February 23, 2021

# Deployment Best Practices of Data Governance Products – Axon, EDC, IDQ, DPM

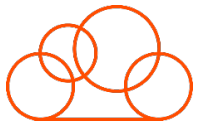
Srinivasa Gopal, Principal Customer Success Technologist

Informatica™

# Housekeeping Tips

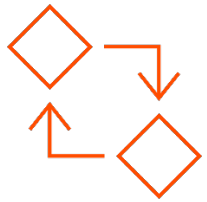
- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available 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.

# Feature Rich Success Portal



Bootstrap trial and POC Customers

Enriched Customer Onboarding experience



Product Learning Paths and Weekly Expert Sessions

Informatica Concierge with Chatbot integrations

Tailored training and content recommendations



Informatica

# More Information

## Success Portal

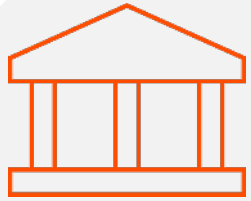
https://success.informatica.com

## 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.

# Agenda

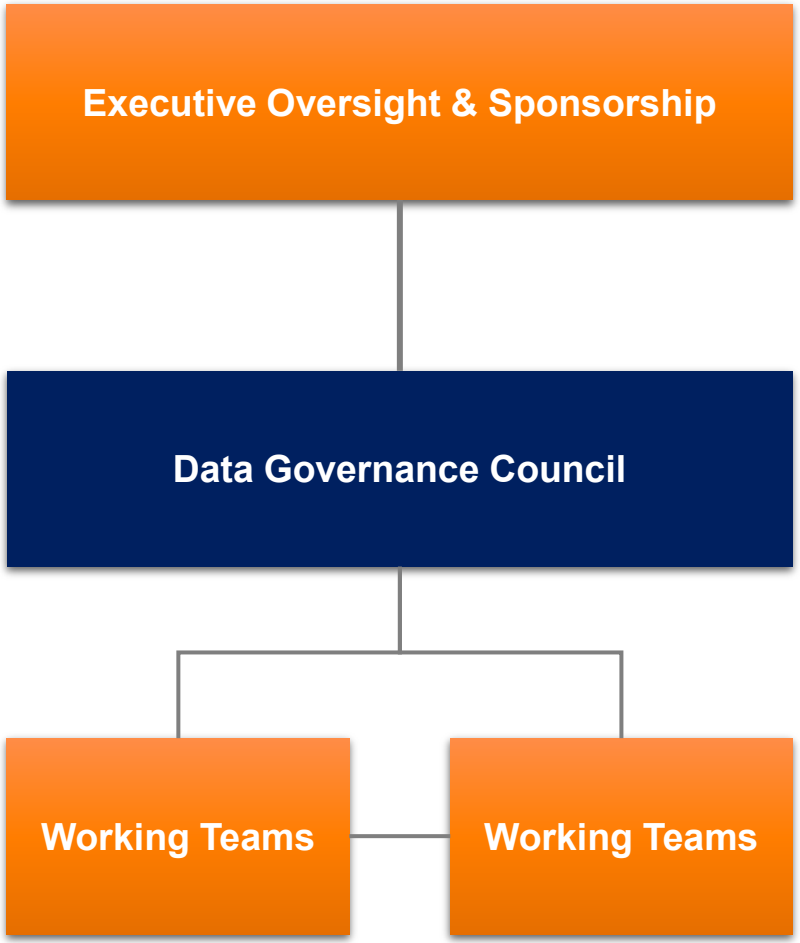
- DG Program Building Blocks
- Platform Architecture – On Premise/Cloud
- Roles and Responsibilities
- Sizing Guideline
- Q&A

# DG Program – Building Blocks

# DG Program Building Blocks

- DG ‘Best Practices’
  - Organization Framework
  - Defined Roles
  - DG Program Charter & Guiding Principles
  - Business Opportunity Prioritization Process
  - Structured Workflows
  - CDE identification process
  - Ongoing maturity assessment process
  - Implementation Roadmap
  - Program Metrics
  - Communications

# DG Organization Framework – Agile and Flexible

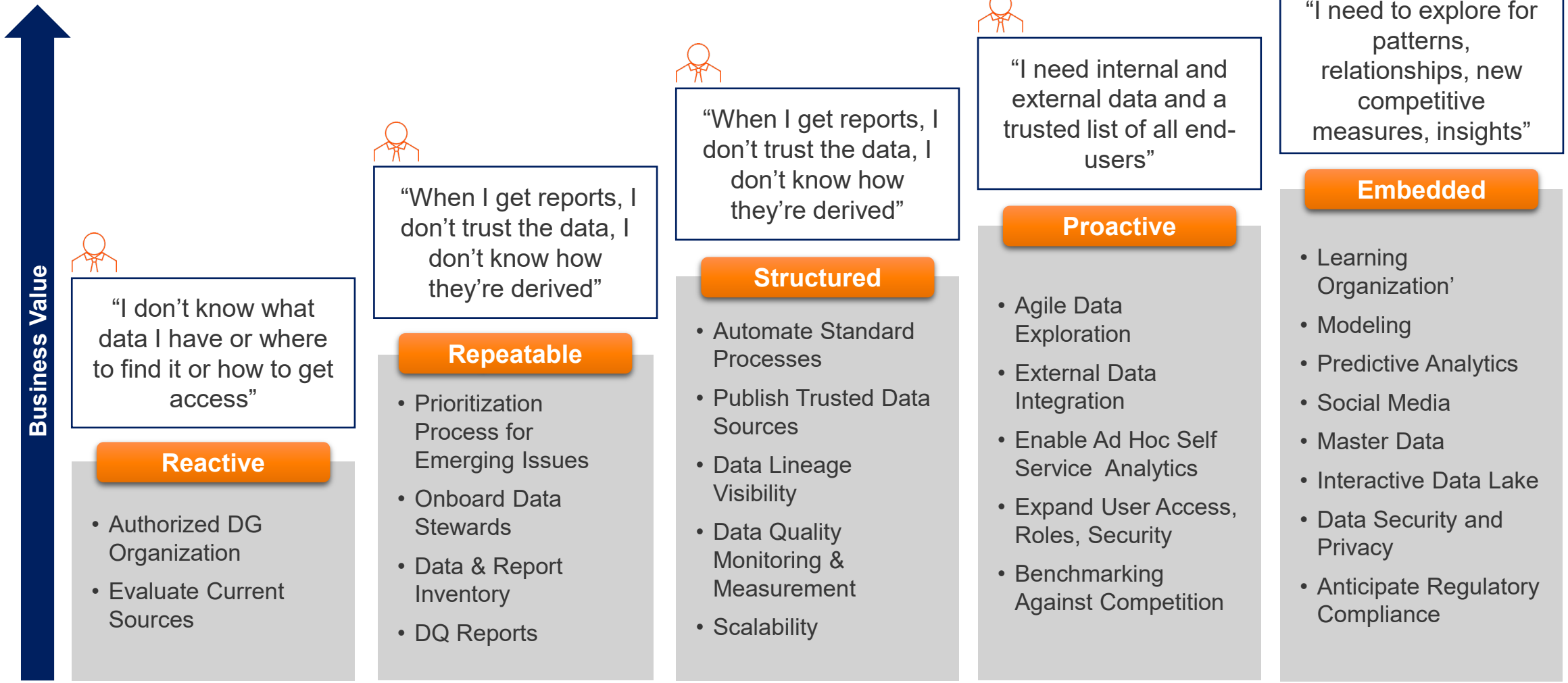



- Align DG with business strategy
- Develop DG charter and guiding principles
- Set program expectations
- Authorize policies and standards
- Monitor program metrics
- Communicate progress

- Collaborative business leadership
- Prioritize data issues
- Approve standards and policies
- Establish program metrics
- Support the working teams
- Enforce compliance

- Data Stewards
- Collaborative business & I.S. SMEs
- Conduct root cause analysis Propose remediation options
- Draft solutions, standards, policies

# 5-Stage Maturity Model

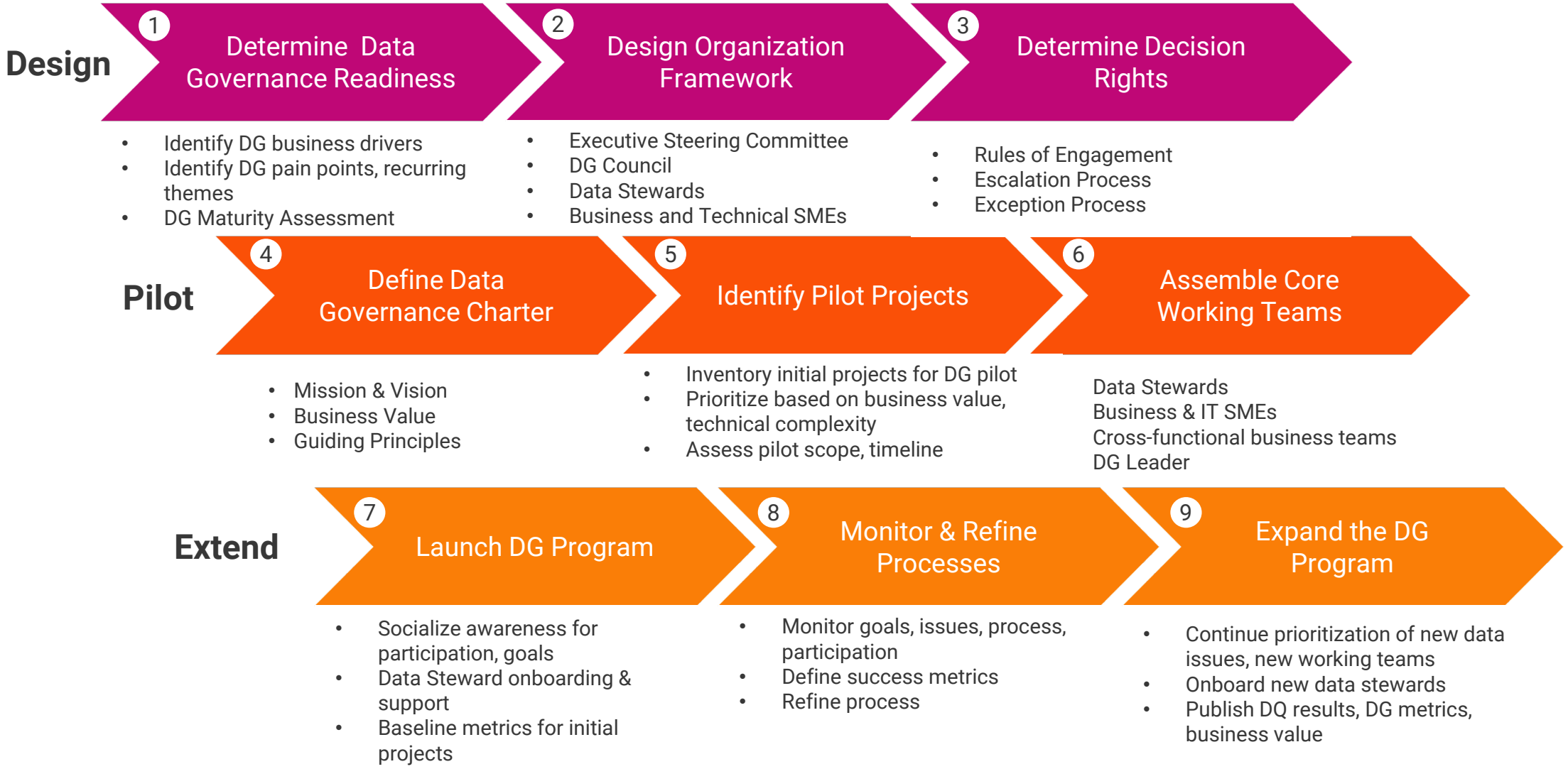






Informatica

# Data Governance Program Development

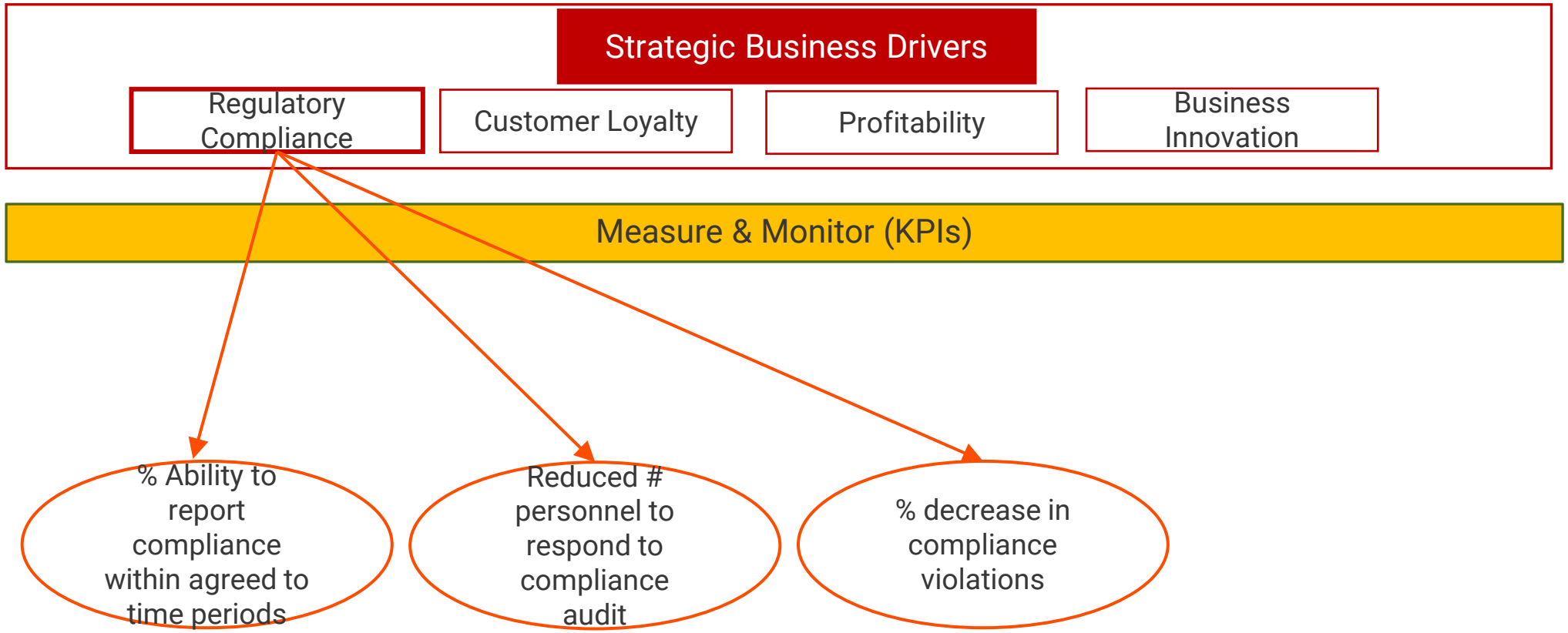






Informatica

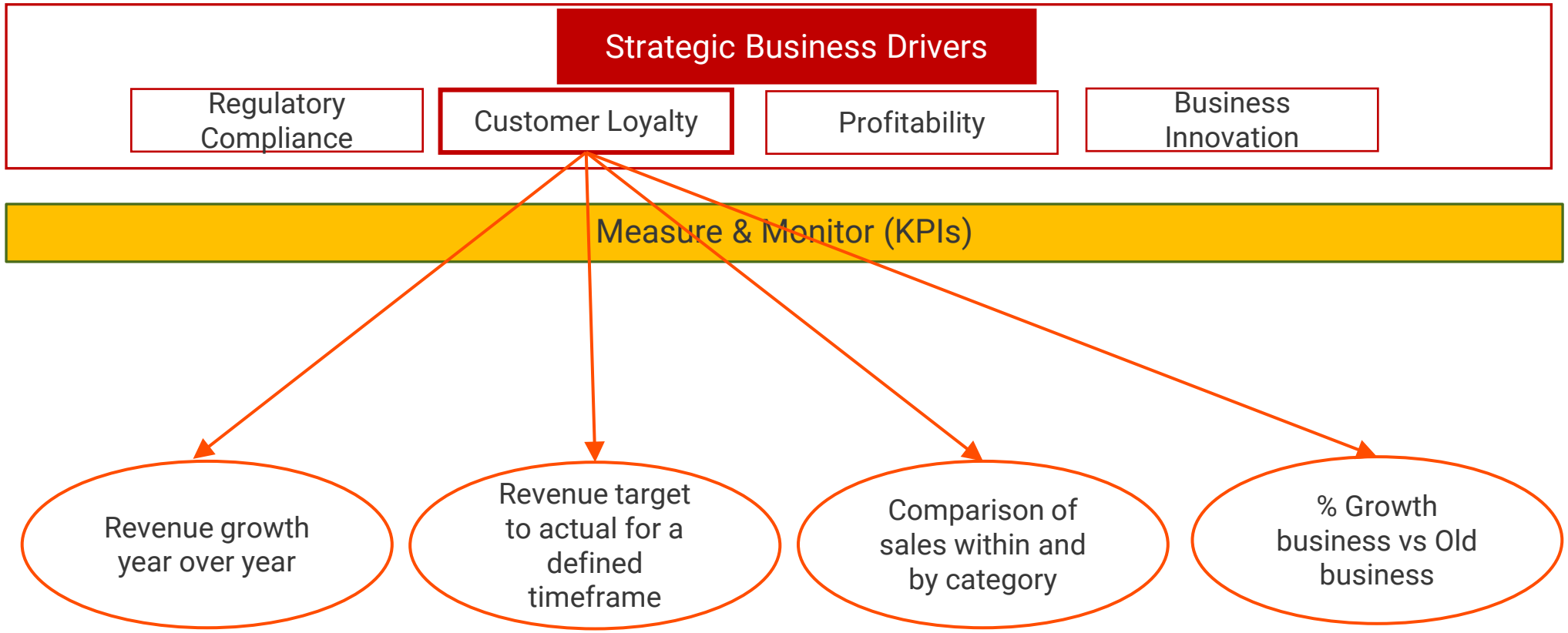
# Identifying KPI's to Measure the Success of the Program






Informatica

# Identifying KPI's to Measure the Success of the Program cont.





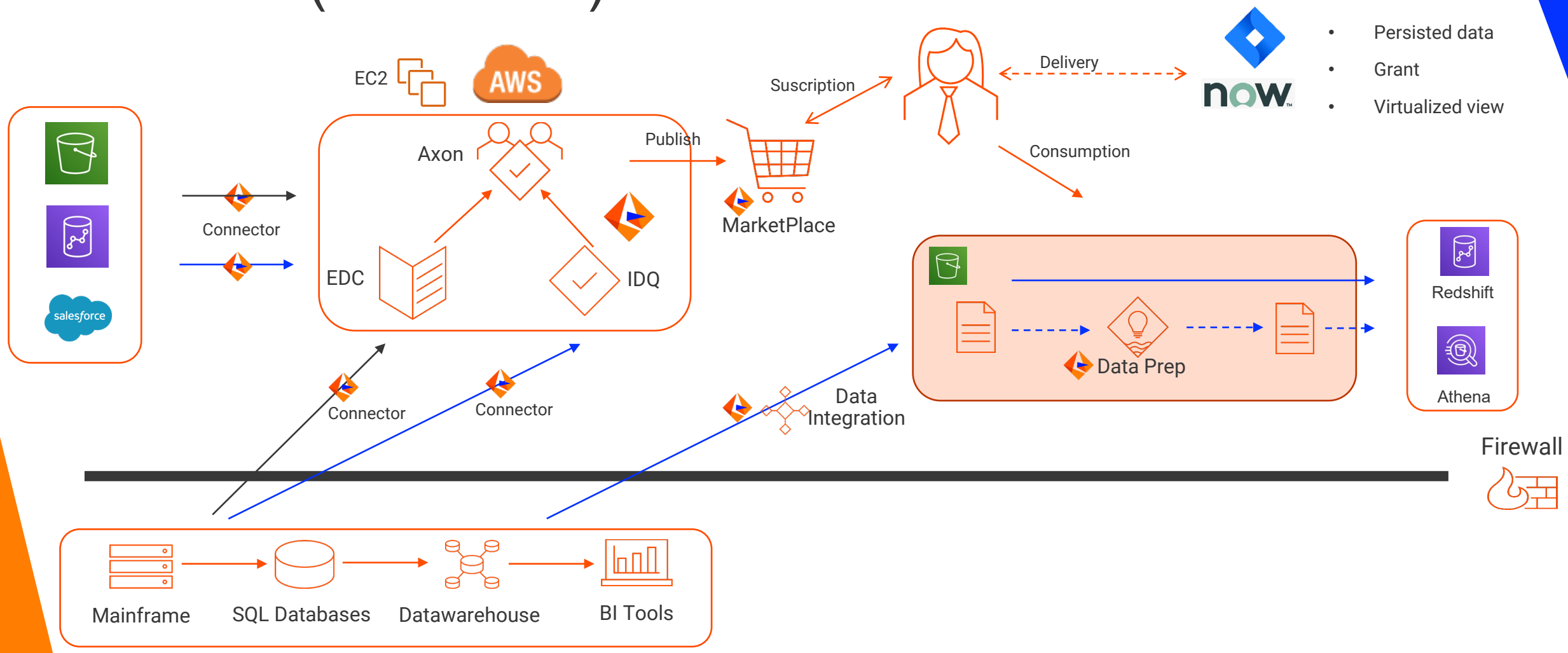

Informatica

# Full List of Initial Measurements and KPI's

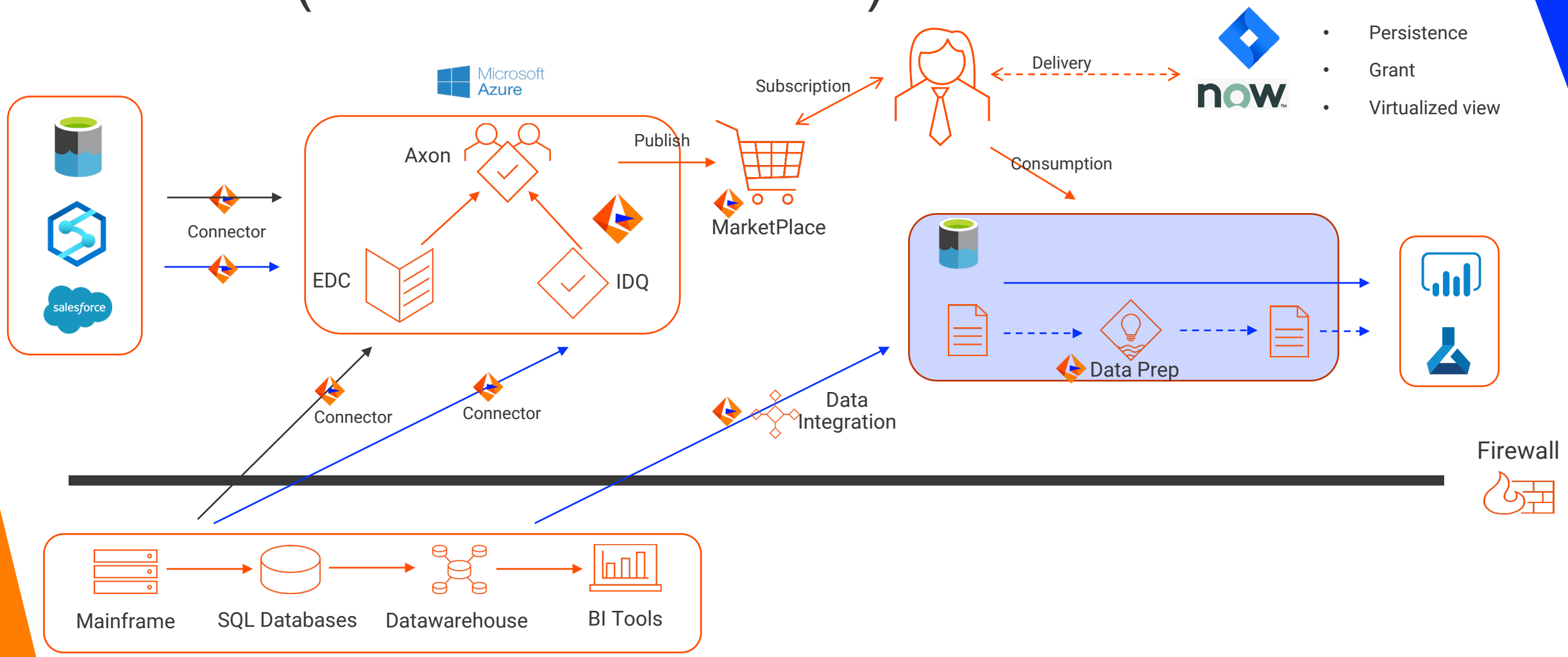
| Data Governance Framework Area | Componet | Measurement/ KPI |
|---|---|---|
| **Strategic Business Drivers** | Business Development | Revenue growth year over year from new Products |
| | Business Innovation | % Growth business vs Old business |
| | Digital Transformation | Number of lines of business, functional areas, IT and governance with committed Stakeholders |
| | Regulatory Compliance | % Ability to report compliance within agreed to time periods |
| | | Reduced # personnel to respond to compliance audit |
| | | % decrease in compliance violations |
| | Risk Management | Appropriate secure treatment of PII, PCI, PHI elements that are covered under regulatory bodies |
| | | Appropriate treatment and accessibility of industry vertical specific elements (e.g. SEC, EPA, OSHA) |
| | | Number and accuracy of security classifications assigned |
| | | Speed and quality of audit responses and eDiscovery requests; Cost of any fines or sanctions |
| | | Ability to comply with CCPA, GDPR, state privacy regulations; number of fines, sanctions |
| | | Number and scope of security breaches |
| | | |
| **Governance Outcomes** | Data Democratization | % of time saved thru. self-service analytics |
| | Data & Process Standardization | %/# of consistent definitions, data elements. CDE's |
| | Improved Business Planning | % Time Reduction to plann for a new business capability |
| | Improved Data Quality | % Complete, Accurate, Valid Customer Contacts |
| | Metadata Documentation | % of databases, integrations and reports captured |
| | | |
| **People** | DG Organization Structure | Org. Framework Documented |
| | Communication | A comminication model and plan documented |
| | Role Formalization | Well defined roles created with individuals identifed to fill |
| | Stewardship | Stewards Identified |
| | | Role Descriptions Approved |
| | | People assigned ownership to data (e.g. Glossary, Data Set) |
| | | |
| **Process** | Change Control | Processes created with workflows to support change |
| | Measure & Monitor KPI's | # of measurements created montotor program |
| | Policy Development | # New/ Existing policies |
| | Process Mapping | % of Policies that have a defined process to support |
| | Workflow Management | # Workflows created to sopport comminication in processes |
| | | |
| Technology | Access Control | |
| | Data Catalog | # of Terms defined |
| | | % of Terms assinged Owners and Stewards |
| | | Number of Glossary elements, Systems, Datasets |
| | Data Discovery | # systems scanned and data mapped to domains |
| | | % of scaned tabled and columns mapped to domains |
| | Data Integration | # systems included in integration detail |
| | Data Lineage | # applications included in data lineage mapping |

# Platform Architecture – AWS/Azure

# Data Flow (with AWS)



# Data Flow (with Microsoft Azure)




Metadata & Data

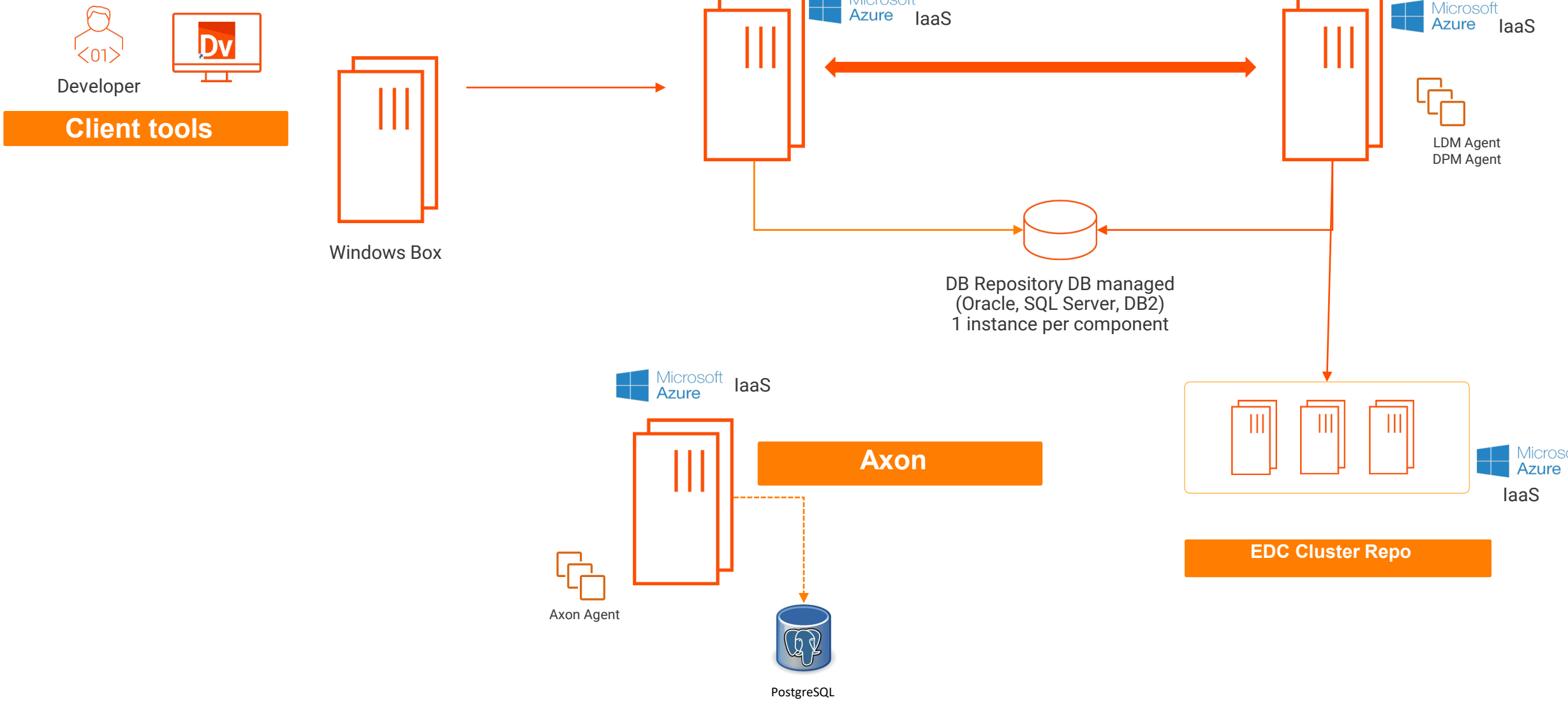
Data



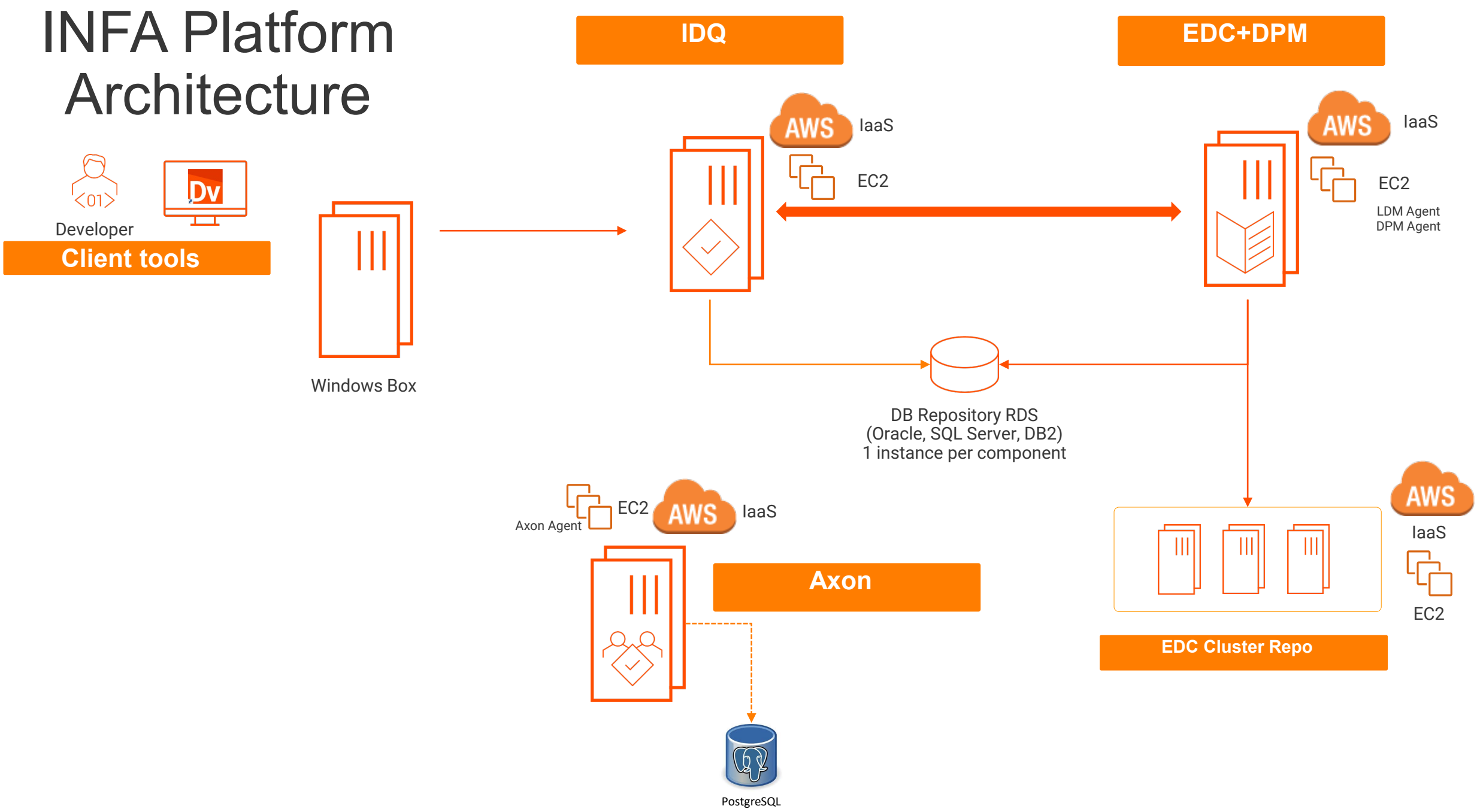


Informatica

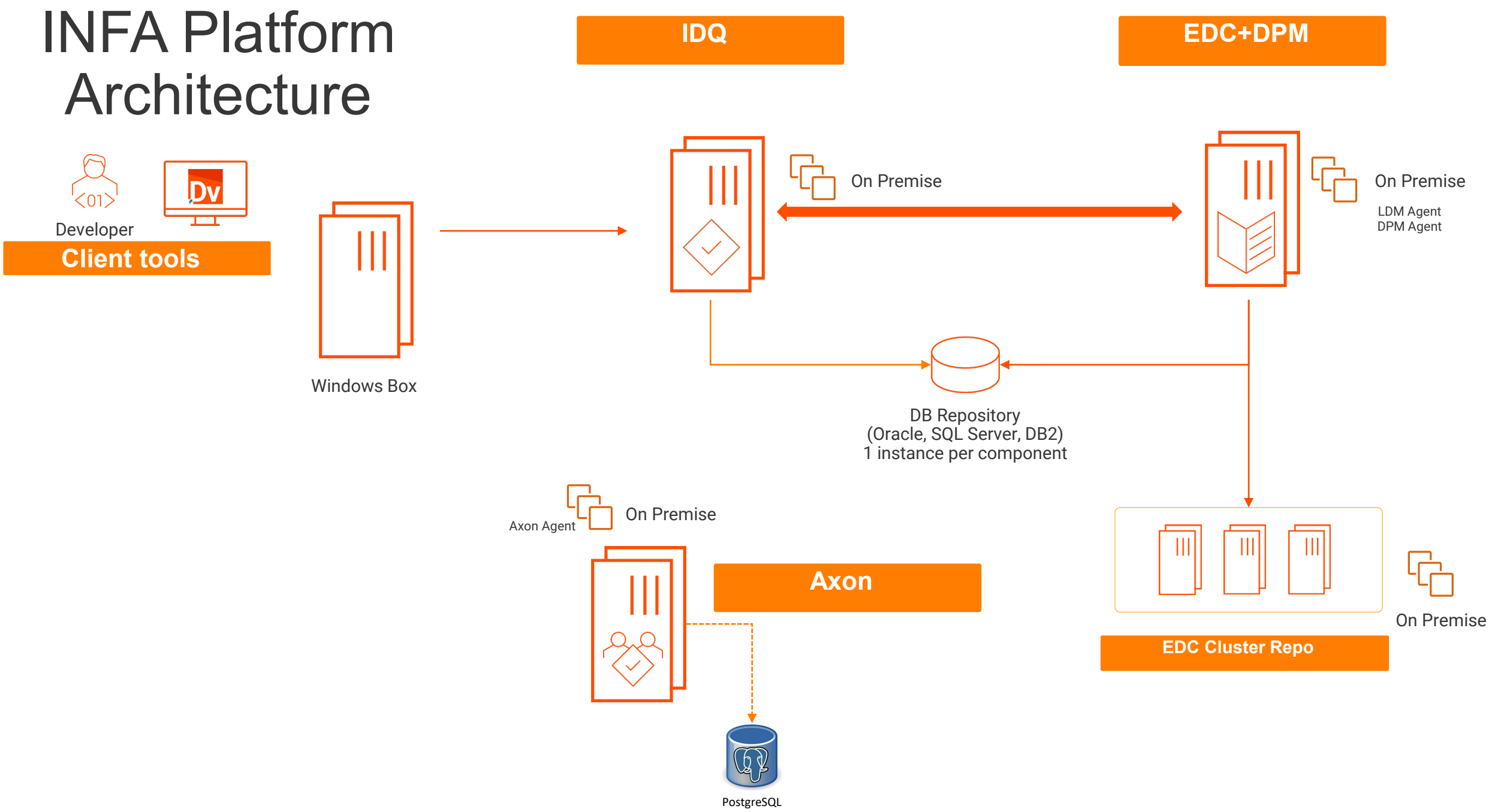
# INFA Platform Architecture



# INFA Platform Architecture



# INFA Platform Architecture

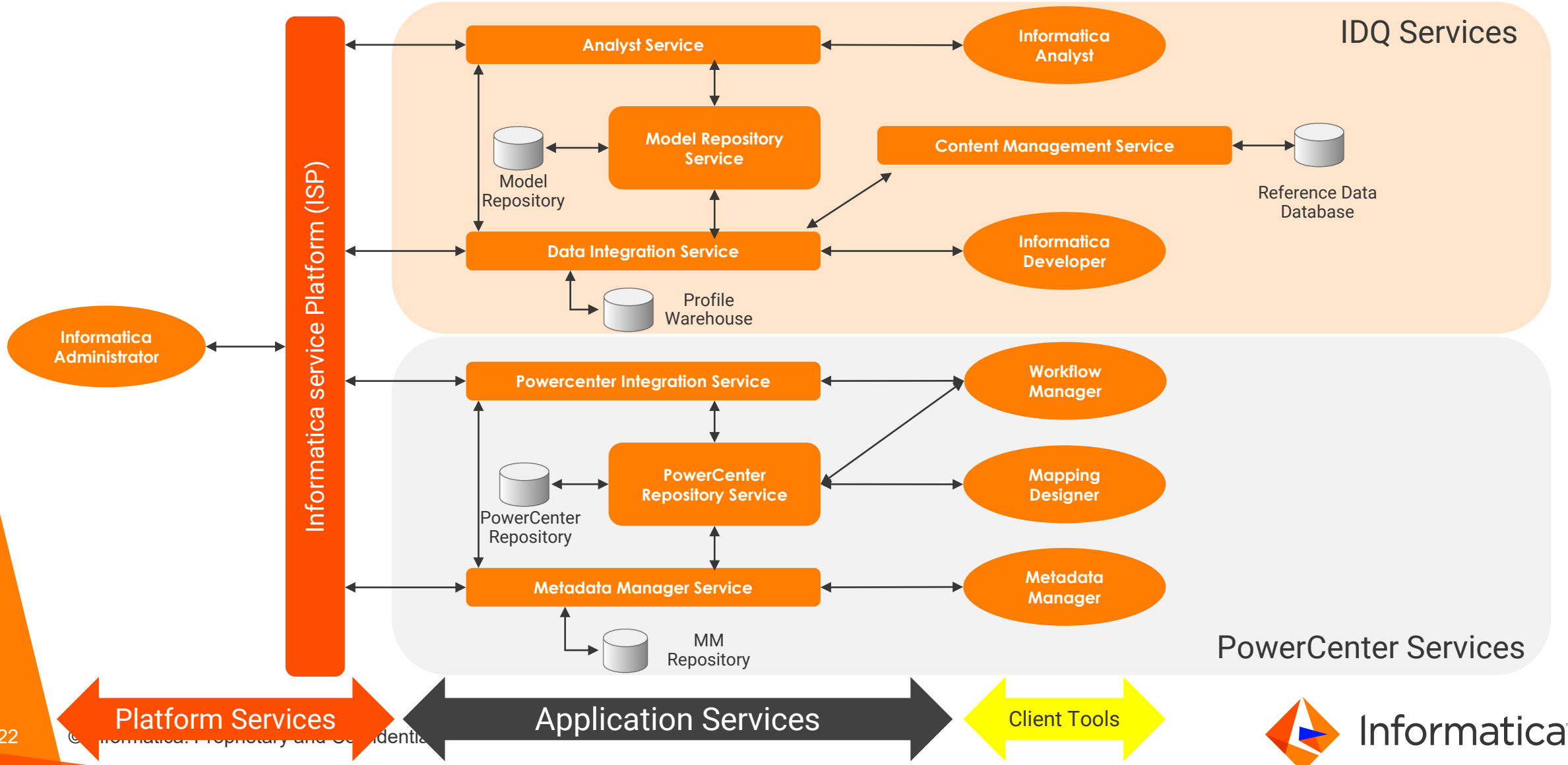


# INFA Split Domain: EDC and IDQ

**Recommendation and Best Practice for EDC and IDQ to be installed in separate Domain, here are pointers:**

- Flexibility applying patches, fixes, upgrades for respective product
- IDQ is higher volume (longer running jobs-less jobs-more operational driven)
- EDC is Metadata (more jobs-less operational driven)
- IDQ licensing is based on number of cores in the machine, whereas EDC licensing is based on number of Resources
- Profiling: Context of Profiling in EDC is for Data Domain Discovery,  Similarity Discovery, Unique Key Inference, CLAIRE on larger set of data, however context of Profiling on IDQ is to perform checks on Data Quality Rules, Scorecards focused on key sets of data.

# Informatica Data Quality - Architecture



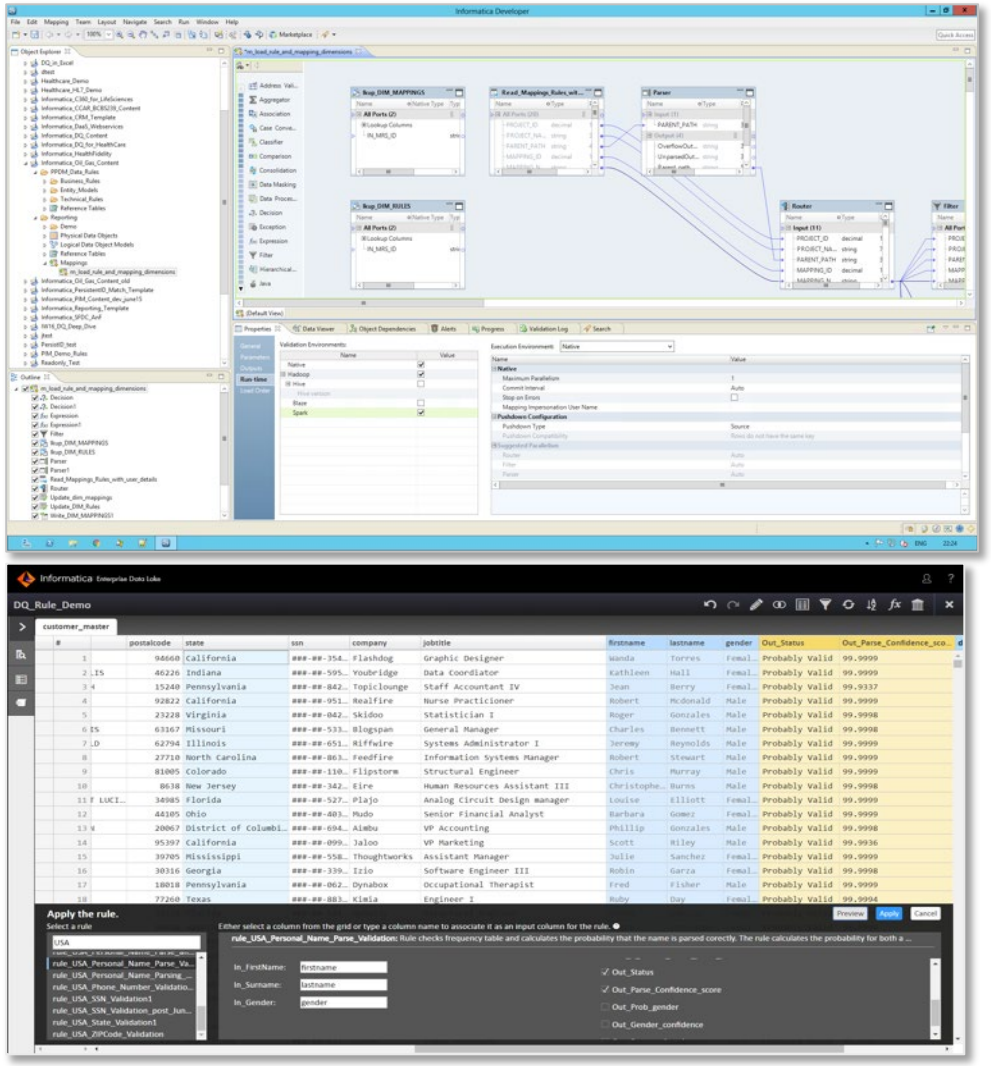




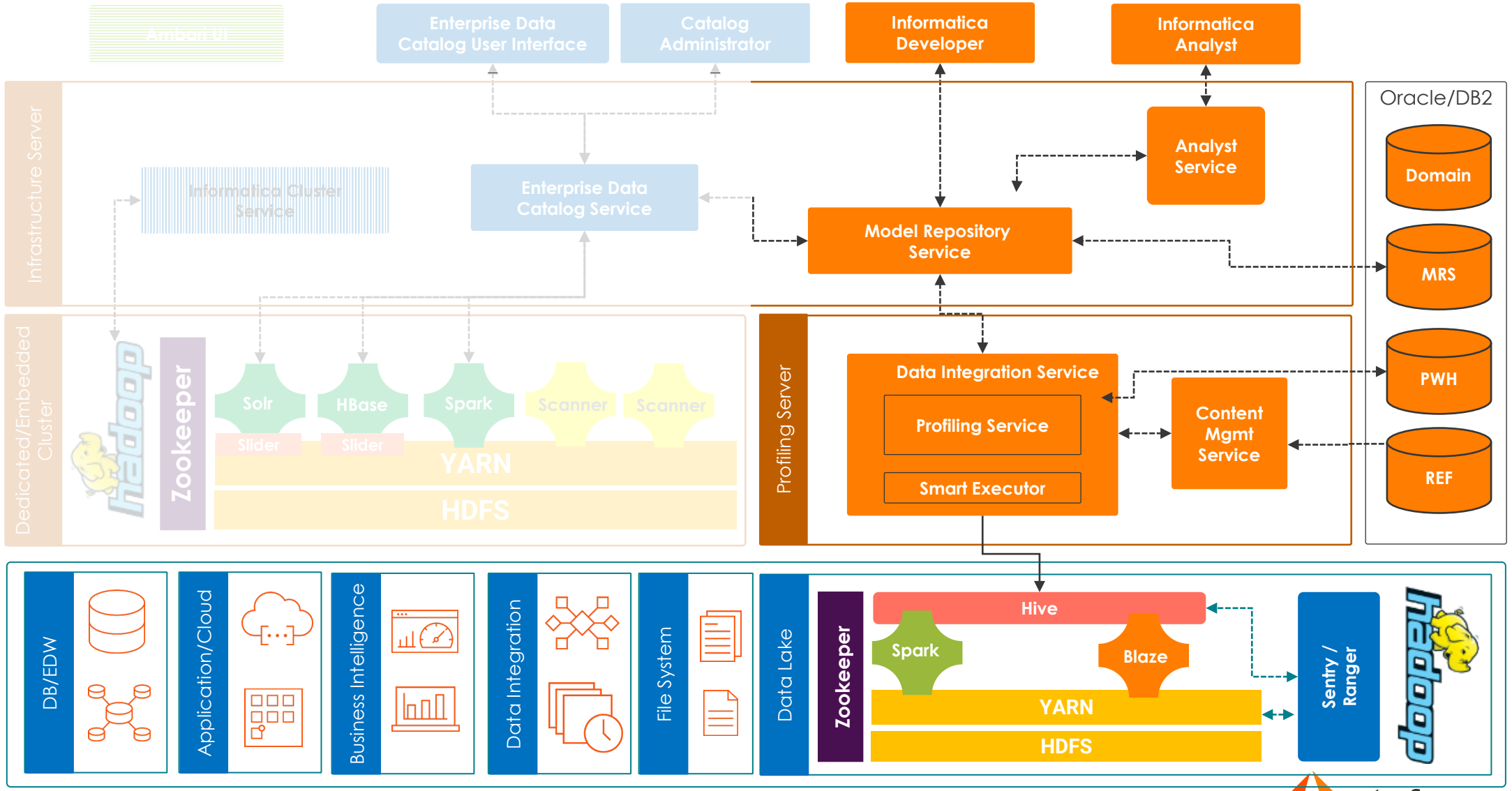

Informatica

# Data Engineering Quality (Pushdown DQ)

- Apply Data Quality Rules within Hadoop Data Flows and pushdown to cluster for execution in Blaze (Informatica), Spark and Databricks
- Integrate across Big Data Use Cases
  - **Data Lakes** – Ingestion and management of all data and its quality on a data lake
  - **Data Preparation** – Self-service data preparation using Enterprise standards and Data Quality Rules directly in data science activities
  - **Data Catalog** – Support custom domain discovery based on Data Quality Standards
  - **Data Streaming** – Apply business rules within data streams from IoT sensors, mobile devices and clickstream interactions



# Data Engineering Quality - Architecture



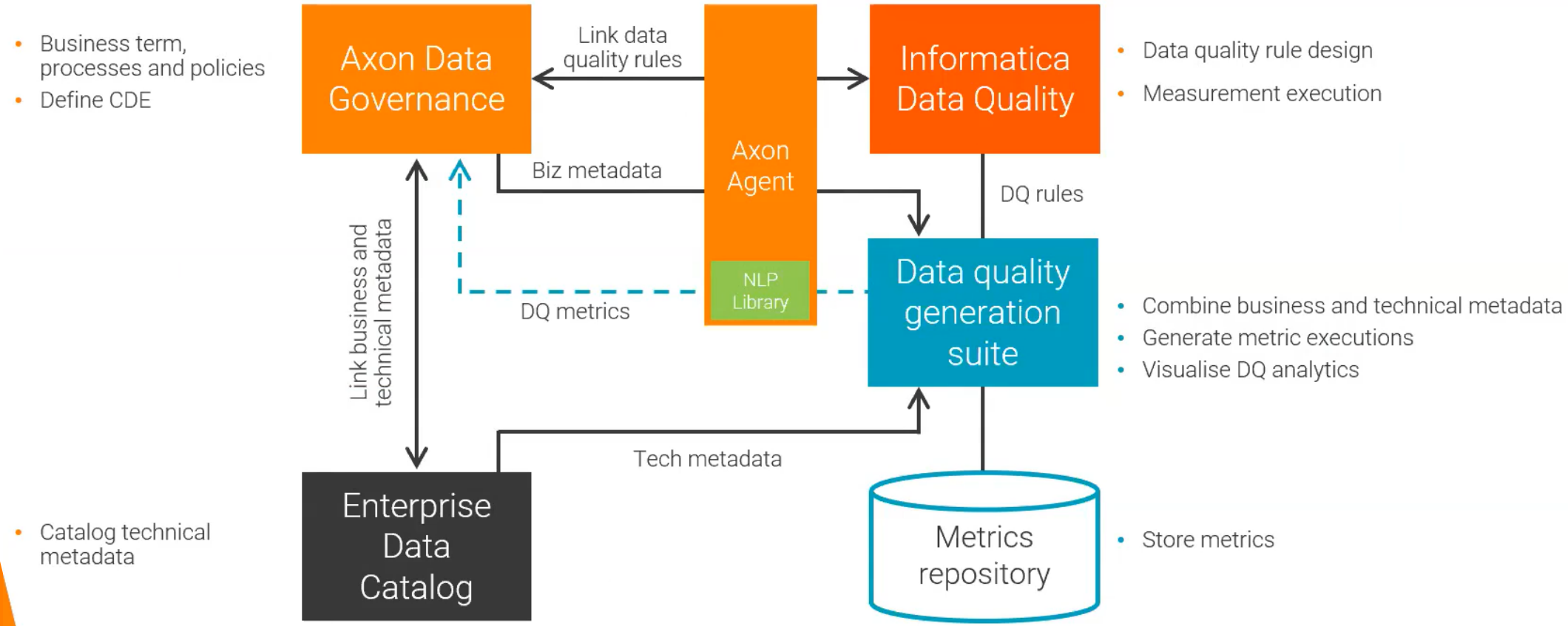





Informatica

# Automation Quality for Data Governance

## End to End Process Automation

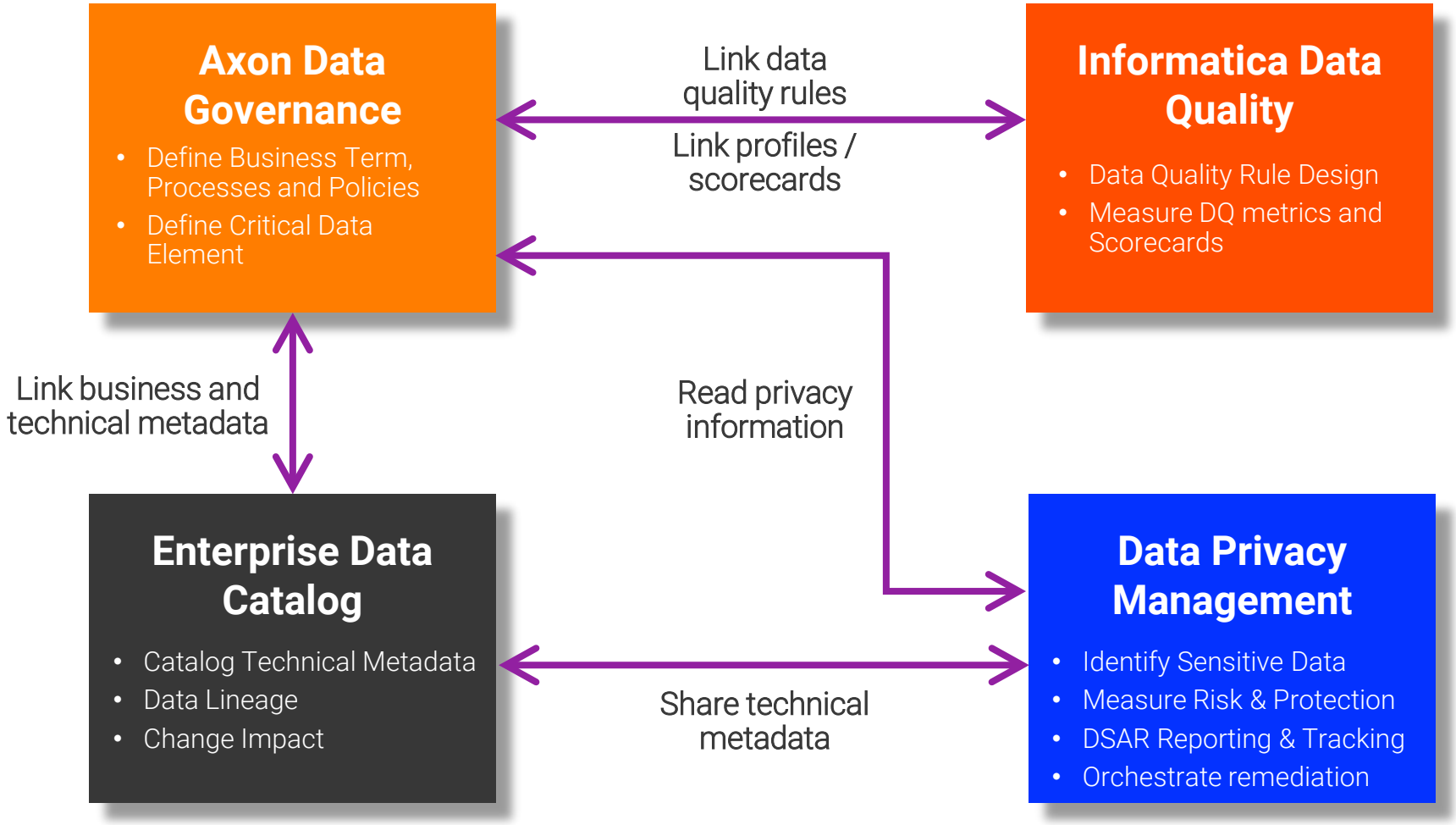




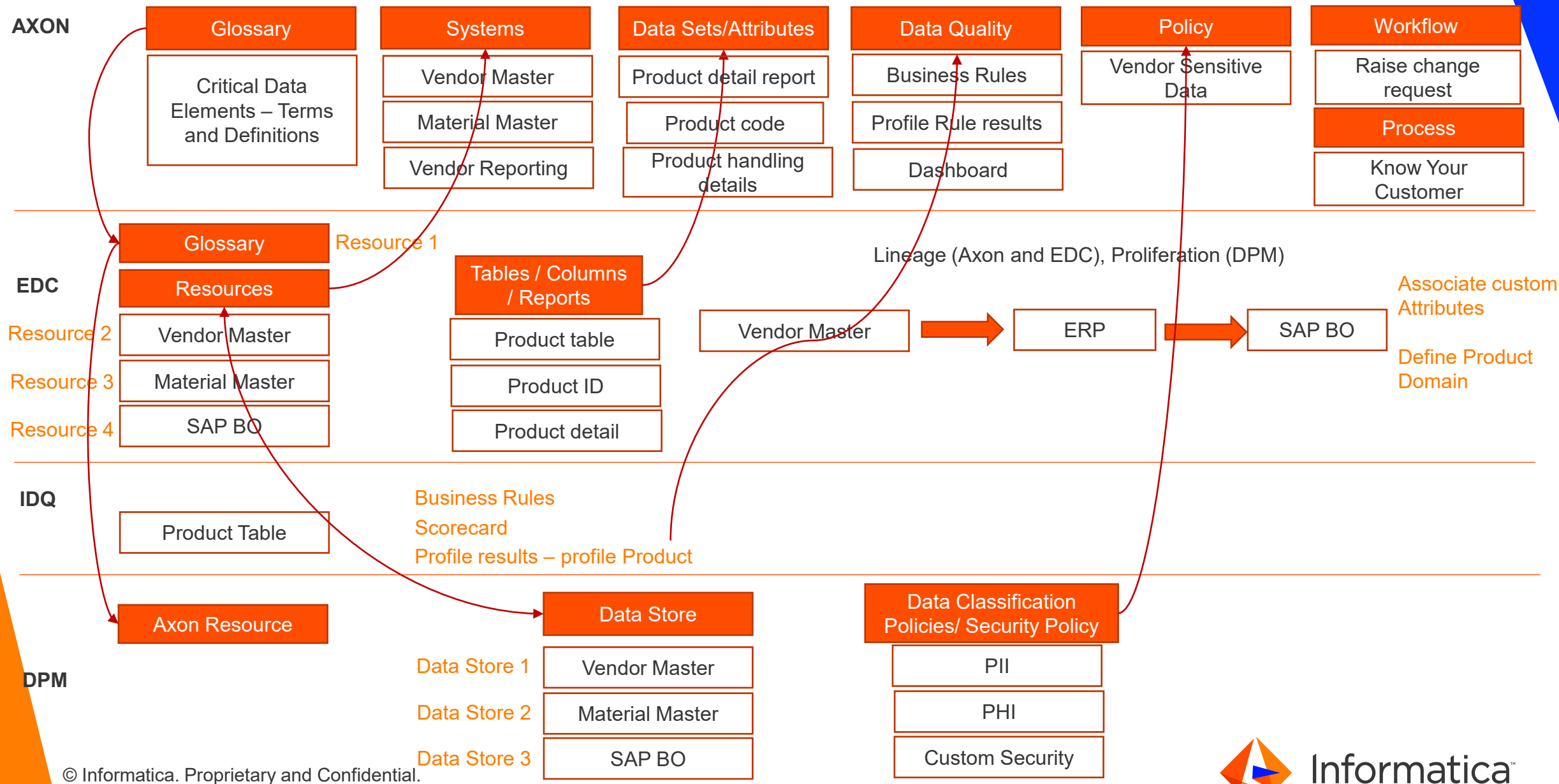


Informatica

# Cross Product View



# DG APPLICATION RELATIONSHIPS

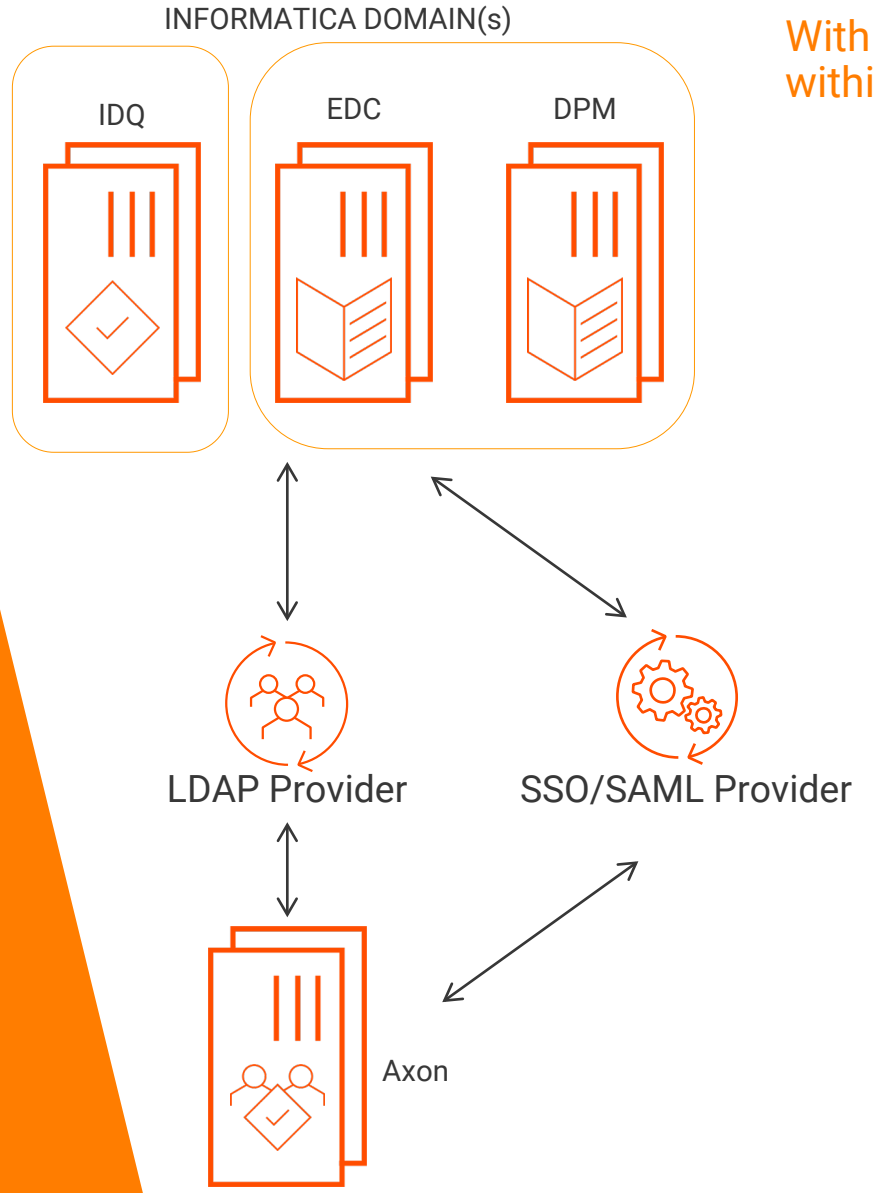





Informatica

# User & Roles Management

# User & Security Management




With Informatica you have three options to configure user access: natively (defined within Informatica platform, with LDAP integration or SSO/SAML access

## LDAP INTEGRATION

- You can configure an Informatica domain to enable users imported from an LDAP directory service to log in to Informatica nodes, services, and application clients
- Axon can connect to the LDAP directory in your organization, retrieve users, assign them user profiles, and display them in the Axon interface. Axon automatically creates the users after it connects to the LDAP server
- LDAP providers certified: MS AD, OpenLDAP etc.
- Refer to PAM for compatibility

## SSO (SAML)

- You can configure Security Assertion Markup Language (SAML) authentication in an Informatica domain and in Axon installation
- For Informatica domain, we support Microsoft Active Directory Federation Services (AD FS) identity provider
- For Axon, we certify Okta, OneLogin and Azure Active Directory. However, Axon does support with all vendors supporting SAML v2.0,
- Refer to PAM for compatibility

# Axon User Engagement - Typical Roles

## Knowledge Roles

### Subject Matter Expert

Subject matter expert with a deep (often localised) understanding of the context in which the data asset operates e.g. requirements, impacts, change horizon etc.

Variants: business SME, technical SME, domain expert, consumer, producer

## Stewardship Roles

### Steward

Advocate of good data governance practices and coordinator of the data knowledge, improvement and alignment efforts.

Variants: domain steward, coordinator, implementer

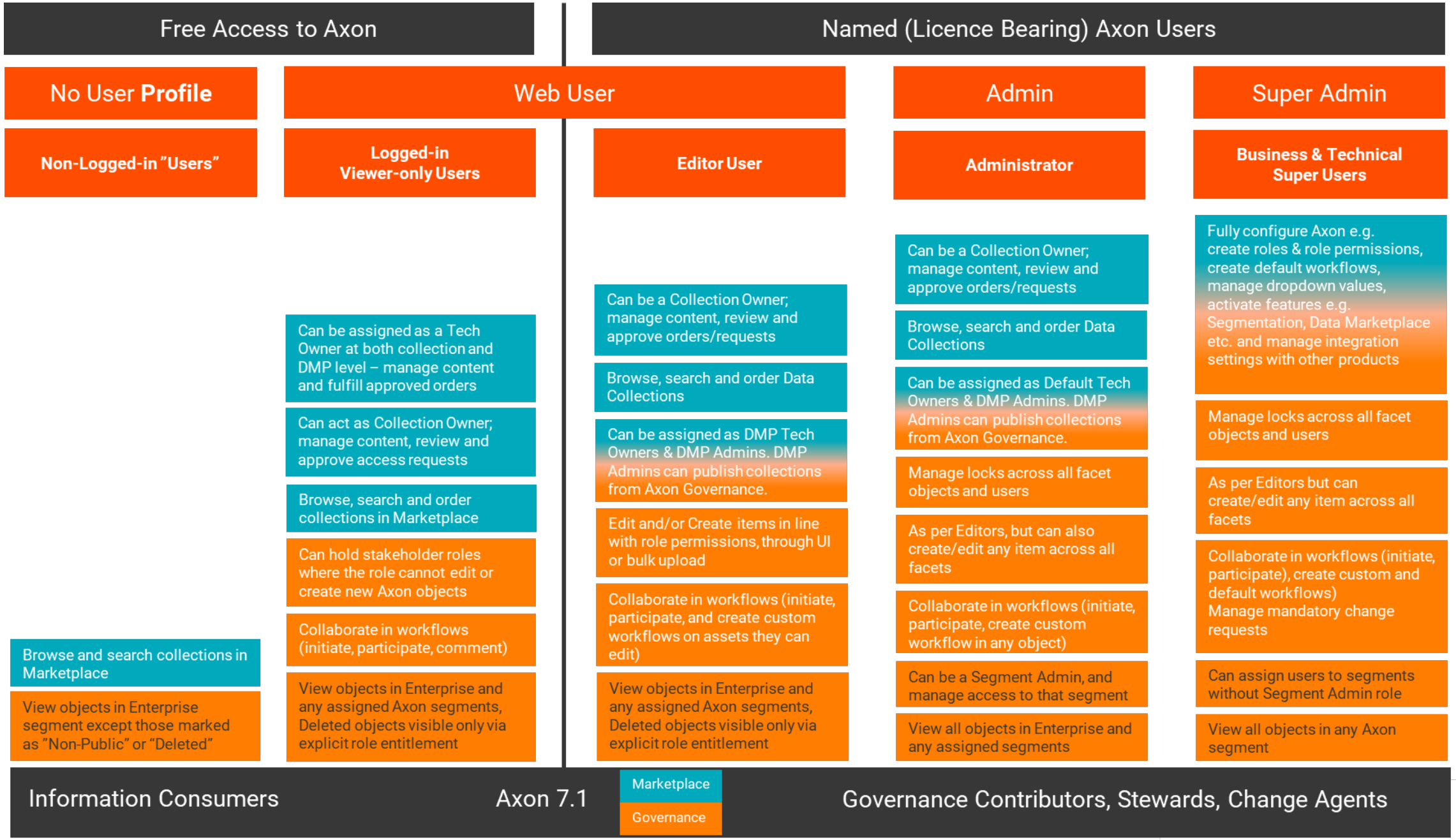
## Authoritative Roles

### Owner

The party responsible for the accuracy of the information captured in the governance catalogue and the alignment with governance policies and practices.

Variants: custodian

The above roles could apply to many of the Axon facets e.g. glossary, system, process


Free Access to Axon
Named (Licence Bearing) Axon Users
No User Profile
Web User
Admin
Super Admin
Non-Logged-in "Users"
Logged-in Viewer-only Users
Editor User
Administrator
Business & Technical Super Users
Can be assigned as a Tech Owner at both collection and DMP level – manage content and fulfill approved orders
Can act as Collection Owner; manage content, review and approve access requests
Browse, search and order collections in Marketplace
Can hold stakeholder roles where the role cannot edit or create new Axon objects
Collaborate in workflows (initiate, participate, comment)
View objects in Enterprise and any assigned Axon segments, Deleted objects visible only via explicit role entitlement
Can be a Collection Owner; manage content, review and approve orders/requests
Browse, search and order Data Collections
Can be assigned as DMP Tech Owners & DMP Admins. DMP Admins can publish collections from Axon Governance.
Edit and/or Create items in line with role permissions, through UI or bulk upload
Collaborate in workflows (initiate, participate, and create custom workflows on assets they can edit)
View objects in Enterprise and any assigned Axon segments, Deleted objects visible only via explicit role entitlement
Can be a Collection Owner; manage content, review and approve orders/requests
Browse, search and order Data Collections
Can be assigned as Default Tech Owners & DMP Admins. DMP Admins can publish collections from Axon Governance.
Manage locks across all facet objects and users
As per Editors, but can also create/edit any item across all facets
Collaborate in workflows (initiate, participate, create custom workflow in any object)
Can be a Segment Admin, and manage access to that segment
View all objects in Enterprise and any assigned segments
Fully configure Axon e.g. create roles & role permissions, create default workflows, manage dropdown values, activate features e.g. Segmentation, Data Marketplace etc. and manage integration settings with other products
Manage locks across all facet objects and users
As per Editors but can create/edit any item across all facets
Collaborate in workflows (initiate, participate), create custom and default workflows)
Manage mandatory change requests
Can assign users to segments without Segment Admin role
View all objects in any Axon segment
Browse and search collections in Marketplace
View objects in Enterprise segment except those marked as "Non-Public" or "Deleted"
Information Consumers
Axon 7.1
Marketplace
Governance
Governance Contributors, Stewards, Change Agents

# Admin vs Super Admin

Super Admins can:

- Configure the application
- Create Roles
- Set Role Permissions
- Create default workflows
- Change dropdown values
- Make any other changes possible via the admin panel

Generally, we would only expect a customer to have a couple of Super Admins as they have such broad access and control.

Administrators on the other hand have full edit rights across objects but will not change the application itself, you therefore may want one for each business area /department involved.

# User Roles and Responsibilities

EDC

## IT Operations

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

## Catalog Program Mgr. & Data Owner

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

## Catalog Administrator

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 RestAPI to extend EDC

## Data Steward & Subject Matter Expert

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

## Catalog User

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




Informatica

# Permissions in EDC

**Read**
View the details of the resource and assets in Enterprise Data Catalog.

**Read and Write**
Allows the user or users included in the user group to enrich the assets in the Enterprise Data Catalog in addition to the read permission. You can enrich assets by assigning custom attributes, business terms, or data domains to the asset. Enriching assets helps you search for the asset using the assigned custom attribute, business term, or data domain.

**Note**: If you configure read or read and write permission for relational sources such as Oracle, you cannot see the following assets for the source till you configure permissions for the assets:

- Tables
- Views
- Synonyms

**Metadata and Data Read**
Allows the user or users included in the user group to view the value frequency for an asset in Enterprise Information Catalog. You can assign this permission to resources that support profiling.

**All Permissions**
Allows the user or users included in the user group to enrich the assets in the Enterprise Information Catalog and view the value frequency for the asset. You can enrich assets by assigning custom attributes, business terms, or data domains to the asset. Enriching assets helps you search for the asset using the assigned custom attribute, business term, or data domain.




Informatica

# Data Governance Administrator Role

Generically these folks should:

- Understand the needs of enterprise governance from a *business* perspective
- Set guidelines for aims, scope and execution of the governance program, and be able to communicate these throughout the organization
  - They might end up doing all the work initially, to iterate and learn how to deliver this
  - Longer term they should be setting direction and guiding/coaching others only
- From an Axon perspective they should also
  - Understand how to structure and deploy the program
  - Be expert users of Axon

NOTE: The Axon User Roles matrix shows mechanically what they can do in Axon.

The Data Governance Administrator Role should have either Administrator/Super Administrator Axon user accounts.

- Both allow super user access to Axon content, to help manage day to day issues, with the SuperAdmin controlling the look, feel and config of the tool. This role should be limited and access to SuperAdmin privileges should be strictly controlled.

# Metadata Lifecycle Considerations

| Metadata Lifecycle | | Data Catalog Manager | Data Steward | Business Domain SME | Data Owner | Source system Owner | Business User | Catalog Admin |
|---|---|---|---|---|---|---|---|---|
| Ingest (new or refresh existing) | Metadata Ingestion Access requests | | | | | | | |
| | Provide security access to data source | A | | | | | | R |
| | Custom connections | | | | | | | R |
| | Ingest technical Metadata | R | S | | C | C | | S |
| | Ingest Business Metadata | R | S | C | C | | | |
| | Creation of Data Domains | A | R | S | | | | |
| | Extract technical metadata and relationships including data lineage | R | | | | | | S |
| Profile | profile data sources to assess data quality, discovery Data Domains, Curation | A | R | | S | | | |
| Classify | Validate Auto tag semantic meaning to technical metadata | S | R | | | | | |
| | Confidentiality of data asset | A | R | | S | | | |
| Index | Index technical metadata to make it searchable | R | S | | | | | C |
| | Index business metadata to make it searchable | R | S | | C | | | |
| Annotate | Connect tech metadata to Biz metadata | A | R | | C | | | |
| | Attach custom annotations | A | R | | C | | | |
| | Managing custom attributes | | R | A | A | | | |
| Consume | Discover metadata | | | | | | | |
| | Find available assets. | | | | | | | |
| | Explore assets to view the quality of data | | | | | | | |
| | View lineage for assets. | | | | | | | |
| | View relationships between assets. | | | | | | | |
| | Enrich assets by tagging them with additional attributes | | R | R | A | | | |
| Govern | Creating user groups | A | R | | | | | |
| | Defining Business glossary for their function / Segment | A | S | R | S | | | |
| | Approve Tagging | | R | C | C | | | |
| | Approve Changes to the Metadata | | R | C | C | | | |
| | Approve requests for access | | R | C | C | | | |
| | Minimum standards on business tagging including custom attributes and data domains | R | S | S | C | | | |
| | Accountable for the quality of meta data for their segment / function | S | R | R | | | | |

# Sizing Guideline

# IDQ

| | Minimum Requirement | Common Enterprise Deployment Examples | | | |
|---|---|---|---|---|---|
| | Single Node (4 Cores & 20Gb Memory) | Small - 4 Nodes (16 Cores & 64Gb Memory) | Medium - 8 Nodes (32 Cores & 128GB Memory) | Large - 16 Nodes (64 Cores & 256Gb Memory) | Extra Large - 32 Nodes (128 Cores & 512Gb Memory) |
| **Analyst User Interface** | | | | | |
| Concurrent Users | 15 | 60 | 120 | 240 | 480 |
| **Data Profiling** | | | | | |
| Concurrent Profiles (avg. 100 Columns * 10M Rows) | 2 | 8 | 16 | 32 | 64 |
| **Data Cleansing & Validation** | | | | | |
| High Intensity (i.e. Address Validation) | 4M rows/hr | 16M rows/hr | 32M rows/hr | 64M rows/hr | 128M rows/hr |
| Medium Intensity (i.e. Parsing, Standardization) | 10M rows/hr | 40M rows/hr | 80M rows/hr | 160M rows/hr | 320M rows/hr |
| Lower Intensity | 20M to 25M rows/hr | 100M rows/hr | 200M rows/hr | 400M rows/hr | 800M rows/hr |
| **Data Matching** | | | | | |
| Matching | 5M to 7M rows/hr | 28M rows/hr | 56M rows/hr | 112M rows/hr | 224M rows/hr |



Informatica

# EDC

| | | 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




Informatica

# Resources

1. Configure Access Axon/IDQ: Click Here
2. Configure Access Axon/EDC: Click Here
3. Configure Access Axon/DPM: Click Here
4. Axon/EDC Automatic Onboarding Workflow: Click Here
5. Automate Data Quality Rules in Axon: Click Here
6. EDC Sizing Guide: Click Here
7. Profiling Sizing Guide: Click Here
8. Integrated Monitoring for Capacity Planning/Resource Utilization: Click Here
9. Product Availability Matrix (PAM): Click Here
10. AWS Informatica Marketplace Offerings: Click Here
11. Azure Informatica Marketplace Offerings: Click Here
12. Deploying DIS on GRID: Click Here

# Thank You!

# Appendix

# DPM 10.4.1 Performance Summary

**Unstructured Domain Discovery:**

- Unstructured throughputs are much better for mixed file types with Agent profiling implementation compared to EDC.
- Achieved throughput of 20 to 25GB/hr. compared to 10.4.0 (3GB/hr) with 16 threads/cores for different file and data set sizes having mixed file types.
- Linear scaling achieved with multiple agents for large volumes of data (Verified up to 1TB data set).

**Subject Registry:**

- At least 20X improvement in performance of Subject Registry scans compared to previous release.
- 10M records was not possible in previous release, successfully. Tests were run for 100M in this release.

**Regression:**

- No regression in Data discovery at DPM.
- No regression in UI performance.
- No regression in DSAR reports performance.

**Performance Improvements in coming releases:**

- High memory usage fix for Unstructured large files.
- Evaluate Policy fix for unstructured scan for large number of files
- Subject registry performance improvements for Exact and multiple matching configurations.

# DPM 10.4.1 Hardware Sizing for Unstructured Domain Discovery - Agent Profiling

## Sizing inputs from Benchmarking results:

❑ A single agent throughput with 16 threads/cores is 20 to 25GB/hr.

❑ Scaling factor from 8 to 16 threads is 1.7X but 16 to 32 threads is about 1.2X. i.e. scaling is good till 16 cores for single agent or multiple agents with in one server.

❑ Scaling is linear with multiple agents across different servers (Scale out/horizontal scaling). It is about 1.8X.

## Profiling Agent Sizing:

❑ Agent can be in the same server or different server based on the resource availability. Agents on different servers is recommended for production.

❑ Number of machines for Agent can be determined based on SLA (throughput to be achieved in GB/hr.).

e.g. If 100GB/hr. is required then we need 5 servers with 16 cores each for Agents.

❑ Recommended configuration for a dedicated Agent server is 16 core and 24GB RAM ( Agent JVM memory setting required is 1G per thread/core).

**Informatica Domain Server Sizing**

| #Cores EDC + DPM | # Cores EDC + DPM + Agent | Memory EDC + DPM | Memory EDC + DPM + Agent | Storage |
|---|---|---|---|---|
| 16 | 32 | 32GB | 64GB | 500GB |

**Hadoop Cluster Sizing:**

| #Hadoop nodes | #Cores per node (EDC + DPM) | Memory per node (EDC + DPM) | Storage (EDC + DPM) |
|---|---|---|---|
| 1 | 8 | 32GB | 200GB |

**** Required only to bring up the services

# DPM 10.4.1 Tuning Guidelines for Unstructured Domain Discovery - Agent Profiling

- ❑ Set Agent JVM max memory setting as 1GB per thread. (e.g. 16 threads, 16GB)
- ❑ If average file size is small (< 500KB), having multiple agents at the same server gives better throughput.
- ❑ Two agents of 8 threads each is better than one agent with 16 threads for large volume of small sized files.
- ❑ Below custom properties need to be set for better performance and to achieve maximum possible throughputs.

| Custom Properties | |
|---|---|
| SatSCustomOptions.SatsAgentProfilingResultFetchInterval | 5000 |
| SatSCustomOptions.SatsAgentProfilingResultPersistersMonitoringInterval | 1000 |
| SatSCustomOptions.SatsAgentProfilingResultPersisterThreadCount | 4 |
| SatSCustomOptions.SatsAgentProfilingUnStructuredFileBatchSize | 1000 |
| SatSCustomOptions.SatsBrowseResultFetchInterval | 1000 |
| SatSCustomOptions.SatsBrowseUnStructuredFileBatchSize | 1000 |
| SatSCustomOptions.SatsEvaluatePolicyForObjectsThreadCount | 4 |
| SatSCustomOptions.SatsSubjectRegistryUnStructuredFileBatchSize | 100 |

# DPM 10.4.1 Tuning Guidelines for Unstructured Domain Discovery - Agent Profiling

## Recommendations for Agent settings based on File sizes:

| File size limit | Number of threads | JVM heap setting |
|---|---|---|
| 50MB | 16 | 16GB |
| 100MB | 8 or 16 | 16 or 32GB |
| 200MB | 8 | 32GB |
| 500MB | 4 | 32GB |
| 1GB | 2 | 32GB |
| 2GB to 5GB | 1 | 32GB |

# Server Sizing for Implementations – Domain Discovery Scan for Structured

| Total number of columns to be processed including all parallel scans/Datastores ( up to 5 ) | Domain Server configuration | | | Hadoop Cluster Configuration | | | | E2E scan time at DPM | Throughput (Tables per hour with 100 columns per table) |
|---|---|---|---|---|---|---|---|---|---|
| | # cores | Memory | Disk Storage | # nodes | CPU per node | Memory per node | Total Storage | | |
| 1M | 16 | 32GB | 200GB | 1 | 8 | 32GB | 500GB | 5hr | 2000 |
| 1M | 32 | 64GB | 200GB | 1 | 8 | 32GB | 500GB | 3hr | 3500 |
| 5M | 32 | 64GB | 200GB | 3 | 8 | 32GB | 500GB | 15hr | 3000 |
| 10M | 32 | 64GB | 300GB | 3 | 8 | 32GB | 700GB | 30hr | 3000 |
| 25M | 32 | 64GB | 500GB | 3 | 8 | 32GB | 1TB | 90hr | 2500 |
| 50M | 32 | 64GB | 1TB | 3 | 8 | 32GB | 2TB | 250hr | 2000 |
| 50M | 64 | 128GB | 1TB | 6 | 8 | 32GB | 2TB | 180hr | 2800 |

**Assumptions:**

- Sizing is considering the number of columns being processed as single scan or up to 5 parallel scans.
- Number of columns per table is 100
- Number of Data domains in policy are 10
- Profile Type is First 10K rows
- Increase in number of cores will help in only improving the profiling time (processing at DIS side).
- Processing time at DPM Job steps increases with increase in volume/sensitive data.
- Above sizing is assuming Domain server having single DIS.
- Each DIS node will add about 30% scalability. E.g. 2 DIS nodes will give about 1.3X more throughput.
- Subject Registry and Domain discovery scans are not run in parallel to avoid contention at Hbase.

# Server Sizing for Implementations – Domain Discovery Scan + Subject Registry

- ❑ No Additional hardware required for Subject Registry for Informatica Domain Server assuming Discovery scans are not run in parallel.
- ❑ Agent sizing for Unstructured Subject Registry scan is same as given in
- ❑ SR scan for structured scales to about 1.7X between 32 and 64 core servers, but overall utilization is less ( < 30% )
- ❑ DPM JVM heap setting for SR is based on number of Golden records. Heap size of about 40GB required for processing linkage scan against 100M Golden records.
- ❑ Below is the sizing for number of fields 8 in the entity.

### Load type definition

| Load Type | # Assets/Objects/ Columns | # Data stores | data stores scan in parallel | Total Number of Records for SR |
|---|---|---|---|---|
| Low | 1M | 30 to 40 | 3 Low | < 10M |
| Medium | 20M | 200 to 400 | 3 Medium | 10M to 100M |
| High | 50M | 500 to 1000 | 5 Medium | > 100M |

### Approximate Disk space required for Subject Registry at HBase

| No of Golden Records | No of Linkage records | Disk Space required for 8 fields in the entity with max filed size 50 | Disk Space required for 16 fields in the entity with max filed size 50 |
|---|---|---|---|
| 10M | 10M | 15GB | 30GB |
| 100M | 10M | 80GB | 160GB |
| 100M | 100M | 150GB | 300GB |

### Informatica Domain Server Sizing Summary

| Load Type | #Cores EDC + DPM | Memory EDC + DPM | #Cores EDC + DPM + SR | Memory EDC + DPM + SR | Storage |
|---|---|---|---|---|---|
| Low | 16 | 32GB | 24 | 32GB | 200GB |
| Medium | 32 | 64GB | 32 | 64GB | 500GB |
| High | 64 | 128GB | 64 | 128GB | 1TB |

### Repositories DB Sizing

| Load Type | #Cores | Memory | Disk |
|---|---|---|---|
| Low | 4 | 8GB | 100GB |
| Medium | 8 | 16GB | 200GB |
| High | 16 | 32GB | 500GB |

### Hadoop Cluster Sizing

| Load Type | #hadoop nodes | #Cores per node (EDC + DPM) | Memory per node (EDC + DPM) | Storage (EDC + DPM) | Storage (EDC + DPM + SR) |
|---|---|---|---|---|---|
| Low | 1 | 8 | 32GB | 200GB | 500GB |
| Medium | 3 | 8 | 32GB | 500GB | 1TB |
| High | 6 | 8 | 32GB | 1TB | 2TB |

******** Agent Sizing and number of Agent servers required mentioned in previous slides.




Informatica

# Server Sizing for Implementations – Domain Discovery Scan + Subject Registry + User Activity + UBA

## User Activity & UBA

- No changes in Sizing for Informatica Domains server with UA, only additional hardware required is for Hadoop cluster.
- Upto10% of E2E scan time impact for Discovery scans time due to sharing of yarn resources when run in parallel with full load of UA events...
- Security Violation Policies only impacts 1% of the user activities / anomalies

## Hadoop Cluster Sizing

| Load Type | #Hadoop nodes | #Cores per node (EDC + DPM) | #Cores per node (EDC + DPM + UA) | Memory per node (EDC + DPM) | Memory per node (EDC + DPM + UA + UBA) | Storage (EDC + DPM) | Storage (EDC +DPM + UA) | Storage (EDC + DPM + SR) | Storage (EDC + DPM + SR+ UA) |
|---|---|---|---|---|---|---|---|---|---|
| Low | 1 | 8 | 16 | 32GB | 64GB | 200GB | 500GB | 500GB | 1TB |
| Medium | 3 | 8 | 16 | 32GB | 64GB | 500GB | 1TB | 1TB | 2TB |
| High | 6 | 8 | 16 | 32GB | 64GB | 1TB | 2TB | 2TB | 4TB |

## Expected Throughputs for UA and UBA

| Number of Events | #Hadoop nodes | Processing Time for all events | Events per second | With security policies | Events per second | Additional Disk Space |
|---|---|---|---|---|---|---|
| 10 million | 1 | 7 hr | 400 | 8 hr | 300 | 15 GB per every 10 million events |
| 10 million | 3 | 3 hr | 1,000 | 3.5hr | 800 | |
| 10 million | 6 | 2 hr | 1,500 | 2.5hr | 1100 | |

# Unstructured Discovery Performance 10.4.1

Hardware Configuration for DPM service and Agent:

CPU:  32 cores
Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz
CPU MHz: 1200
Memory: 64GB

Scenario:

- Data sets of different sizes and mixed file types
- End to end testing from DPM
- File System: NFS

| Volume | | | No of Data domains/policy | Time taken at Agent profiling | Number of cores/threads | Throughput (GB/hr) | % Sensitive | Comments |
|---|---|---|---|---|---|---|---|---|
| #files | Size | Avg size | | | | | | |
| 10K | 21GB | 2MB | 8 DD | 52min | 16 | 24 | 1 | |
| 10K | 21GB | 2MB | 8 DD (EDC scan) | 6h 40m | 20 | 3 | 1 | Done with EDC for same dataset and DD's |
| 10K | 21GB | 2MB | PII (14 DD's) | 51min | 16 | 24 | 1 | |
| 10K | 21GB | 2MB | GDPR (49 DD's) | 5hr 11min | 16 | 4 | 1 | |
| 10K | 21GB | 2MB | GDPR (EDC scan) | Failed | 20 | | | Most of the files failed and job hanging for long time at EDC. Need to rerun |
| | | | | | | | | |
| 100K | 200GB | 2MB | PII | 7hr 16min | 16 | 27 | 1 | |
| 100K | 200GB | 2MB | GDPR | ~ 50hr (Projected) | 16 | 4 | 1 | |
| | | | | | | | | |
| 1M | 179GB | 200KB | 8 DD | 12hr 49min | 32 | 14 | 0.9 | 4 agents  with 8 threads each |
| 1M | 179GB | 200KB | 8 DD | 5hr 45min | 32 | 30 | 0.33 | 4 agents with  8 threads each |
| 1M | 179GB | 200KB | 8 DD | 7hr 42min | 16 | 23 | 0.33 | 4 agents with 4 threads each |

- Browse time for 1M files is 20min
- Throughputs given are for Profiling step.

# Unstructured Discovery Performance 10.4.1

## Performance of Large files

| Each File Size | Number of files | No of threads | Heap/ Memory Used | Agent JVM Heap Setting | Time taken | No of Domains matched | Domain Impressions | Policy Impressions |
|---|---|---|---|---|---|---|---|---|
| 55MB | 16 | 16 | 16GB | 16GB | 6 min | 3 | 15.3M | 10.4M |
| 1GB | 1 | 1 | 22GB | 32GB | 28 min | 3 | 18.4M | 12.5M |
| 1GB | 2 | 2 | 22GB | 32GB | 28 min | 3 | 36.8M | 25M |
| 2GB | 1 | 1 | 32GB | 32GB | 57 min | 3 | 36.8M | 25M |
| 5GB | 1 | 1 | 32GB | 32GB | 3 hr | 3 | 85M | 57.8M |
| 5GB | 1 | 1 | 49GB | 64GB | 2hr 20min | 3 | 85M | 57.8M |
| 370MB | 1 | 1 | 9GB | 16GB | 7 min | 1 | 19.9M | 19.9M |

Performance also depends on the impression count. Performance degrades with increase in impression count for same file sizes.

# Subject Registry Performance for 10.4.1

**Hardware Configuration 1 for DPM :**
CPU:  32 cores,
Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz
CPU MHz: 1200
Memory: 64GB

**Hardware Configuration 2 for DPM :**
CPU:  64 cores
Intel(R) Xeon(R) Gold 6142 CPU @ 2.60GHz
CPU MHz: 3300
Memory: 128GB

**10.4.1**

| Volume and scan type | Time taken on 32 core box | Time taken on 64 core box |
|---|---|---|
| 10M Golden scan | 18min | 14min |
| 10M raw scan with 10M Golden records | 3hr 21min | 2hr 13min |
| 10M each two Data stores Golden scan | 3hr 15min each | 2hr 12min each |
| 100M Golden scan | 4h 40min | |

**10.4.0**

| Volume and scan type | Time taken on 32 core box | Time taken on 64 core box |
|---|---|---|
| 5M Golden scan | 16hr | |
| 10M Golden scan | 40hr+ (not completed) | 07hr 46min |
| 1M Raw scan with 1M GR | 2hr 21min | 1hr |
| 1M each two data stores golden scan | 1st DS scan - 2hr 6min | 1hr 11min |
| | 2nd DS scan - 6hr 35min | 2hr 34min |

- Significant improvement in SR performance in10.4.1 compared to 10.4.0
- Large volumes were not working in 10.4.0 due to exponential increase in process time with increase in volume. Up to 100M tested in 10.4.1 for Golden scan.
- Performance degradation compared to earlier builds due to additional steps in scan to fix dashboard functional issues.
- No notable overhead due to encryption implementations
- No regression in DSAR performance.
- No regression in unstructured SR scans.

# Role Assignment in Axon

- As a glossary steward (Web user) when I create a glossary, I have 3 roles that are required (Glossary steward, Glossary SME & Business Data Owner). Currently all these roles get assigned to me (glossary steward) by default. Is there a way I can assign the other roles (SME/Data Owner) to specific people by default and not manually?

- We'll assume that you have set all these 3 roles are default in the Admin Panel and hence the user/creator gets associated with all of these roles. There is no way to associate certain roles to other people. We are looking at role inheritance for next year where these roles can be inherited from the parent. Now, for the Mandatory Approval scenarios, we do have a way of inheriting the roles from the top level but this is only if the Mandatory Approval is enabled for that facet.

# DQ rule stakeholders in Axon

- I have two stakeholder roles for the Data Quality Facet.  Both roles can view/edit and one role can also Create.  When I go to any Data Quality rule as a user with either stakeholder role for that rule, the edit button is not available.  Nor can either user see Data Quality rule in the Create menu.  Works fine for roles I've done on glossary, but not DQ.

- Local DQ rules in the DQ facet, and permissions ran off the data set that held the attributes aligned to the local DQ rules. Standard rules ran off the Glossary facet, but there's still some work to do to align the permissions.
- Axon functionality is that local rules are managed by Data Set stakeholders, Std rules by glossary stakeholders, and that's what we are going to return to when the permissions issues are addressed
- There has never been an intent to give DQ stakeholders control of the object, they do have the ability to have workflow and CRs… but only to discuss things  - by the way, if you raise a CR on any DQ object it is the Local Rule workflow that is raised - we do not have workflow/CR for standard rules, another thing that needs addressed.

# Role Permissions

- A client is asking to decide their roles & responsibilities model by referencing the default Axon one.  Is there any document for how each role is mapped to which Facet and the permission assigned to each role?  Or is such metric table in any documentation?

- As a general rule Axon ships as follows:
  - Most facets have 2 roles, an Owner and a Steward
    - Owner has View
    - Steward has New, Edit and View
  - If the facet has only one role then it has all three permissions
- We do have a section for how to define/configure roles and define the permission in Axon Administration Guide/User Guide (refer User roles section).  We can look at the details in UI by navigating to  “Roles Permission” section under Admin Panel

# Questions?

# Thank You