Informatica

Jan 09, 2024

How Cloud Mass Ingestion (CMI) Helps to Build Real-Time Analytics Layer in Cloud?

- Dhirendra Sinha, Principal Product Manager

Where data & AI come to LIFE

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 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

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

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

Build your real-time analytics layer in cloud for AI using Cloud Mass Ingestion(CMI)

Dhirendra Sinha

Principal Product Manager

Where data & AI come to LIFE

Agenda

- Introduction: Need for Real-time Analytics
- Informatica Cloud Mass Ingestion (CMI)
- Change Data Capture Use Cases
- What's New in CMI
  - **Cross agent failover - Support for High availability**
  - **Integration with Databricks Unity Catalog**
  - **Query based CDC for Oracle Source**
  - **CLI support for task edit and redeployment**
  - **Combined load support for Query based CDC in case of SQL Server source**
  - **Azure SQL DB incremental load support**
  - **Timestamp column on Snowflake target**
  - **Replication using Superpipe**

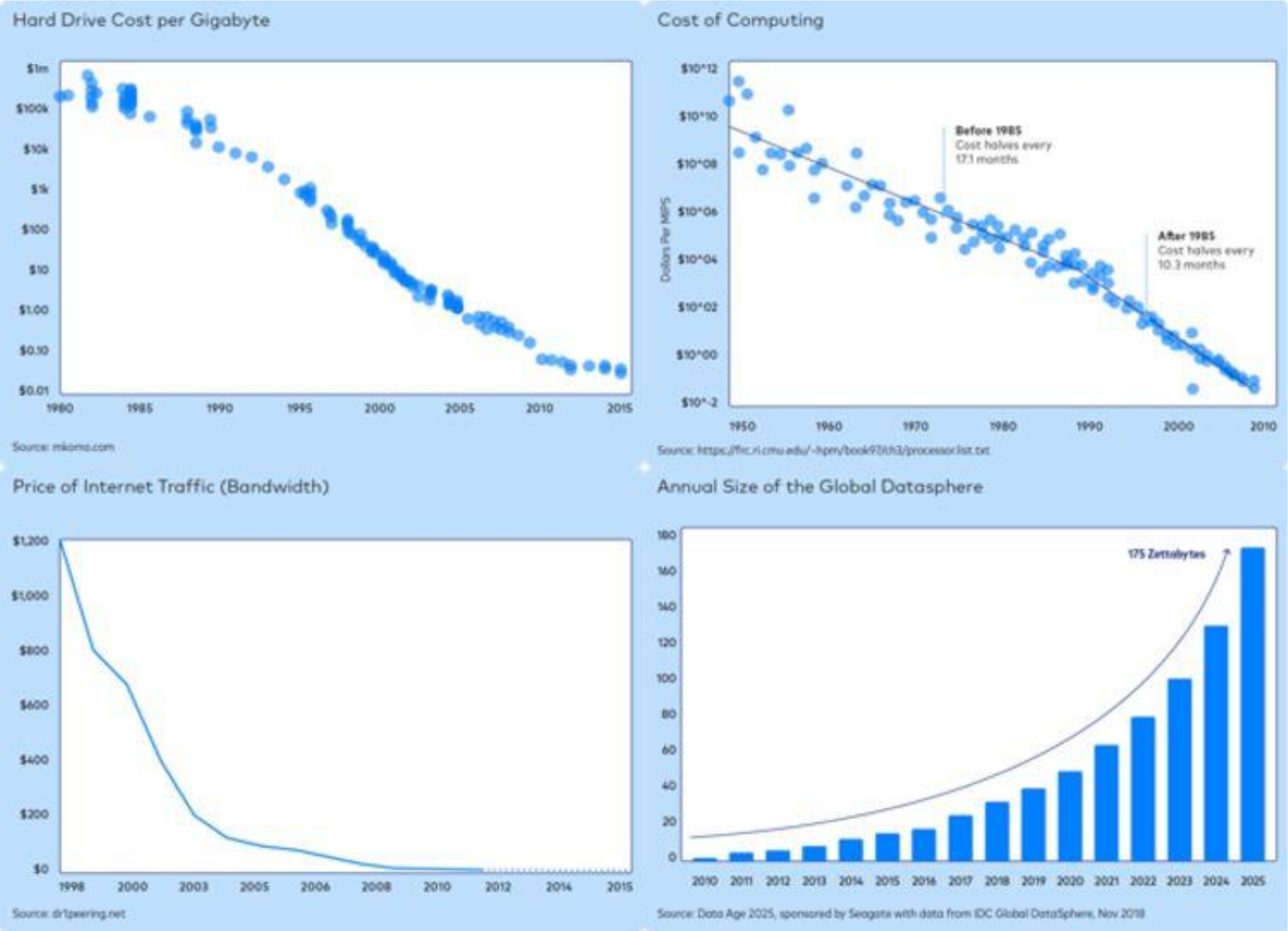
“Gartner predicts that by 2025, GenAI will be a workforce partner for 90% of companies worldwide.”

Source: Gartner Press Release, “Gartner Says CIOs Must Prioritize Their AI Ambition and AI-Ready Scenarios for Next 12-24 Months,” October 16, 2023

Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner’s research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose

Cloud – Purpose built for Analytics and AI

Flexibility, cost savings, scalability make cloud an ideal choice



Hard Drive Cost per Gigabyte

Source: mkomo.com

Cost of Computing

Before 1985
Cost halves every
17.1 months

After 1985
Cost halves every
10.3 months

Source: https://frc.ri.cmu.edu/~hpm/book97/ch3/processorlist.txt

Price of Internet Traffic (Bandwidth)

Source: dr1peering.net

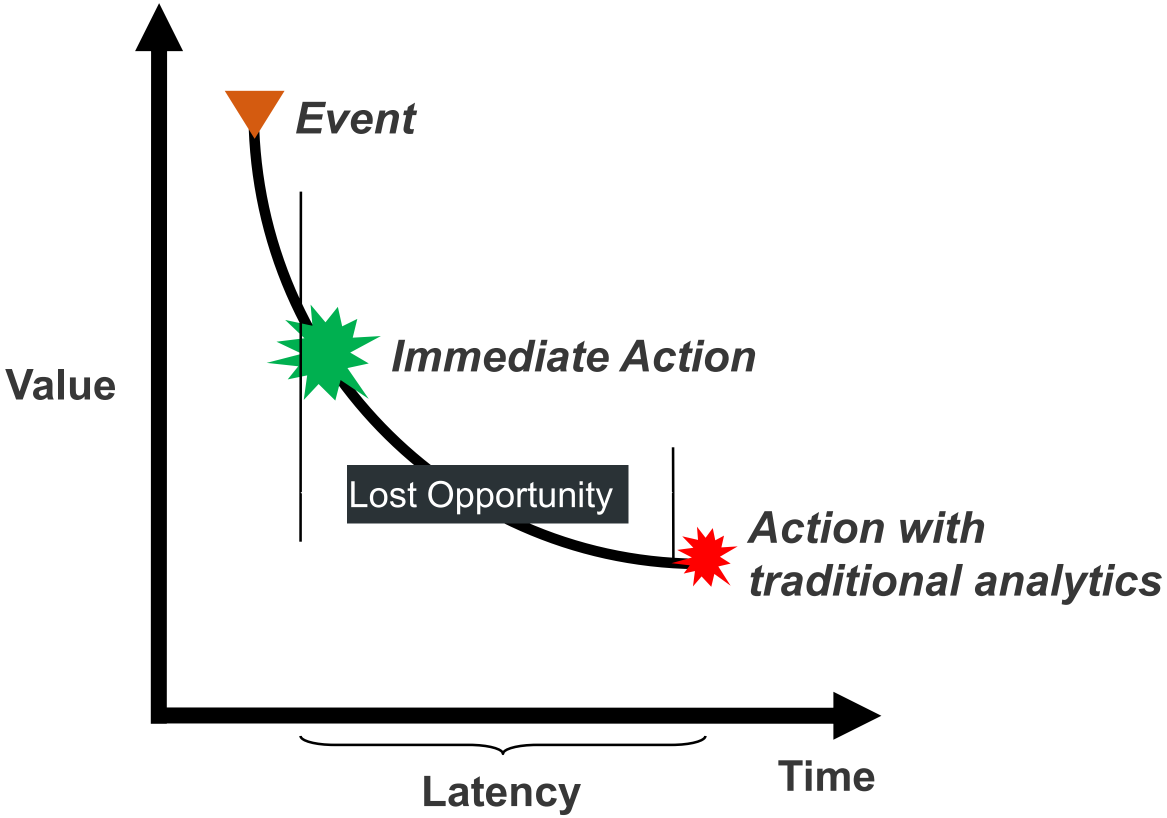
Annual Size of the Global Datasphere

Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Informatica®

Real-time analytics & AI demand up-to-date data

Gaining value from immediate insights



Source: "The BI Watch: Real-Time to Real-Value", Richard Hackathorn

#PerishableInsights

Insights that can provide exponentially more value than traditional analytics but the value expires and evaporates once the moment is gone

Forrester: Mike Gualtieri, Principal Analyst

Data Integration Challenges

Point solutions and complex integration tools, acute shortage of resources, out of control costs

85%

COMPLEXITY

of enterprises made bad decisions, driven by hand-coding and unreliable DIY data pipelines

40M

SKILL SHORTAGE

is the shortage of qualified developers/engineers worldwide, expected to reach 85.2M by 2030

80%

OUT OF CONTROL COST

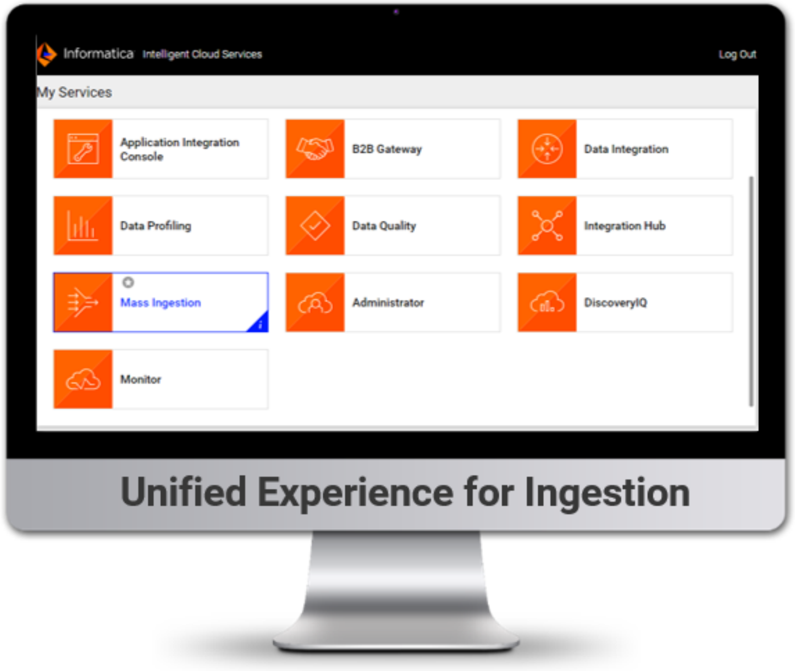
of data management professionals cited difficulty predicting and managing costs

Informatica

Need for a simple, no-code, cost-effective tool which helps in real-time data replication by capturing changed data to cloud for Analytics and AI

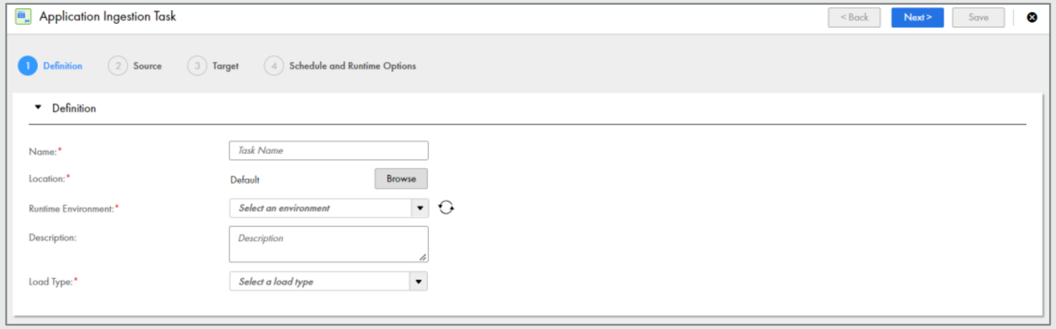
Where data & AI come to LIFE

Informatica’s Data Ingestion and Replication Solution – Cloud Mass Ingestion



Unified Experience for Ingestion

✓ Step-by-step wizard for designing and creating an ingestion task



Name:* Task Name

Location:* Default

Runtime Environment:* Select an environment

Description: Description

Load Type:* Select a load type

✓ Deployment, scheduling, real-time monitoring and lifecycle management

Ingest in Real-Time

Real-Time Monitoring

✓ Versatile out-of-the-box connectivity to sources and targets

Databases & CDC

Streaming Sources

Files

Applications

Mass Ingestion – Replication from SAP, SaaS and DBs

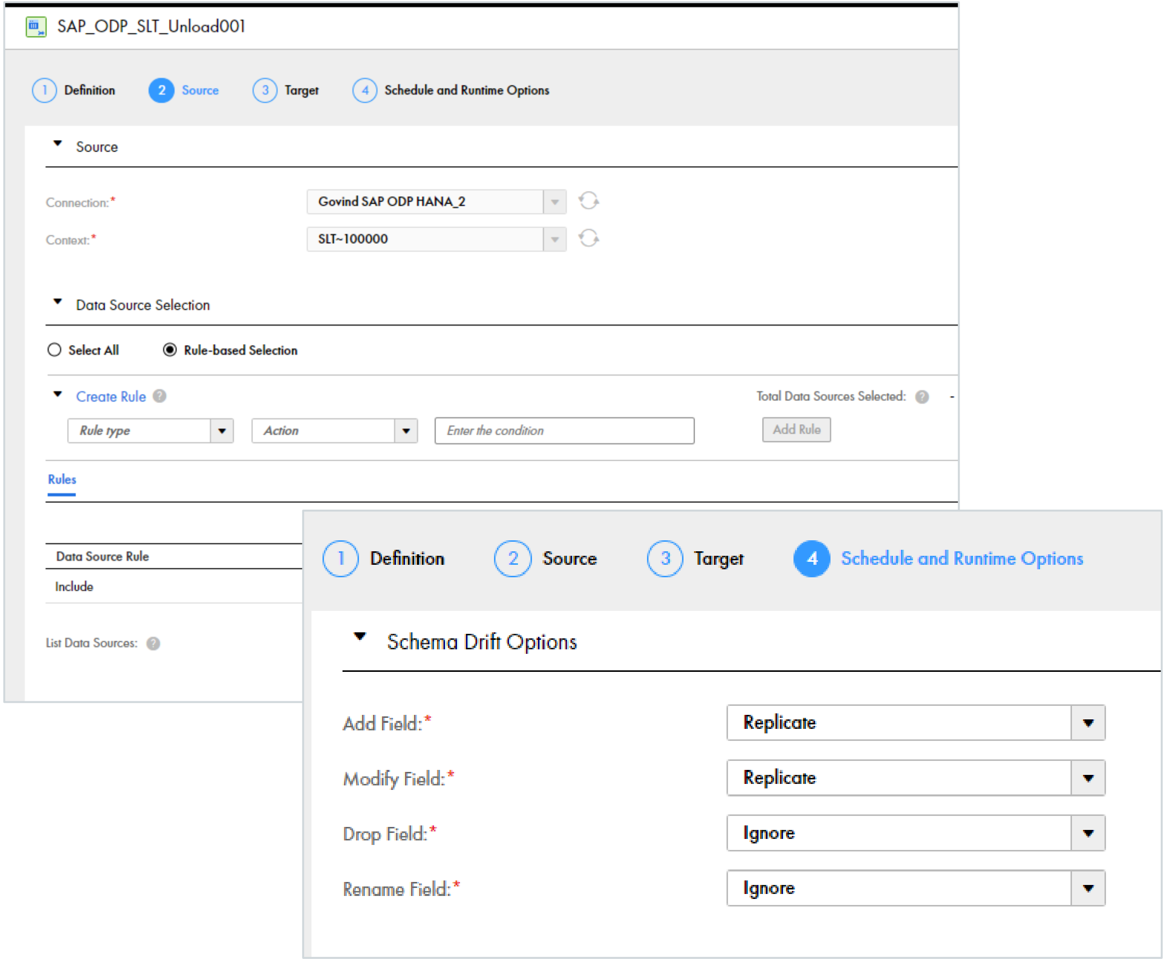
High performant real time replication solution for enterprise use cases

Key Highlights

- Efficient bulk load and CDC capture from various sources including SAP, SaaS apps, relational DB and Mainframes
- Performant way of ingesting and applying CDC onto CDW, CDL and messaging systems
- Automatic schema drift handling– for source schema changes
- Alerting and monitoring of the replication pipeline using

Benefits

- Single platform for ingestion and replication from variety of sources
- Simple and easy way to build and manage ingestion and replication pipeline with wizard driven interface
- No breakage of pipelines in case of schema changes at the source



SAP_ODP_SLT_Unload001

Source

Connection:* Govind SAP ODP HANA_2

Context:* SLT~100000

Data Source Selection

○ Select All ● Rule-based Selection

Create Rule

Total Data Sources Selected: -

Data Source Rule
Include

List Data Sources:

Schema Drift Options

Add Field:* Replicate

Modify Field:* Replicate

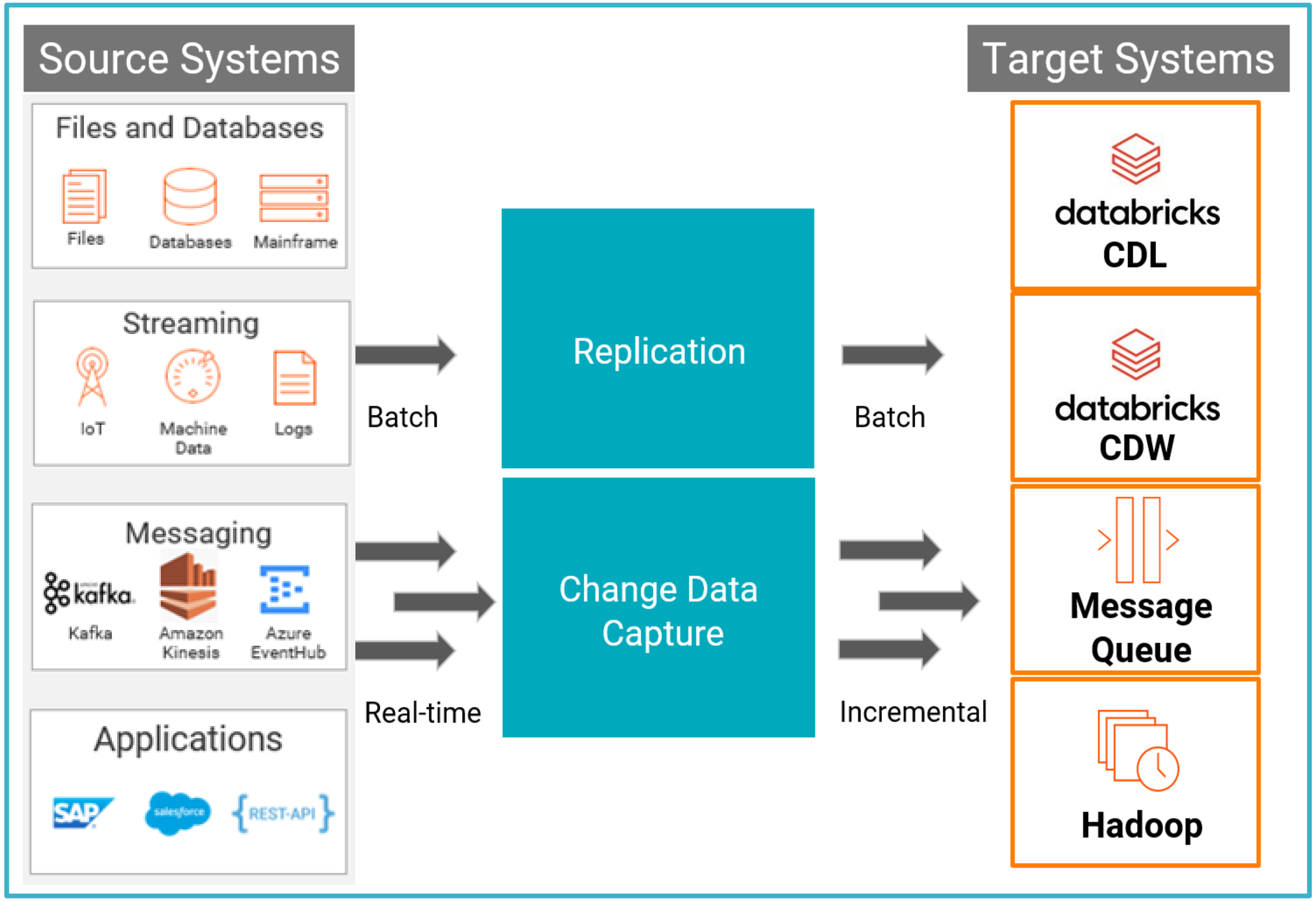
Drop Field:* Ignore

Rename Field:* Ignore

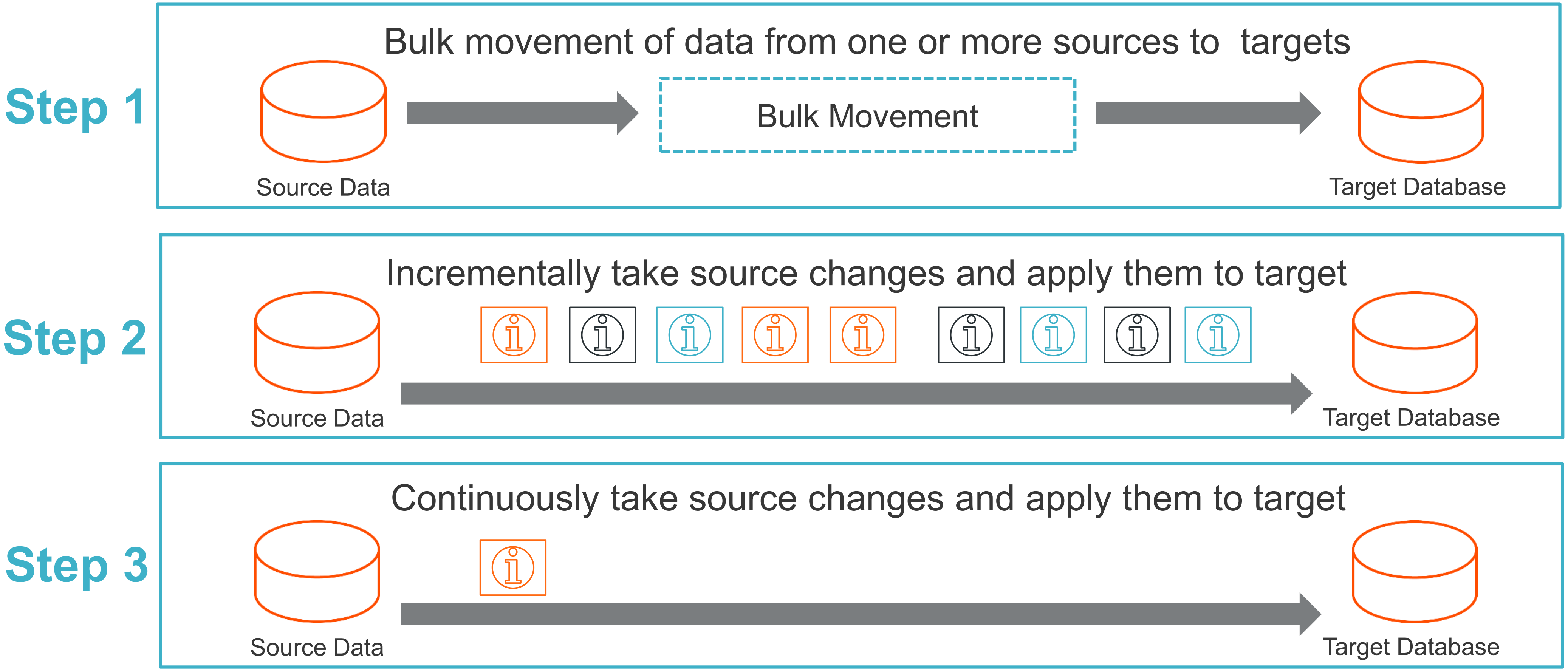
Informatica

Use cases with Change Data Capture

- Real time Data Synchronization/Replication
- Real-time Data Lake Ingestion
- Real-time Streaming Analytics Real-time



Data Synchronization between source & target- Typical flow



Step 1

Bulk movement of data from one or more sources to targets

Source Data

Bulk Movement

Target Database

Step 2

Incrementally take source changes and apply them to target

Source Data

Target Database

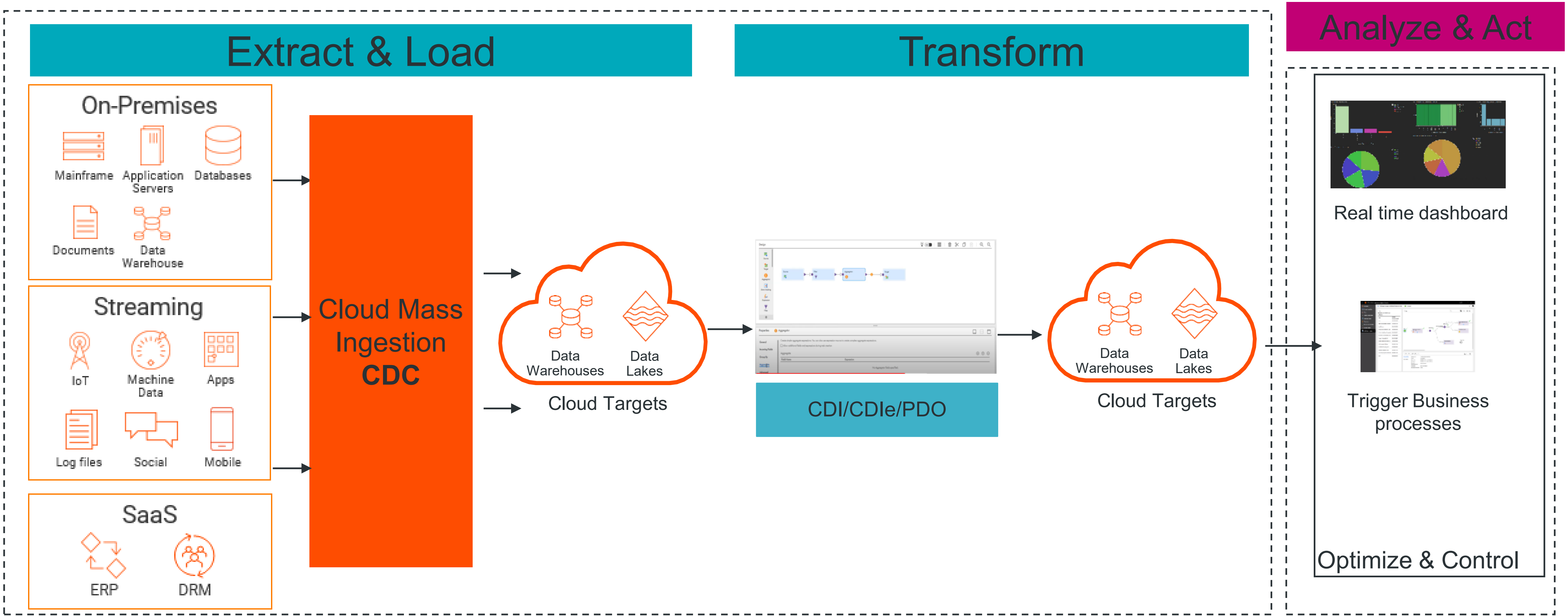
Step 3

Continuously take source changes and apply them to target

Source Data

Target Database

Supporting the Analytics use case



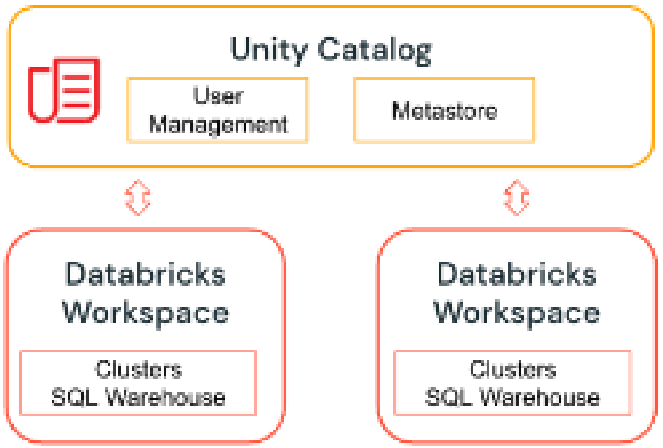
What’s New with CMI

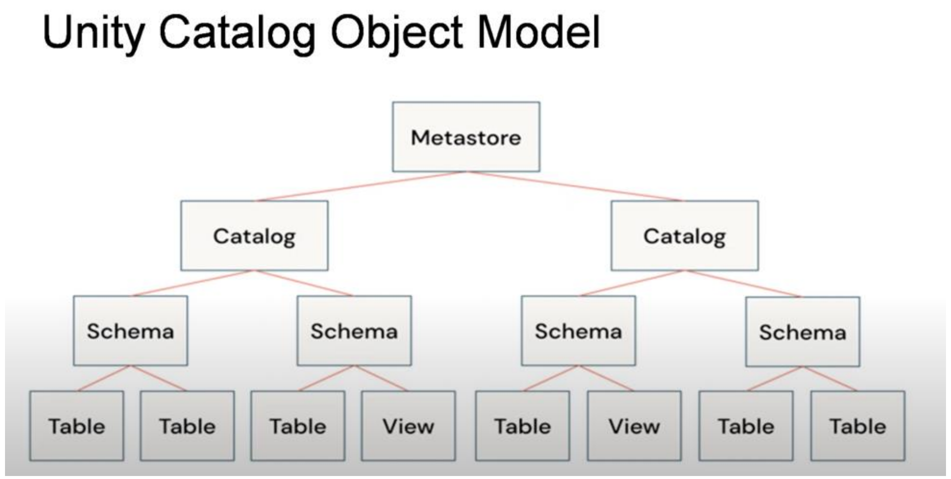
What's new in CMI

- **Cross agent failover - Support for High availability**
- Added support for cross agent failover in case of Oracle and SQLserver sources for incremental load and combined load jobs
- Currently, we need to restart the incremental job afresh when an agent goes down
- With addition of cross agent failover, once the agent goes down the job can restart on another agent with the restart point
- Support added for Oracle and SQLserver source
- As part of next release, we are looking to add support for other sources which doesn’t have dependency on persistence storage

What's new in CMI

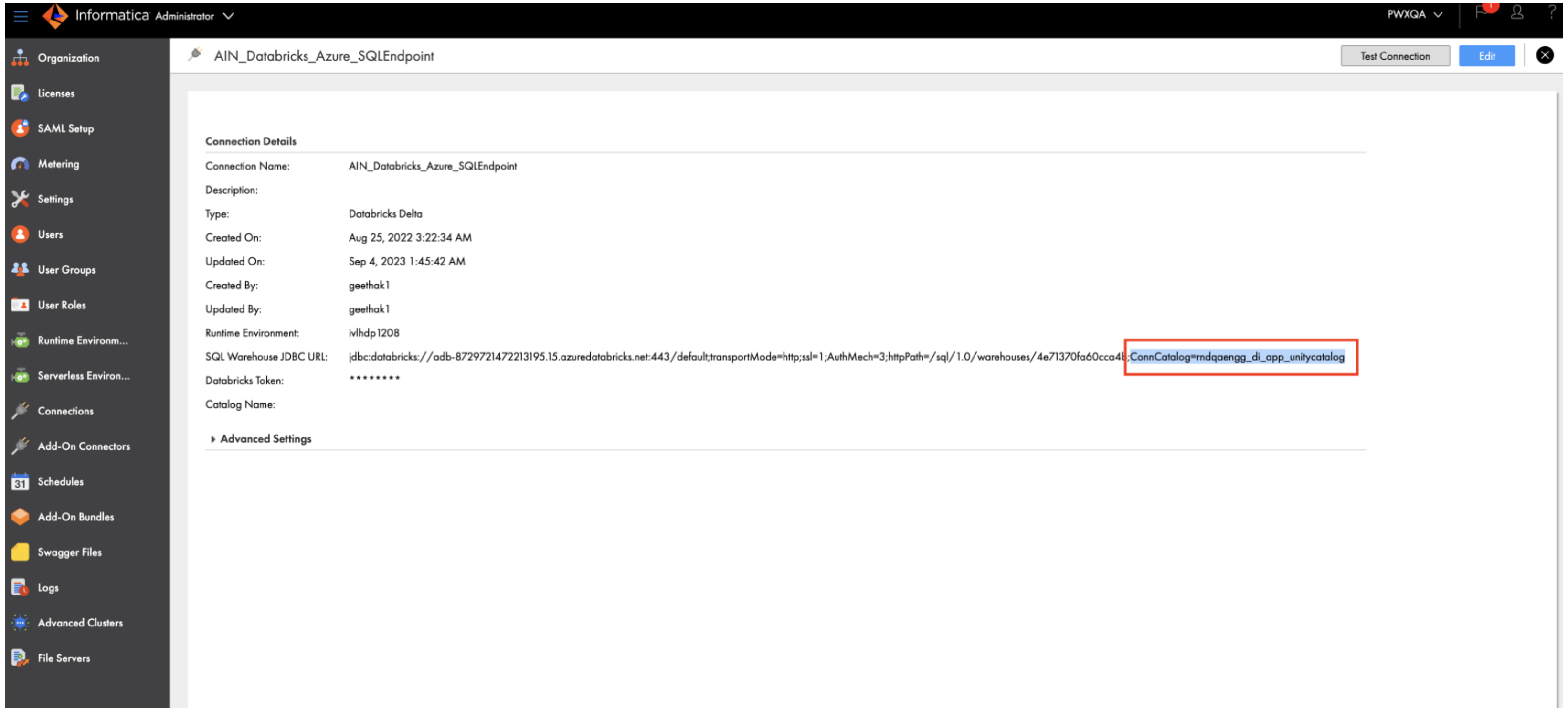
- **Integration with Databricks Unity Catalog**
  - Unity catalog – Databricks unity catalog is intended to unify the governance for data on the lakehouse.
  - Centralized metadata layer simplifies the permission model for data and AI
  - Helps in streamlining monitoring and observability
  - Open data sharing – Easy to share data across cloud regions and platforms





Unity catalog name passed as string

- We need to pass the name of the catalog as part of the JDBC url in the connections page for Databricks



AIN_Databricks_Azure_SQLEndpoint

Connection Details

Connection Name:	AIN_Databricks_Azure_SQLEndpoint
Description:	
Type:	Databricks Delta
Created On:	Aug 25, 2022 3:22:34 AM
Updated On:	Sep 4, 2023 1:45:42 AM
Created By:	geethak1
Updated By:	geethak1
Runtime Environment:	ivlhdp1208
SQL Warehouse JDBC URL:	jdbc:databricks://adb-8729721472213195.15.azuredatabricks.net:443/default;transportMode=http;ssl=1;AuthMech=3;httpPath=/sql/1.0/warehouses/4e71370fa60cca4b;ConnCatalog=rndqaengg_di_app_unitycatalog
Databricks Token:	********
Catalog Name:	

▸ Advanced Settings

What's new in CMI

Query based CDC for Oracle Source

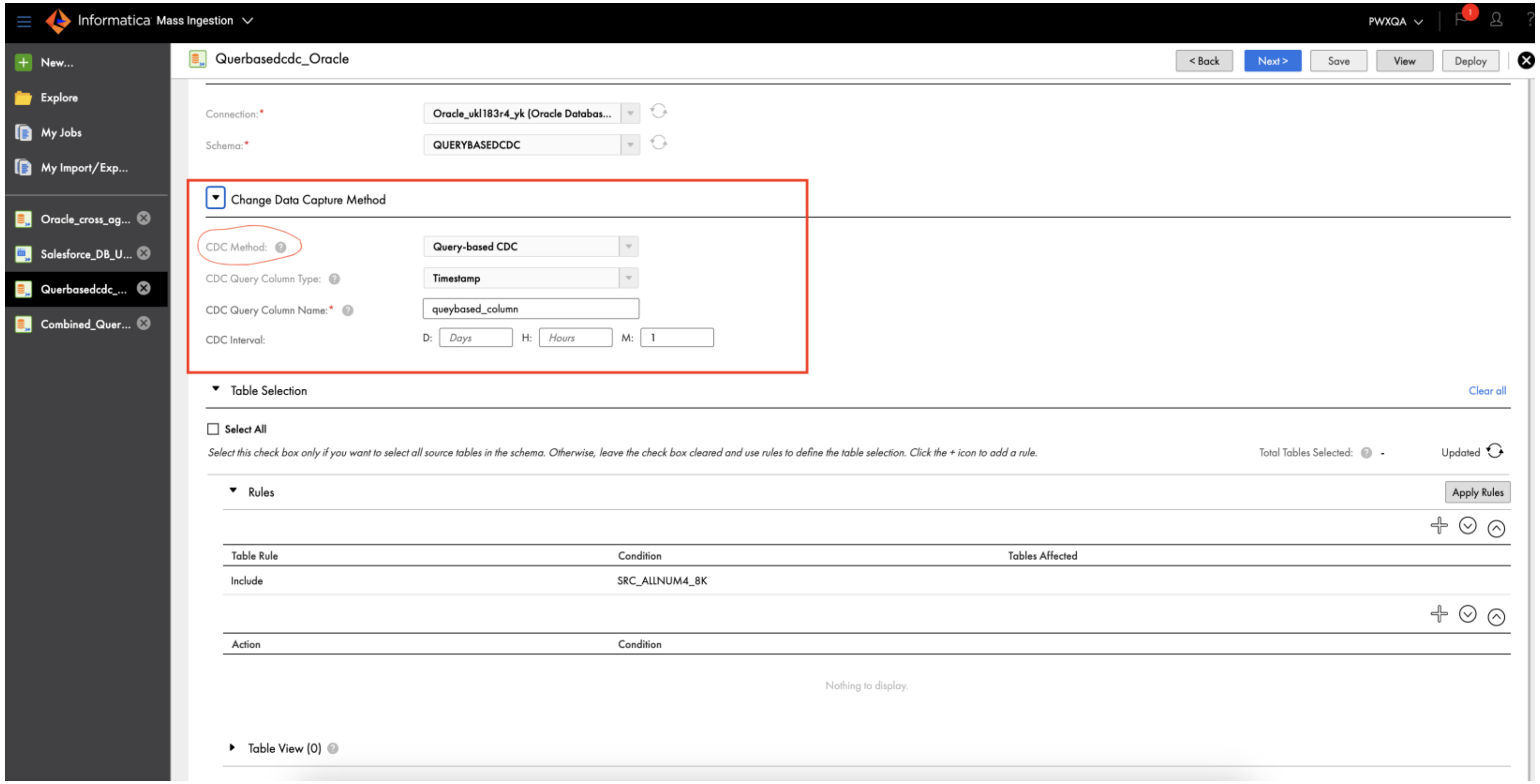
- **For customers who don’t have access to the DB log files, this provides a CDC capability using a SQL based approach**
- We already have support for query-based CDC for SQL Server sources.
- This adds support for Oracle sources for query-based CDC as well.
- This is enabled for Snowflake and SQLserver target

Type	Endpoint
Source	Oracle/SQLserver
Target	Snowflake and SQLserver

Task definition and Screenshot

Source tab

- Select the CDC method as Query based CDC
- Provide the timestamp column
- Define the CDC interval



Querbasedcdc_Oracle

Connection:* Oracle_ukl183r4_yk (Oracle Databas...

Schema:* QUERYBASEDCDC

Change Data Capture Method

CDC Method: Query-based CDC

CDC Query Column Type: Timestamp

CDC Query Column Name:* queybased_column

CDC Interval: D: Days H: Hours M: 1

Table Selection

Clear all

□ Select All

Select this check box only if you want to select all source tables in the schema. Otherwise, leave the check box cleared and use rules to define the table selection. Click the + icon to add a rule.

Total Tables Selected: -

Updated

Rules

Apply Rules

Table Rule	Condition	Tables Affected
Include	SRC_ALLNUM4_8K	

Action	Condition

Nothing to display.

Table View (0)

What's new in CMI

- **CLI support for task edit and redeployment**
- **Use case-**
  - One of our customer wanted to have support from CLI for editing the task definition and redeploy it
- If the task is **not yet deployed**, then **any** value can be updated in the UI
- If the task is **deployed**, then only a **limited set** (**aka. editable**) of properties can be updated from the UI
- If the task is **undeployed**, then as with #1, **any** value can be changed from the UI
- This is available to customers only based on request, to be enabled through environment variable (for the task which is already deployed)

What's new in CMI

Combined load support for Query based CDC in case of SQL Server source

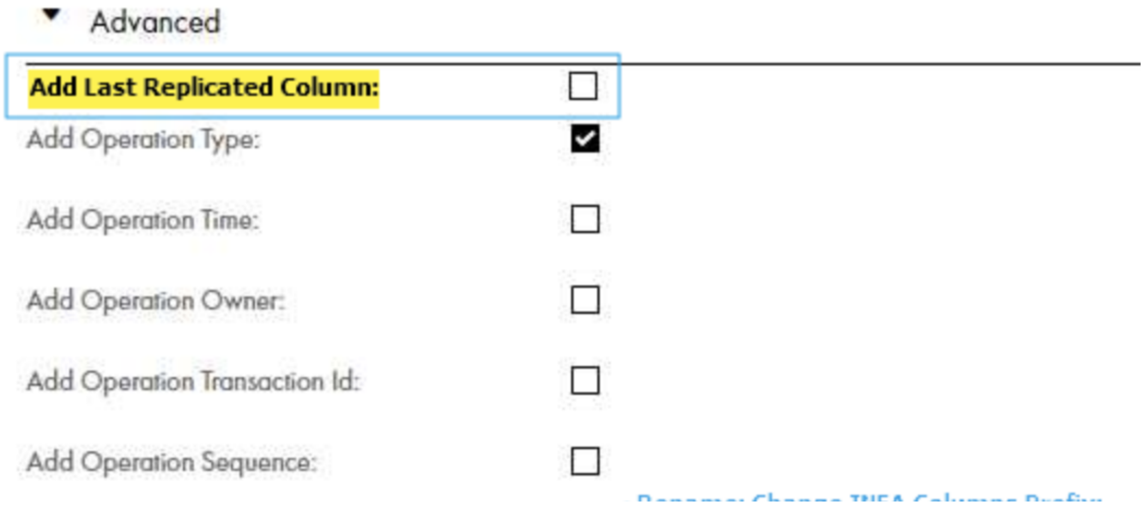
- This improves the ease of use and additional capabilities for ingesting from SQL Server
- Both incremental and combined load are supported now for Query based CDC in case of SQLserver
- This is enabled for all the targets

Azure SQL DB incremental load support

- Certification of Azure SQL DB source for change data capture support
- We certified Azure SQL DB source using the CDC table approach

Timestamp column on Snowflake target

- **Addition of timestamp column** on Snowflake target having the time when the data is written
- New checkbox added to support it for all different apply modes
- This is the timestamp written based on the target end
- Different combination supported –
  - Source – All (DBMI/AppMI)
  - Target – Snowflake



▾ Advanced

Add Last Replicated Column: □

Add Operation Type: ☑

Add Operation Time: □

Add Operation Owner: □

Add Operation Transaction Id: □

Add Operation Sequence: □

Application Ingestion -

- ServiceNow : Support change data capture for objects which are inheriting fields from a Parent object
- Oracle Fusion BICC : Combined load support
- SAP: Combined load support for pool and cluster table

Replication using Superpipe

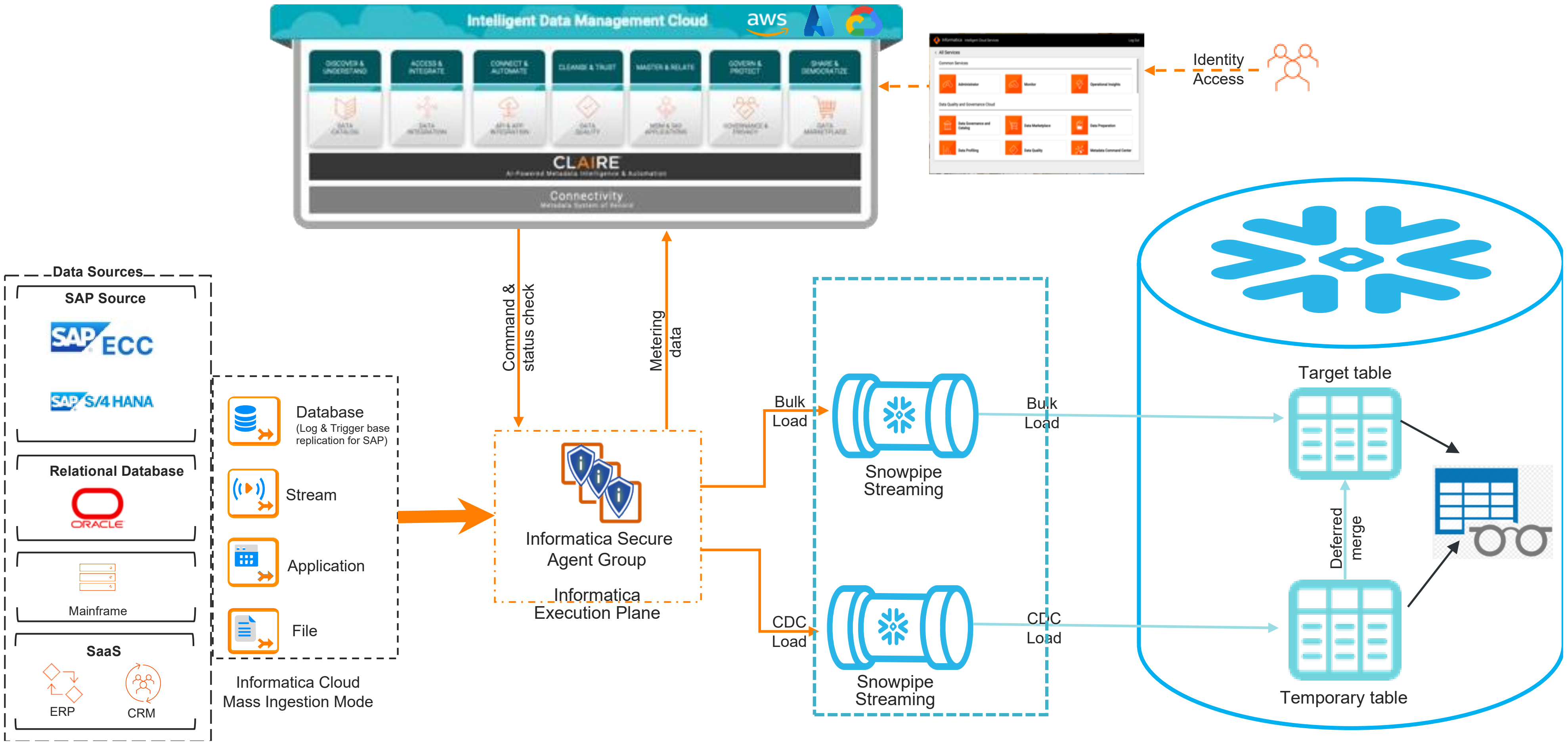
- Superpipe is a feature leveraging Snowflake's **Snowpipe streaming** and **Deferred Merge** for high-performance real-time ingestion into Snowflake Data Cloud.
- Real-time data view is always available regardless of CDC's deferred merge interval set.

✓ Up to 3.5 times performance gains
✓ Up to 40% lesser Snowflake credits consumed

Use case (Tech)	Optimization	Benefit
Real-time Ingestion	Use Snowpipe Streaming to bulk load and CDC on to target table and change table	Snowpipe streaming streams the changes. View created on the final target table and change table helps user have a near real time view always.
Deferred merge	Merge/Apply CDC changes on periodic basis instead of doing on transactional boundaries	Reduces TCO by doing merge in batches while providing real time replicated view of change data

Informatica

Architecture



Informatica

Presenter name / Date

Thank You

Where data
& AI come to LIFE