Mar 22nd, 2022

Enterprise Data Catalog 10.5.x Architecture

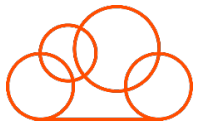
- Sugi Narayana, Senior Manager, Customer Success Technologist
- Avanish Srivastava, Principal Customer Success Technologist

Informatica

Housekeeping Tips

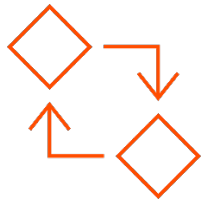
- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our INFASupport YouTube channel and Success Portal - where you can download the slide deck for the presentation. The link to the recording will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Feature Rich Success Portal



Bootstrap trial and POC Customers

Enriched Customer Onboarding experience



Product Learning Paths and Weekly Expert Sessions

Informatica Concierge

Tailored training and content recommendations

Informatica

More Information

Success Portal

https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com

University

https://www.informatica.com/in/services-and-training/informatica-university.html

Informatica

Safe Harbor

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

Statements made today are based on currently available information, 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.

Informatica

Agenda

1. Introduction to EDC 10.5.x Tech Stack
2. Architecture Deep Dive
3. Scanner Process Overview
4. Security, High Availability & Disaster Recovery
5. Sizing & Deployment
6. Q&A

Scope

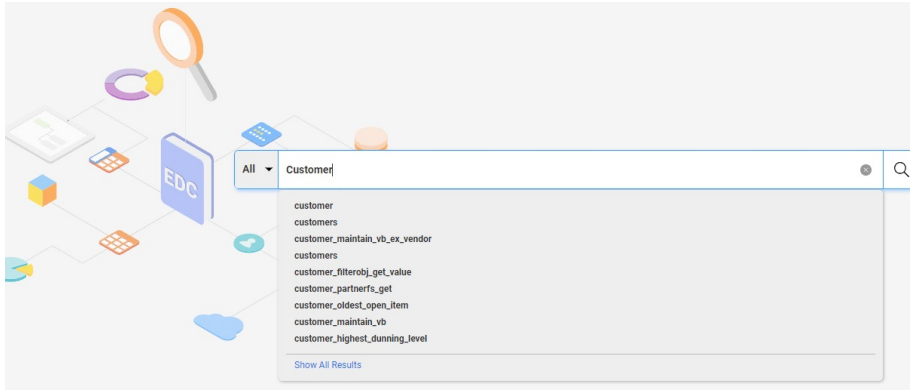
- Enterprise Data Catalog (EDC) version 10.5.x is considered for the discussion.
- The content is not specific to a particular Cloud Ecosystems, but the architectural concepts of EDC will remain the same.

Enterprise Data Catalog - Vision

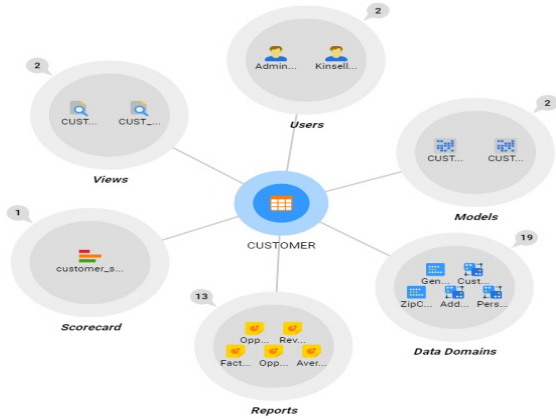
Enterprise Data Catalog enables Business and IT users to unleash the power of their enterprise data assets by providing a unified metadata view that includes technical metadata, business context, user annotations, relationships, data quality and usage

Enterprise Data Catalog - Vision

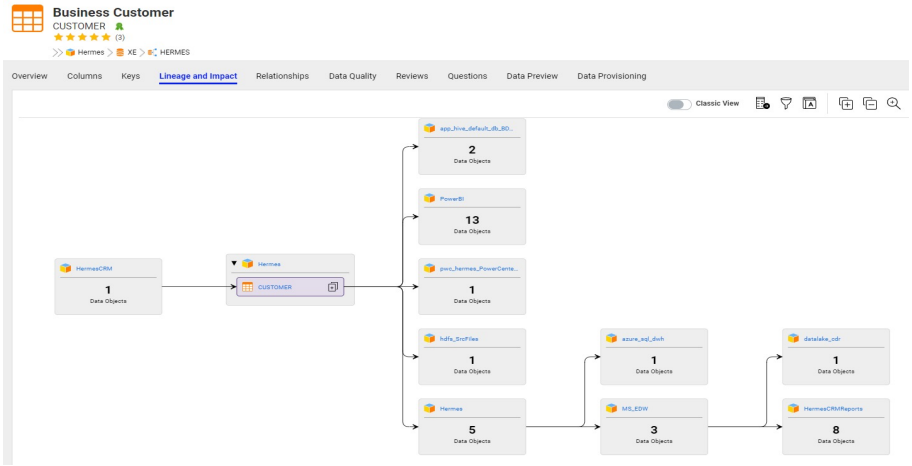
Easily find data with simple, powerful semantic search and faceting



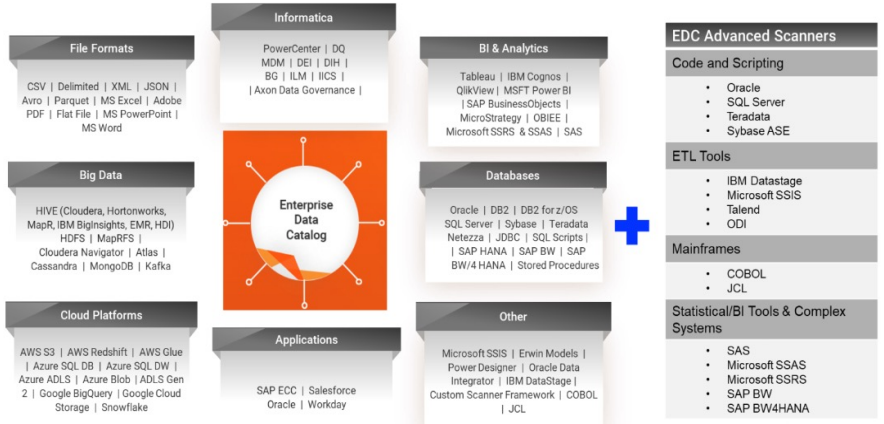
AI-powered automatic entity discovery, classification, data similarity, suggestion



Understand & trust data with data profiles, lineage, certification & user rating



Enterprise scale with broad metadata scanners and open API



Informatica

EDC 10.5.x Architecture Deep-dive

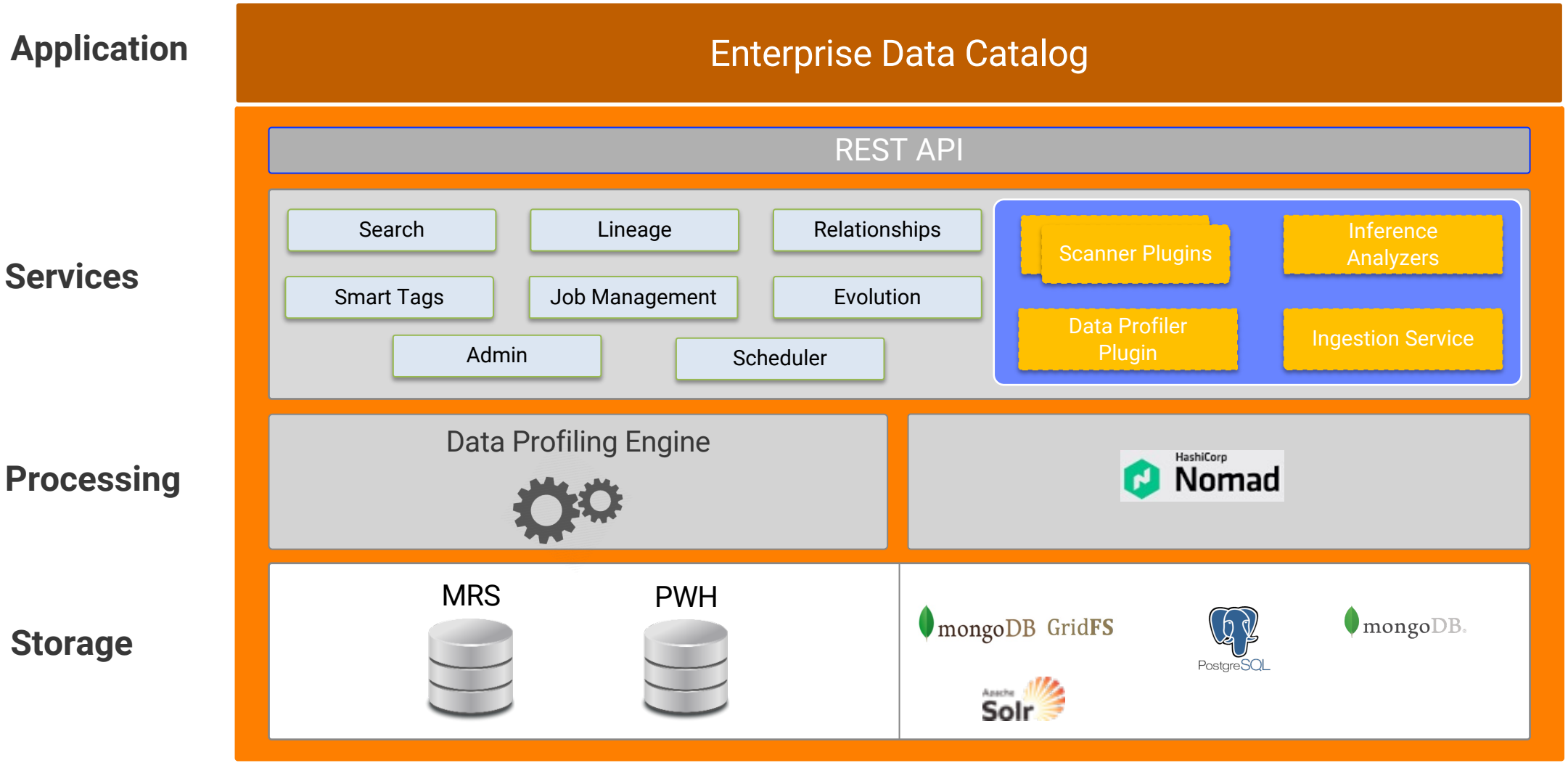
Informatica

EDC 10.5.x – New Tech Stack

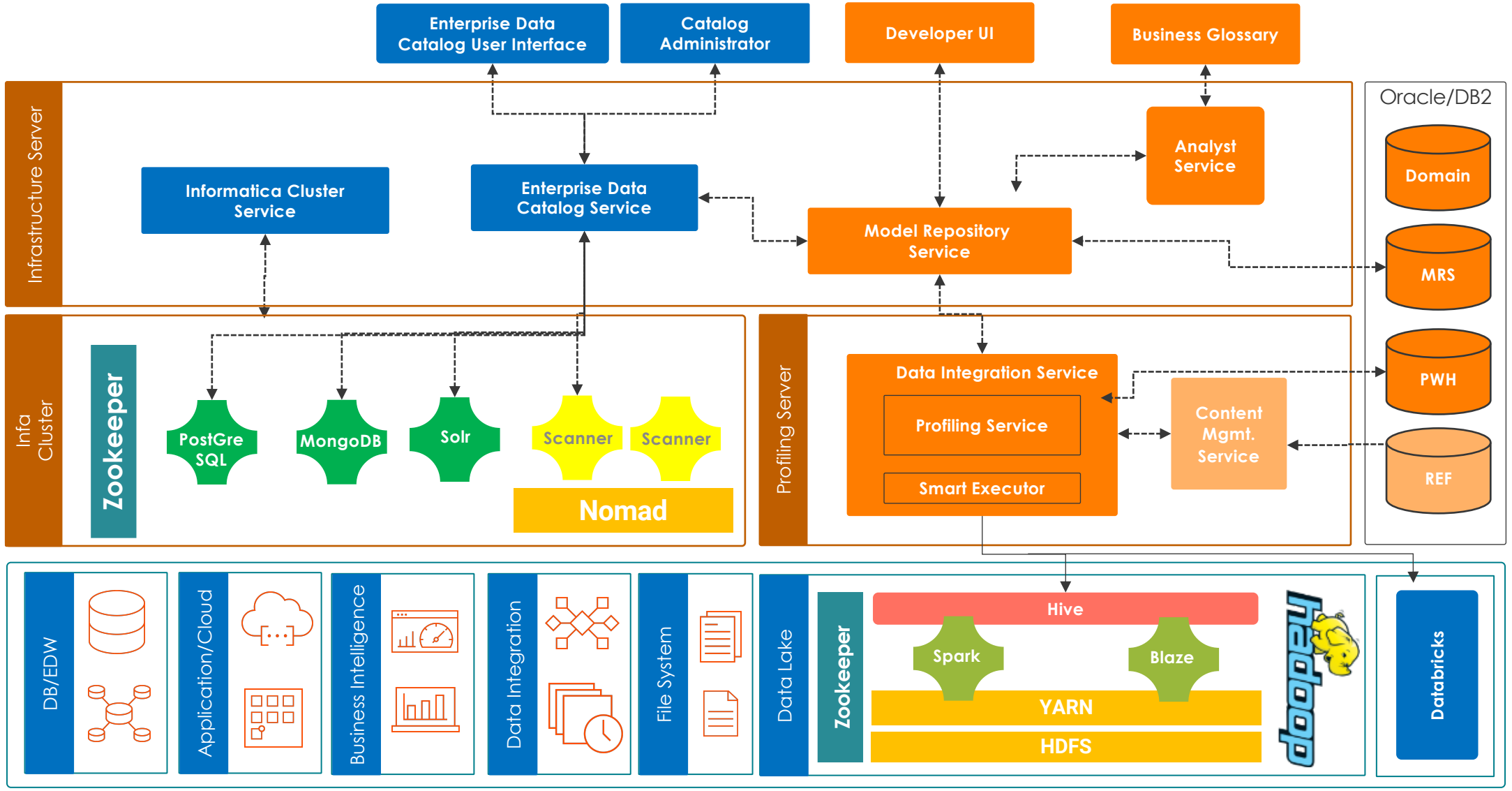
- Enterprise Data Catalog 10.5.x will use:
  - For Storage
    - MongoDB, MongoDB GridFS
    - PostgreSQL
    - SolR
  - For Orchestration and Security
    - Nomad
  - For Service Coordination
    - Zookeeper
  - For Compute
    - Native Java (Scanners, Ingestion service)

Store/Engine	EDC 10.5.x Tech stack	How is it used?
Asset Store	MongoDB	Storage for Metadata scanned.
Graph Store	MongoDB	Storage for lineage and relationship.
Stage Store	MongoDB	Staging area for scan content.
Monitoring Store	MongoDB	Monitoring stats from job runs.
Scan Content Store	MongoDB GridFS	Long term scan content storage.
Similarity Store	PostgreSQL	Storage for similarity profiling - similar columns inferred from other data source.
Event Store	Relational DB	Event store for Data Asset Analytics.
Config Store	Relational DB	Config store for Resources to scan. Uses the same Model Repository Service DB.
Index Store	SolR	Indexer for faster search and analytics.
Compute	Native on Nomad	Native jobs on Nomad.

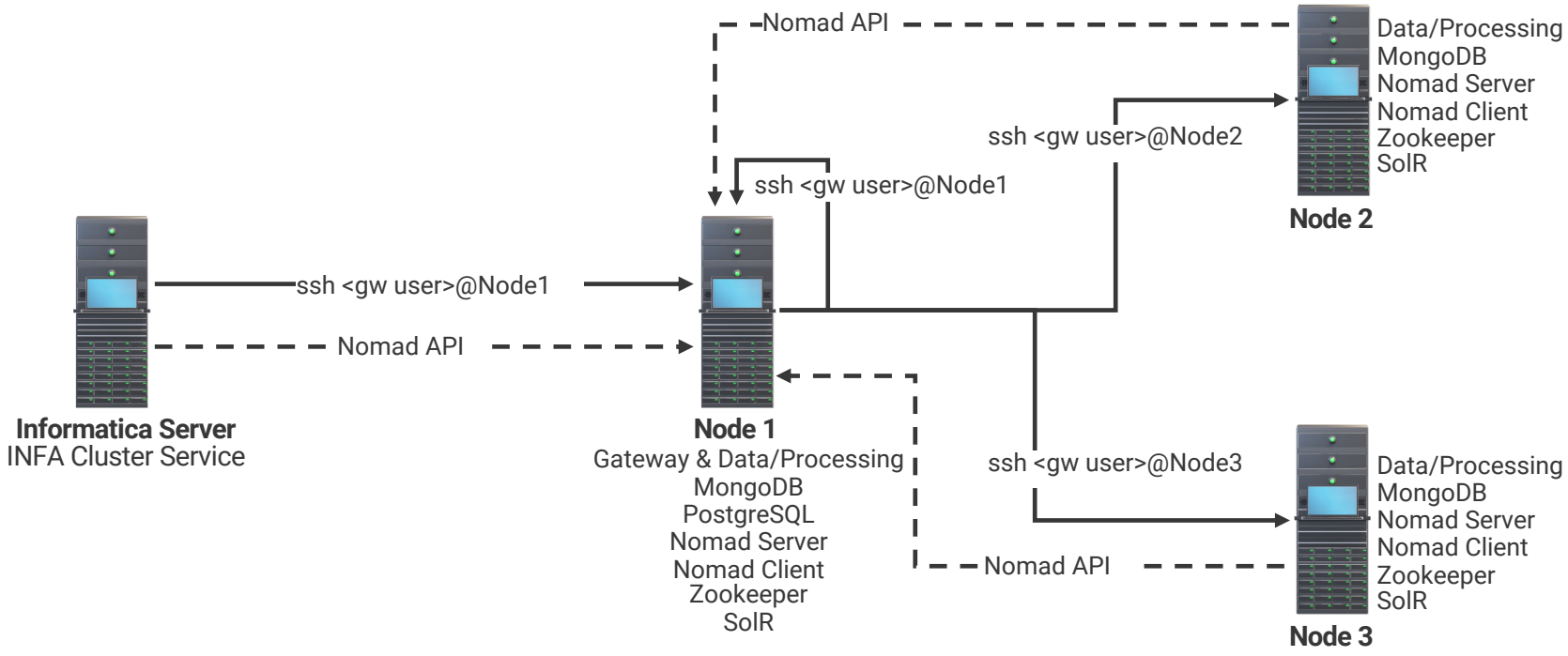
EDC 10.5.x – Application Stack



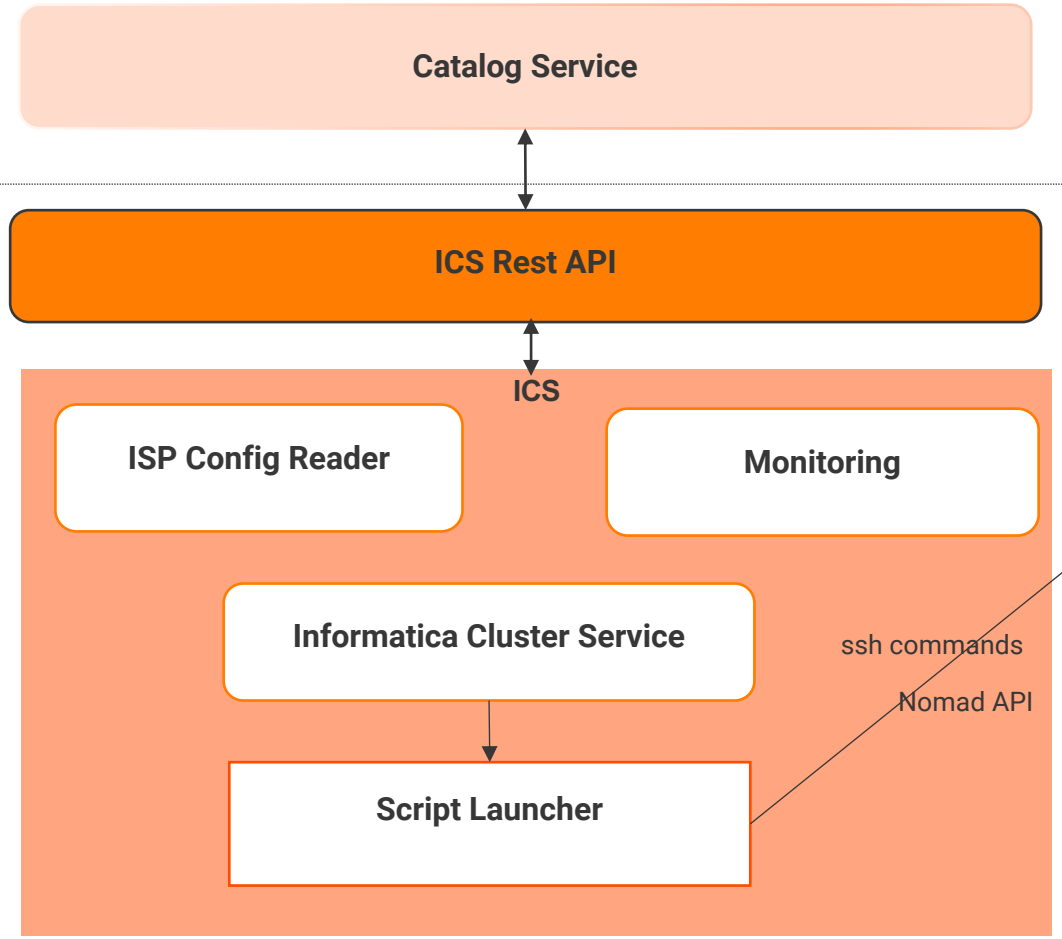
EDC 10.5.x – Services Architecture

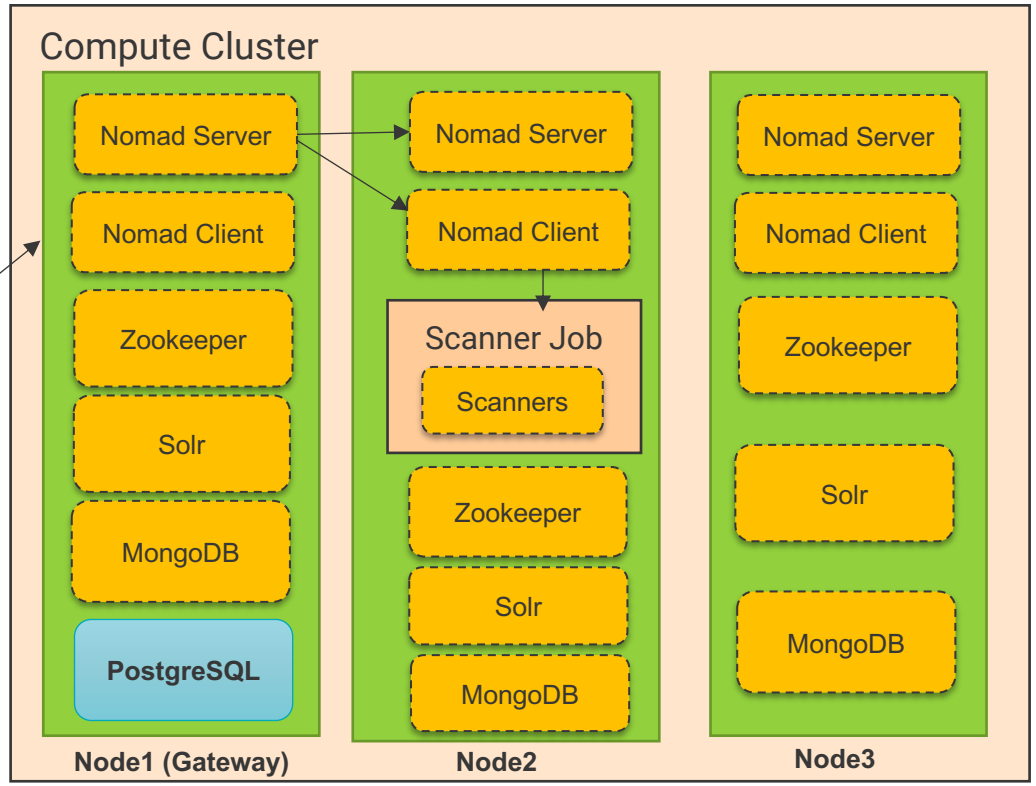


Informatica Cluster Service Architecture



Informatica Cluster Service Architecture





Scanner Execution Overview

- **Scan / Metadata Load**

  - Launches a scanner to scan source system, produce exchange documents to file system.

- **Stage**

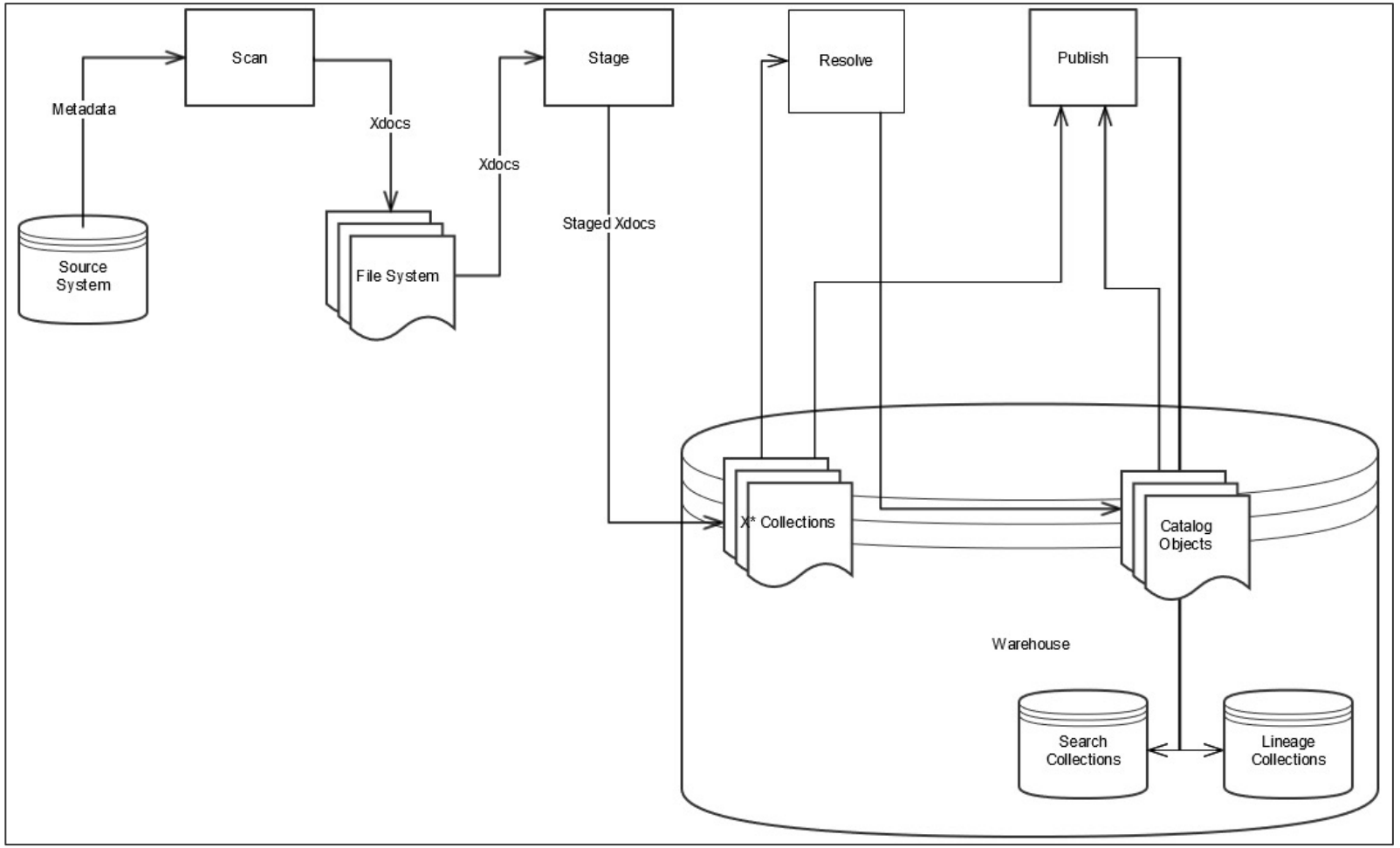
  - Runs in parallel with scan task, monitors file system, read exchange documents, transform & uploads to warehouse.

- **Resolve**

  - Works on all staged documents for a single resource and helps to resolve parameter values, create reference objects etc.

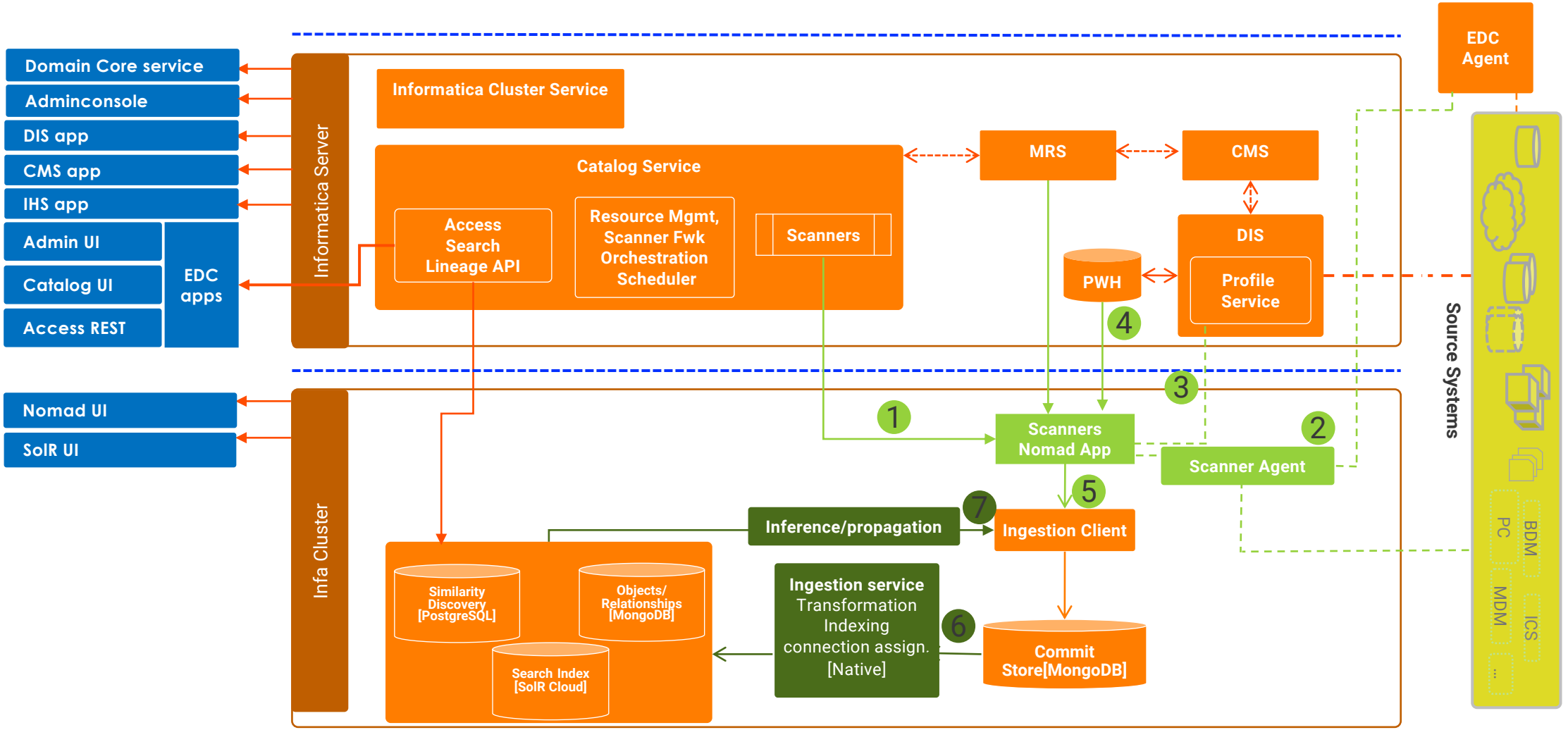
- **Publish**

  - Ingest search attributes, indexes and lineage into their respective storage in the metadata warehouse.



Informatica

Scanner Process Internals



EDC High Availability & Disaster Recovery

Informatica

EDC High Availability vs Disaster Recovery

High availability is often confused or intermixed with disaster Recovery

Platform High Availability

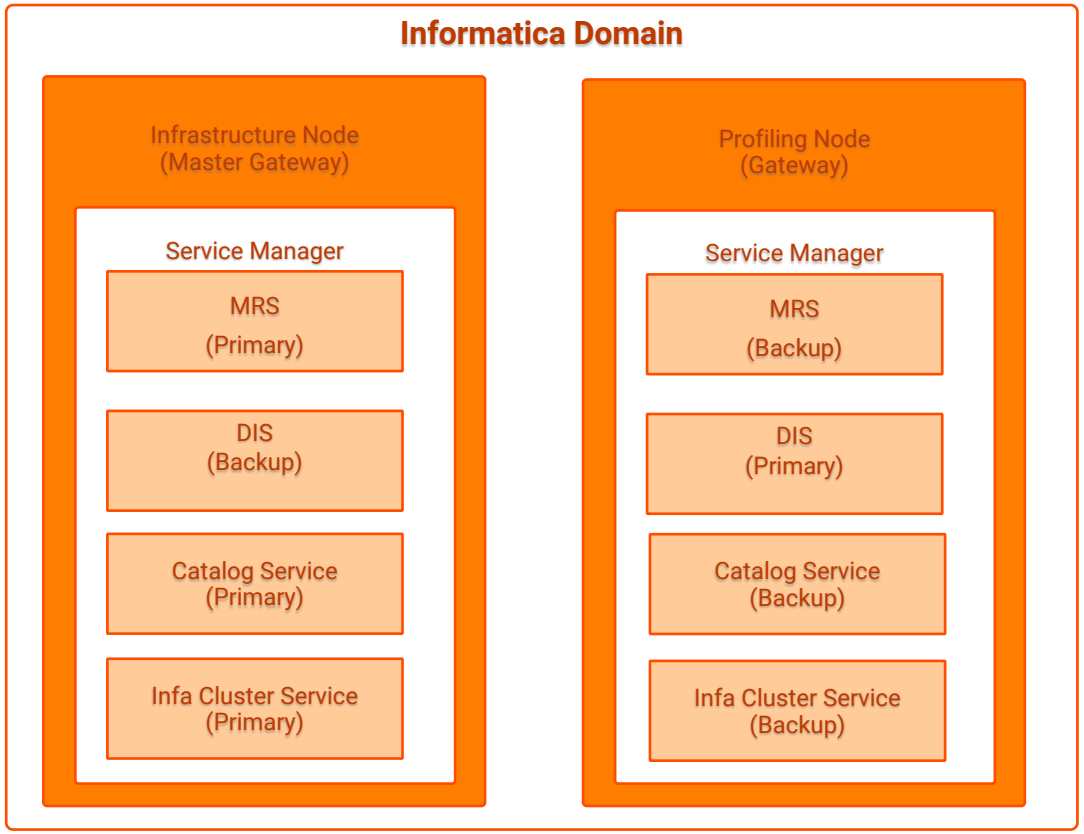
- Avoid Single point of Failure (SPOF)
- Single Site Installation
- Automatic Recovery (Some Services)
- Automatic fail over in case of
  - Hardware Failover
  - Application Failover
  - Service Failover

Platform Disaster Recovery

- Avoid operations interruptions
- IT extension for Business Continuity
- Recovery / Restart ability Possible
- Manual Fail over / Recovery in Case of
  - Earthquake
  - Hurricane,
  - War

EDC Services High Availability

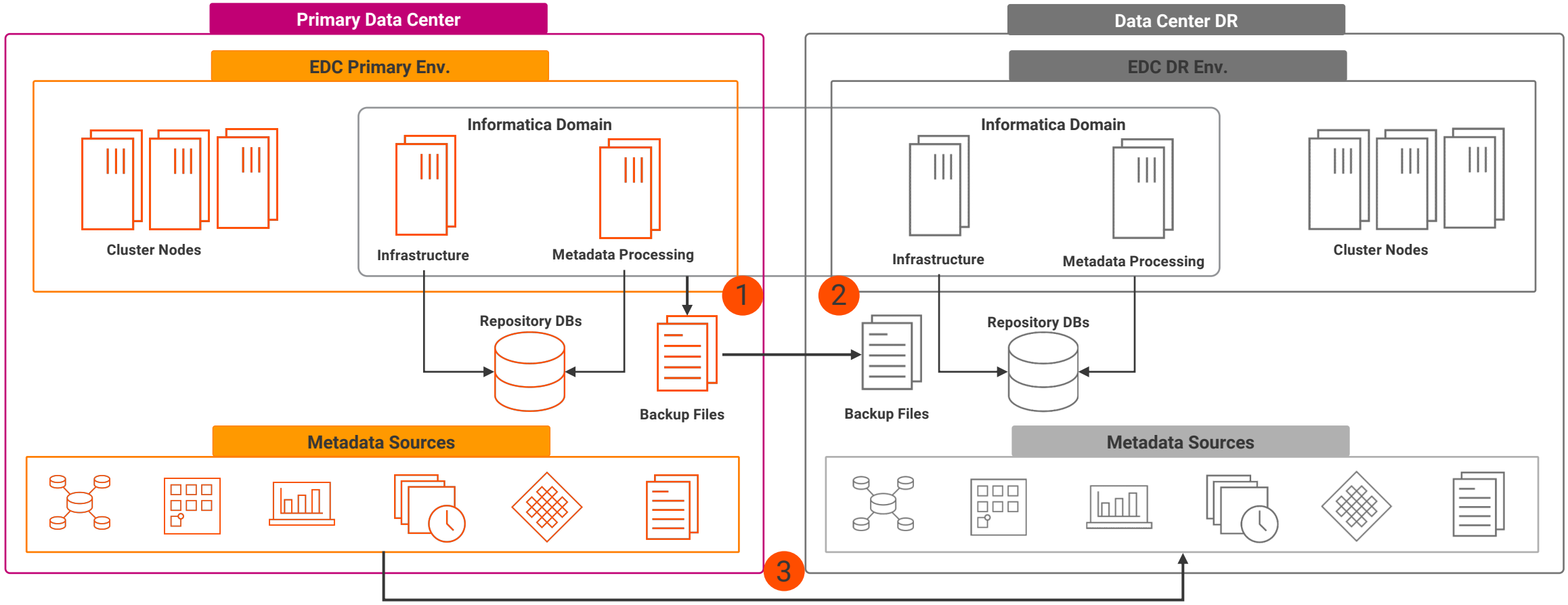
- EDC benefits from Informatica Platform HA
  - In a domain with 2 or more nodes, the service can have a backup node
  - It is recommended to have a multi-node domain
    - Allow high availability to be configured
    - Allow segregation of Infrastructure and profiling services on 2 distinct machines
- EDC Services can be configured for HA
  - Domain gateway services automatic failover
    - Model Repository Service
    - Data Integration Service
    - Catalog service
    - Informatica Cluster Service



Cluster High Availability

- When Informatica Cluster service is deployed on 3 node or more
  - Zookeeper is deployed on all cluster nodes.
  - Solr is deployed on all cluster nodes.
  - Mongo DB can be deployed multiple/all nodes.
  - Nomad can be deployed multiple/all nodes.
  - PostgreSQL DB is deployed on a single node.
- Known limitations
  - PostgreSQL DB is not highly available OOTB. (External PostgreSQL DB support is on roadmap which can be highly available and thus removing this limitation.)

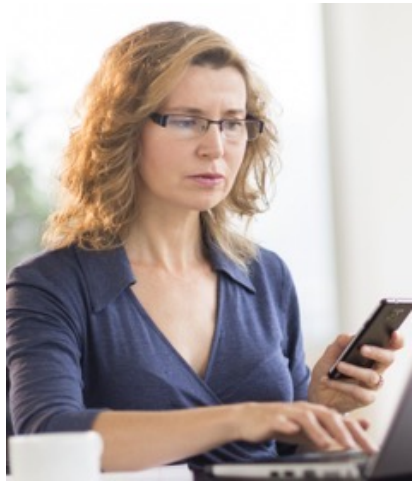
EDC Disaster Recovery – Architecture Overview



1. Perform regular backup and copy the backup files to the DR env. Leverage EDC's Hot backup to ensure seamless experience to users.
2. At DR time, start services and restore service backup
3. Ensure dependent services and other applications as restored

EDC Security considerations

How to make EDC secured ?



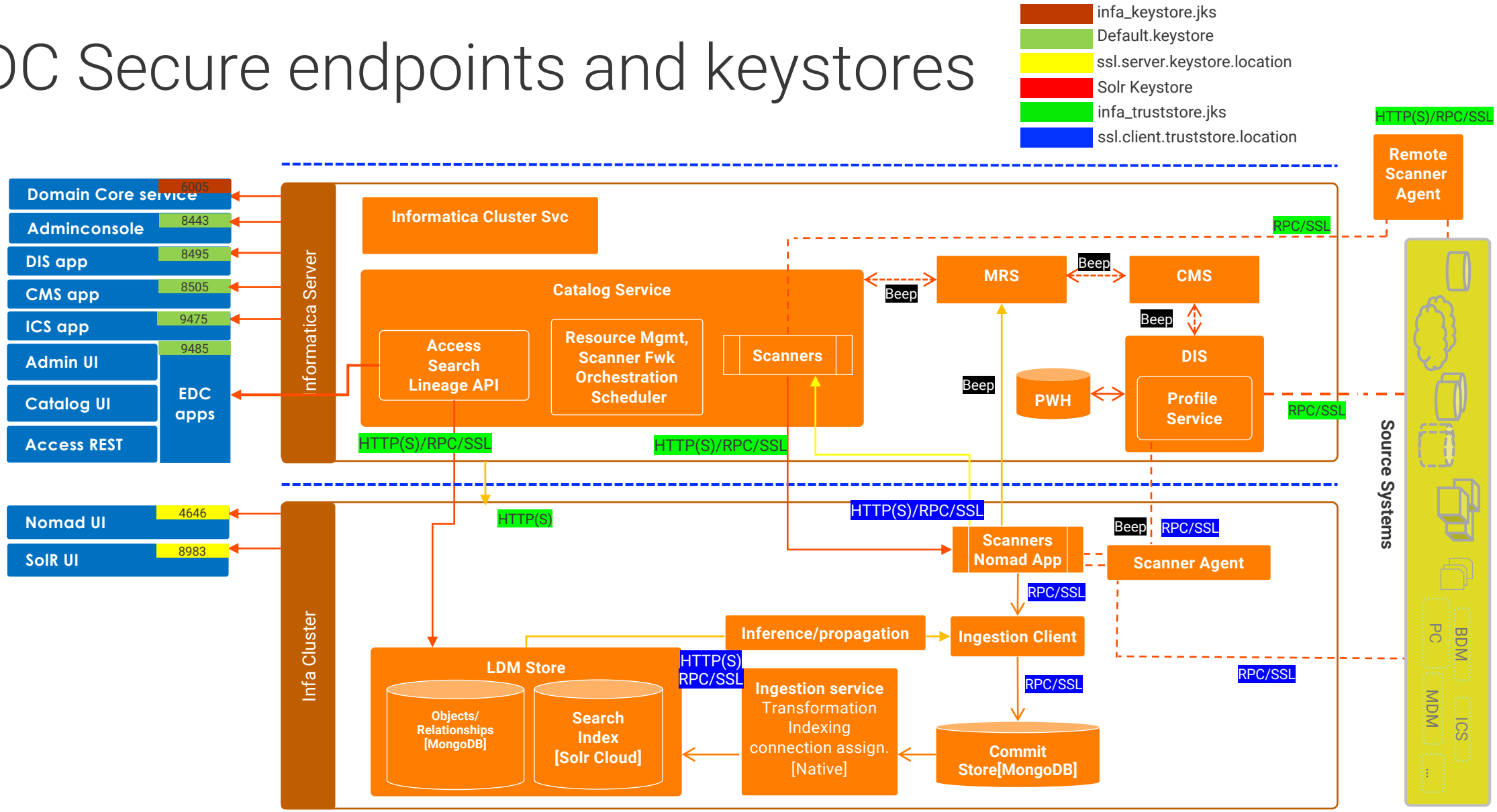
Catalog Administrator

Catalog metadata may be treated with high risk

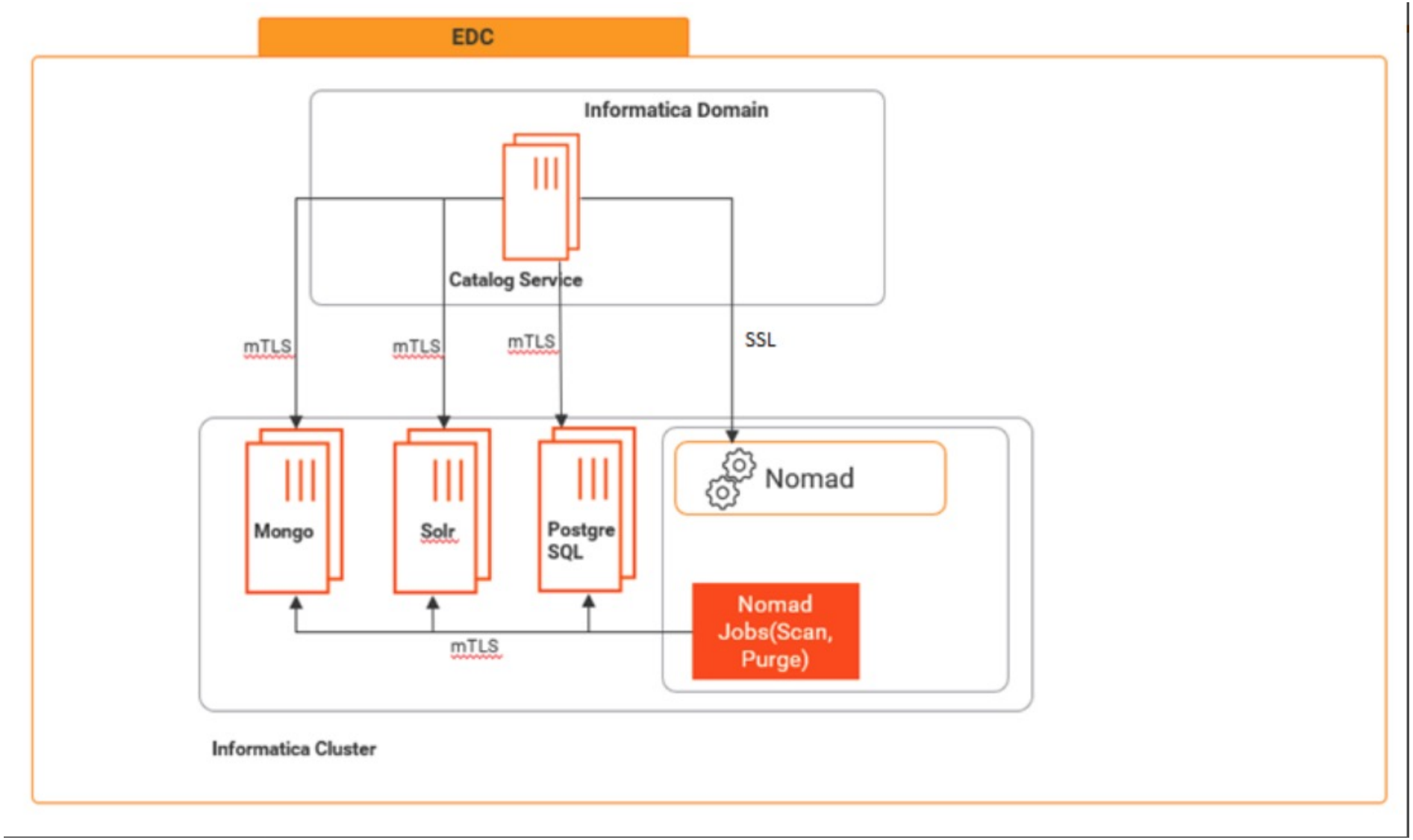
- **Communication level** encryption (Metadata and data in transit)
  - EDC support SSL for all external endpoint (Catalog UI / REST API)
  - EDC support SSL for internal communication
- **Storage level** access control (Metadata and data at rest)
  - Passwords in scanner configuration encrypted using siteKey provided while domain creation.
  - Value Frequency data stored in Catalog is AES-256 encrypted by default.
  - Encryption for MRS & Domain DB is maintained by the platform (siteKey)
- **Application's level** metadata and data access protection through privileges and permissions
  - EDC provides control over who can access/modify functionalities
  - EDC provides control over who can access/modify specific sources for both metadata and data accessible in the catalog

Informatica

EDC Secure endpoints and keystores



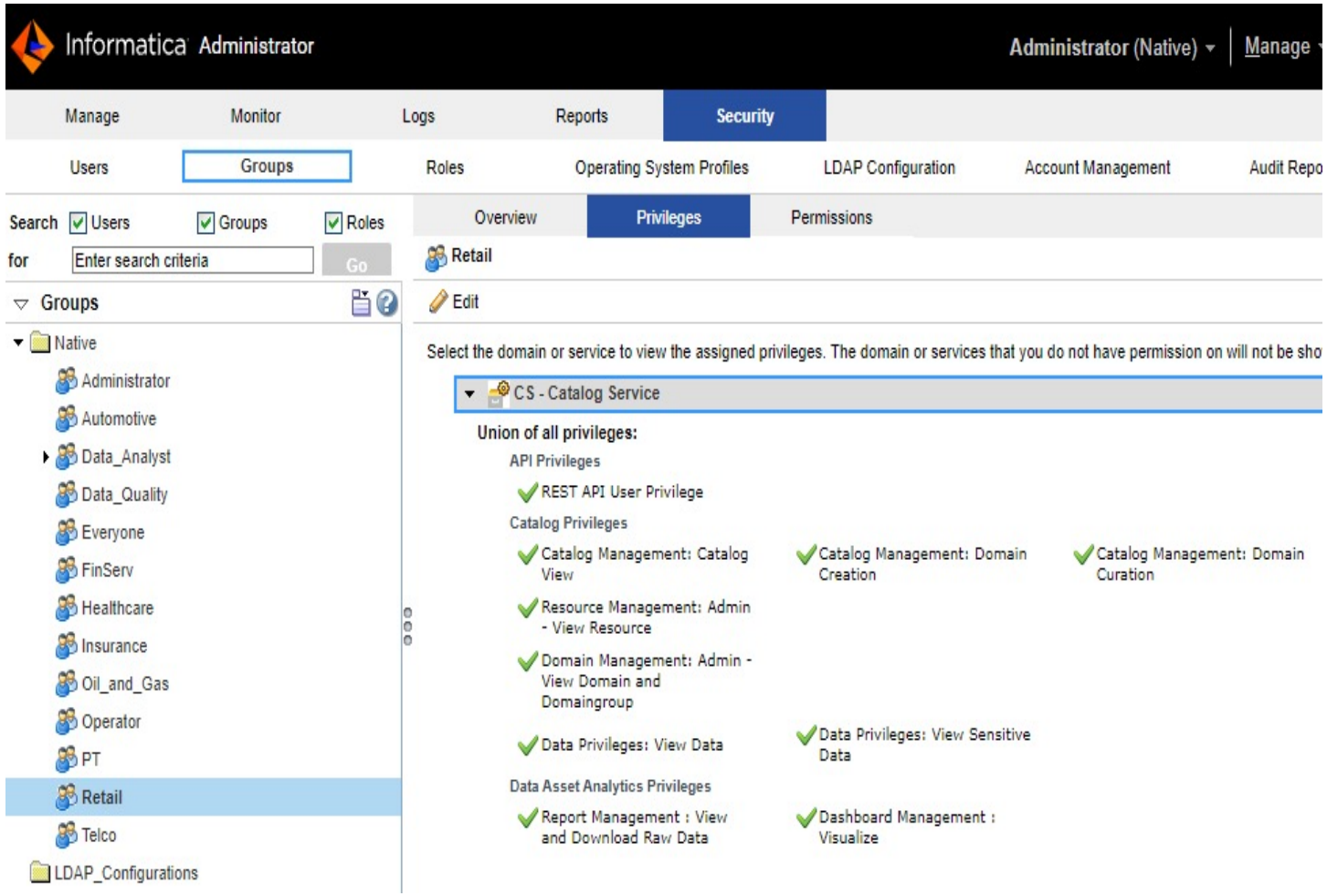
ICS mTLS/SSL Architecture (Mutual TLS for Authentication)



Informatica

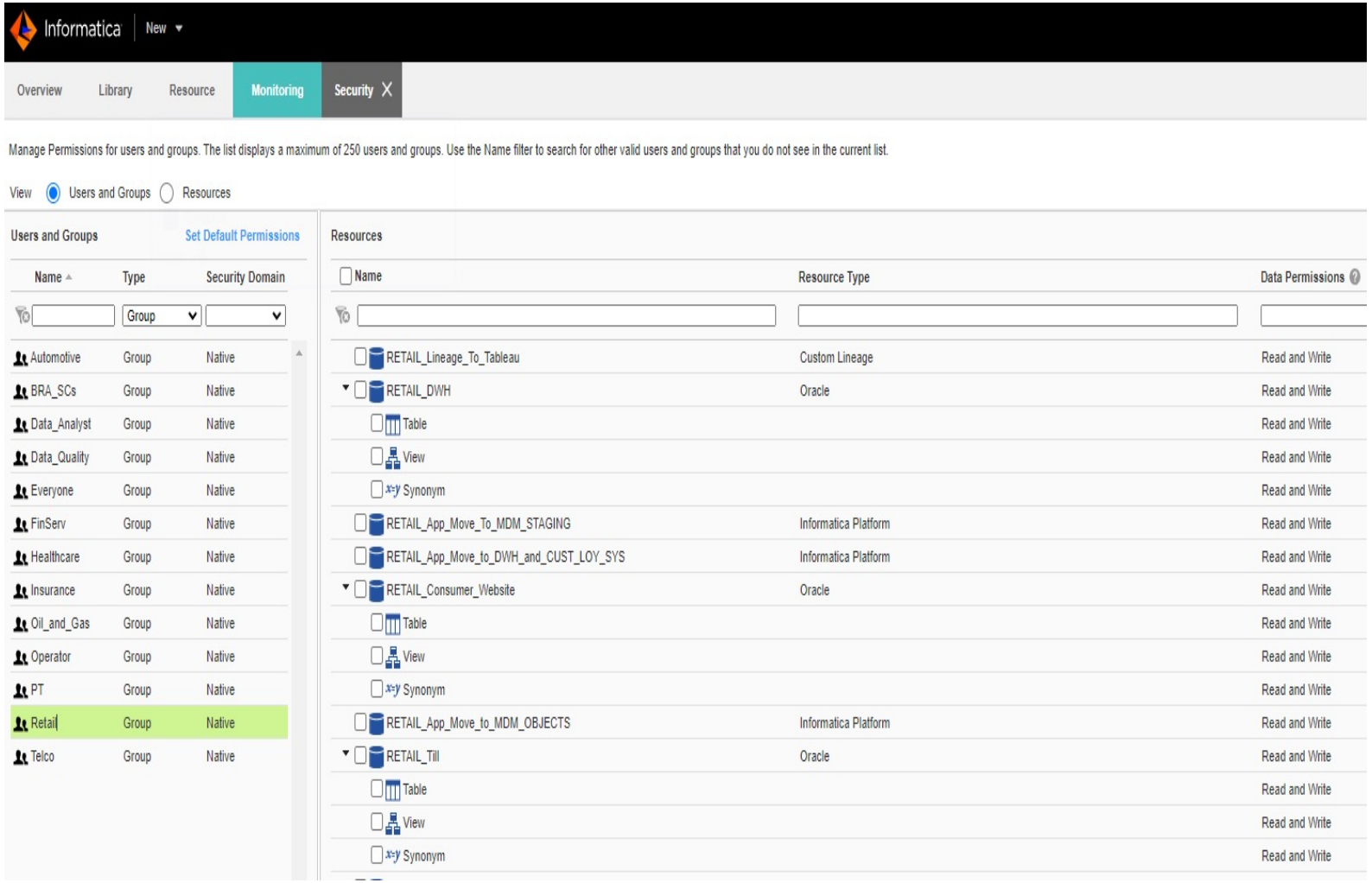
Privileges - Informatica Admin Console

- Privileges are granted at the service level
- Catalog Service access
  - View metadata (minimum to access the Catalog UI)
  - View data and sensitive data
  - Edit metadata / curation
- Catalog Administration
  - Resource management
  - domain and attributes management
  - monitoring
- Development – REST API
  - API access for user / full access



Permissions – Catalog Administrator

- Permission assigned at resource level
  - Read only
  - Read and Write
  - Metadata and data read
  - All permissions
- Granularity down to the object type for RDBMS only (tables, views, synonyms)



Informatica

EDC Roles / User Profiles

EDC-administrator	Created one common LDAP group , these are technical folks who take care of applications/servers etc. Full EDC tool privileges including data source configuration and ingestion scheduling
EDC-steward-<LOB>	EDC Domain editing and curation for their specific Lines of business , full permissions to the assets,will have READ permissions to everything else in catalog.
EDC-data-analyst	General user access for Metadata read-only and lineage (no access to sensitive data or curation capabilities). Most users will fall under this category
EDC-privileged-data-analyst-<LOB>	Metadata and lineage access plus authorized privilege to read sensitive data for specific resources, will have READ permissions to everything else in catalog.

Platform Admin	Catalog Admin*	Data Steward	Data Owner	Catalog user
Responsible for the infrastructure that supports deployment and smooth operations of EDC • Size and configure servers • Install, upgrade, configure of the Informatica Platform, Enterprise Data Catalog and supporting software • Perform tasks in the Informatica domain admin console or the command line • Create all the necessary application services needed for EDC to connect to the data sources	Responsible for implementing the requirements defined by the Data Catalog Owner • Configure and schedule resources on EDC Catalog Administrator to ingest metadata • Configure automatic glossary association, data profiling, domain discovery, and custom attributes to enrich technical metadata • Configure users, user groups and permissions to resources and column profile data • Work with Data architects & Developers to leverage EDC custom scanner framework, and Open RestAPIto extend EDC	Responsible for curating the data catalog with business contextual information • Define and associate business glossary terms to technical assets • Maintain all the synonyms, data domains, and custom attributes for their sources in Catalog Administrator. • Work with Developers to create rules for a data domain rule or manage DQ reference tables. • Approved/reject tags • Certify data assets • Moderate reviews, ratings and Q&A on the catalog	Responsible for delivering business outcomes defined in the program strategy • Identify the technical metadata sources, define business contextual enrichment requirements, define user training requirements, and define user access permissions needed to support the business use cases • Work with Data Owners in business functional areas in understanding and defining business adoption requirements • Train and engage catalog users • Monitor and track usage metrics • Drive business adoption	Responsible for increasing business value of the data catalog with their feedback and data knowledge • Accomplish their tasks related to data analytics, data governance or data asset management much more efficiently and effectively with EDC • Collaborate on data assets through reviews, ratings and Q&A

Sizing & Deployment

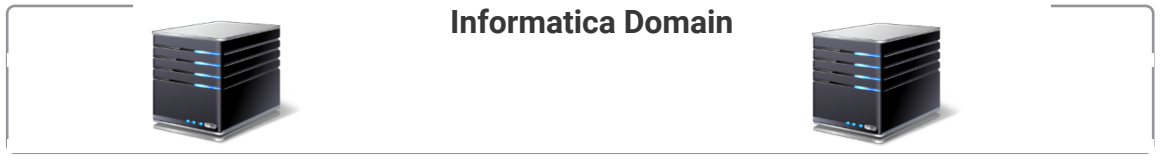
Informatica®

EDC Sizing guideline (summary)

Updated requirements for 10.5.x

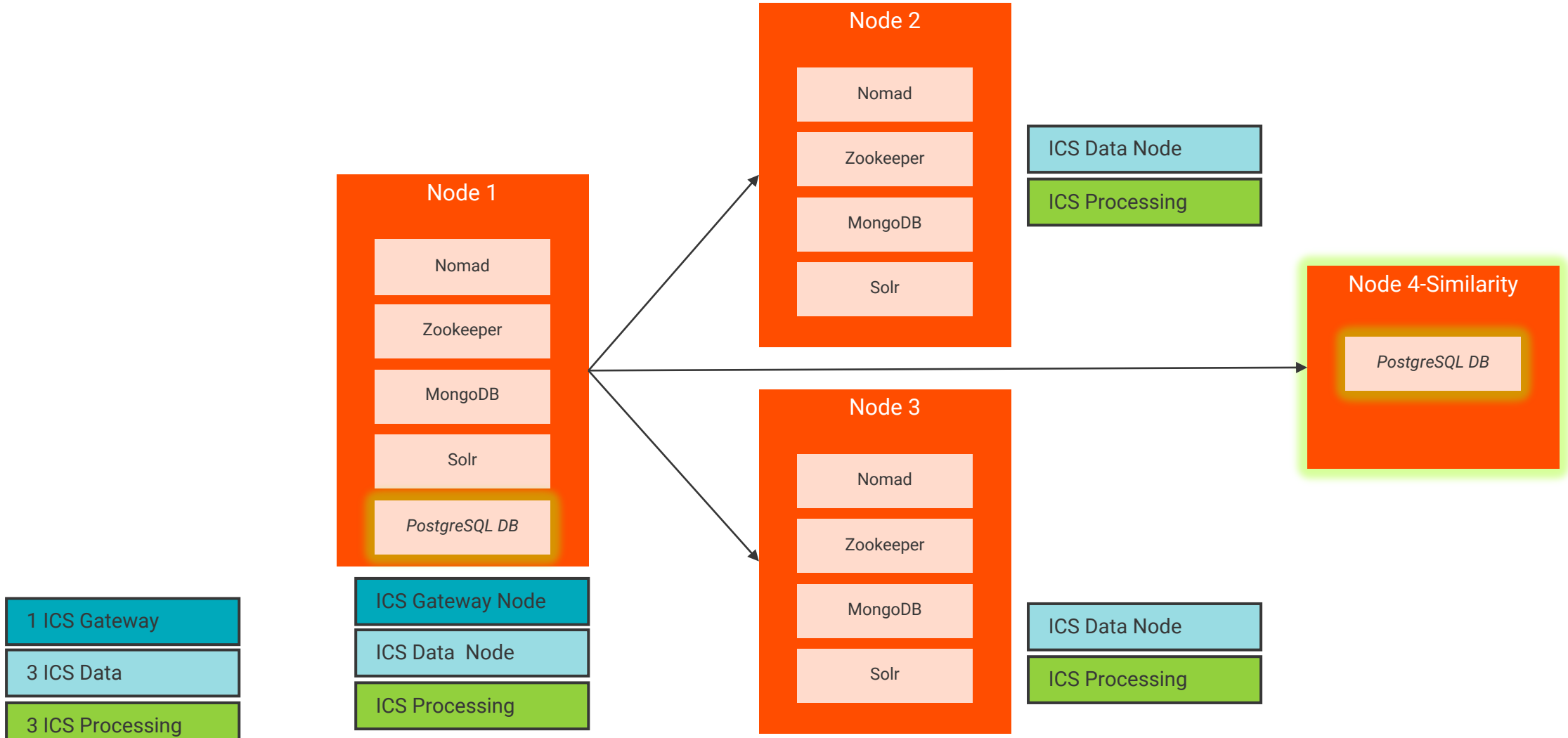
- Refer to Sizing and Performance Tuning Guide for sizing recommendations, parameter tuning and more.

			Infrastructure			Metadata processing						Infa Cluster Services						
			Catalog Service			Discovery			Adv. Scanners*			Data/Compute (MongoDB, Nomad)				Similarity Store (PostgreSQL)		
Env. Size	# of conc. (total) users	# of objects	CPU	RAM (GB)	Disk (GB)	CPU	RAM (GB)	Disk (GB)	CPU	RAM (GB)	Disk (GB)	# of nodes	CPU	RAM (GB)	Disk (GB)	CPU	RAM (GB)	Disk (GB)
Small	20 (200)	2 Million	16	32	200	8	16	20	4	12	50	1	8	24	120	8	16	200
Medium	50 (500)	20 Million	24	32	200	32	64	100	4	16	100	3	24	72	2 TB	8	16	500
Large	75 (750)	50 Million	48	64	300	32	64	500	4	32	200	3	48	144	6 TB	8	16	1 TB
X-Large	100 (1000)	100 Million	48	64	300	64	128	500	4	32	200	6	96	288	12 TB	8	16	2 TB



- Can be merged with Discovery node
- If you use Similarity Discovery - Separate node is recommended for Similarity Discovery (PostgreSQL DB)

ICS – Cluster Nodes (Reference Recommendation)

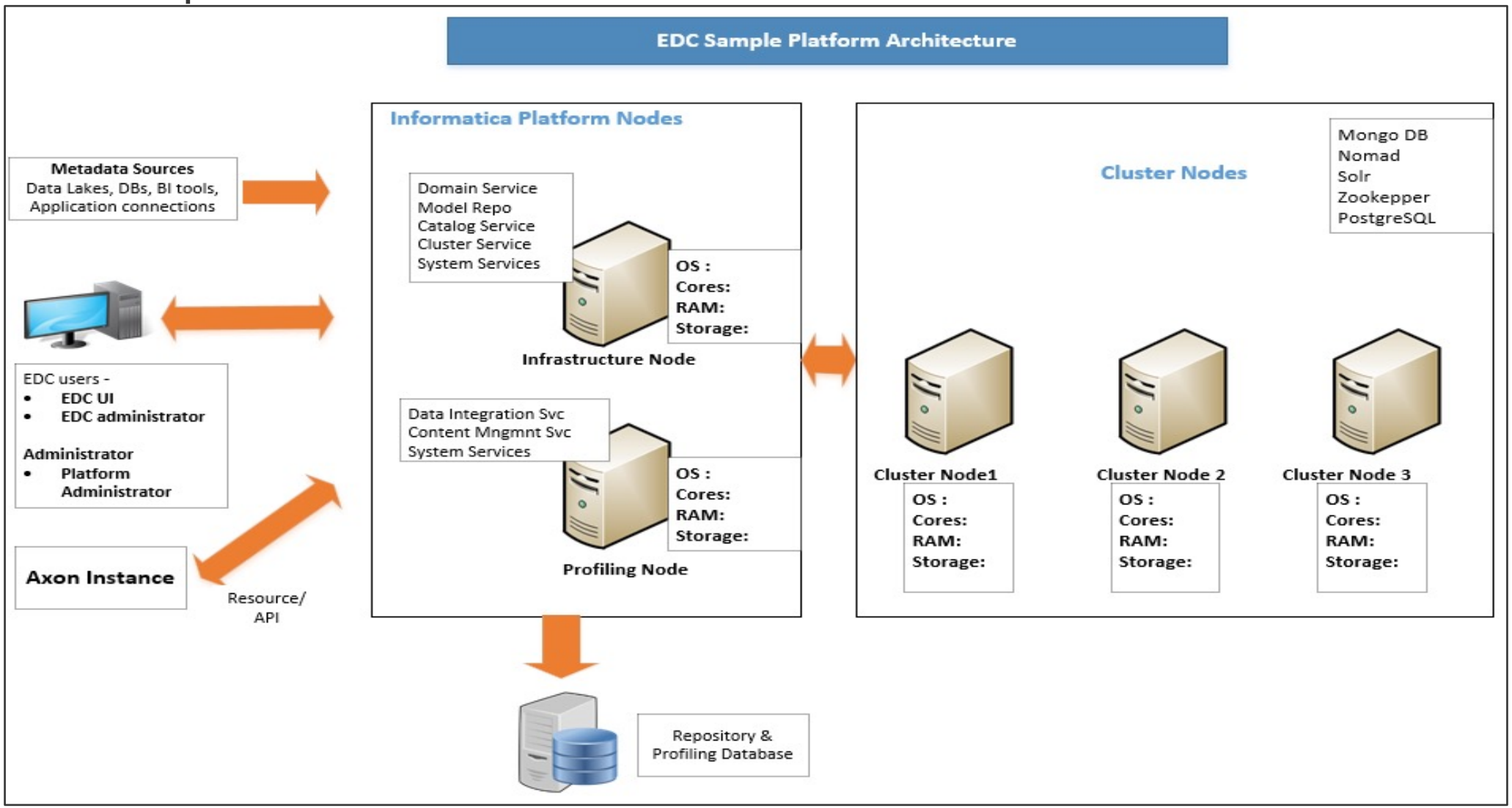


Similarity Store(PostgreSQL DB) is recommended to be on a separate node

but it can be deployed alongside gateway node if sizing/requirement permits.

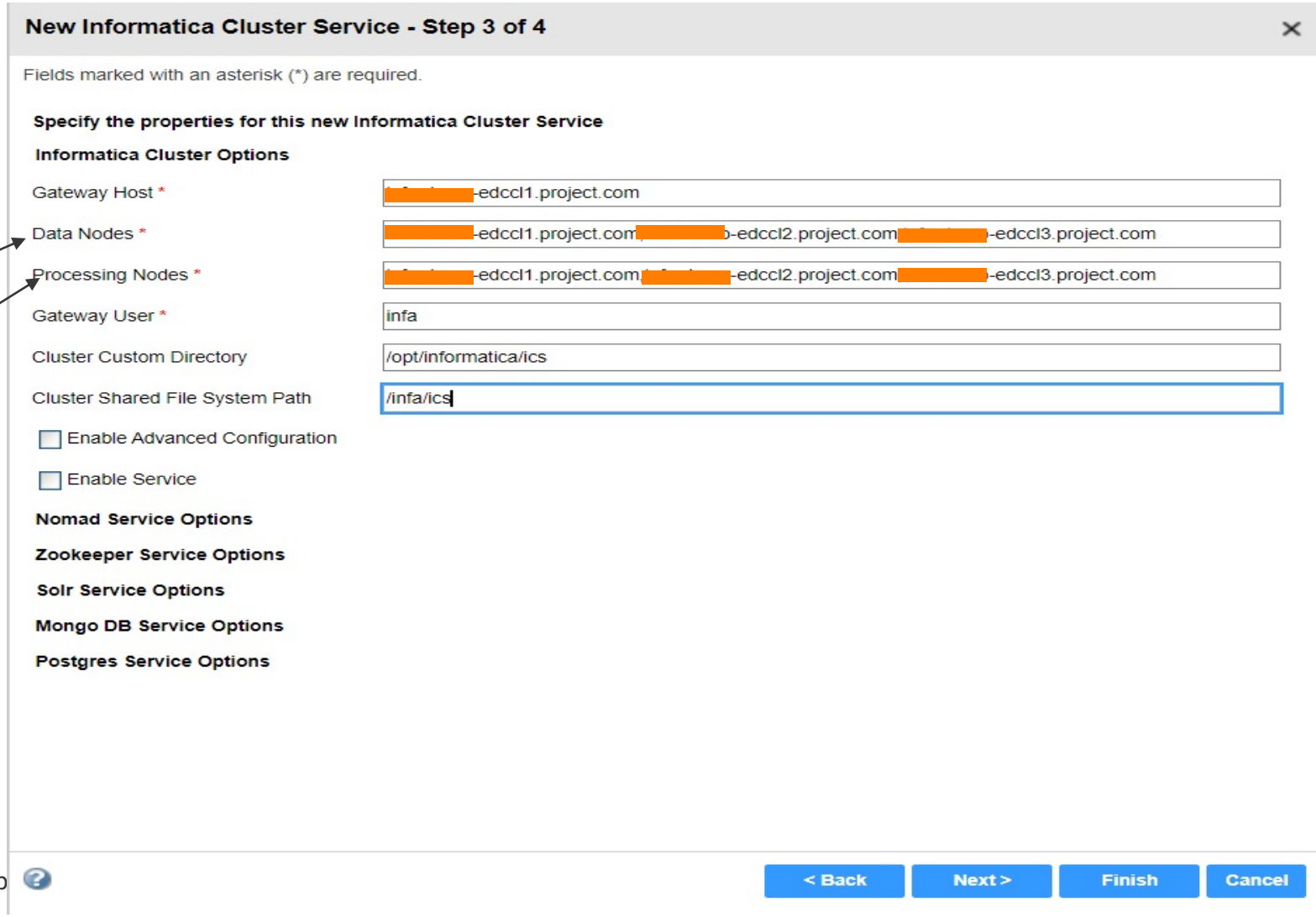
Vertical scale up of gateway node should be reviewed to accommodate Postgre.

EDC Sample Platform Architecture Post Install



Informatica Cluster Service - architecture

- Deployment Nodes



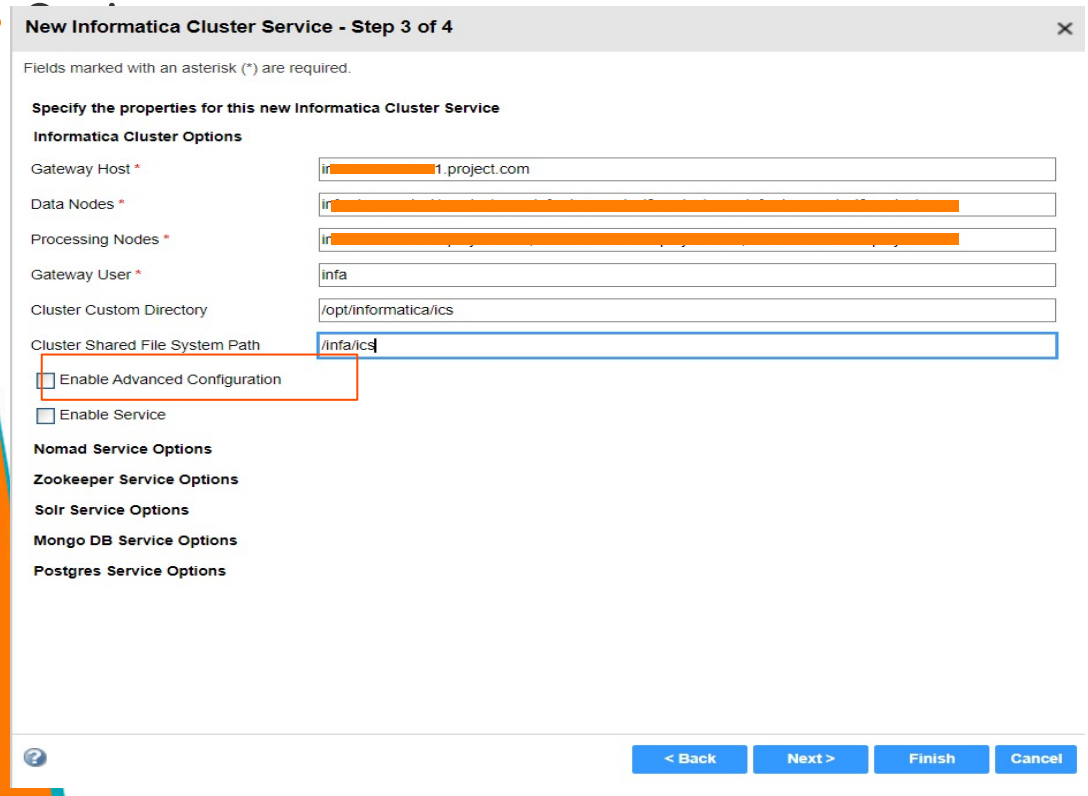
MongoDB

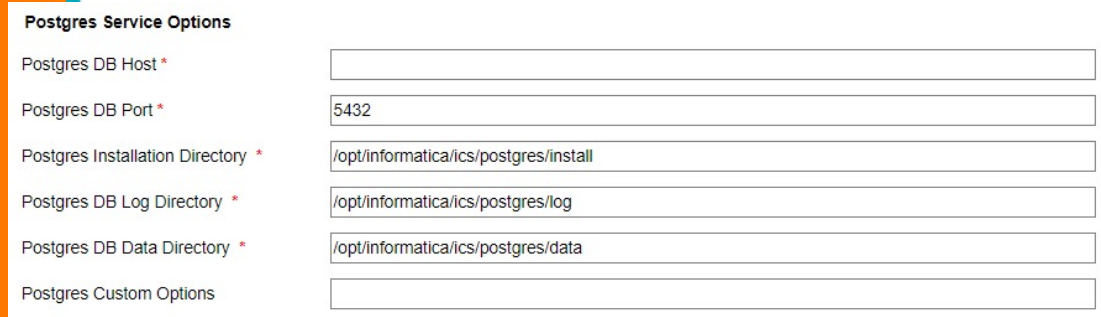
Nomad

Solr and Zookeeper is deployed on all nodes

Informatica

ICS creation





Processing Nodes

Zookeeper and Solr should be on All Nodes

Data Nodes

Advance Configuration

Nomad Service Options

Nomad Server Hosts *	
Nomad HTTP Port *	4646
Nomad Serf Port *	4648
Nomad RPC Port *	4647
Nomad Server Working Directory *	/opt/informatica/ics/nomad/nomadserver
Nomad Client Working Directory *	/opt/informatica/ics/nomad/nomadclient
Nomad Custom Options	

Zookeeper Service Options

Zookeeper Hosts *	
Zookeeper Port *	2181
Zookeeper Peer Port *	2888
Zookeeper Leader Port *	3888
Zookeeper Installation Directory *	/opt/informatica/ics/zk/install
Zookeeper Data Directory *	/opt/informatica/ics/zk/data
Zookeeper Custom Options	

Solr Service Options

Solr Hosts *	
Solr Port *	8983
Solr Installation Directory *	/opt/informatica/ics/solr/install
Solr Data Directory *	/opt/informatica/ics/solr/data
Solr Custom Options	

Mongo DB Service Options

Mongo DB Hosts *	
Mongo DB Port *	27017
Mongo DB Log Directory *	/opt/informatica/ics/mongo/log
Mongo DB Data Directory *	/opt/informatica/ics/mongo/data
Mongo Custom Options	

References

Customer Success EDC 10.5 Upgrade Planner: https://success.informatica.com/upgrade-kit-10-5/Enterprise-Data-Catalog.html

EDC 10.5 Performance Tuning Guide & Sizing:

https://docs.informatica.com/data-catalog/enterprise-data-catalog/h2l/1565-tuning-enterprise-data-catalog-performance-in-10-5/tuning-enterprise-data-catalog-performance-in-10-5/enterprise-data-catalog-sizing-recommendations.html

EDC 10.5.1 Documentation Set: https://docs.informatica.com/data-catalog/enterprise-data-catalog/10-5-1.html

Performance Tuning and Sizing for Data Asset Analytics (DAA):

https://docs.informatica.com/data-catalog/enterprise-data-catalog/h2l/1565-tuning-enterprise-data-catalog-performance-in-10-5/tuning-enterprise-data-catalog-performance-in-10-5/performance-tuning-parameters-for-data-asset-analytics.html

Performance Tuning for EDC Profiling:

https://docs.informatica.com/data-catalog/enterprise-data-catalog/h2l/1565-tuning-enterprise-data-catalog-performance-in-10-5/tuning-enterprise-data-catalog-performance-in-10-5/performance-tuning-parameters-for-profiling.html

EDC Sizing:

https://docs.informatica.com/data-catalog/enterprise-data-catalog/h2l/1565-tuning-enterprise-data-catalog-performance-in-10-5/tuning-enterprise-data-catalog-performance-in-10-5/enterprise-data-catalog-sizing-recommendations.html

Additional information regarding External CA Certificates configuration:

https://knowledge.informatica.com/s/article/HOW-TO-Configure-custom-SSL-for-Enterprise-Data-Catalog-with-external-CA-signed-certificates-using-custom-SSL-scripts?language=en_US

Q&A