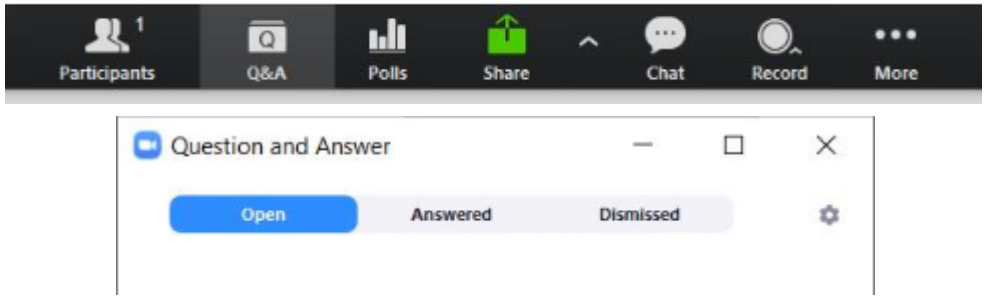
02.11.2020

Enterprise Data Catalog Architecture

Sugi Narayana
Principal Technologist – Customer Success

Informatica™

Housekeeping Tips

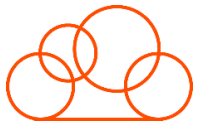


- Todays Webinar is scheduled to last 1 hour including Q&A
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to “All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available to view on our INFASupport YouTube channel and Success Portal. The link will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Success Portal

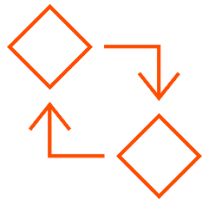
https://success.informatica.com

Learn. Adopt. Succeed.



Bootstrap product trial experience

Enriched Onboarding experience



FREE Product Learning Paths and weekly Expert sessions

Informatica Concierge with Chatbot integrations

Tailored training and content recommendations

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

- EDC Architecture
- EDC Deployment Options
- EDC Security Considerations
- EDC High Availability
- Walk-thru EDC Services
- Q&A

Scope

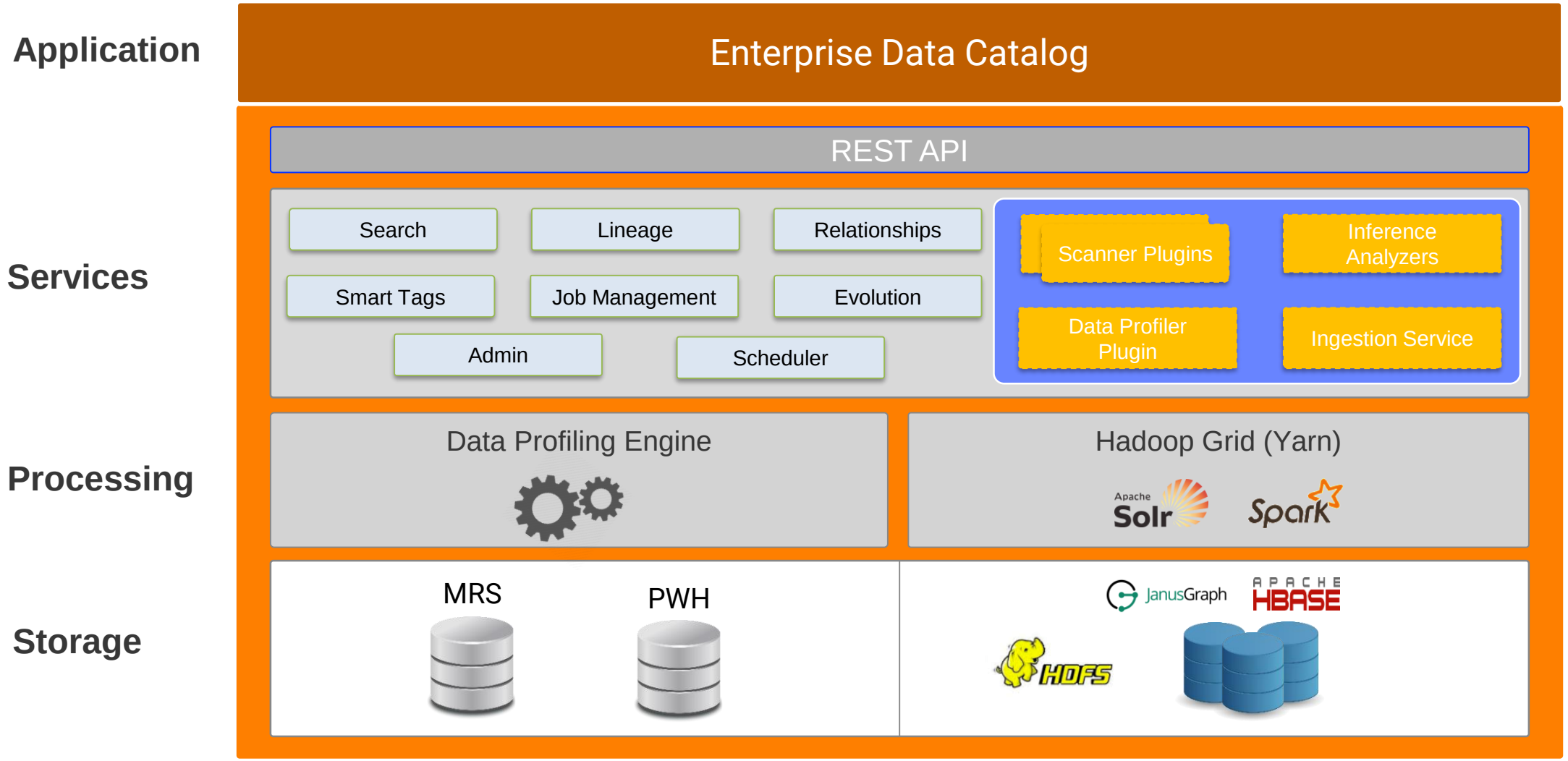
- The latest EDC version 10.4 is considered for the discussion.
- EDC on Cloud Ecosystem is not covered as part of the discussion.

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

EDC Architecture

Enterprise Data Catalog - Application Stack



Informatica

EDC Deployment Options on Hadoop

Supported Hadoop Distributions	Cloudera	Hortonworks	Azure HDInsight

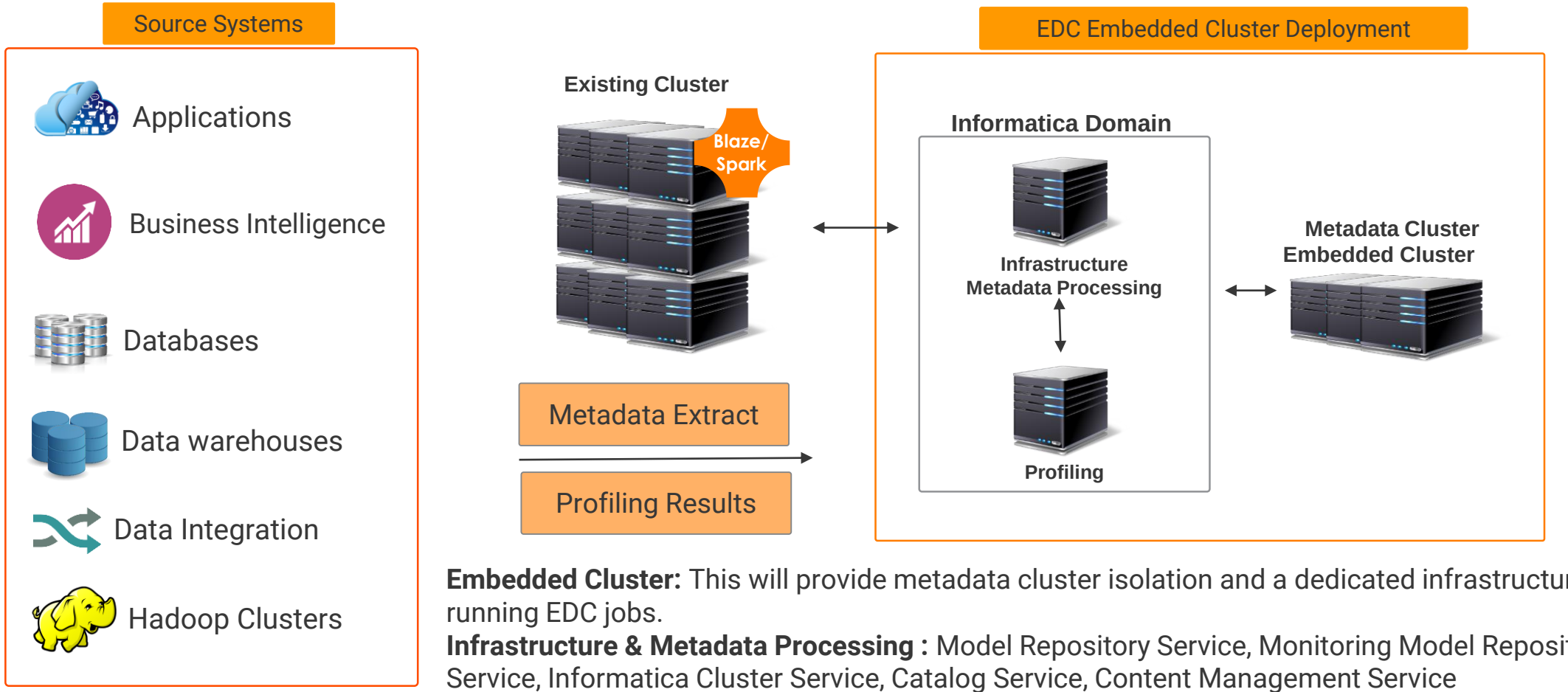
Existing Cluster

- EDC is deployed on an existing cluster on a specified set of Hadoop nodes. It will support specific version/vendor of Hadoop. EDC deploys its own HBase, Solr and Spark instances as Yarn applications.

Embedded Cluster

- EDC deploys its own Hadoop cluster(Hortonworks) on a given set of servers (Linux) along with HBase, Solr and Spark instances as Yarn applications

Embedded Cluster Deployment



Embedded Cluster: This will provide metadata cluster isolation and a dedicated infrastructure for running EDC jobs.
Infrastructure & Metadata Processing : Model Repository Service, Monitoring Model Repository Service, Informatica Cluster Service, Catalog Service, Content Management Service
Profiling : Data Integration Service
*if existing Hadoop cluster to be scanned, pushdown cluster resource profiling jobs on Blaze (or Spark from 10.4) to the existing Hadoop cluster

Deployment Option Comparison

Existing Cluster

EDC is deployed on an existing cluster with its own HBase, Solr and Spark instances as Yarn applications.

- Metadata and data processing jobs are run in one cluster
- Supports specific CDH/HDP/HDInsight versions
- Additional cluster hardware is not required.
- Recommended for customers who are planning to have all data processing in the one cluster

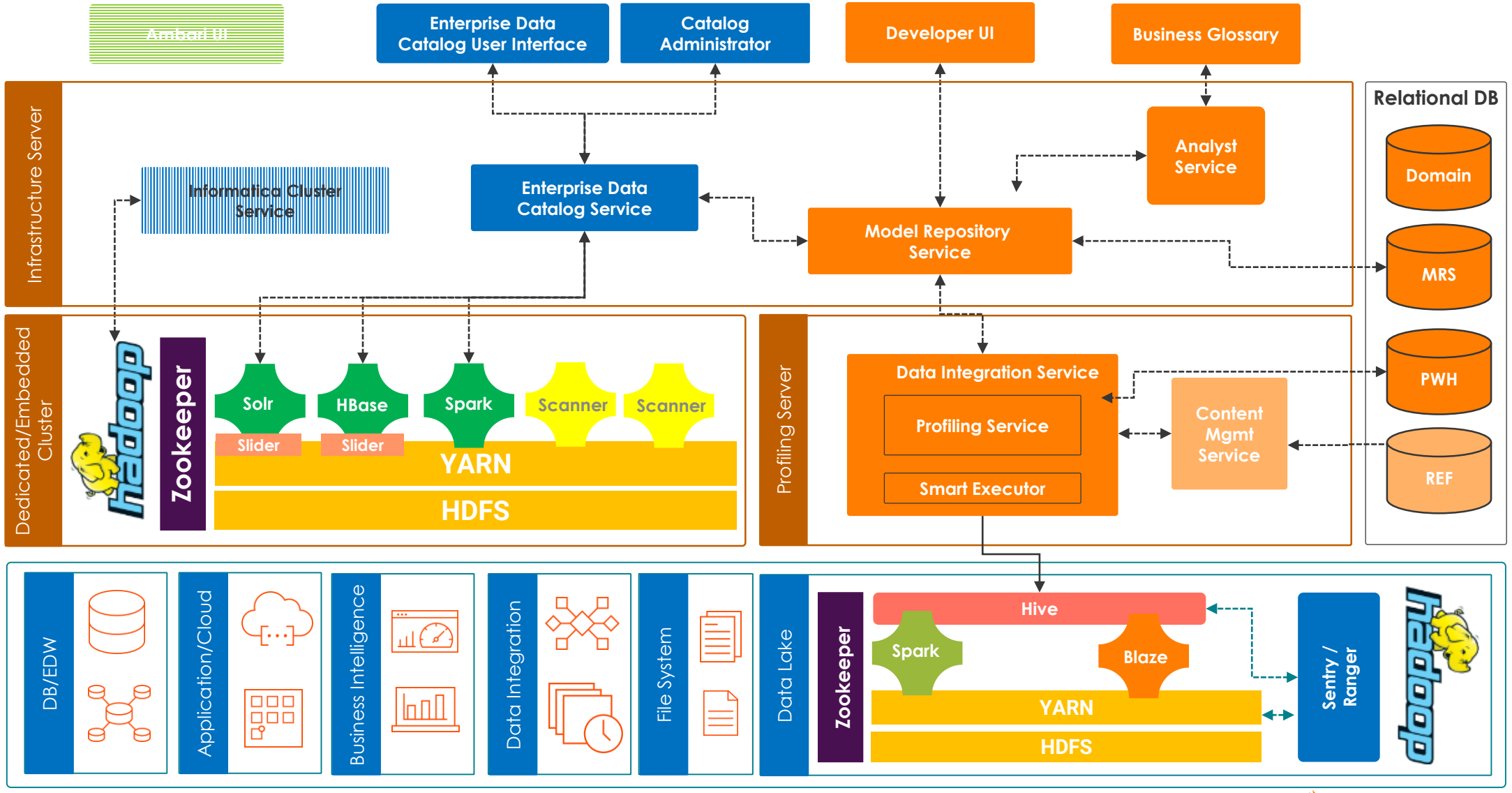
Embedded/Metadata Cluster

EDC is deployed on its own cluster on a given set of Linux servers along with HBase, Solr and Spark instances as Yarn applications

- EDC jobs will not compete for the same resources as data processing jobs which enables Metadata process Isolation
- No dependency for existing cluster upgrades
- Additional cluster hardware is required.
- Recommended for
  - Customers looking for isolated environment with optimized performance
  - Customers with unsupported cluster distributions
  - Customers who don’t have a Hadoop cluster

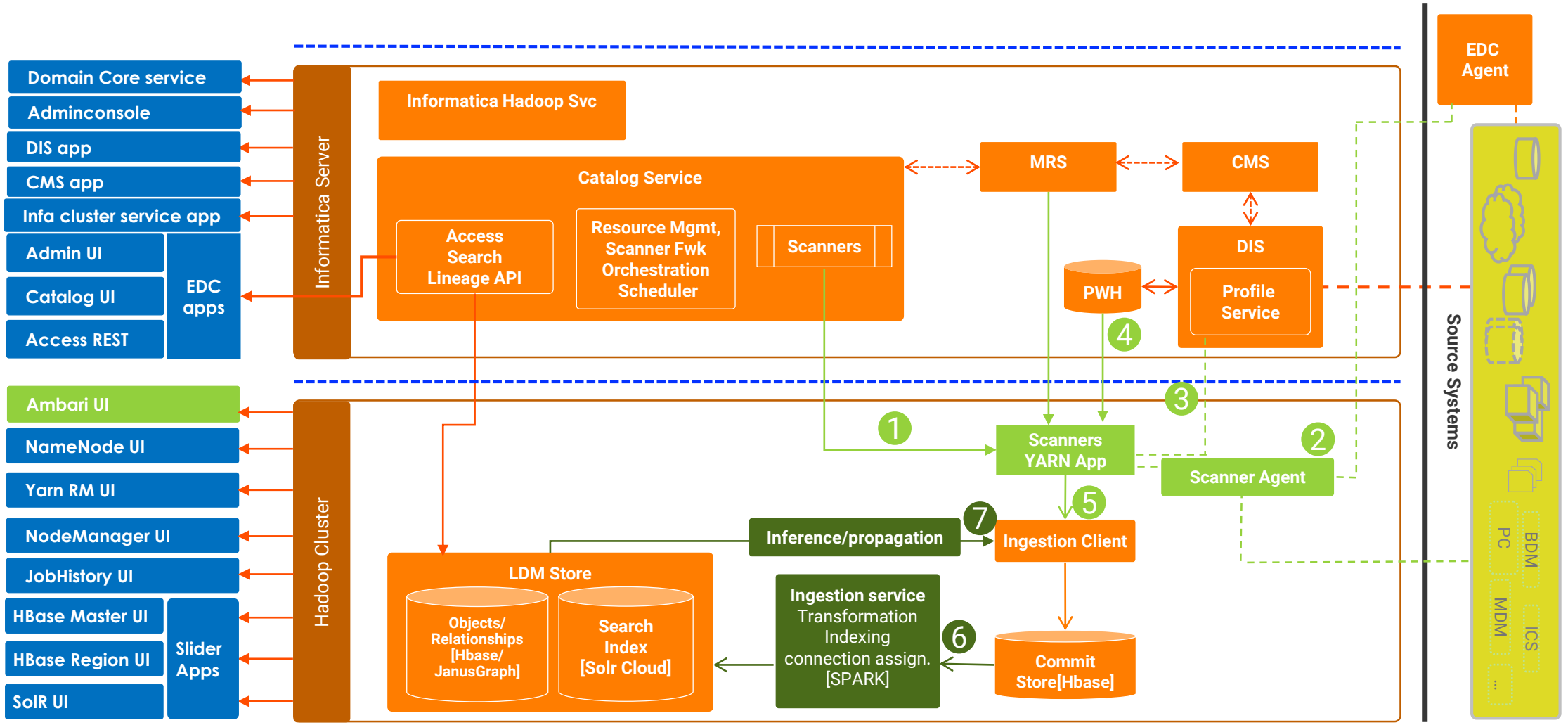
Informatica™

EDC Services Architecture



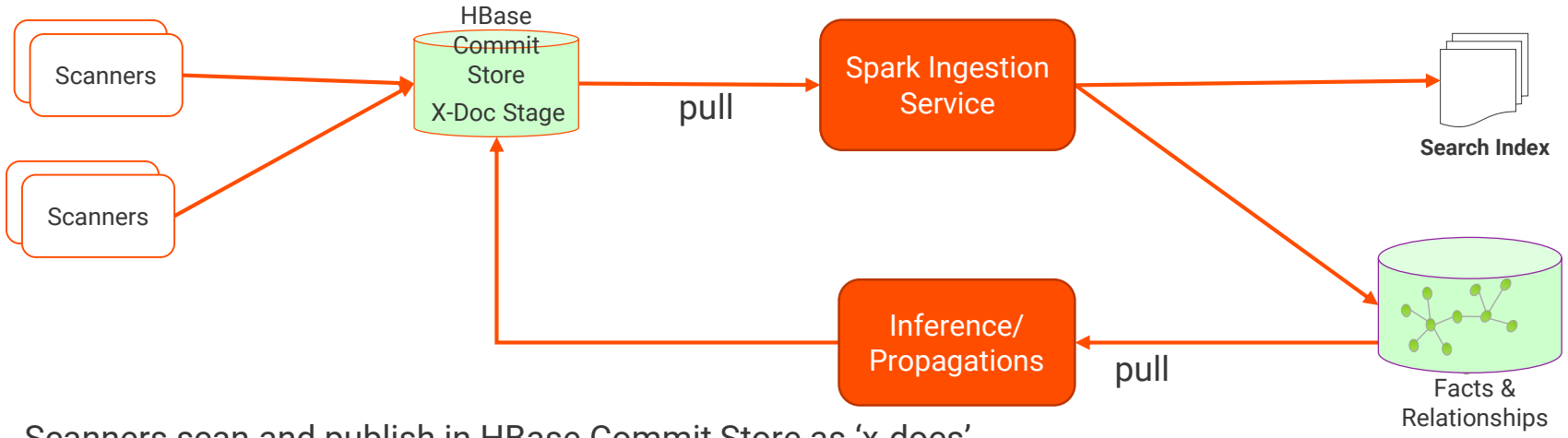
Informatica

EDC Internals – Scanner process



Informatica

EDC Scanner - Ingestion Flow

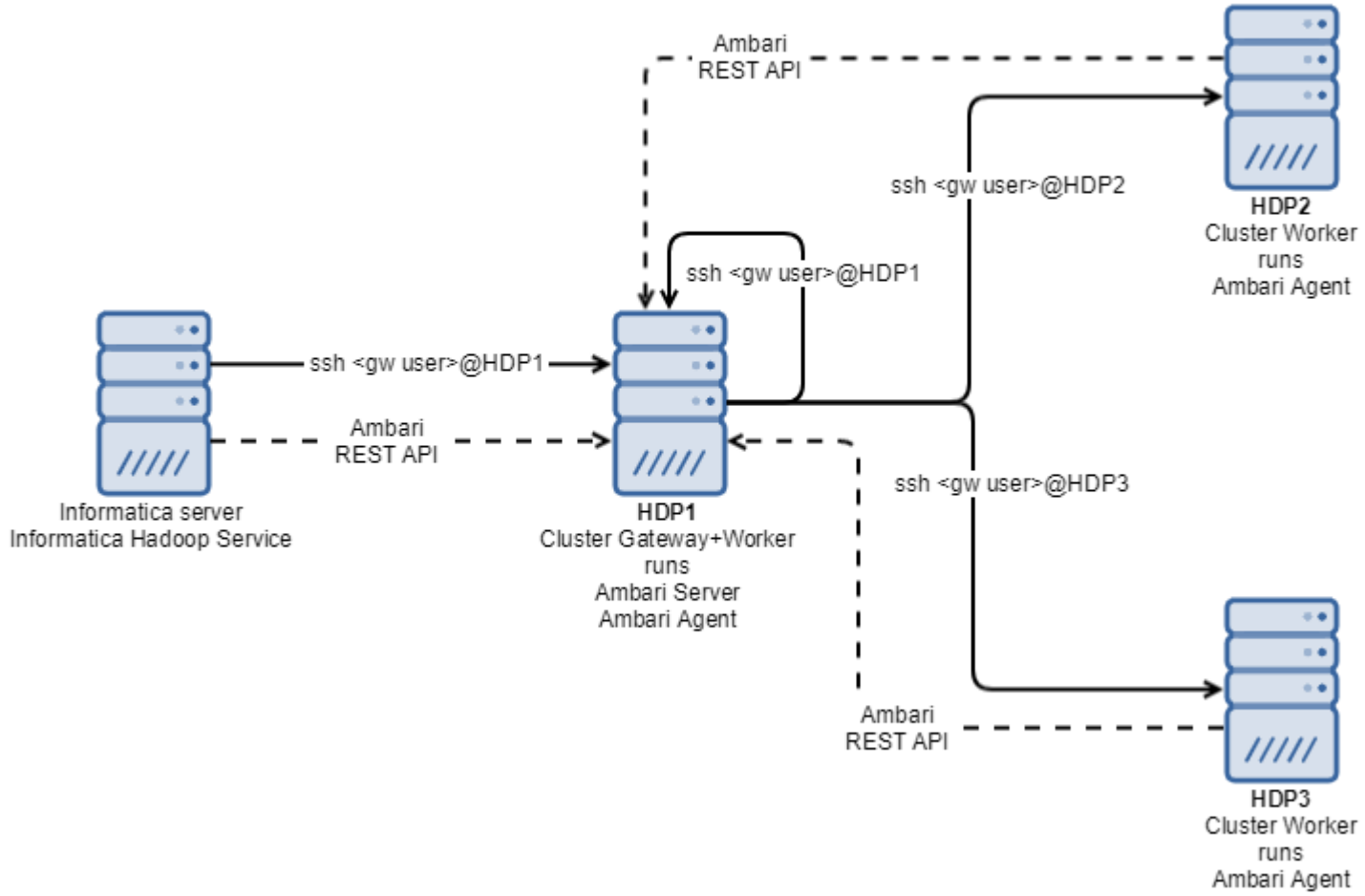


1. Scanners scan and publish in HBase Commit Store as ‘x-docs’
2. Spark Ingestion Service picks a batch of documents and processes them
3. Spark Ingestion Service updates the Graph & search index
4. Propagation/Inference service retrieves facts and infers new facts based on some rules
5. Submits new facts to HBase Commit Store for Spark ingestion service to pickup and process

Embedded Cluster Internals

Deployment

- Informatica Hadoop/Cluster Service issues command to connect to the gateway
- Commands are then issued from the gateway to each node
- In most cases, the gateway also act as a worker.
- Password less ssh is required for installation and runtime
- Sudo privileges are required for installation only

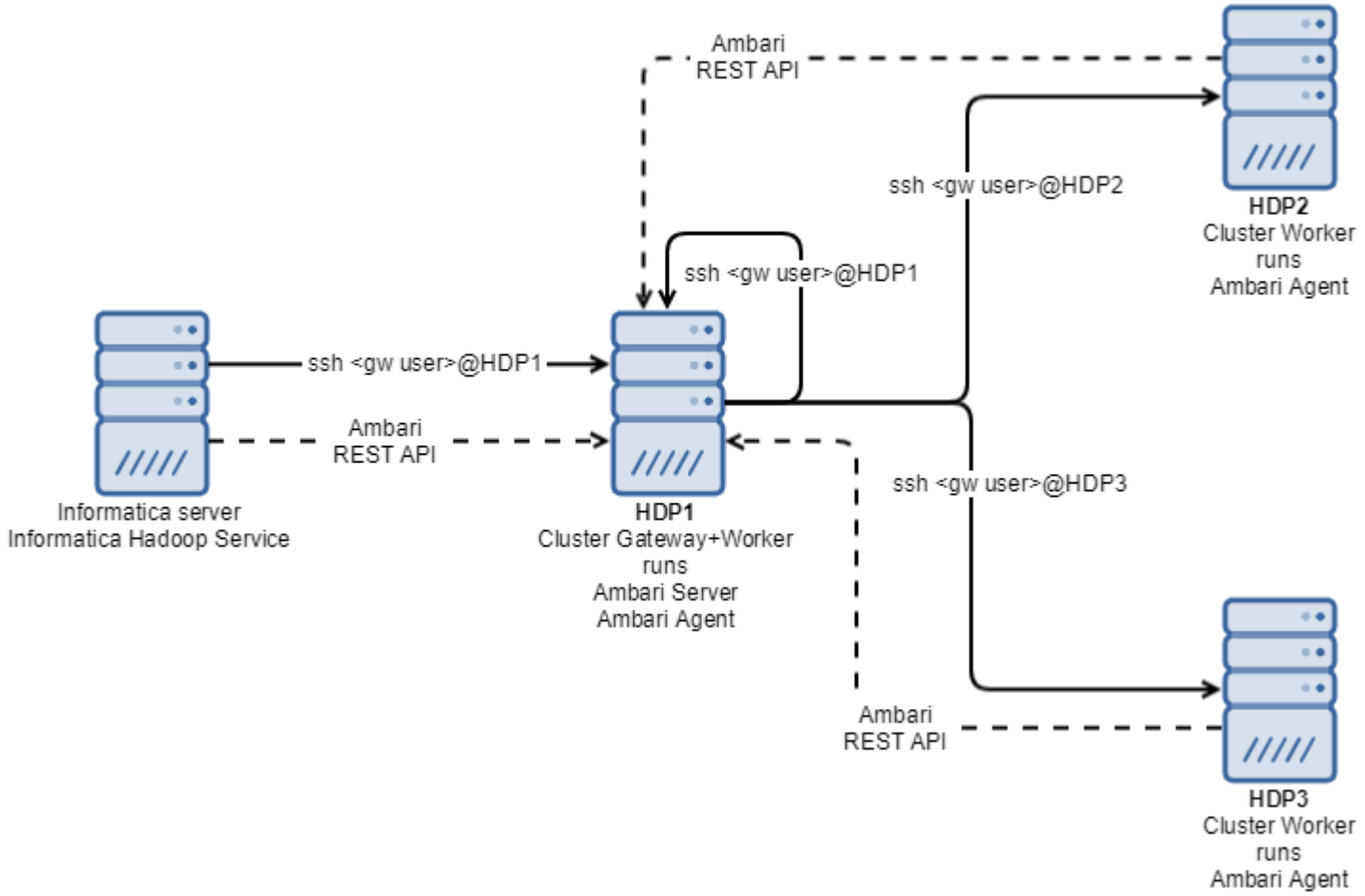


Informatica

Embedded Cluster Internals

Runtime

- At runtime, Informatica cluster service start/stop the Hadoop services using the Ambari REST API
- Cluster service monitor the health of the Hadoop services using the Ambari REST API
- Ambari provide status of the services via the Ambari Metrics service



EDC Security considerations

How to make EDC secured ?

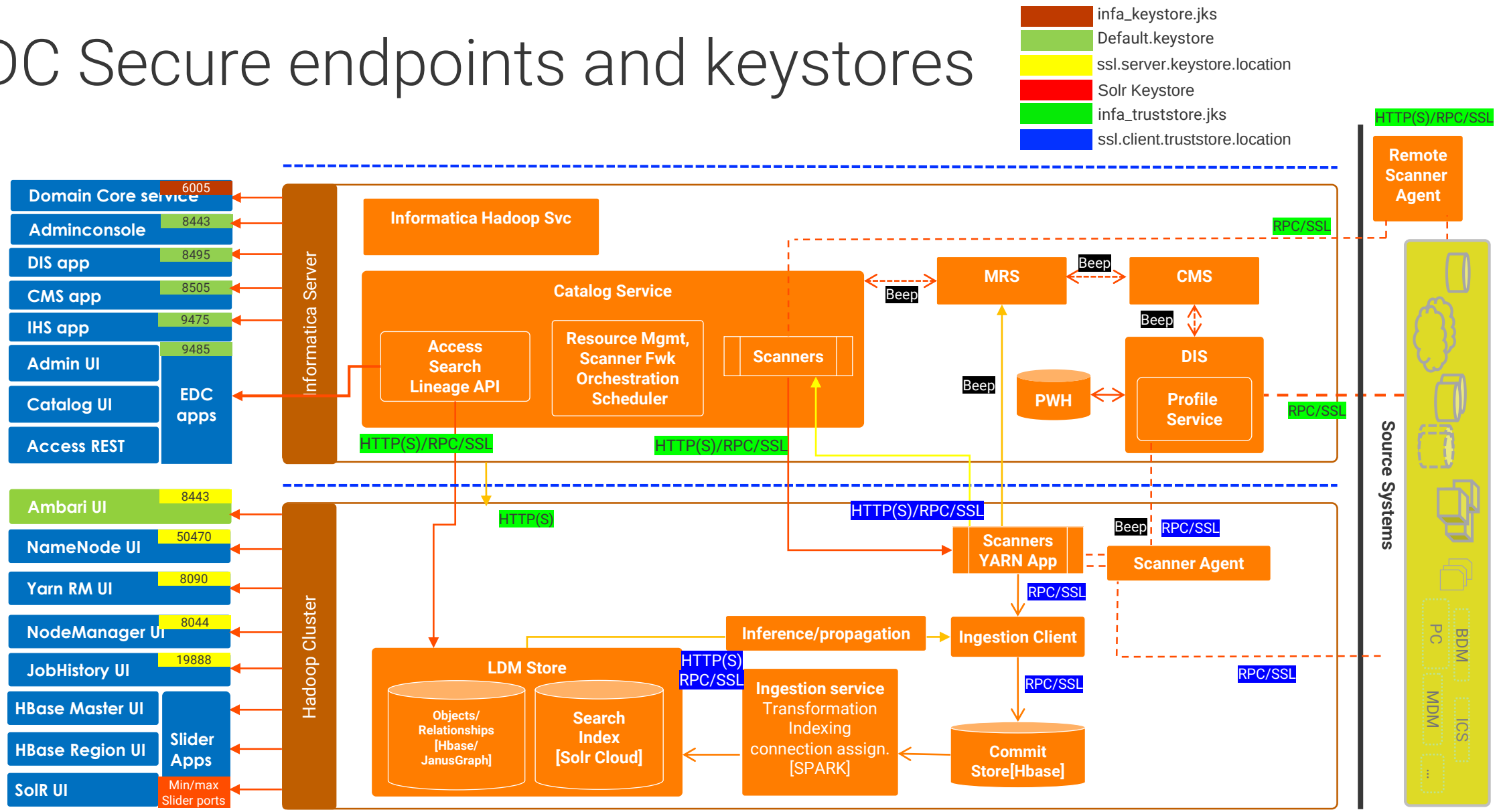
Catalog Administrator

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)
  - Catalog data stored in HDFS is AES-128 encrypted by default.
  - Passwords in scanner configuration encrypted using siteKey provided while domain creation.
  - EDC support Kerberos enabled cluster and SolR access can be restricted thru Kerberos.
- **Application 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

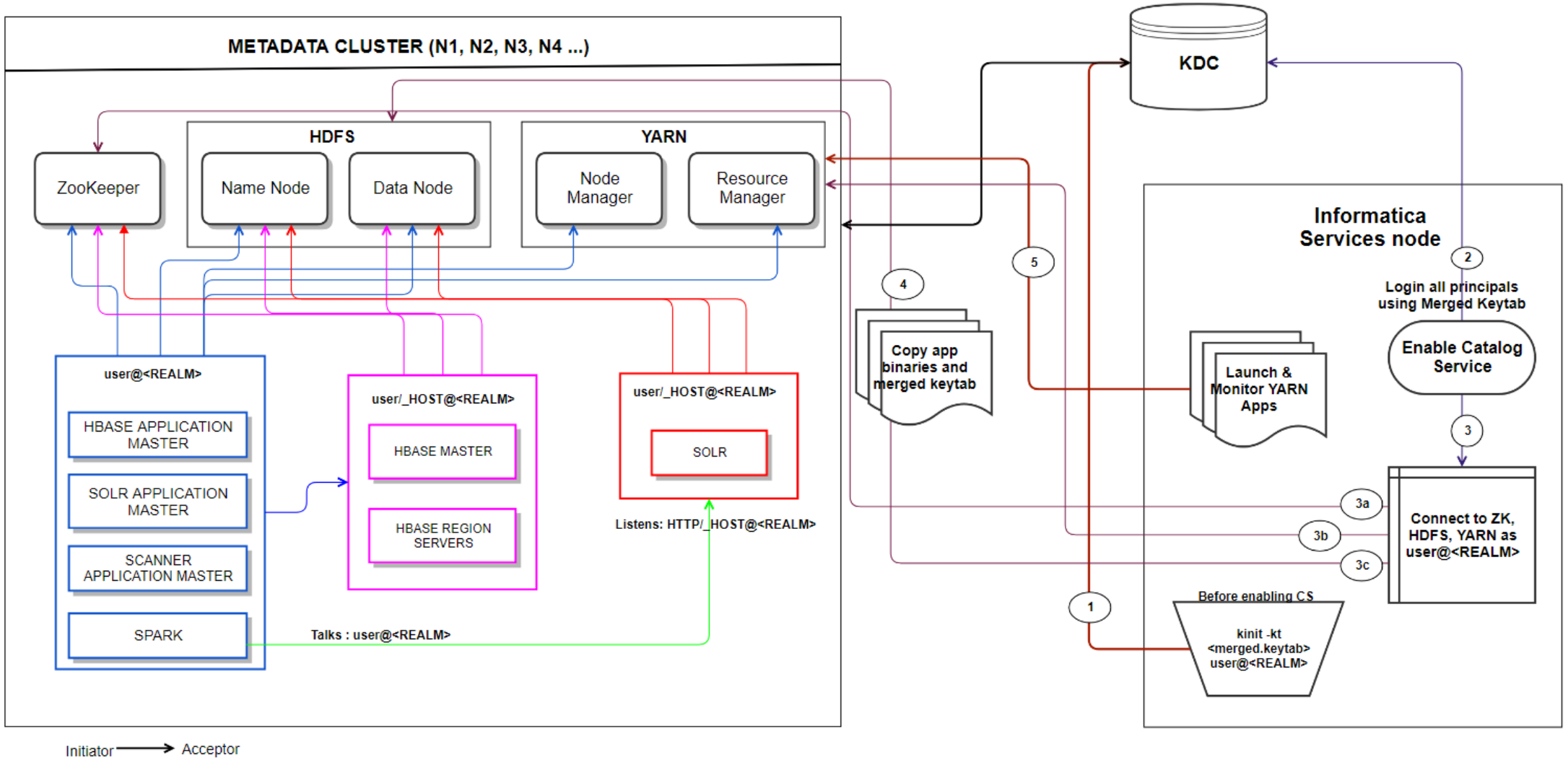
Informatica

EDC Secure endpoints and keystores



Informatica

EDC Security with Kerberos



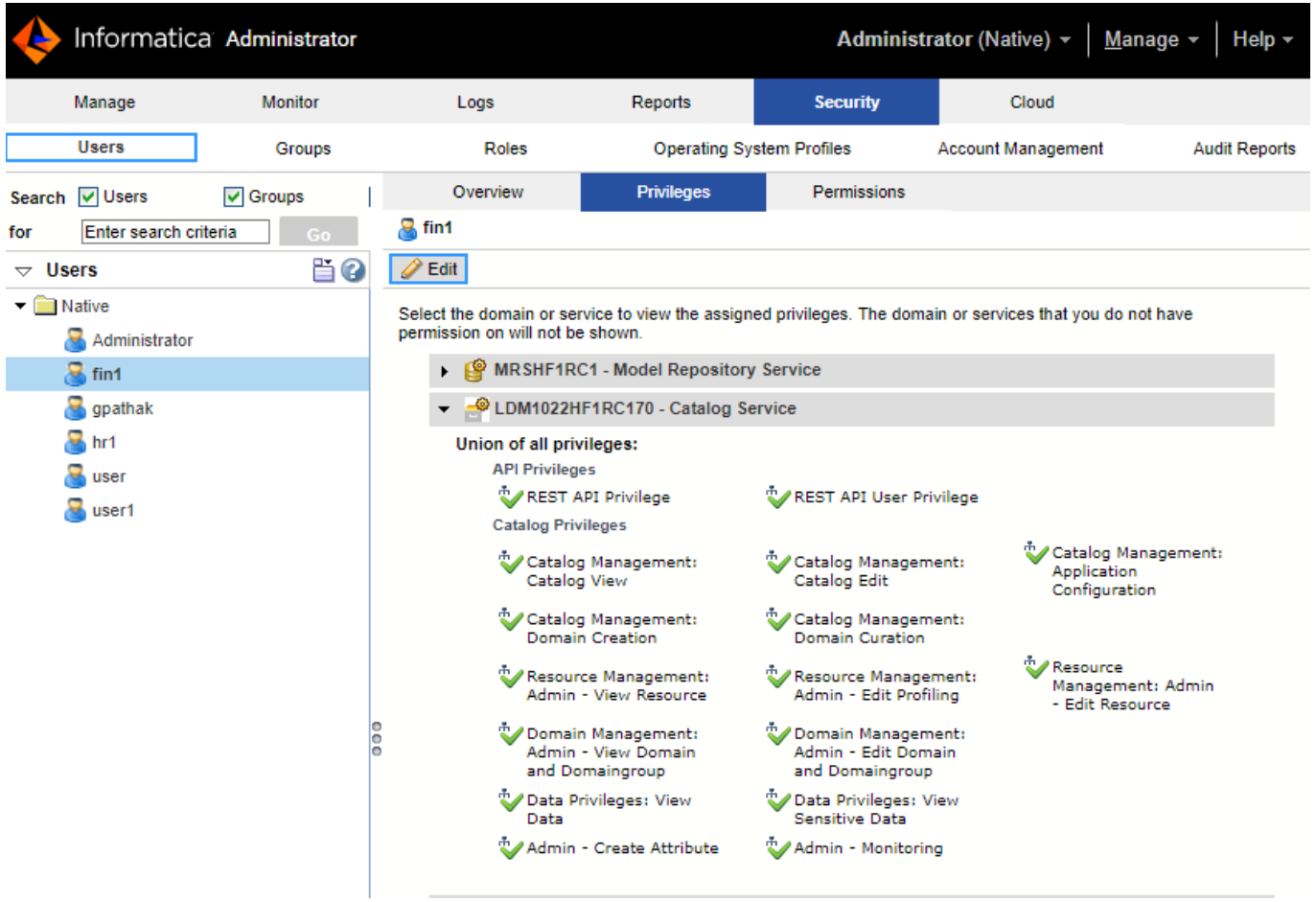
EDC Security behavior with Kerberos

- EDC services can be deployed in Kerberos enabled Hadoop cluster
  - Access to HDFS directories restricted to Service Cluster Name user that you provide.
  - Services Keytab contains credentials for Service Cluster Name user as the Service principal.
  - HBase, Solr, Spark Ingestion services, Scanner jobs run under the Service Cluster Name user on each data node.
  - EDC is not supported on Kerberos Enabled Informatica domain yet.

- EDC can scan Kerberos enabled data sources
  - Scanners Keytab contains credentials to connect to the target applications
  - Must be placed on informatica node (owned by informatica user) and the data nodes (owned by Service Cluster Name user).

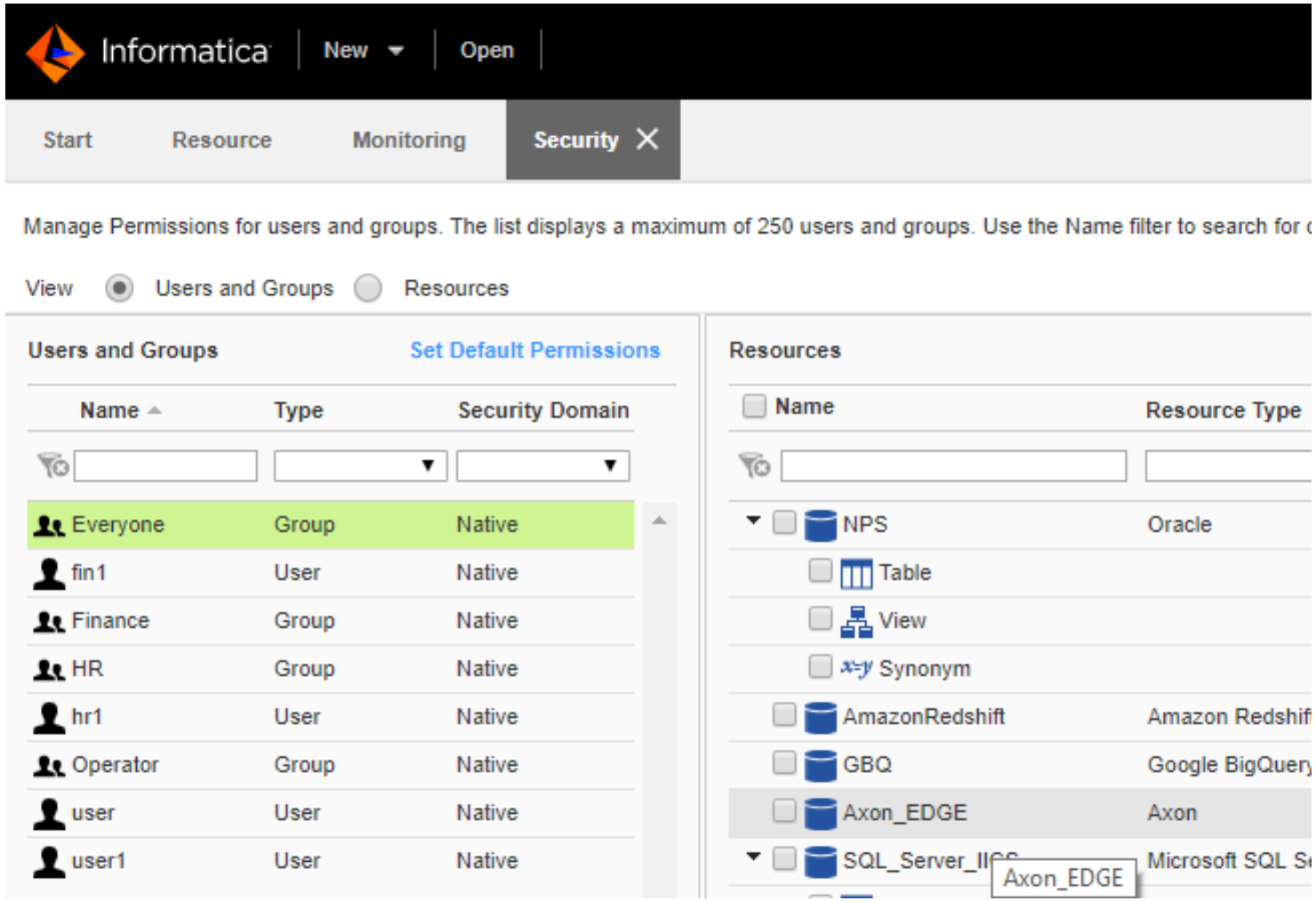
Privileges - Informatica Admin Console

- 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

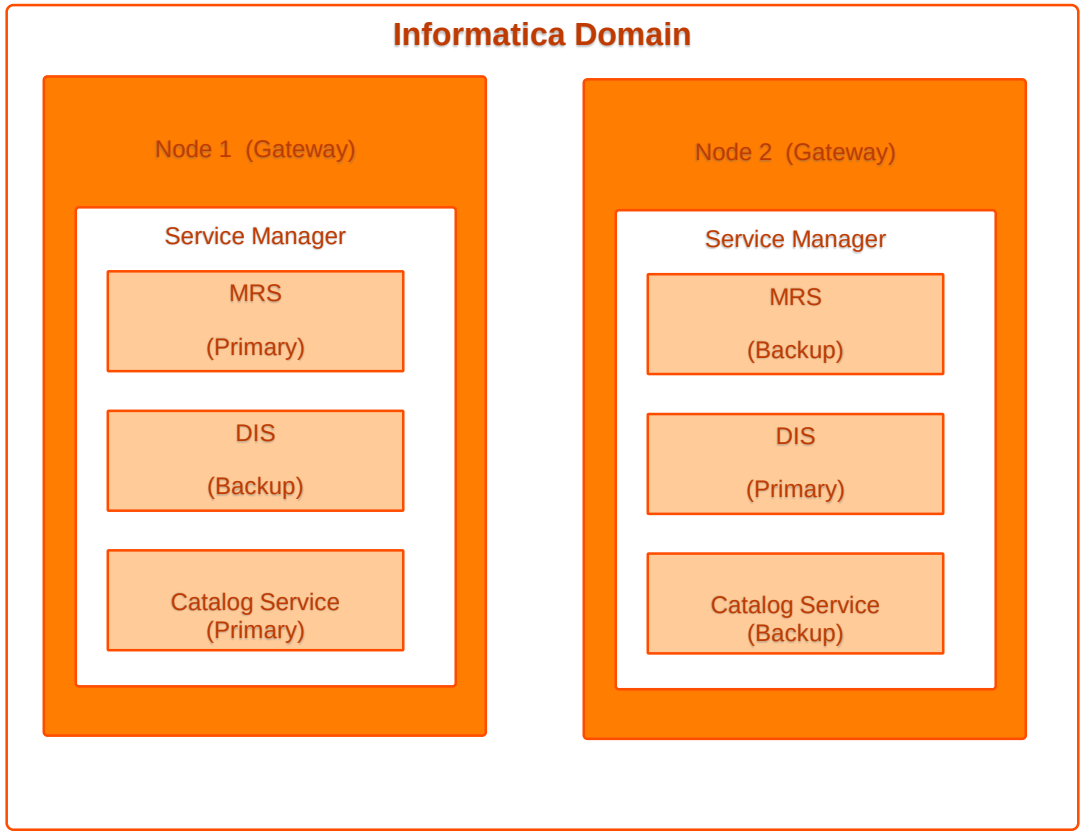
- Permission assigned at resource level
  - Read only
  - Read and Write
  - Metadata and data read
  - All permissions
- Granularity down to the object type for RDBMS only (tables, views, synonyms)




EDC High Availability

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
    - Content Management Service
    - Catalog service
    - Informatica Cluster Service



Embedded Cluster High Availability

- When Informatica Cluster service is deployed on 3 node or more
  - Zookeeper is deployed on all Data nodes
  - HDFS is setup as with Name node HA, replication factor is set to 3 by default.
  - YARN is setup with Resource manager HA
  - If one of the services fail or node goes down, the service application will be restarted on another node by YARN/Slider
- Known limitation: Ambari Server is a single point of failure (SPOF)
  - Ambari server remain non HA as this is not supported by Hortonworks.
  - Informatica Cluster Service relies on Ambari to monitor the Hadoop services
  - If Ambari server or the entire gateway node goes down, the Informatica Cluster service and the Catalog service will go down as well.

Walk-thru EDC Services

Thank You

Sugi Narayana
Principal Technologist – Customer Success

Informatica™

References

- EDC Performance and tuning guide
  - https://kb.informatica.com/h2l/HowTo Library/1/1300-TuningEnterpriseDataCatalogPerformance-H2L.pdf
- Profiling Sizing Guidelines
  - http://kb.informatica.com/h2l/HowTo%20Library/1/0545-Profile-Sizing-Guidelines-H2L.pdf
- Generate and configure custom keystore
  - https://kb.informatica.com/howto/6/Pages/20/511374.aspx
- Configure Kerberos and SSL
  - https://kb.informatica.com/h2l/HowTo%20Library/1/1086-ConfiguringEnterpriseInformationCatalogonaKerberosenabledCluster-H2L.pdf
- Ports configuration for Enterprise Data Catalog
  - https://kb.informatica.com/h2l/HowTo%20Library/1/1071-ConfiguringPortsforEnterpriseInformationCatalog-H2L.pdf
- AWS Marketplace Quick Start templates for EDC Deployment
  - https://aws.amazon.com/marketplace/pp/prodview-qm4jjwykj4yxy?qid=1580925870615&sr=0-1&ref_=srh_res_product_title
- Azure Marketplace for EDC
  - https://azuremarketplace.microsoft.com/en-us/marketplace/apps/informatica.enterprisedatacatalog_10_2_2_hf1?tab=Overview
- EDC Roles and Privileges template
  - https://kb.informatica.com/howto/6/Pages/23/616459.aspx