

Advanced Scanners Architecture, Installation, Capabilities and Updates

Presented by:

Puneet Bhardwaj

Sr. Principal

Topics:

- What are Advanced Scanners
- Why Advanced Scanners
- Overall Architecture with key components
- Advanced Scanners installation and configuration requirements
- Advanced Scanners Capabilities with examples (in slides not demo)
- Advanced Scanners new or future updates
- Q&A

What are EDC Advanced Scanners

1. Automatic lineage extraction from code - both static and dynamic code

Competitors stumble at just scanning static code. EDC Advanced Scanners support scanning both static and dynamic code to provide automated data lineage for use cases such as regulatory compliance, impact analysis and root cause analysis. Examples of dynamic code are parameterized stored procedures and dynamic SQL generated at runtime.

2. Automatic lineage extraction from code embedded in ETL Processes

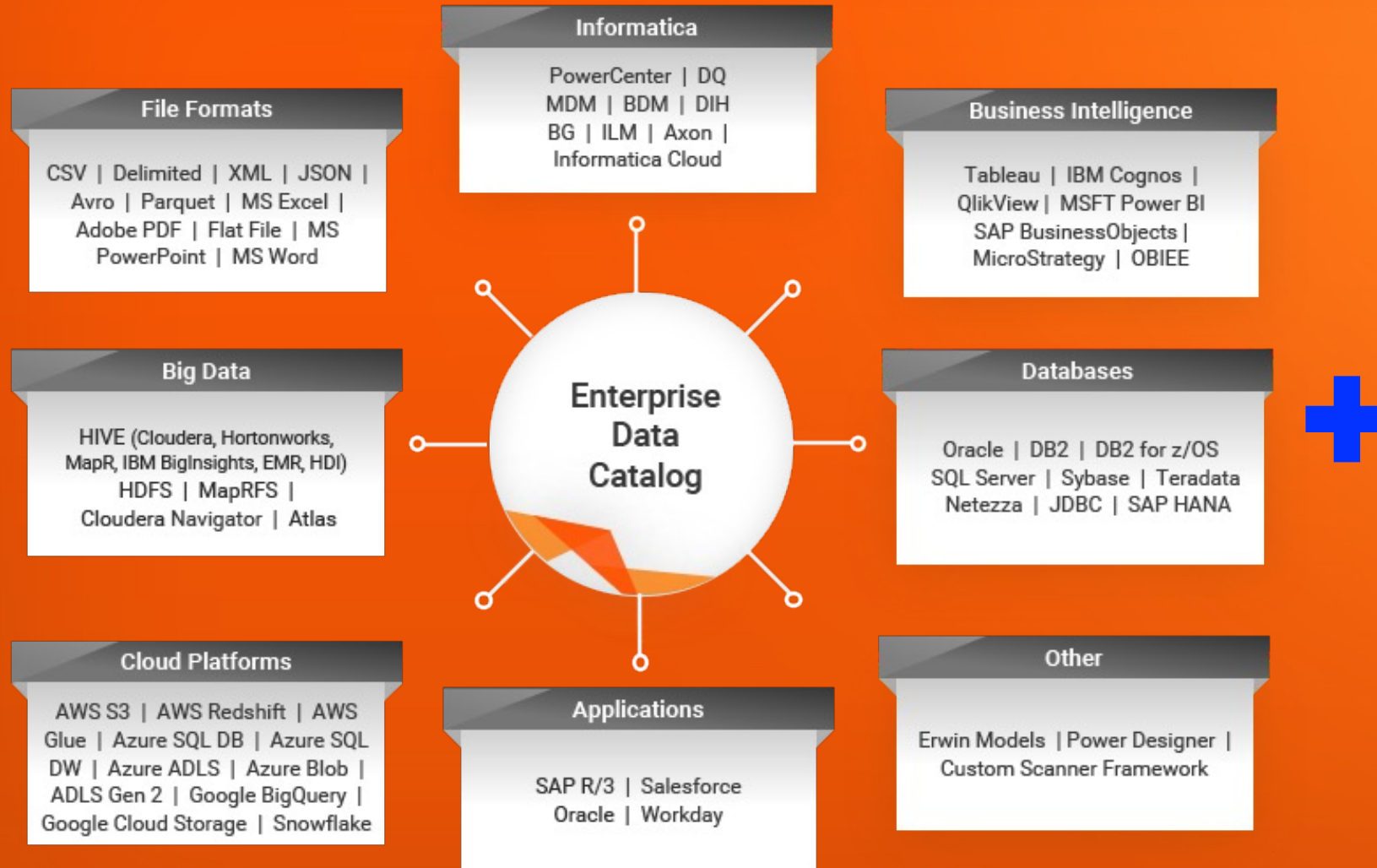
Many customers have a large amount of SQL embedded in their ETL code. EDC Advanced Scanners support automatically extracting lineage from embedded SQL – not just from Informatica PowerCenter, but also from a wide range of multi-vendor ETL tools.

3. Metadata extraction from complex sources

EDC Advanced Scanners can extract metadata from complex and legacy data sources like SAS, SAP BW and Cobol/JCL among others. These sources have remained metadata silos due to the complexity of extracting metadata from their respective repositories. By providing metadata connectivity for these sources, EDC Advanced Scanners fills a big void in the market.

Why Advanced Scanners

Broadest and Most Complete Metadata Connectivity



EDC Advanced Scanners

Code and Scripting

- Oracle
- SQL Server
- DB2
- Teradata
- Netezza
- Sybase ASE

ETL Tools

- IBM Datastage
- Microsoft SSIS

Mainframes

- COBOL
- JCL

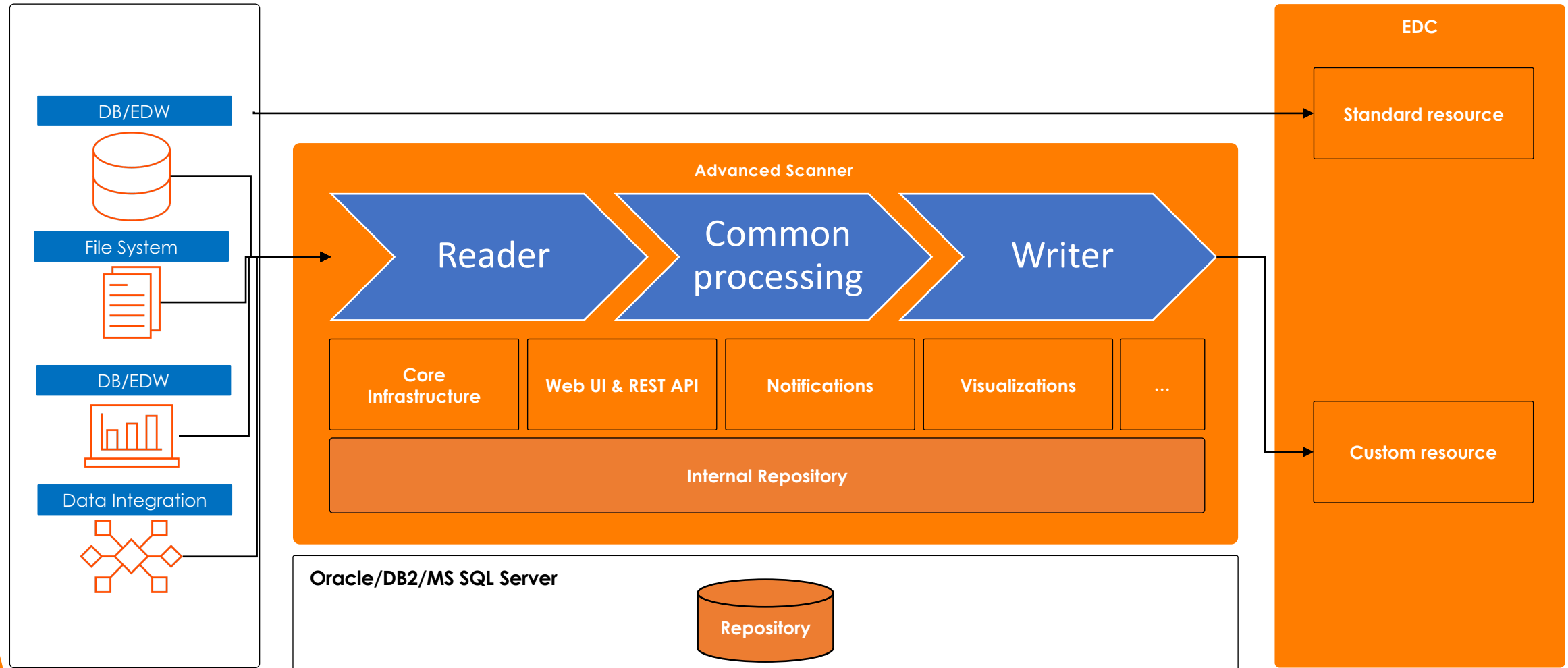
Statistical and BI Tools

- SAS
- Microsoft SSAS
- Microsoft SSRS
- SAP BW
- SAP BW4HANA

Overall Architecture

- Advanced Scanner is a Java based application
- Advanced Scanner has its own UI to configure the scanners
- Can be installed
 - On Linux (can be in the same machine as EDC or not)
 - On Windows (can be on the same machine as the EDC agent)
- Load metadata in EDC through custom scanner framework automatically
- Can be online (direct connection to EDC) / offline (need manual steps to load the metadata)

Advanced Scanner Key Components



Advanced Scanners Installation and Configuration

High Level Installation Requirements

OS: Linux or windows

2 CPU CORE minimum

4GB per configuration
exec

- rare case of complex system require up to 16GB

5GB of disk space

Repository

- Oracle DB 11g, 12c, 18c
- IBM DB2 10.6
- MS SQL Server



Informatica®

Advanced Scanners Scan Capabilities

PowerCenter with Stored Procedures

PowerCenter Base Support

PowerCenter Scanner scans mappings, mapplets and workflows

- Input files
- Calculations
- Transformation Details including column level lineage

Supports Stored Procedure lineage executed via PowerCenter

- Connects to database metadata tables to get the stored procedure lineage

SAS

SAS base support

SAS Scanner brings metadata from SAS programs for

- SAS data sets
- Calculations
- Transformation Details including column level lineage

Supports reading from SAS files directly

- No direct connect to SAS server.

Can also extract SAS program execution from logs

SSAS/SSRS

SSAS/SSRS Scanners

Provide end to end lineage from analytical sources to reports

Extract Cube definition from SSAS

- SSAS Projects
- Data Sources, Cubes, Measure Groups, Hierarchies, Dimensions, Calculations, KPIs, Perspectives

Extract Reports definition from SSRS

- SSRS Projects,
- DataSets, Reports including lineage at field level

DATASTAGE

DataStage Scan Requirements



Provide access to IBM DataStage jobs in the scope including all dependencies (i.e. parameter sets and containers) in form of exported .isx files on the Advanced Scanner server.

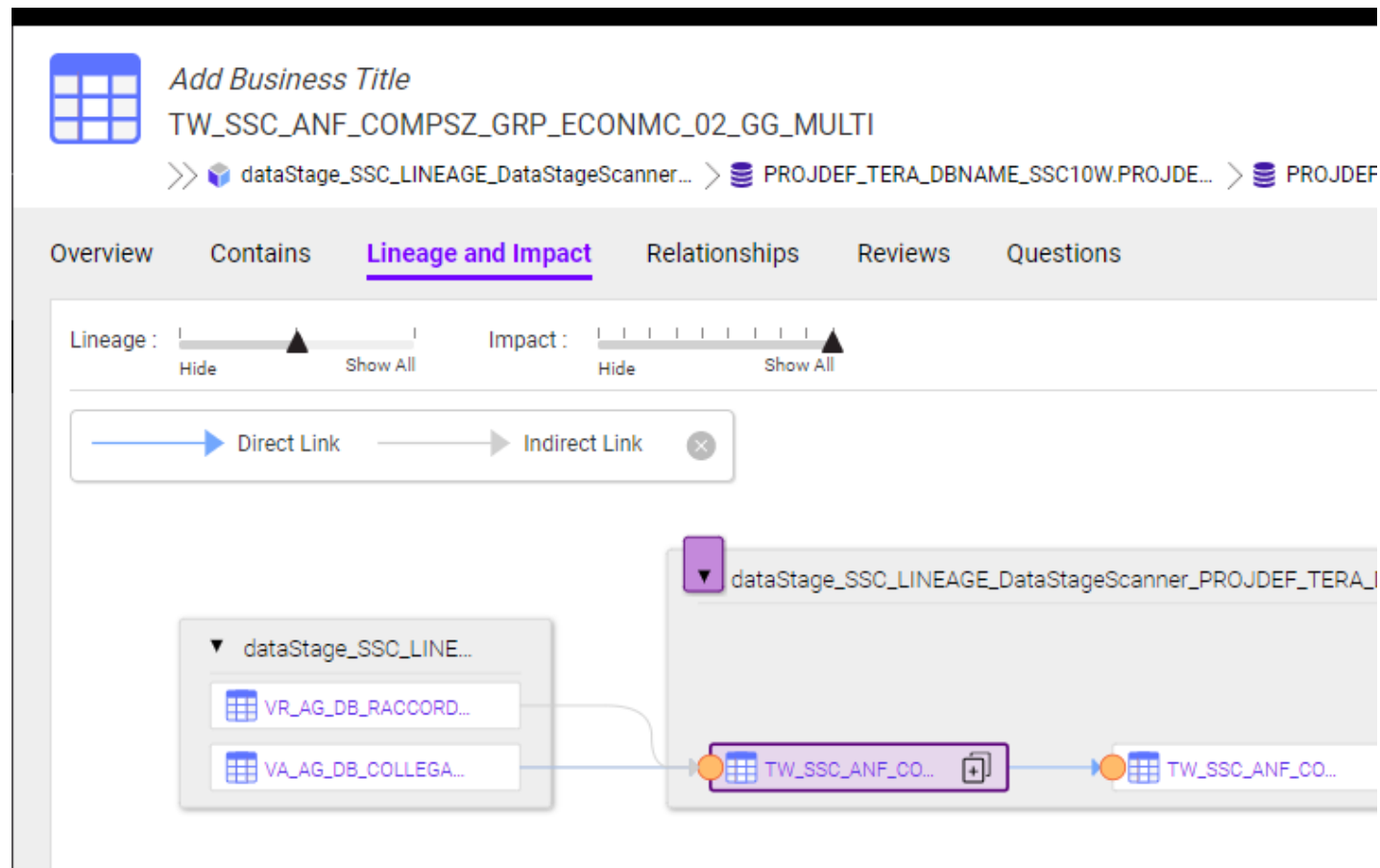


Export IBM DataStage jobs using IBM IS tool.

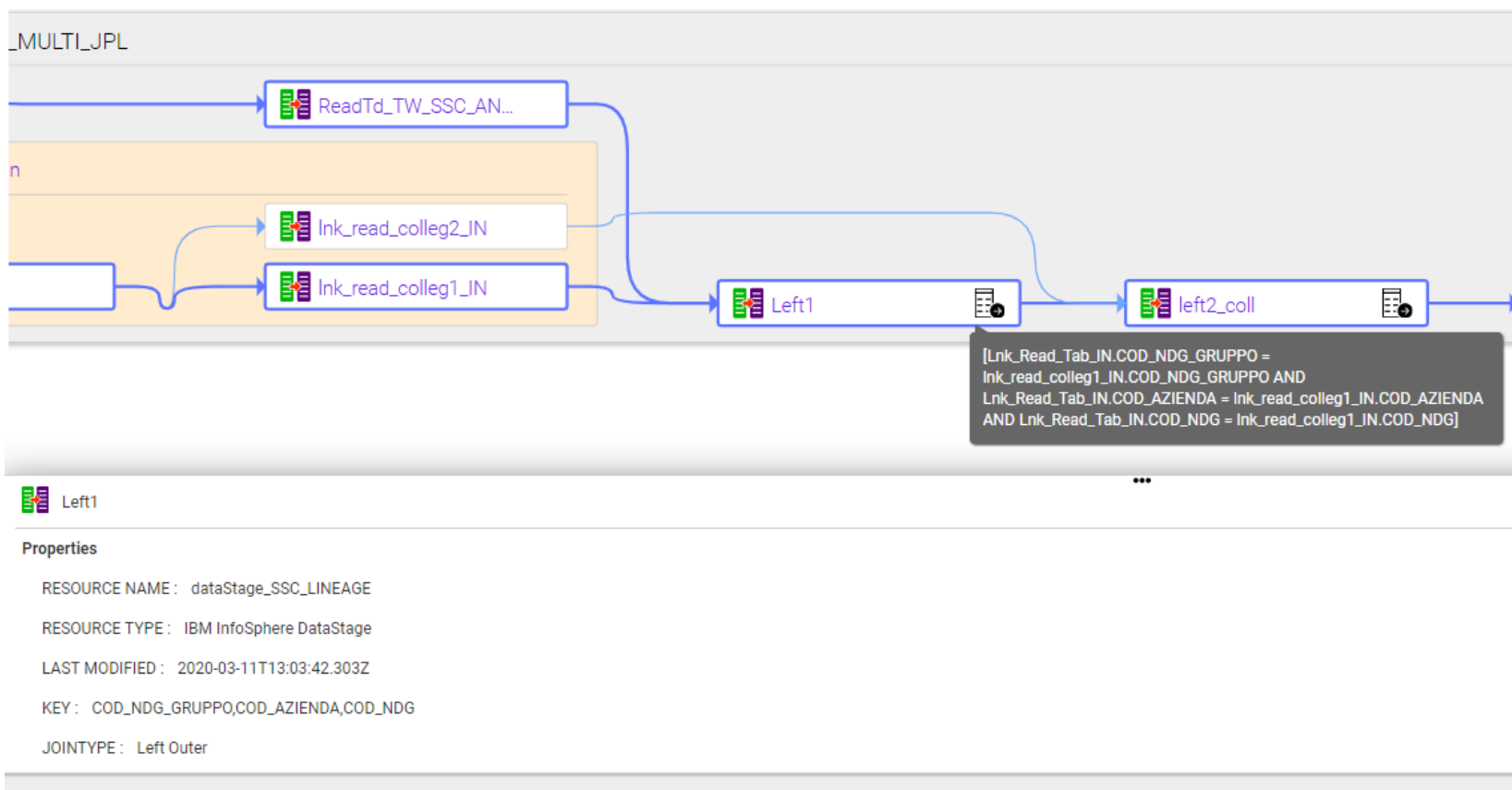


Provide access to all the databases referenced by the IBM DataStage jobs in the scope.

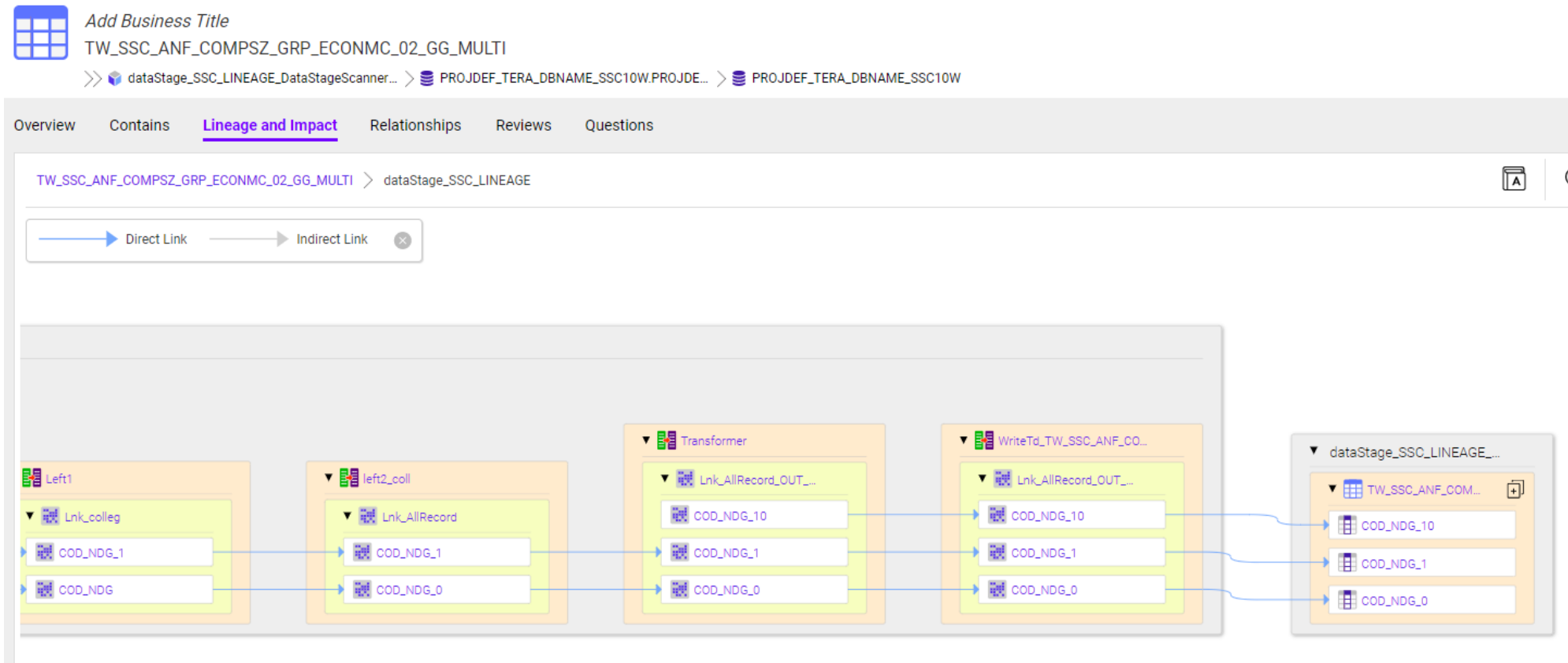
DataStage Transformation Details



DataStage Join Details



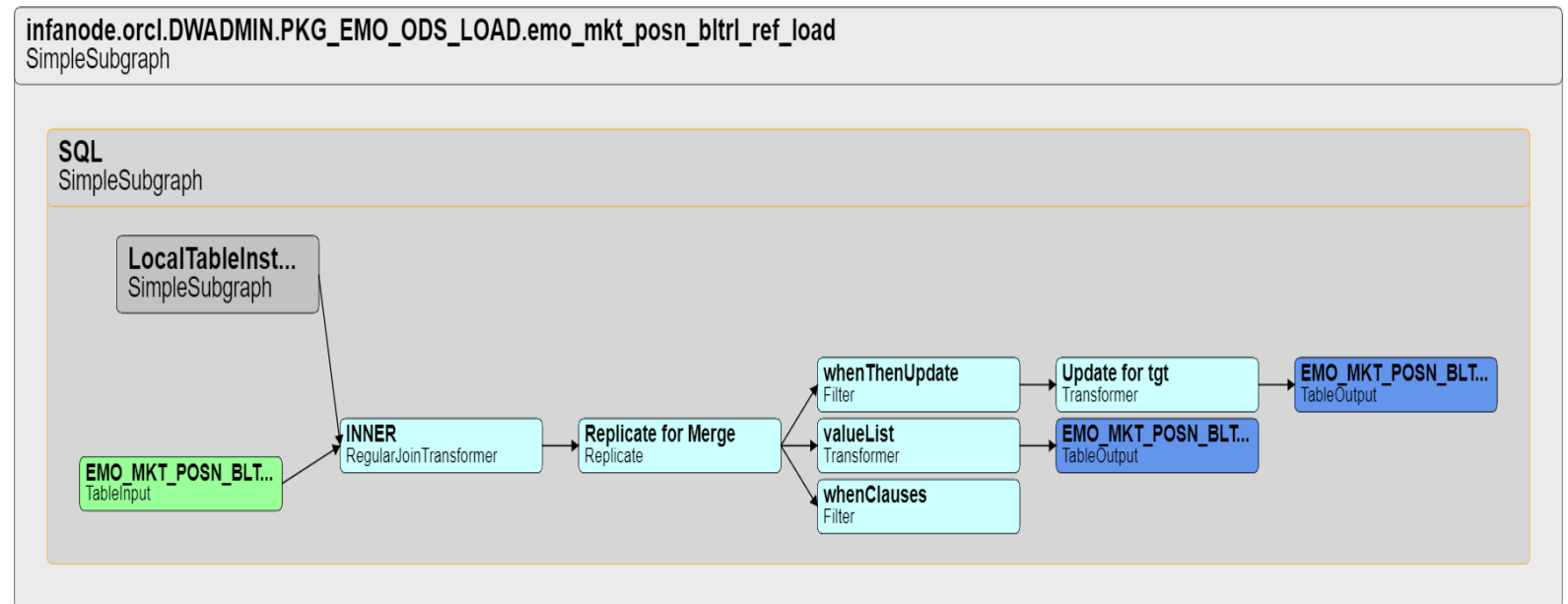
DataStage Attribute Level Details



Advanced Scanners Advanced Capabilities

Advanced Visualizations

- Static visualizations
- Decompose the SQL statements as individual transformations
- Made Available as a HTML link from EDC
- Mapping report



Advanced Logging capabilities

Tree view of the extracted object


Filter by log level



Easy navigation to log message related to a particular object

Log Analysis

- Breakdown analysis of issue types per source container (database, schema, etc.)

Oracle_SP > ora_ptnt_db


Processing 39902  Mar 3, 2020, 4:38:00 PM - Mar 3, 2020, 4:38:38 PM

Logs	Log analysis	Output	Snapshot	Generated objects
Q	INFO	Log code		 Download
16:38:31.550	O	ERROR	Could not resolve SQL text due to unknown value. Potential metadata data needed: psvlxmlfapp1.orcl.	
16:38:31.551	O	WARN	Unable to process iteration: 101 since maximum number of iteration is set to: 100	
16:38:31.555	C	INFO	Analysing lineage for Graph: psvlxmlfapp1.orcl.TEST_DB_SCRIPT.*__orcl.TEST_DB_SCRIPT.RUN_ETL()	
16:38:31.555	C	INFO	Lineage analysis finished.	
16:38:31.561	C	INFO	Executable hierarchy entity not found for executable /psvlxmlfapp1/orcl/TEST_DB_SCRIPT/RUN_ETL()	
16:38:35.777	C	INFO	No image to copy for HTML.	
16:38:35.785	C	INFO	Writing HTML to /opt/outputFiles/html/psvlxmlfapp1.orcl.TEST_DB_SCRIPT.*__orcl.TEST_DB_SCRIPT.RUN_ETL()	
16:38:35.787	C	INFO	Finished writing /opt/outputFiles/html/psvlxmlfapp1.orcl.TEST_DB_SCRIPT.*__orcl.TEST_DB_SCRIPT.RUN_ETL()	

Processing 39902

- psvlxmlfapp1
 - orcl
 - TEST_DB_SCRIPT
 - PKG_ETL
 - RUN_ETL
 - RUN_ETL()

Oracle_SP > db_exp_GAA

Processing 15909  Aug 26, 2019, 1:48:48 PM - Aug 26, 2019, 1:49:03 PM [Edit configuration](#) [Delete](#)

Logs	Log analysis	Output	Snapshot	Generated objects
WARN				
Unknown text		See logs	2	50.00%
Host: ASW201601			2	50.00% 100.00%
Database: PDB			2	50.00% 100.00%
Schema: GAA_ADMIN			2	50.00% 100.00%
Function: FMTWRP_APPLICATION_RUNTIME.fn_truncate			1	25.00% 50.00%
Variable: in_vschema_name			1	25.00% 100.00%
Function: FMTWRP_APPLICATION_RUNTIME.fn_fmtwrp_sp_invoke			1	25.00% 50.00%
Variable: in_vproc_name			1	25.00% 100.00%
Other			2	50.00%





Security

Supports both build-in and LDAP users for authentication

Permission can be granted to the individual accounts or groups







System-level privileges ("Global roles")

- Admin
 - Permission to create, edit and delete projects
 - Allows for projects roles assignment
 - Permission to upload, edit and delete files in the allowed server location
 - ...
- Repository Viewer
 - Allows viewing Advanced Scanner repository content
 - Allows accessing built-in HTTP server

	Assigned	Effective
Admin		
Repository Viewer		

Security

- Project-level privileges (“Project roles”)
 - Viewer – reads only access
 - Operator – executes configurations and defines email notifications
 - Developer – manages and executes configurations
 - Owner – full permission including project level details

Project	Viewer ?	Operator ?	Developer ?	Owner ?	
	Set to all	Set to all	Set to all	Set to all	
Mainframe Demo					
ODI Demo					
SAS Demo					

Advanced Scanners Updates

Core Update

Security and usability improvements

- Lots of library updates:
 - Java version updated to 8.0.322
 - Multiple updates related with detected CVEs and others (Spring, Tomcat, DataDirect drivers, Angular and other smaller)
 - H2 database version updated to 2.1.x line (changed jdbcUrl syntax, behavior and parameters)
 - Please remember to use Microsoft driver for internal repository always! (shipped with package, supports domain login also from linux through NTML)
- Additional optimization of log handling for big executions
 - Prevents "HikariCP-LogRepository - Connection is not available" error
- Security: limit uploaded file types
 - By extension, property scanners.web.security.restrictedFileExtensions, default .exe,.jsp,.js
- Security: automatic encrypting of sensitive global variables, project variables, system properties and environment properties
 - Additional patterns may be configured: scanners.security.additional-keys-to-sanitize

Core Update

Security and usability improvements

- Generated Objects are now decorated with execution status (warns/errors)
- Handling of column names ending with a space with Sql Server as internal repository
 - "some column " becomes "some column␣"
- Trimming invisible chars from project and blueprint name (were visible only when exported)
- Verification for duplication of core jar(s) file
 - Customers often update installation by copying new files onto old ones instead of replacing them entirely
 - Currently will work only in some situations (updated version loaded first), but with time all versions will have that check inside

Core Update

Upcoming changes (10.5.2)

- Product rename from Advanced Scanners to MetaDex
- Execution rename from Processing to Job
- Increase default server memory limit from 2GB to 4GB
 - Thanks to that bigger graphs can be visualized out of the box
- Execution report available as Excel file
 - Like HTML reports manual turning them on required
- Finished with warns and finished with errors will be converted to finished (so MetaDex doesn't exited unexpected) with objects statistics (finished successfully/with warns/with errors)

CATALOGS Update

- Internal improvements
 - Models development
- A very small but hopefully very helpful log message improvement

```
Jobs for resource SomeResource could not be completed. The following tasks did not  
end successfully: SCAN_JOB={Staging=FAILED}.
```

changed to:

```
Jobs for resource SomeResource could not be completed. The following tasks did not  
end successfully: SCAN_JOB={Staging=FAILED}. To resolve this error, see the job logs in  
the Catalog Administrator tool.
```

ETL

ETL: Salesforce

Salesforce

- Authentication mode (salesforceAuthMode)
Choose authentication mode to read metadata from s...
- REST API base URI* (salesforceRestApiBaseURI)
The baseURI that is used to connect to Salesforce ...
- REST API user (salesforceRestApiUser)
The user that is used to connect to Salesforce RES...
- REST API password (salesforceRestApiPassword)
The password of user that is used to connect to Sa...
- REST API Security Token (salesforceRestApiSecurityToken)
The Security token that is used to connect to Sale...
- REST API Consumer Key* (salesforceRestApiConsumerKey)
The ConsumerKey of user that is used to connect to...
- REST API Consumer Secret* (salesforceRestApiConsumerSecret)
The ConsumerSecret of user that is used to connect...
- Salesforce Application Filters* (salesforceFilter)
Salesforce Application filters that is used to app...
- REST API Refresh Token* (salesforceRefreshToken)
The Refresh token generated in Salesforce using th...

OAuth 2.0 Username-Password Flow

OAuth 2.0 Refresh Token Flow

ETL: Salesforce: EDC Resource



Add Business Title

Test_Salesforce_MDX-26224

Overview

Lineage and Impact

Relationships

Change Summary

▼ Description

No description found. Click the "✎" icon to add a description for this asset.

▼ Resource Contains



818

Objects



149

List Views



4

Triggers



1

Organization



ETL: Salesforce: EDC List View Properties

Add Business Title
All Customers
>> Test_Salesforce_MDX-26224 > INFA

Overview Contains Lineage and Impact Relationships

▼ Description

No description found. Click the icon to add a description for this asset.

▼ Contains(5/6) [Show All](#)

Name	Business Title	Asset Type
Created Date		List View Column
Customer ID		List View Column
Last Modified Date		List View Column
Name		List View Column
System Modstamp		

SELECT TotalLifeTimeValue, Name, Id, CreatedDate, LastModifiedDate, SystemModstamp FROM Customer ORDER BY TotalLifeTimeValue ASC NULLS FIRST, Id ASC NULLS FIRST

▼ People

Data Owner
not assigned

Data Steward
not assigned

▼ System Attributes

Id
00B5g00000VrVSn

Last Modified
Feb 21, 2022, 05:28 PM

Query
SELECT TotalLifeTimeValue, Name, Id, CreatedDate, LastModifie...

Resource Name
Test_Salesforce_MDX-26224

Resource Type
Salesforce Advanced Scanner

Scope
everything

Asset Type
List View

ETL: ODI Exporter

Odi Reader Module

■ Method for providing input files (odiInputMode)

How input files are provided to the reader...

Exporter

■ Input files (inputFiles)

The paths of files that should be available during...

■ Oracle_Home folder* (odiExporterOracleHome)

Path to Oracle_Home folder of ODI Studio installat...

/opt/Oracle/Middleware/Oracle_Home

■ ODI repository Connection registry entry* (odiExporterRepositoryConnectionRegistryEntry)

The 2 segment entry pointing to Connection registr...

■ ODI Studio user* (odiExporterOdiUser)

The name of the Odi Studio user...

■ ODI Studio password* (odiExporterOdiPassword)

The user password to Odi Studio installation...

■ WORK REPO name* (odiExporterWorkRepo)

The name of the WORK_REPO with user projects...

■ Export directory* (odiExporterExportDir)

The directory where the export file will be saved...

■ Exporter timeout in minutes* (odiExporterTimeoutInMinutes)

The timeout for Odi Exporter to prepare input file...

10



ETL: IICS

Informatica Reader

<div><div>IICS Cloud Url (iicsCloudUrl)</div><div>IICS Cloud Url...</div></div>	<input type="text"/>
<div><div>IICS Cloud User (iicsUser)</div><div>IICS Cloud User...</div></div>	<input type="text"/>
<div><div>IICS Cloud Password (iicsPassword)</div><div>IICS Cloud Password...</div></div>	<input type="password"/>
<div><div>IICS Filters List (iicsFilters)</div><div>Case sensitive include/exclude filters that allow ...</div></div>	<input type="text"/>

<div><div>Informatica file variables (informaticaParamFilesMap)</div><div>List of Informatica files mapped to local files, a...</div></div>	From	To
	<input type="text"/>	<input type="text"/>
<div><div>Informatica connections (connections)</div><div>List of Connections obtained from Informatica Powe...</div></div>	Database Connection key	Resolved MDREPO entry
	<input type="text"/>	<input type="text"/>
<div><div>Explicit parameters file (informaticaOverrideParamsFile)</div><div>Additional Informatica parameter file with overrid...</div></div>	<input type="text"/>	
<div><div>Session instances to be included during processing (includeSessionInstances)</div><div>Expression which matches full path to session inst...</div></div>	<input type="text" value="FL:*.WF:*.SI:*.MP:."/>	
<div><div>Session instances to be excluded during processing (excludeSessionInstances)</div><div>Expression which matches full path to session inst...</div></div>	<input type="text"/>	
<div><div>Informatica MDREPO database (HOST.DATABASE.SCHEMA) (informaticaDatabase)</div><div>Host, db name and schema in MDREPO under which Inf...</div></div>	<input type="text"/>	

BIMS Features

- SSIS
 - passing dynamic variable values between components
 - support for System:: parameters
 - support for \$Package:: parameters
 - support for SSIS variables used in SqlCommand of source component
 - support for direct extraction of packages from SSISDB
- ADF
 - support for datasets of type SFTP
 - support for stored procedures in Lookup activity
- Power BI
 - add Include Personal Workspace and Modified Since filter options
 - support for character escaping in filter expression
 - using GetGroupsAsAdmin API to filter out workspaces on the server side
- Security
 - disable XML entity resolving in ADF XML reader

LANG

Lang Features

- Hardening multiple scanners:
 - ADF: expressions + Databricks Evaluator
 - Databricks + python
 - Improved output model in 2022.03
 - Databases: Athena, Hive, Oracle , PostgreSQL family, Snowflake, SQL Server
 - Mainframe
 - JCL
 - improved parameters resolution
 - COBOL
 - PowerQuery
 - QlikView
 - SAS



Thank You



Questions