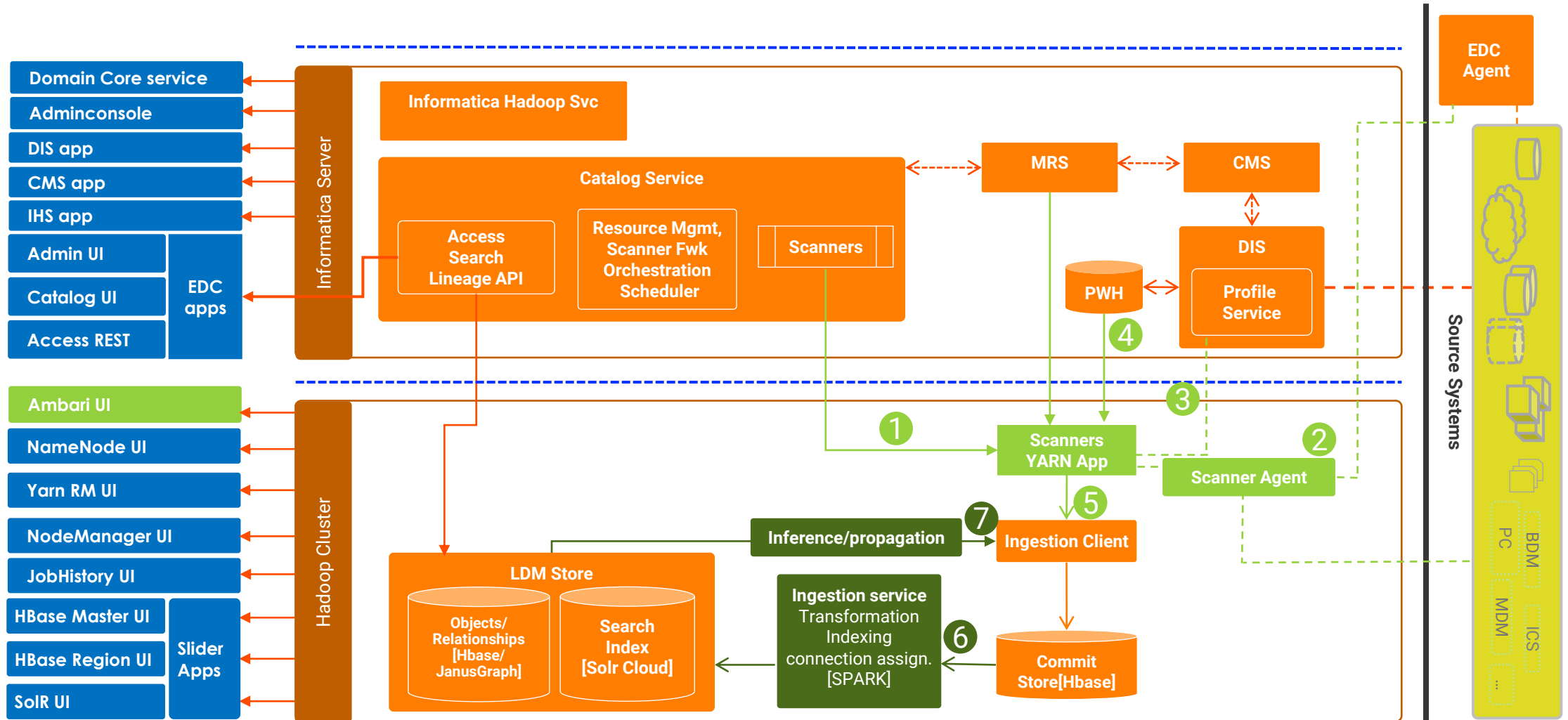


# EDC Architecture 10.5 – Services

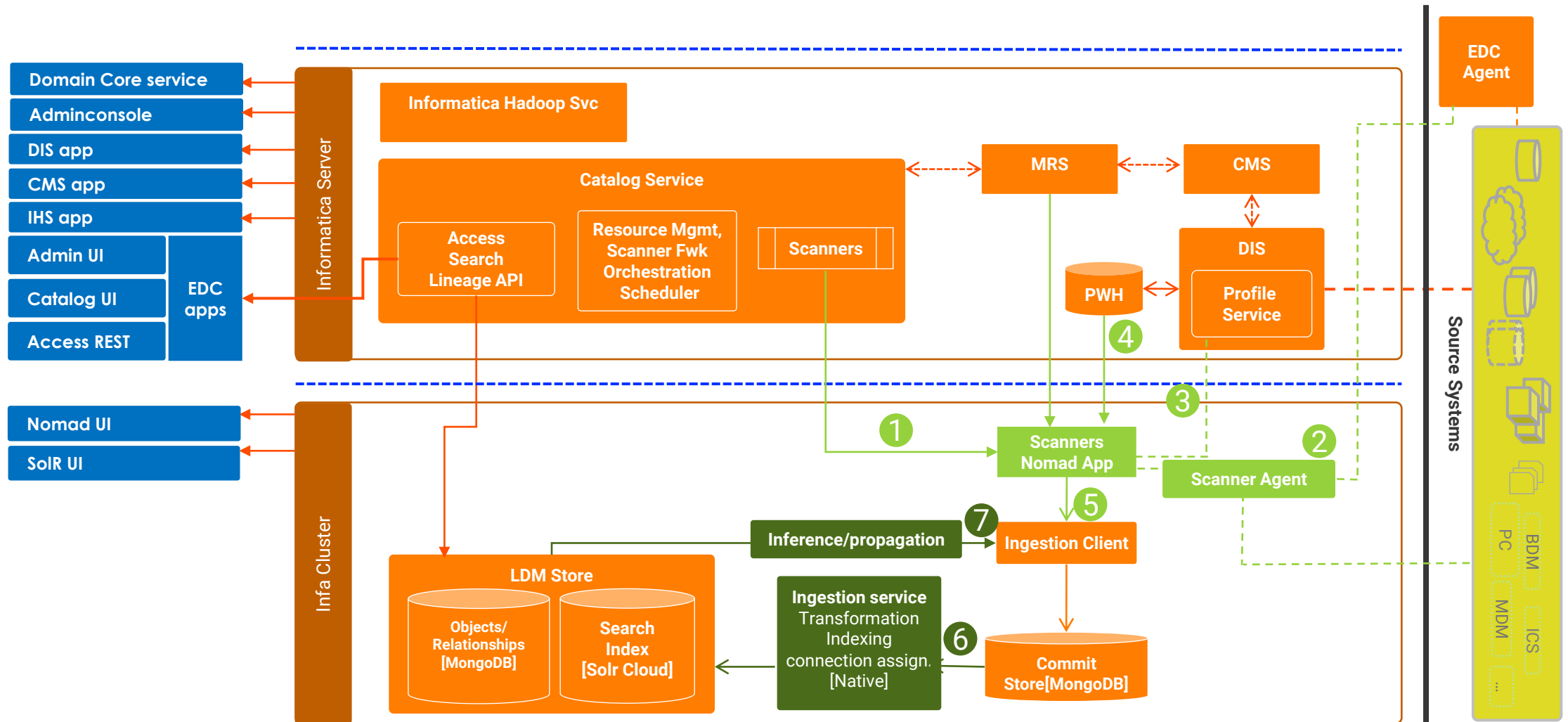




# EDC Internals 10.4 – Scanner process



# EDC Internals 10.5 – Scanner process



# 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

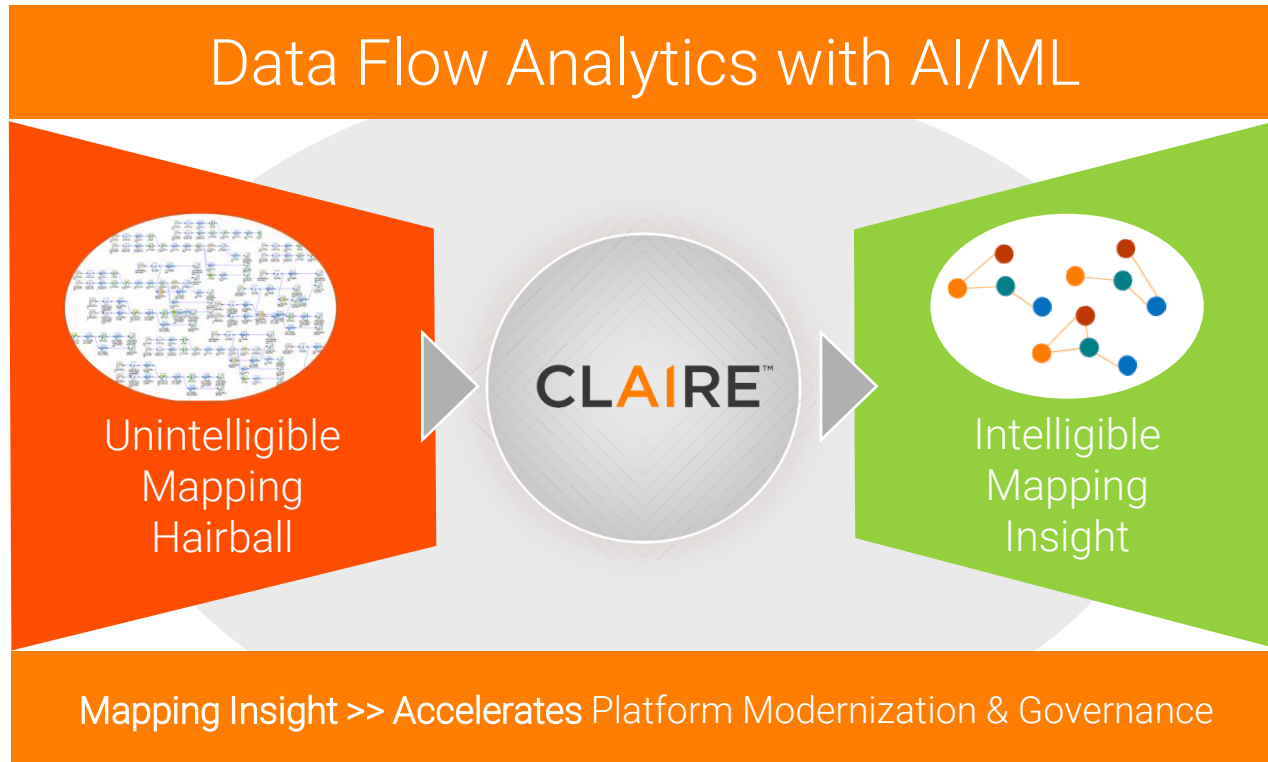
# Data Mapping Complexity Inhibits Digital Modernization

- 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

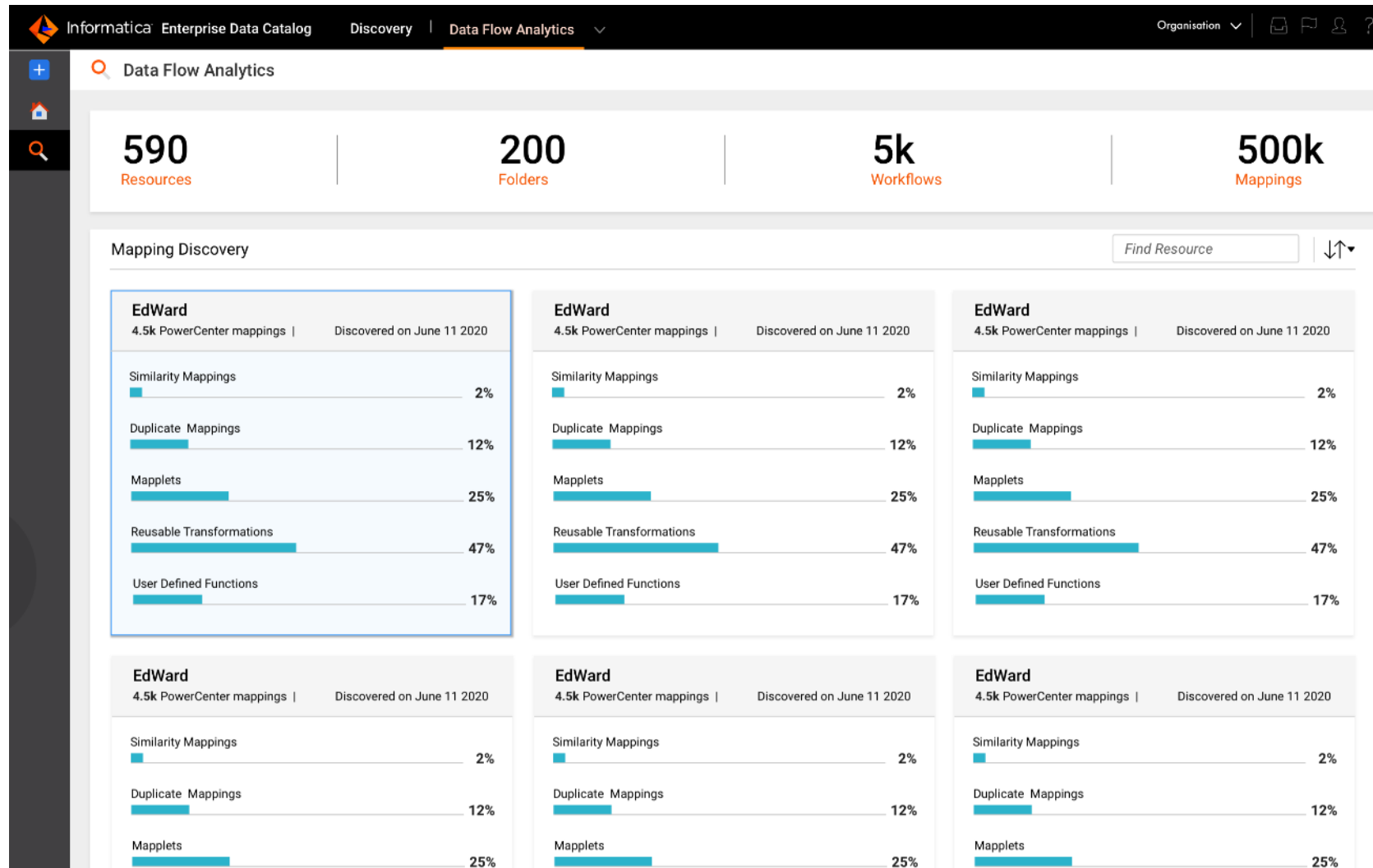
- Similar Mappings
- Duplicate Mappings
- Reusable Mapplets, Transformations, & User Functions for complex expressions

## 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



# Mapping Cluster Discovery

Informatica Enterprise Data Catalog | Discovery | Data Flow Analytics | Organisation

Data Flow Analytics

EdWard | Folders: 200 | Type: PowerCenter | Scan Date: June 24, 2020, 2:03 PM

Mapping Discovery

Similar Mappings | Duplicate Mappings | Mapplets | User defined Functions | Reusable Transformations

4,532 Mappings in EdWard | 2998 Similar Mappings | 30 Groups within Similar Mappings

Similar Mappings: 53% | Non Similar mappings: 2245 | Other Mappings: 47%

Similar Mapping Groups (30)

Showing top 50 of 3839 Mappings

Mapping Name	Folder Name	Similarity Score
Value	Value	Value
Value	Value	Value
Value	Value	Value
Value	Value	Value
Value	Value	Value
Value	Value	Value
Value	Value	Value
Value	Value	Value
Value	Value	Value
Value	Value	Value



Mapping Resource



Mapping discovery metrics for the resource



Mapping discovery group details



# Mapping Details

**m\_circle\_checks\_dump** *(Updated on June 24, 2019, 2:03 PM)*
Resource: EdWard | Folder: Lorem ipsum nir tue nach din | Type: PowerCenter  
 Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor...
 Sessions: 5 | Last Executed: June 24, 2019, 2:03 PM

---

**m\_circle\_checks\_dump** discovered in

01

Similar Mapping Group

01

Duplicate Mapping Group

06

Mapplet Group

01

Reusable Transformation Group

02

User Defined Function Group

**Mapping Group** ↓↑

Group No : 11111

Total Mappings : 3839 (4.5%)

Showing 1 of 1 mappings

Mapping Name	Folder Name	Similarity Score	
Value	Valuevalue ...	Value	<a href="#">View Mapping</a>

```

    graph LR
      SOURCE[SOURCE.intake_diagnosis_adl] --> SQ[SQ_intake_diagnosis_adl]
      SQ --> fil1[fil_processed_records]
      fil1 --> fil2[fil_extra_records]
      fil2 --> exp[exp_data]
      exp --> rtr[rtr_insert_update]
      rtr --> srt[srt_remove_duplicate_reclds]
      rtr --> upd[upd_insert]
      srt --> agg[agg_min_scd_end_dt]
      upd --> TARGET[TARGET.cim_intake_diagnosis_adl]
      agg --> TARGET
      
```

← Mapping Discoveries to which a mapping belongs

↓ Mapping Diagram

# Thank You

Devashish Sharma

*[dsharma@informatica.com](mailto:dsharma@informatica.com)*

Louis-Noel Trapadoux

*[ltrapadoux@informatica.com](mailto:ltrapadoux@informatica.com)*

