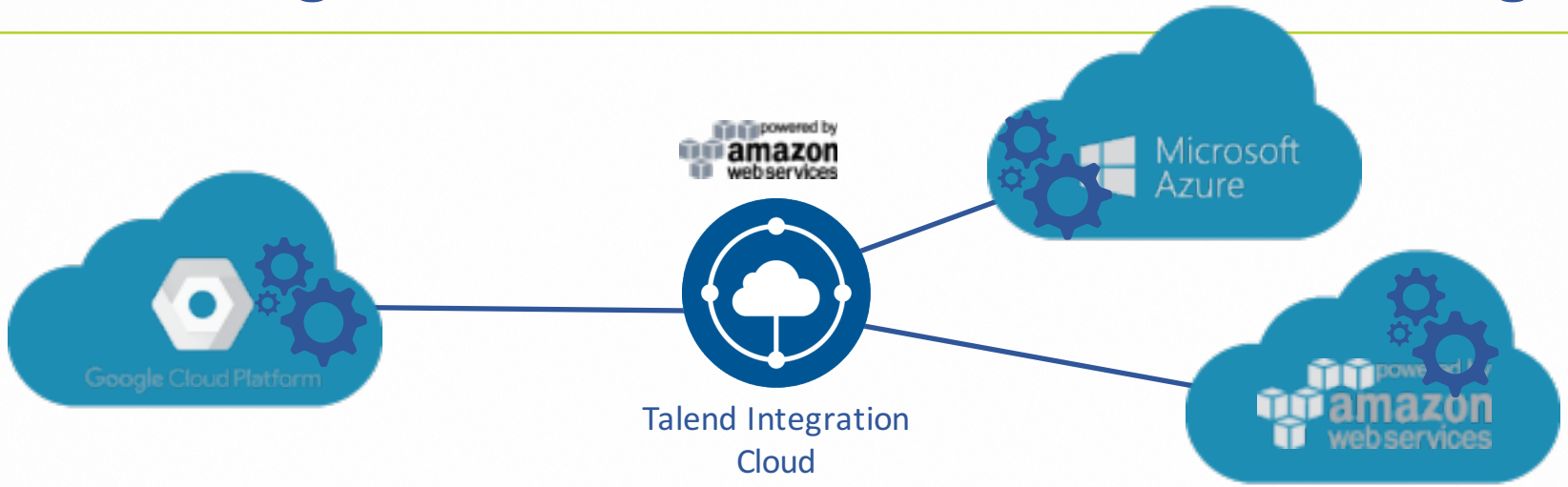
Talend Summer ‘17

Technical Overview

Talend Integration Cloud

Talend Integration Cloud: Multi-Cloud Remote Engine



- Support of Remote Engine on Google, Azure, AWS
- Unrestricted number of parallels jobs execution in Remote Engine

Summer ‘17

- Remote Engine limitation has been removed

➔ You can now run unlimited number of flows in parallel

- Support of Remote Engines on AWS, Azure and Google Cloud Platform

➔ You can run job in your own cloud environment

- No maintenance window planned for the upgrade

➔ Upgrade will be transparent

Google Cloud Connectors

Google Service	Data Integration	Spark	Spark Streaming
Google Dataproc	N/A	Supported	Supported
Google BigQuery	Supported	Supported	Supported
Google Pub/Sub	Not Supported	Not Supported	Supported
Google Cloud Storage	Supported	Supported	Supported
Google Bigtable	Not Supported	Not Supported	Not Supported
Google Cloud SQL	Compatible	Compatible	Compatible
Google Drive	Supported	N/A	N/A

https://cms.talend.com/display/ST/Integration+Cloud+Toolkit

Azure Cloud Connectors

Azure Service	Data Integration	Spark	Spark Streaming
Azure Blob Storage	Supported	Supported	Supported
Azure SQL Database	Supported	Supported	Supported
Azure Document DB	Supported	Supported	Supported
Azure Storage Table	Supported	Compatible	Compatible
Azure SQL Data Warehouse	Supported	Supported	Supported
Azure Data Lake Store	Not Supported	Supported	Supported
Azure HD Insight	N/A	Supported	Supported
Azure Storage Queue	Supported	Compatible	Compatible
Azure Storage File	Supported	N/A	N/A
Azure SQL Server Stretch	Compatible	Supported	Supported

https://cms.talend.com/display/ST/Integration+Cloud+Toolkit

talend

Big Data

Talend Summer'17 - Big Data: In a nutshell

01 **More Cloud !**
Cloudera Altus, Google, AWS, Azure & more…

02 **More Spark !**
Spark 2.1, Natural Language Processing with Spark DQ, Performances…

03 **(Even) Better Developer Experience**
Spark Properties, Contextualization, Metadata through JobServer…

04 **Q/A**

AND OF COURSE, UPGRADED DISTRIBUTIONS !

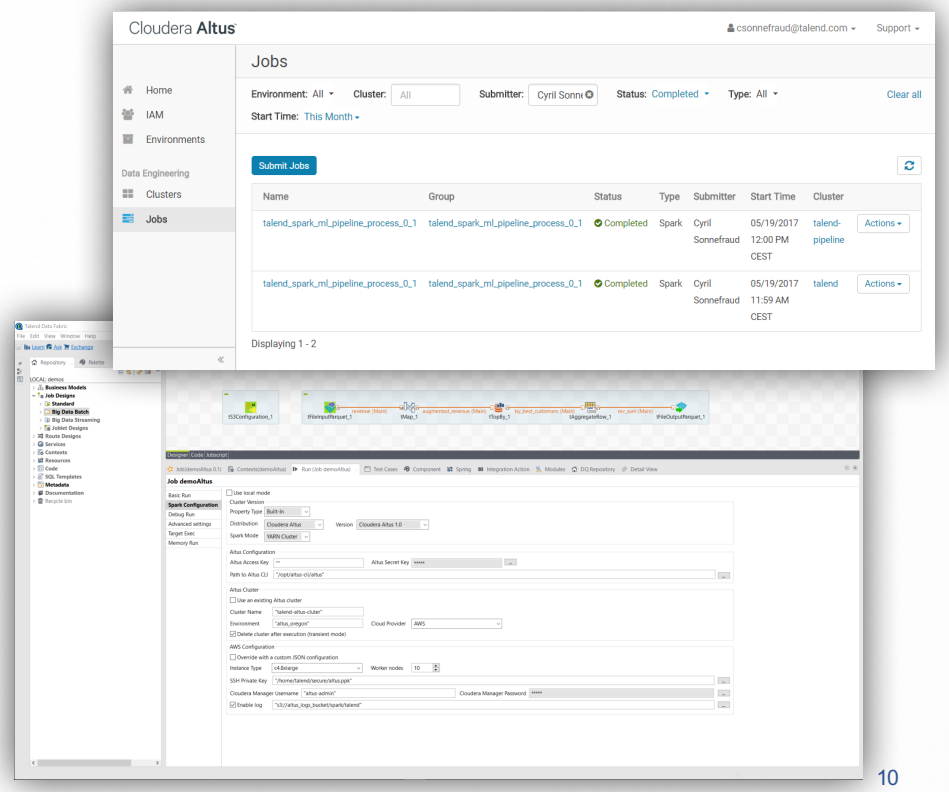
Talend Summer ’17: New supported distributions

Vendor	Distribution	Hadoop version	Spark version
Hortonworks	HDP 2.6	Hadoop 2.7.3	1.6.3/2.1
Microsoft	HDInsights 3.6	Hadoop 2.7.3	2.1
Cloudera	CDH 5.10.1	Hadoop 2.6.0 (~2.7)	1.6.1/2.1
Cloudera	Altus 1.0	Hadoop 2.6.0 (~2.7)	1.6.1/2.1
Amazon	EMR 5.5	Hadoop 2.7.3	2.1
Google	Dataproc 1.1	Hadoop 2.7.3	2.0.2

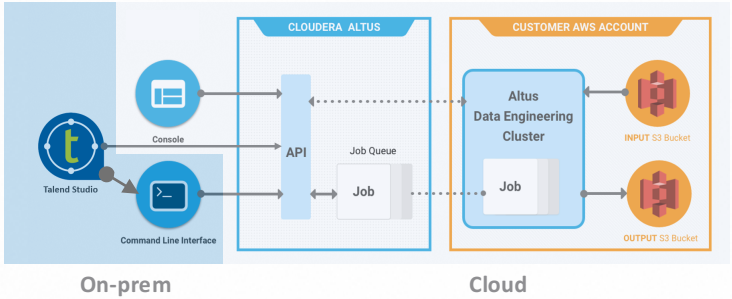
More Cloud! – Introducing Cloudera Altus

Talend is the first tool to integrate with Cloudera Altus. *But what is it ?*

- Cloudera Altus is a PaaS offering, managed Big Data service
- Cloudera 5.11 on AWS (for the first release)
- "*Bring your own AWS account*", and Altus does the rest: provisioning, setup & management
- "*Pay per use*" for Cloudera, eventually with transient clusters

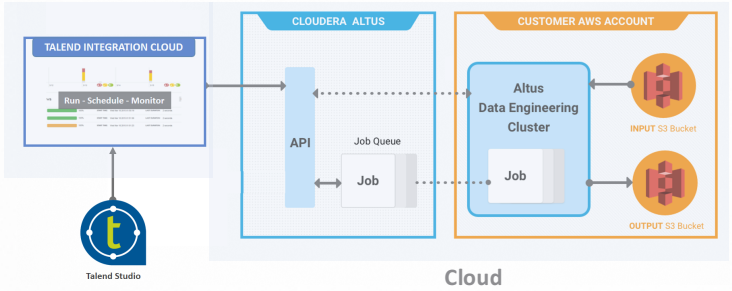


More Cloud ! – Introducing Cloudera Altus



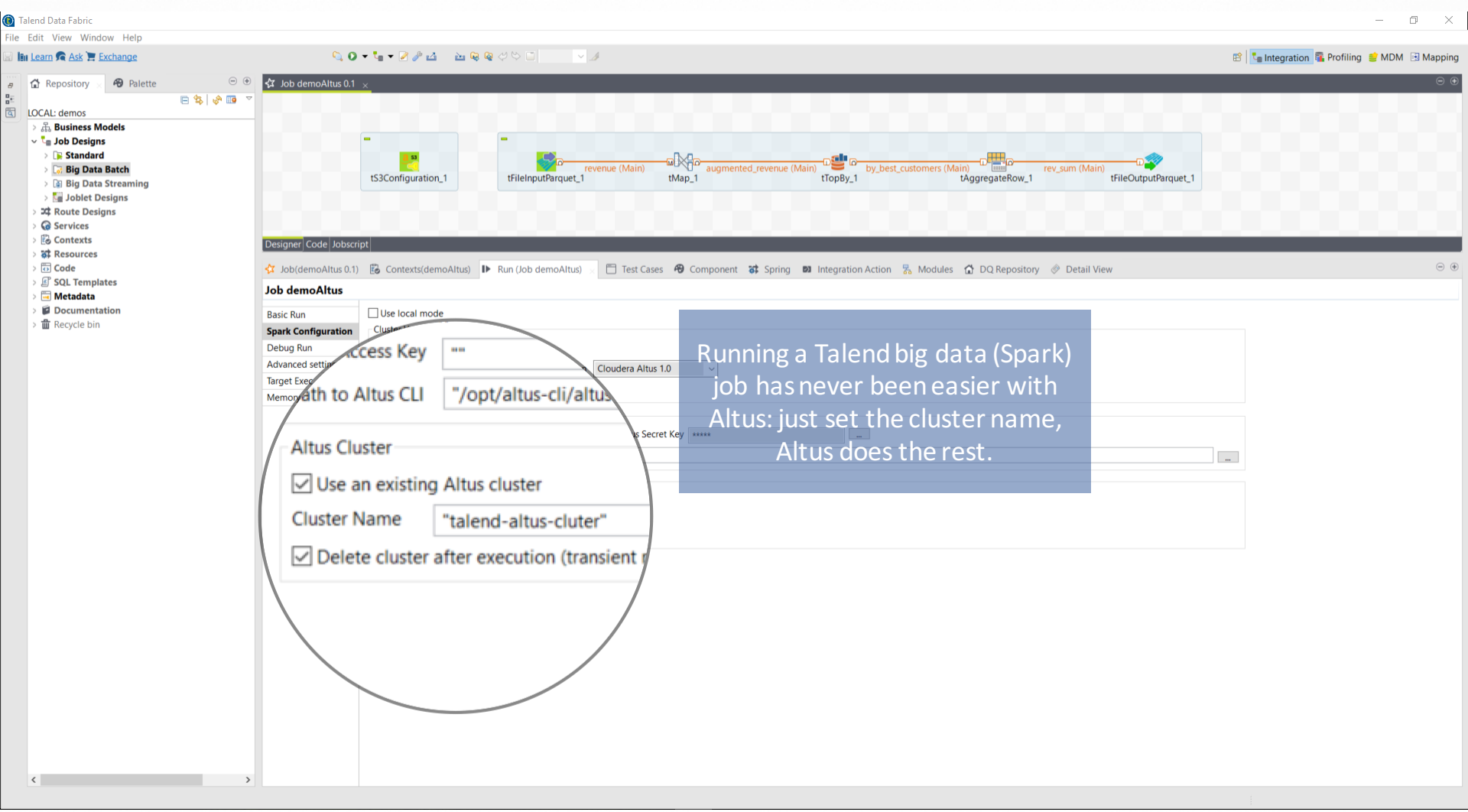
Talend can easily integrate with Cloudera Altus by uploading, submitting a Spark job from on-premises to the remote cluster

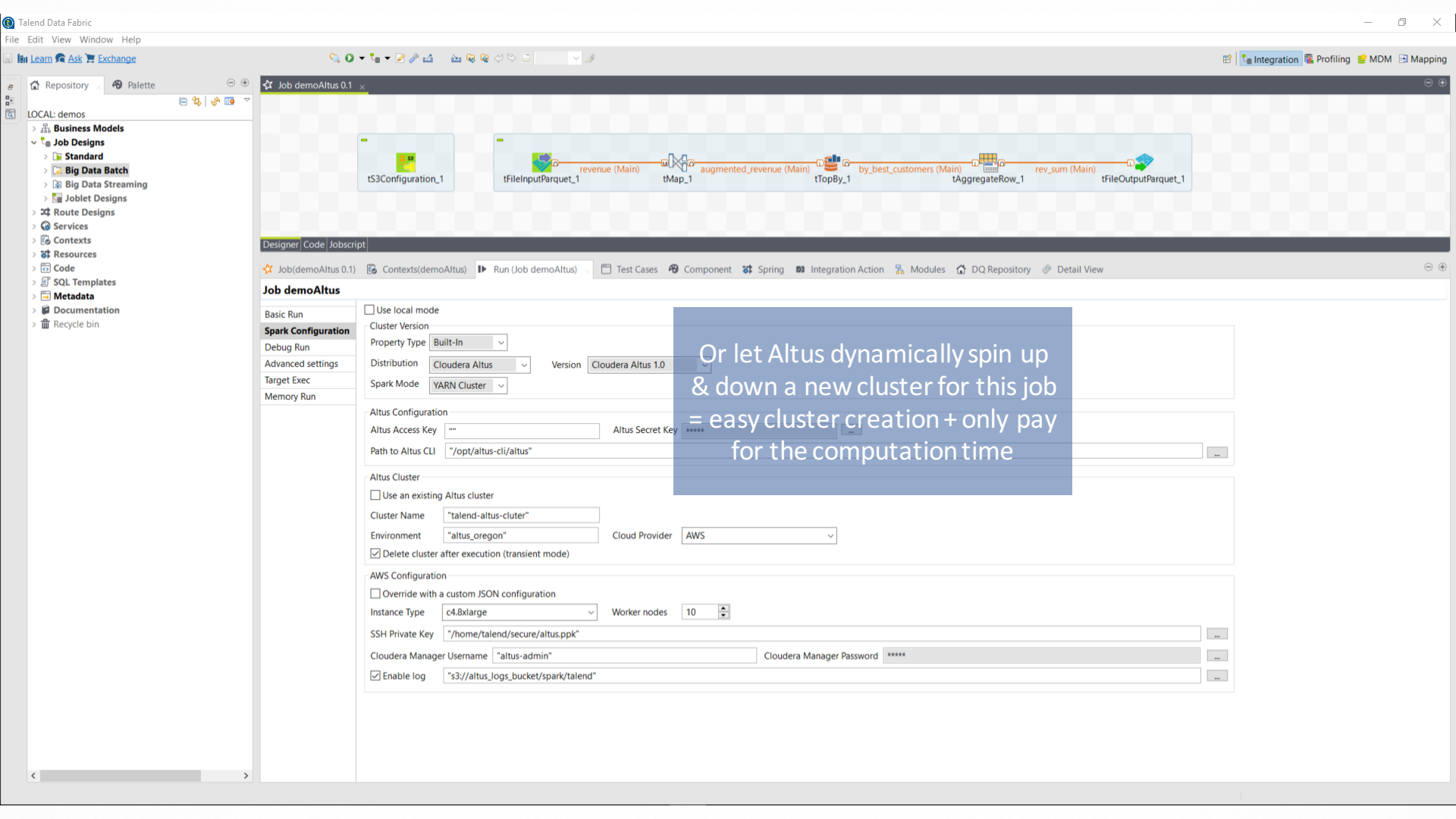
Benefits: Use Cloudera Altus as any other distribution. Hybrid Cloud for Big Data.



It is also possible to have fully managed Big Data integration by leveraging Talend Integration Cloud and Cloudera Altus

Benefits: Fully managed Big Data in the Cloud = Zero-Ops Big Data management.


Talend Data Fabric
File Edit View Window Help
Learn Ask Exchange
Integration Profiling MDM Mapping
Repository Palette
LOCAL: demos
Business Models
Job Designs
Standard
Big Data Batch
Big Data Streaming
Joblet Designs
Route Designs
Services
Contexts
Resources
Code
SQL Templates
Metadata
Documentation
Recycle bin
Job demoAltus 0.1
tS3Configuration_1
tFileInputParquet_1
revenue (Main)
tMap_1
augmented_revenue (Main)
tTopBy_1
by_best_customers (Main)
tAggregateRow_1
rev_sum (Main)
tFileOutputParquet_1
Designer Code Jobscript
Job(demoAltus 0.1) Contexts(demoAltus) Run (Job demoAltus) Test Cases Component Spring Integration Action Modules DQ Repository Detail View
Job demoAltus
Basic Run
Spark Configuration
Debug Run
Advanced settin
Target Exec
Memor
Use local mode
Cluster
ccess Key
""
Cloudera Altus 1.0
ath to Altus CLI
"/opt/altus-cli/altus
s Secret Key
*****
Altus Cluster
Use an existing Altus cluster
Cluster Name
"talend-altus-cluter"
Delete cluster after execution (transient
Running a Talend big data (Spark) job has never been easier with Altus: just set the cluster name, Altus does the rest.

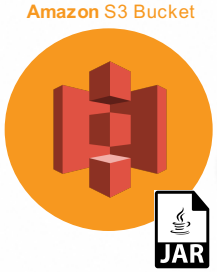

Talend Data Fabric
File Edit View Window Help
Learn Ask Exchange
Integration Profiling MDM Mapping
Repository Palette
LOCAL: demos
Business Models
Job Designs
Standard
Big Data Batch
Big Data Streaming
Joblet Designs
Route Designs
Services
Contexts
Resources
Code
SQL Templates
Metadata
Documentation
Recycle bin
Job demoAltus 0.1
tS3Configuration_1
tFileInputParquet_1
revenue (Main)
tMap_1
augmented_revenue (Main)
tTopBy_1
by_best_customers (Main)
tAggregateRow_1
rev_sum (Main)
tFileOutputParquet_1
Designer Code Jobscript
Job(demoAltus 0.1) Contexts(demoAltus) Run (Job demoAltus) Test Cases Component Spring Integration Action Modules DQ Repository Detail View
Job demoAltus
Basic Run
Spark Configuration
Debug Run
Advanced settings
Target Exec
Memory Run
Use local mode
Cluster Version
Property Type Built-In
Distribution Cloudera Altus
Version Cloudera Altus 1.0
Spark Mode YARN Cluster
Altus Configuration
Altus Access Key ""
Altus Secret Key *****
Path to Altus CLI "/opt/altus-cli/altus"
Altus Cluster
Use an existing Altus cluster
Cluster Name "talend-altus-cluter"
Environment "altus_oregon"
Cloud Provider AWS
Delete cluster after execution (transient mode)
AWS Configuration
Override with a custom JSON configuration
Instance Type c4.8xlarge
Worker nodes 10
SSH Private Key "/home/talend/secure/altus.ppk"
Cloudera Manager Username "altus-admin"
Cloudera Manager Password *****
Enable log "s3://altus_logs_bucket/spark/talend"
Or let Altus dynamically spin up & down a new cluster for this job = easy cluster creation + only pay for the computation time

More Cloud ! Talend and Altus: Under the hood

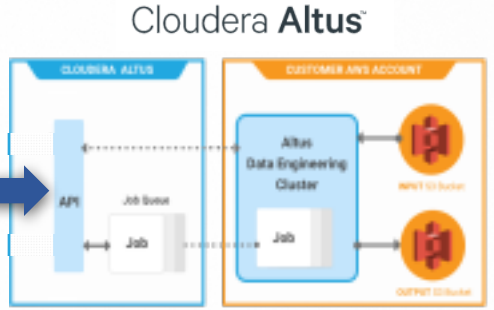
Generate a Java program from a Graphical Job in Talend Studio

Talend builds the Job as a JAR file

Talend uploads the Job on Amazon S3 Bucket

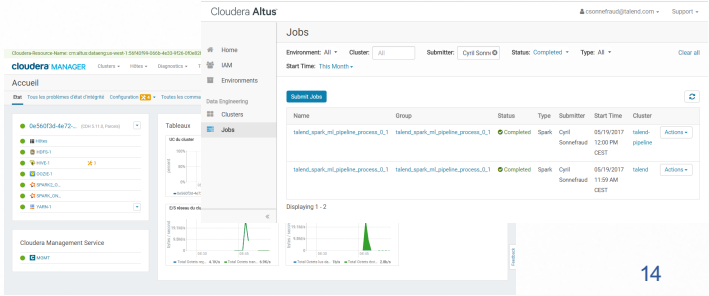


Talend can spin an new cluster, and submit the Job to Cloudera Altus



Monitor the execution in the Cloudera Altus Console.

Access to Cloudera Manager monitoring, metrics, and job history servers



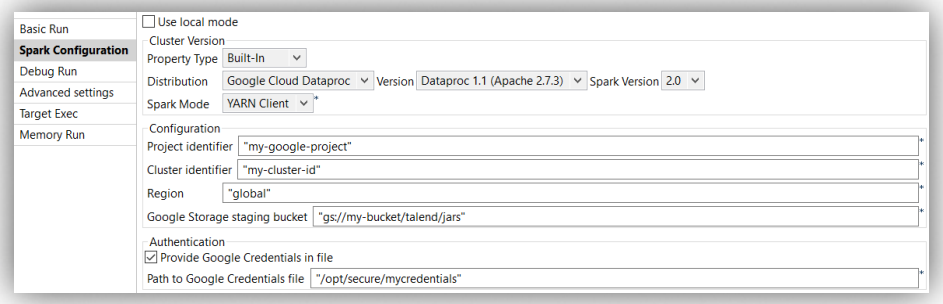
talend

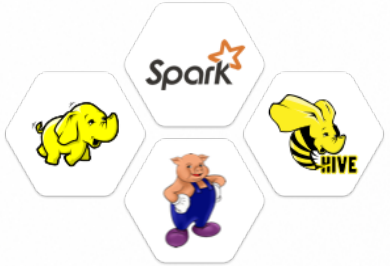
More Cloud ! – New supported distribution: Google Dataproc

Dataproc is Google's managed Cloud offering competing with Amazon EMR / Microsoft Azure HDInsight

- It is a Hadoop distribution with Spark / Hive / Pig
- No HBase: Google BigTable instead
- Data storage is done on Google Cloud Storage

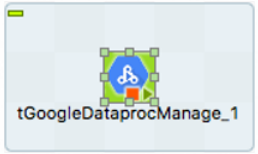
Supported as a new distribution in the Studio



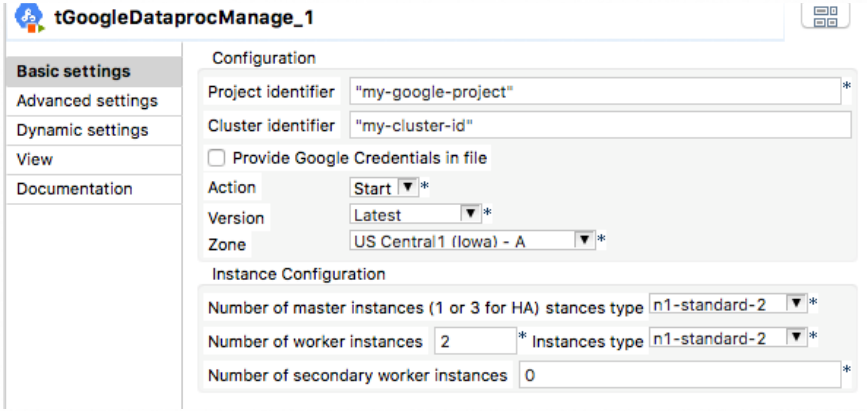


Benefits: Easily integrate with Google DataProc just like another distribution

More Cloud ! – More Google components



Like for Amazon EMR, it is possible to spin-up a Google Dataproc cluster using tGoogleDataprocManage

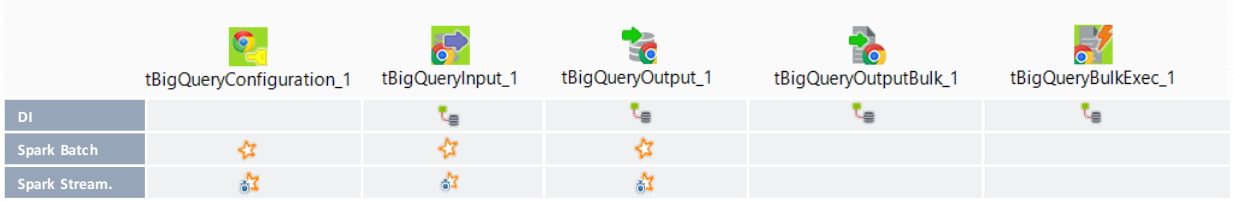


Also enables to stop a cluster = define execution plans that will only use Google Cloud for the computation needs

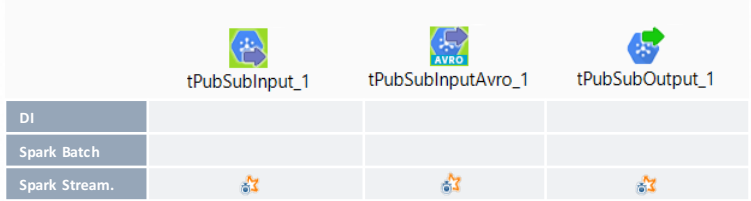
Benefits: “Pay per use” using Google DataProc.

More Cloud ! – More Google components

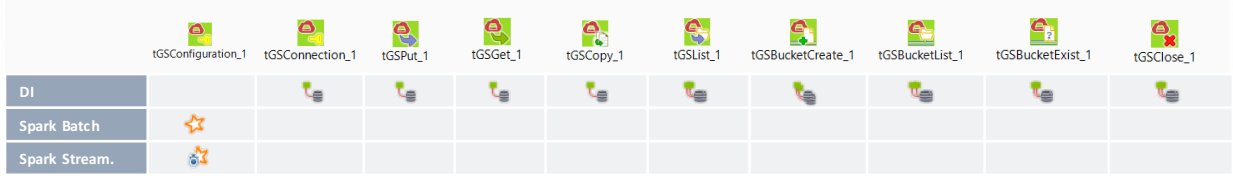
Google BigQuery components



Google PubSub components



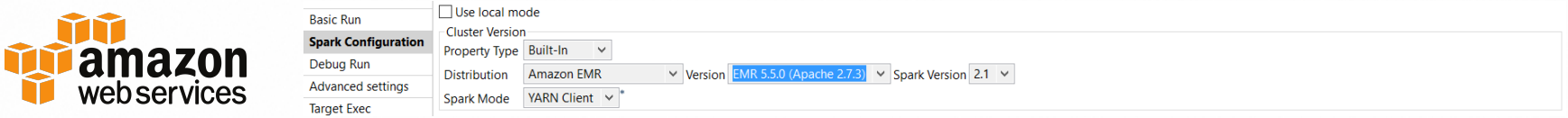
Google Cloud Storage components



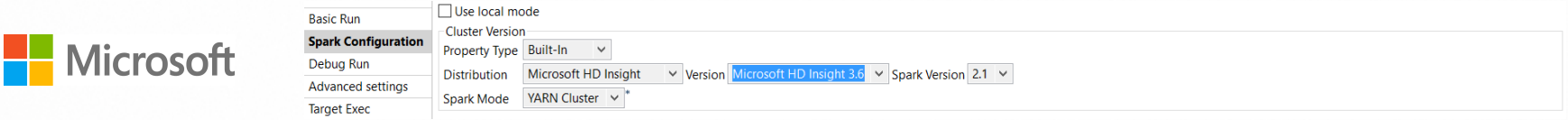
Benefits: Best-in-class connectivity to Google Cloud ecosystem.

More Cloud ! – Amazon EMR & Microsoft HDInsight Upgrade

- Studio has been upgraded to support Amazon EMR 5.5.0 (with Spark 2.1)



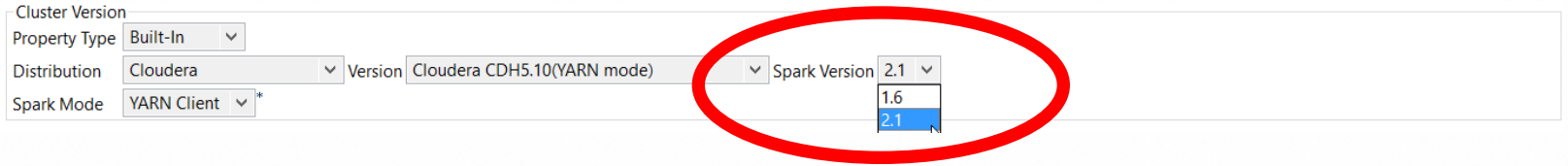
- Studio has been upgraded to support Microsoft Azure HDInsight 3.6 (with Spark 2.1)



Benefits: Keep-up with the pace of the Cloud providers. Leverage latest versions.

More Spark ! – Spark 2.1 now available in the Studio !

- Spark 2.1 is the second release on the 2.x line
- Aimed towards 2.x production readiness
- Focus on usability, stability, with a lot of polishing

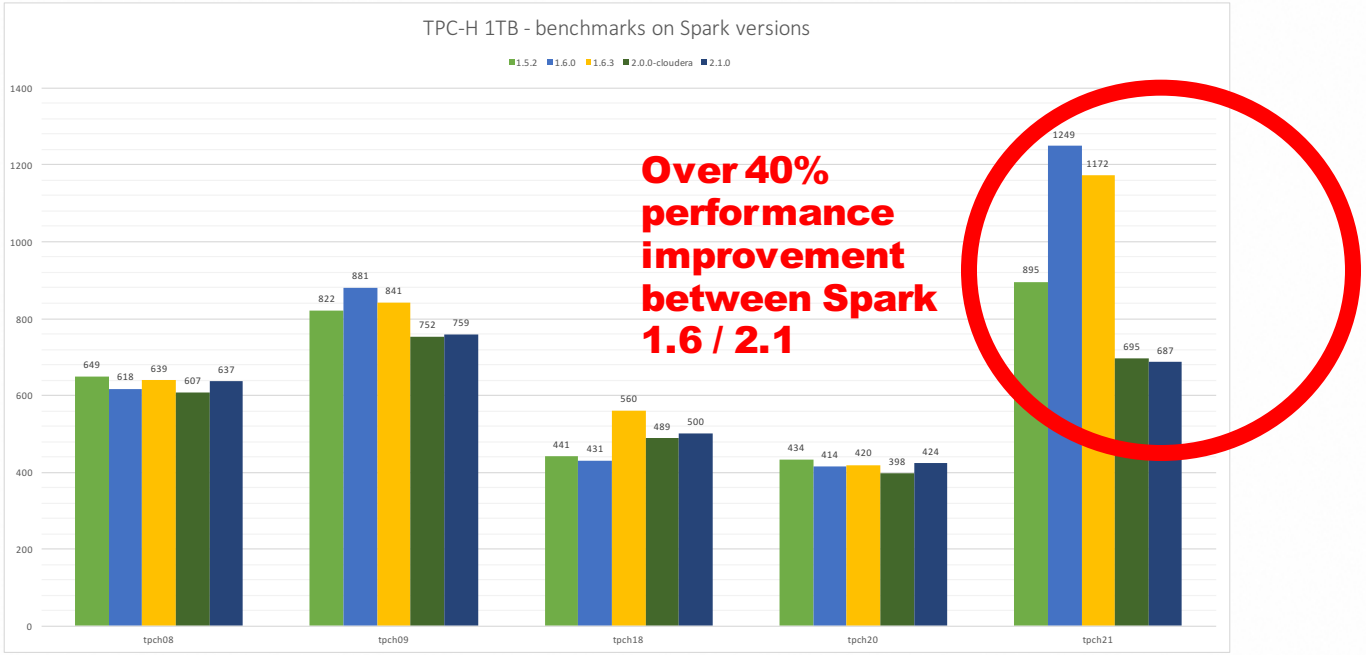


Supported with:

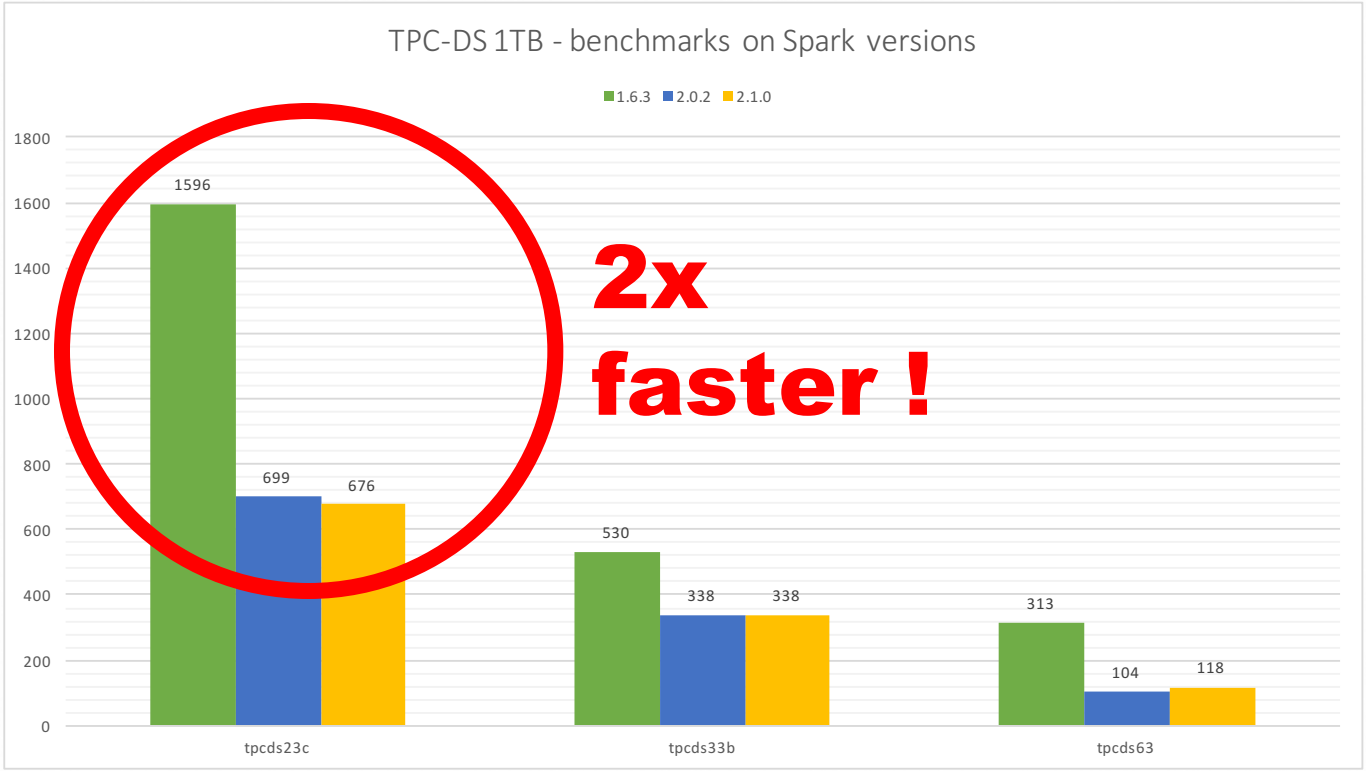
- Hortonworks 2.6
- Cloudera 5.10
- Cloudera Altus
- Amazon EMR 5.5
- Local Spark

Benefits: Production-ready with Spark 2.1. Enterprise-class robustness.

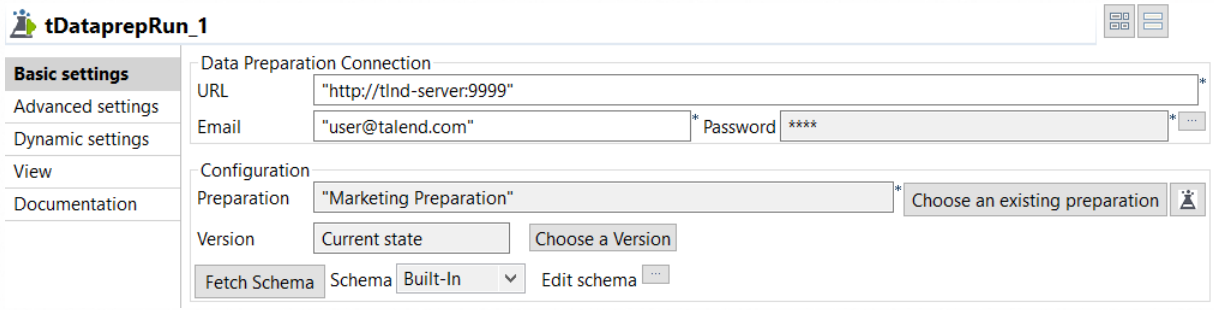
More Spark ! – Better performance with Spark 2.1



More Spark ! – Better performance with Spark 2.1



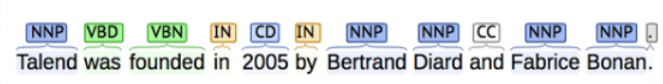
More Spark ! – DataPrep in Spark Streaming

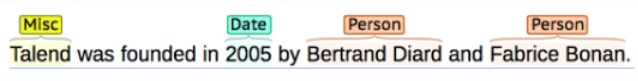


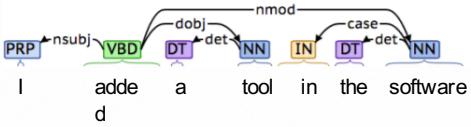
tDataPrep run is now available as a component for Spark Streaming

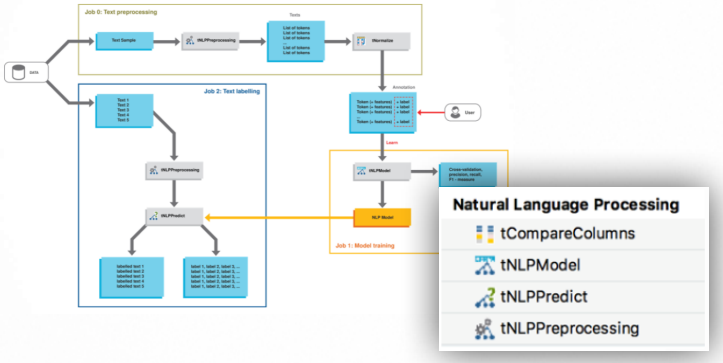
Benefits: Leverage governed self-service within your Real-Time Big Data pipelines.

More Spark ! – Natural Language Processing (NLP) with Spark







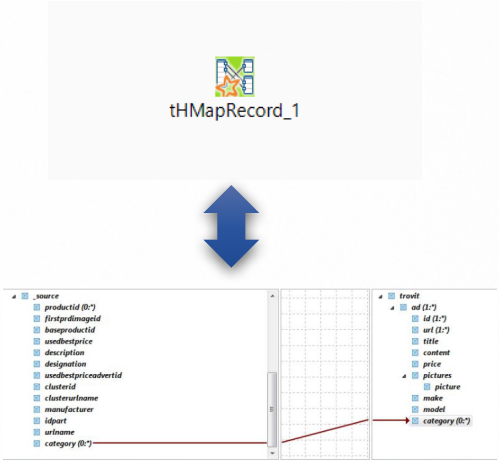


- New Natural Language Processing (NLP) components are available in Spark Batch
- What does it do exactly ?
  - Extract useful information from textual resources (people names, companies, tools…)
  - Classify discussions by topics (group discussions together, find discussions where people are mentioned)
  - Entity linking (e.g. persons and organizations linking, links between persons and any other information that may be used for reidentification)
- What are the use cases ?
  - Intelligent Search
  - Sentiment Analysis
  - Marketing Personalization
  - GDPR
  - …

Benefits: Make your integration intelligent. Create new data-driven insights.

More Spark ! – Updated TDM with Spark

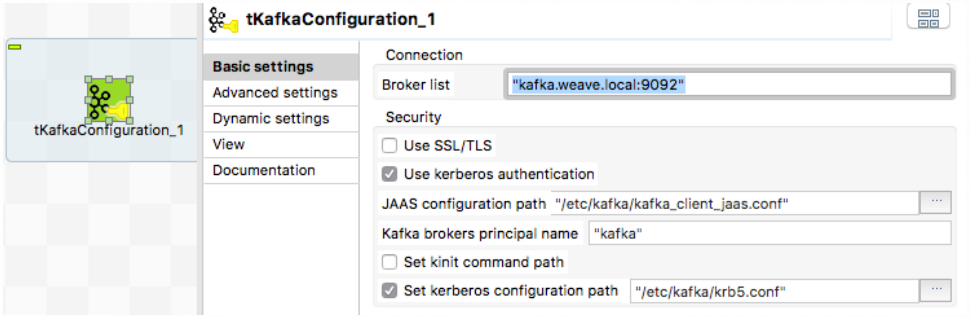
- Spark Streaming now features TDM with the new tHMapRecord component
- This will help for advanced parsing / formatting of hierarchical data coming from streams
  - Typical use-case is AVRO data coming from a Real-Time stream (e.g. Kafka)
- Works with any component in the Spark Streaming palette



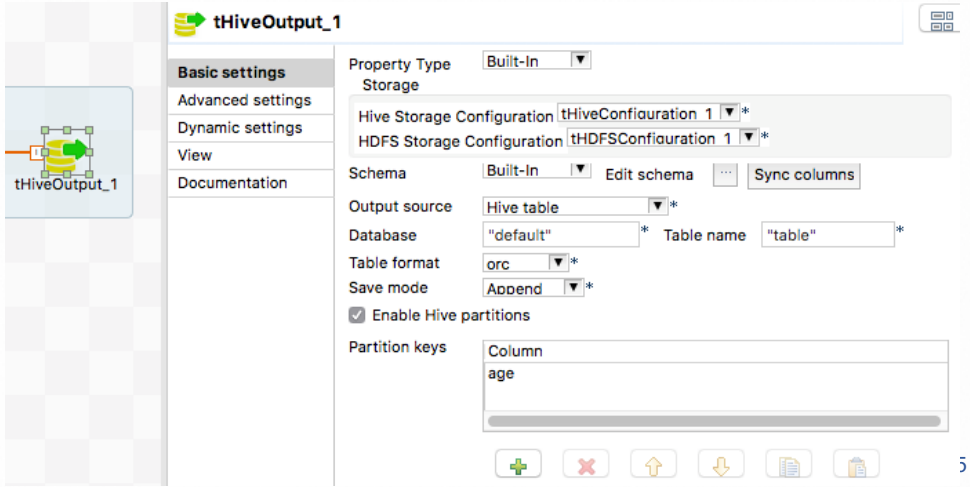
Benefits: Advanced mapping for Real-Time Big Data.

More Spark ! – Miscellaneous other extra enhancements

- Support for Kafka 0.10 with Kerberos



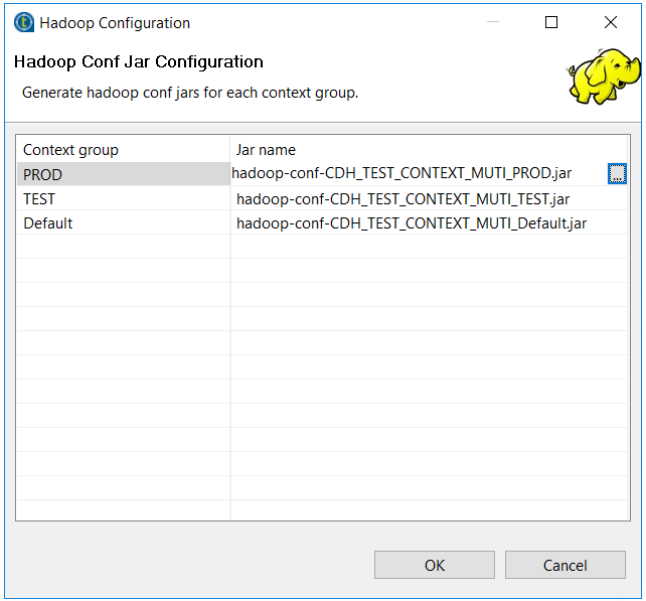
- Hive partitioning for Spark Batch & Streaming



talend

(Even) Better developer experience – Contexts management

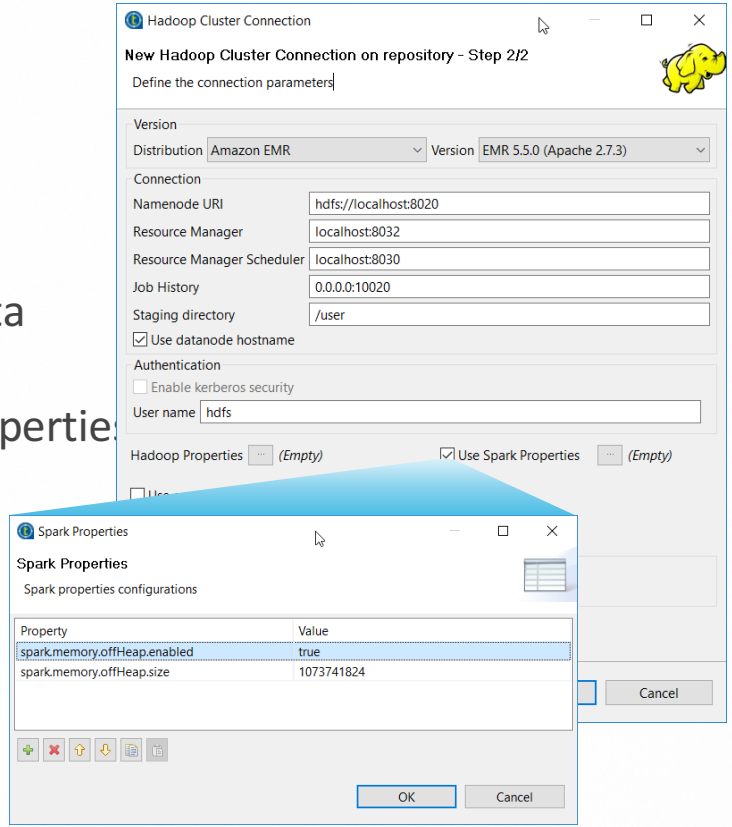
- It is now possible for the developer to define different context groups for a Big Data job (like DEV, UAT, PROD)
- This enables to define artifacts that aren't going to be tied to a particular environment
- From a technical standpoint, configuration is bundled in dedicated jars, and then injected at runtime



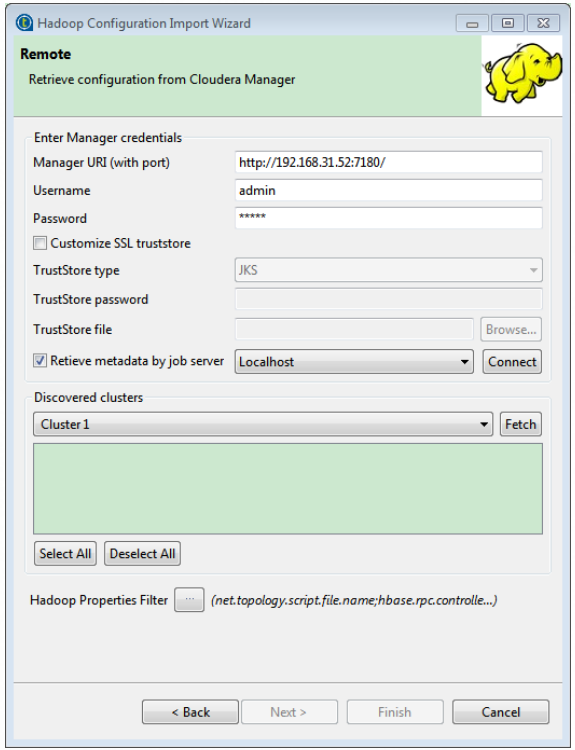
(Even) Better developer experience – Spark properties

- Spark properties can be tied to a Hadoop metadata, in the Repository

- This helps to:
  - Centralize Spark properties within a Big Data project
  - Make less errors when setting up Spark properties
  - Setup fine-tuning in Spark projects



(Even) Better developer experience – Retrieve metadata through Jobserver



- Jobserver can now be used to fetch metadata from Hadoop cluster
- Useful in secured contexts where the gateway to the Hadoop cluster is the Edge node
- First steps towards removing direct connections between Studio and Hadoop

Data Services / ESB

ESB 6.4 in a nutshell

REST API Design

- Swagger Spec and Swagger UI
- OpenID Connect Clients and Servers

Developer Productivity

- Debugging Data Services and Routes from Studio in the Talend Runtime
- Sharing custom jars with team members (cConfig)
- cREST and cSOAP

Design Patterns

- cTalendJob – 'Fast Job Invocation' Option

Service Registry

- Service Metadata Support

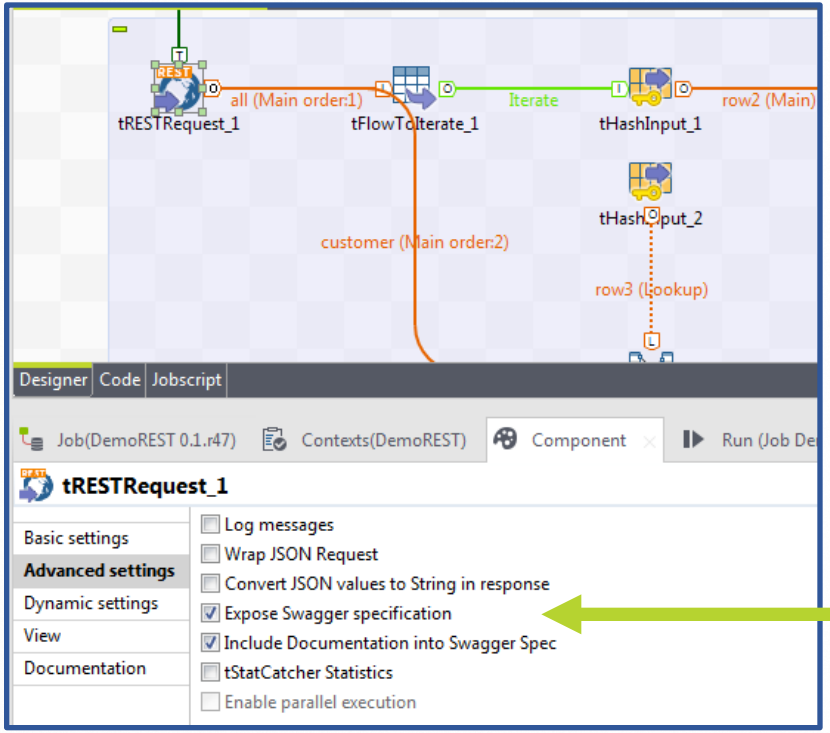
Core Framework Updates

OSS Project	6.3.1	6.4.1
Apache CXF	3.1.7	**3.1.12**
Apache Camel	2.17.3	**2.17.6**
Apache Karaf	4.0.7	**4.1.1**
Apache ActiveMQ	5.14.1 (5.14.2 WebConsole)	**5.14.5**
Jetty	9.2.1	**9.3.14**
Syncope	1.2.9	**2.0.2 (via Talend IAM)**
Spring-Boot	1.3.7	**1.3.7**

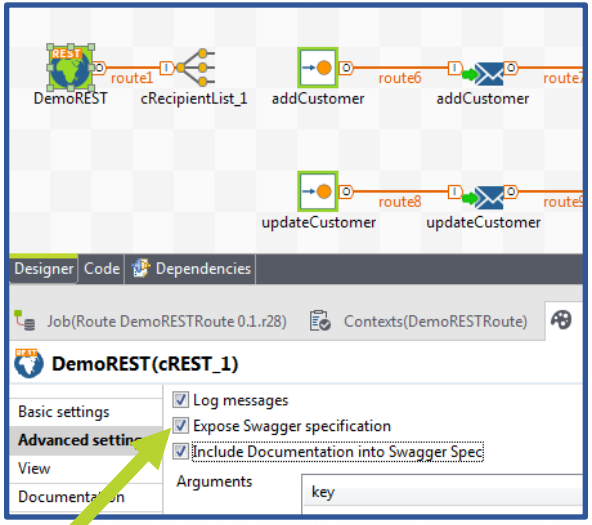
REST API Design

Swagger: How to enable it for REST Provider

REST DS - tRESTRequest



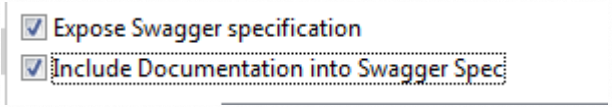
REST Routes – cREST (Provider)

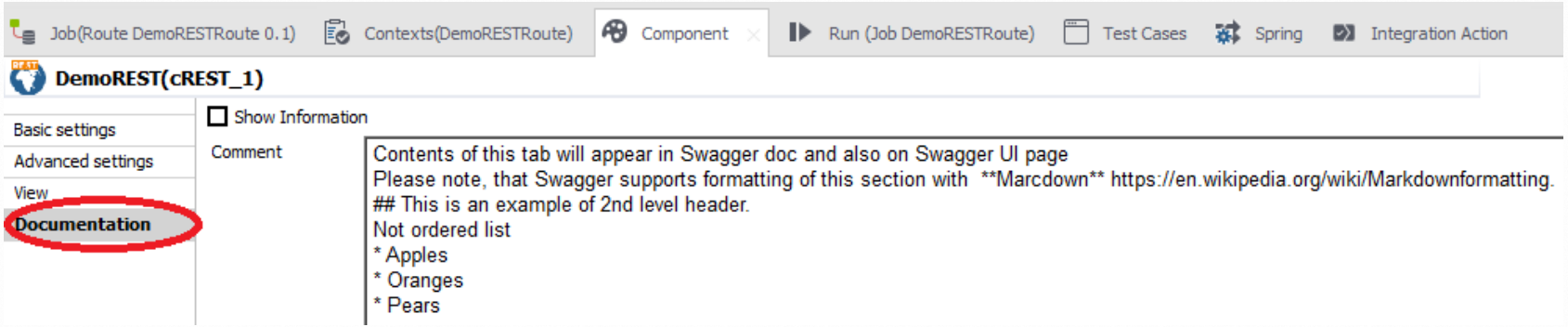


Just select: ‘Expose Swagger specification’ and optionally ‘ Include Documentation into Swagger Spec’

Swagger: Studio Documentation in Spec

- Swagger supports(at least, partially) Markdown formatting https://en.wikipedia.org/wiki/Markdown . Below is an example of such formatting:





Swagger: URL to access Spec & UI

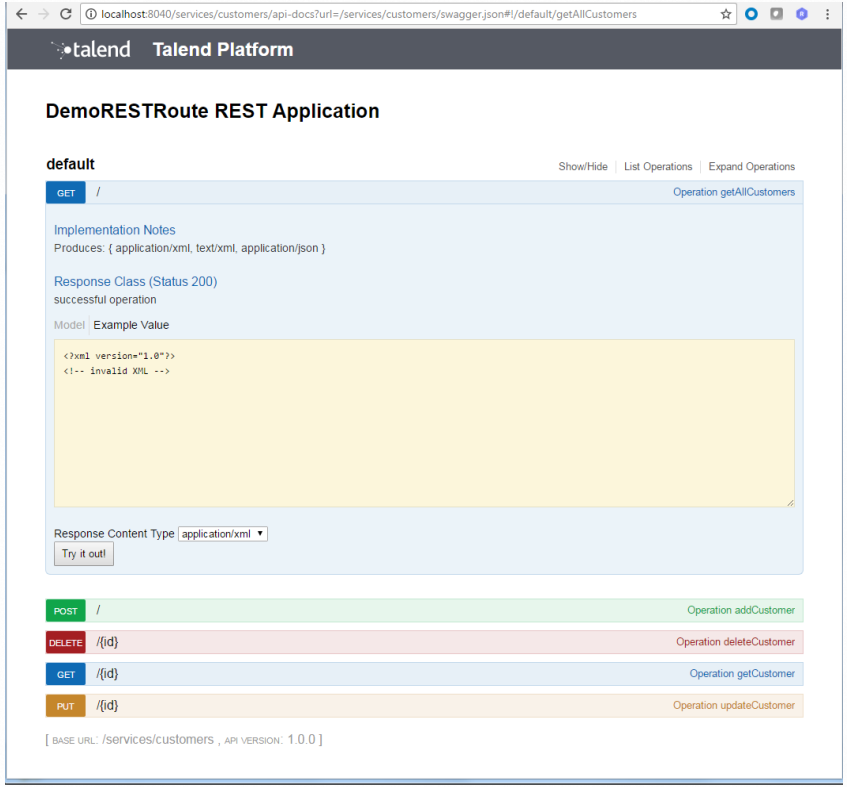
- Access the Swagger Spec
  - http://localhost:8040/services/ENDPOINT/swagger.json
  - http://localhost:8040/services/ENDPOINT/swagger.yaml
  - E.g. for ESB Demo Examples – DemoRestRoute
    - http://localhost:8040/services/customers/swagger.json
    - http://localhost:8040/services/customers/swagger.yaml

- Access the Swagger UI
  - http://localhost:8040/services/ENDPOINT/api-docs?url=/services/ENDPOINT/swagger.json
  - E.g. for ESB Demo Examples – DemoRestRoute
    - http://localhost:8040/services/customers/api-docs?url=/services/customers/swagger.json

talend

Swagger: UI – explore the API

- Swagger UI is prepackaged with the Talend Runtime Container (and ESB Microservices)
- Swagger UI is a 3rd party standard Web Application which can be use to
  - Visualize the Swagger Spec of the Service
  - To execute API Calls (e.g. for some quick ad hoc testing)
- The Swagger UI is a standard way to allow developers to explore and use REST API's



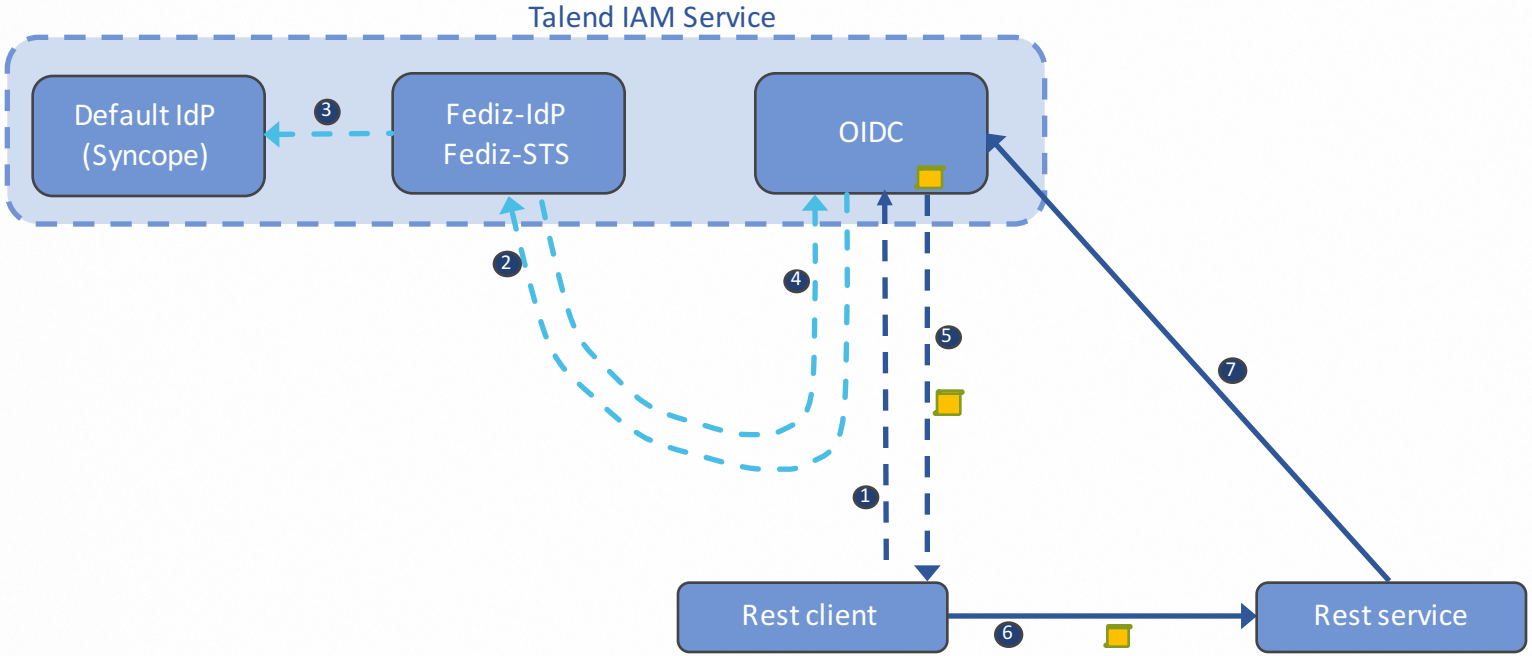
Talend IAM - Syncope, Fediz, STS OIDC

- What is it?
  - User Management – Syncope 2.0.2
  - Token Mapping and Federation && OpenID Connect - Fediz 1.4.0
  - Full server-side support of OpenID Connect (OAuth 2.0, SAML 2.0)
- Becomes common module vs. ESB module
- Benefits
  - SSO for Talend Data Preparation and Stewardship
  - Enables OIDC for ESB Data Services and Routes (REST)
  - Enables JavaScript / AJAX clients to talk to Talend Data Services

IAM Service - Introduction

- With 6.4 platform, we are introducing a new IAM service. This IAM service is based on standard specifications and protocols like OpenID Connect (OIDC), OAuth2, WS-Federation and WS-Trust. In 6.4 it provides
  - SSO between TDP and TDS based on OIDC & WS-Federation standards.
  - Authentication to ESB Rest clients based on OAuth2 standards.
  - SAML authentication to ESB clients based on STS and WS-Trust.

Open ID Connect (OIDC) - OAuth flow for ESB

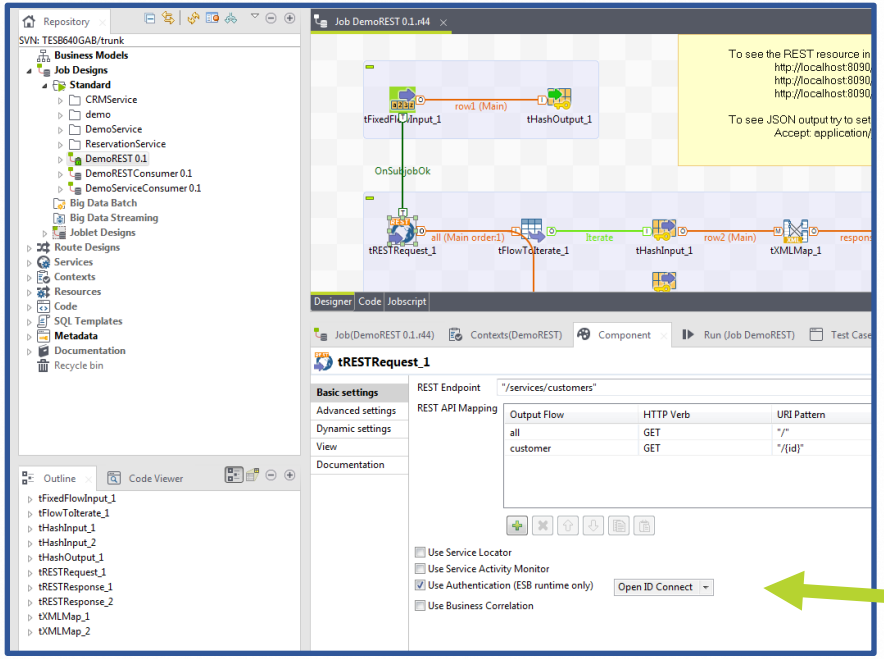


OpenID Connect / OAuth2

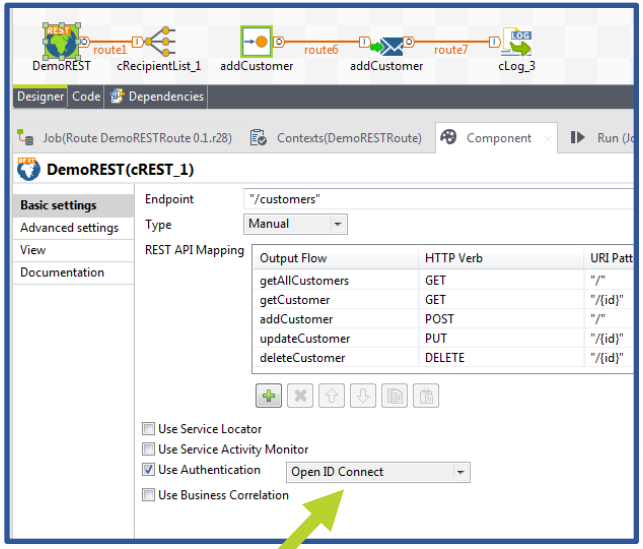
WS-Federation

OIDC: ESB Rest – Provider

REST DS - tRESTRequest



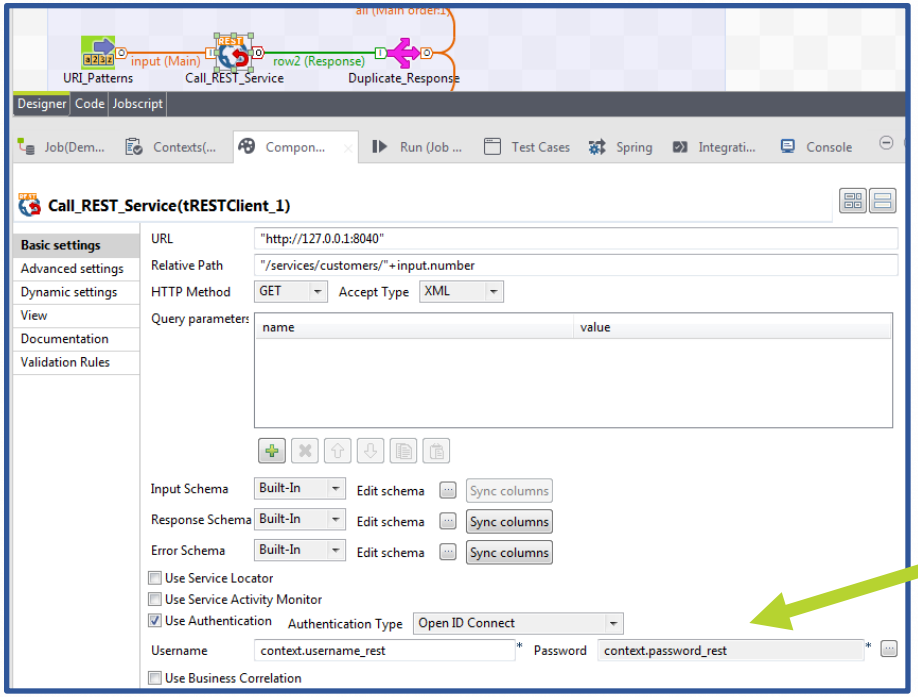
REST Routes – cREST (Provider)



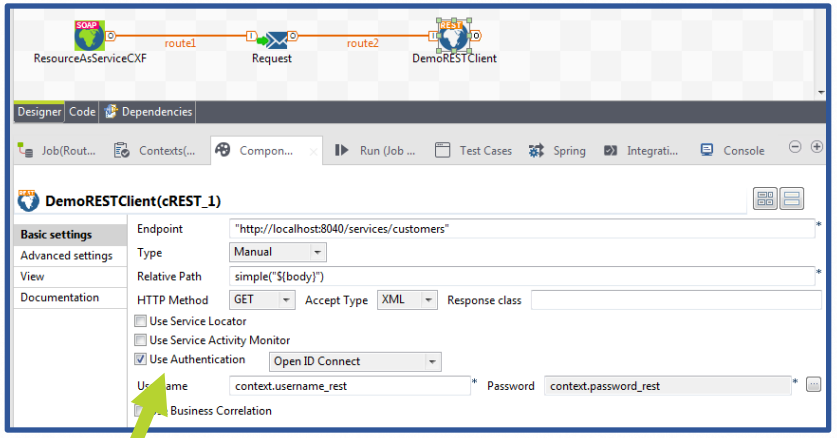
Just select: 'Open ID Connect as the Authentication Options – That's it!

OIDC: ESB Rest – Client

REST DS - tRESTClient



REST Routes – cREST (Client)

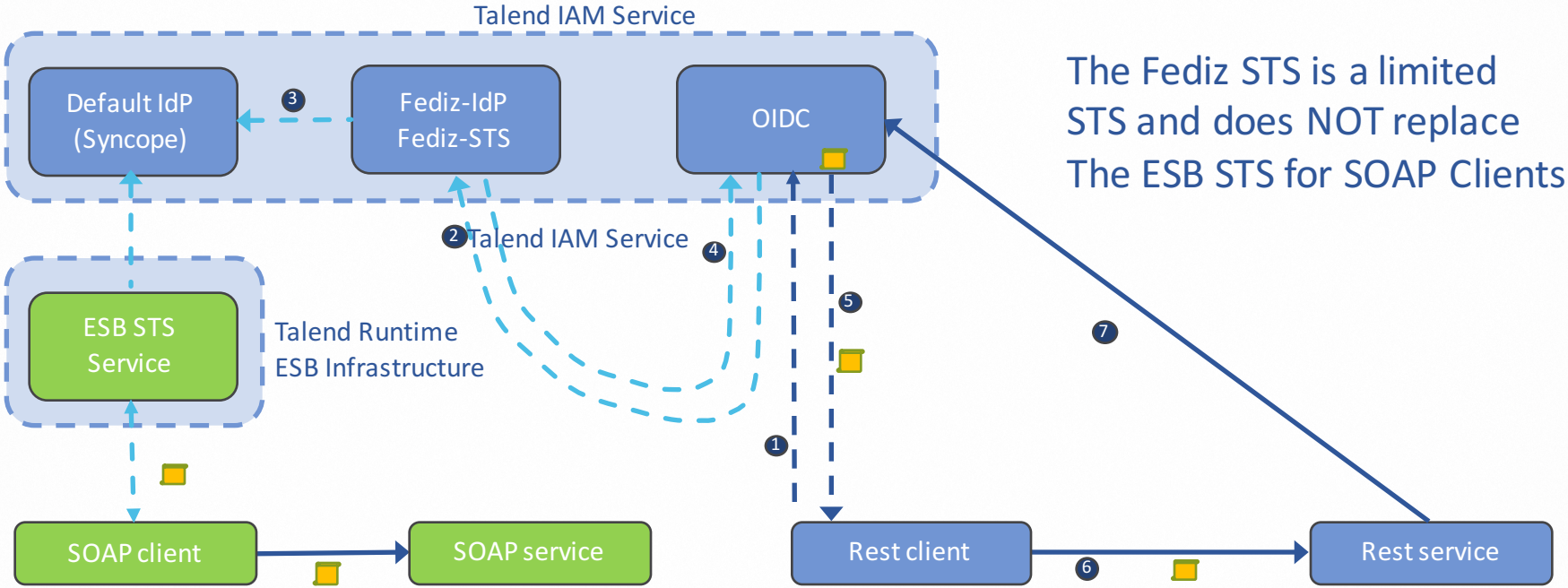


Just select: 'Open ID Connect' as the Authentication Options – and provide a Username / Password (which is defined in Talend IAM (Syncope)

OIDC: Runtime Configuration (Jobs)

- `org.talend.esb.job.oidc.cfg` – is the new config file we added for OIDC
- 4 Properties
  - `org.talend.esb.job.oidc.token.endpoint= \`
    `http://localhost:9080/oidc/oauth2/token`
  - `org.talend.esb.job.oidc.validation.endpoint= \`
    `http://localhost:9080/oidc/oauth2/introspect`
  - `org.talend.esb.job.oidc.public.client.id=aFSloIZSXHRQtA`
  - `org.talend.esb.job.oidc.scope=openid`
- Note:
  - The public client id and the scope are fixed and should not be changed in Talend IAM
  - It is technical possible to register own clients but we not expose this in this version

STS Service: ODIC vs SAML Use Case



OpenID Connect / OAuth2

WS-Federation

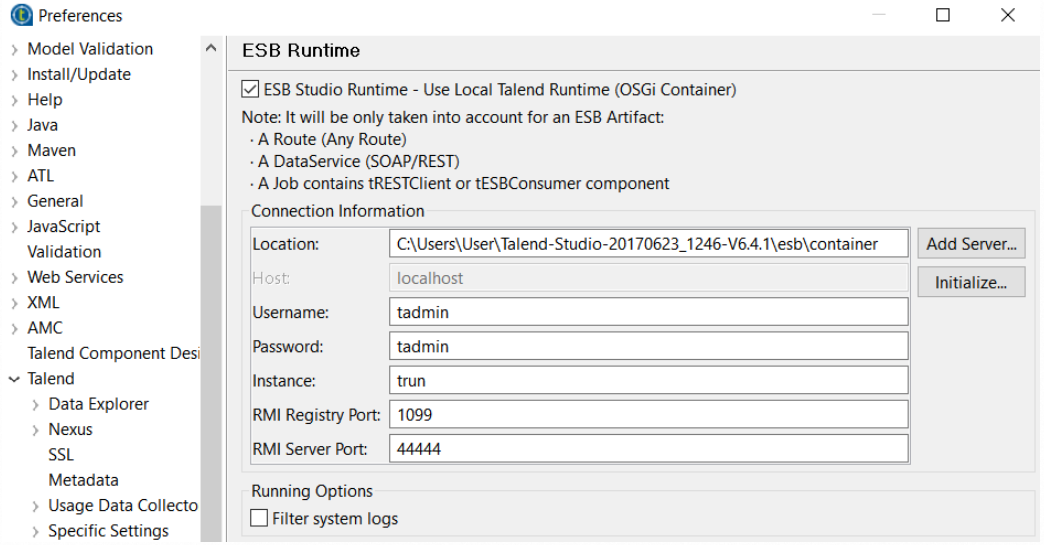
Developer Productivity

Talend Runtime packaged with Studio

- ESB Capabilities testable from Studio now
  - Data Sources (connection pooling and alias)
  - Service Locator
  - Service Registry
  - Service Activity Monitoring
  - Authentication and Authorization
  - Transport Security (TLS)
  - Concurrent routes (e.g. cVM calls)
  - Concurrency of WebServices

No more: “It works in Studio but not on Runtime!

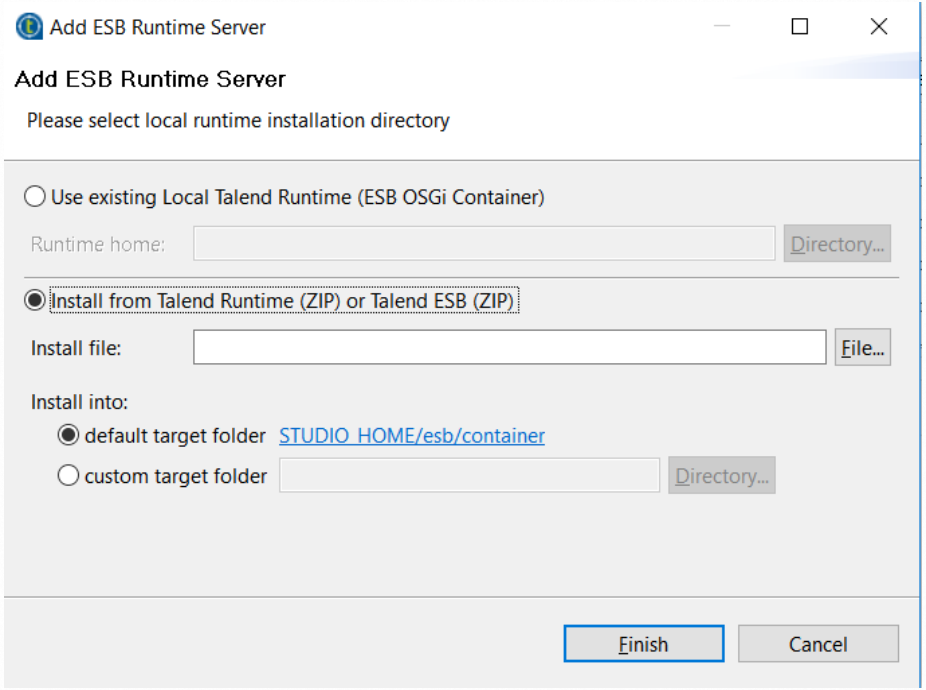
Local Runtime: Setup/Preferences 1



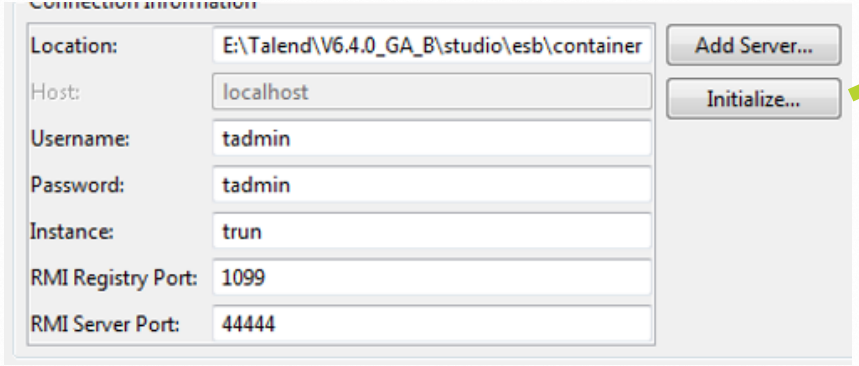
- Activate Local Runtime
- Preferences > Talend > Run/Debug > ESB Runtime
- Enable ‘ESB Studio Runtime’, ‘Add Server’ and ‘Initialize…’

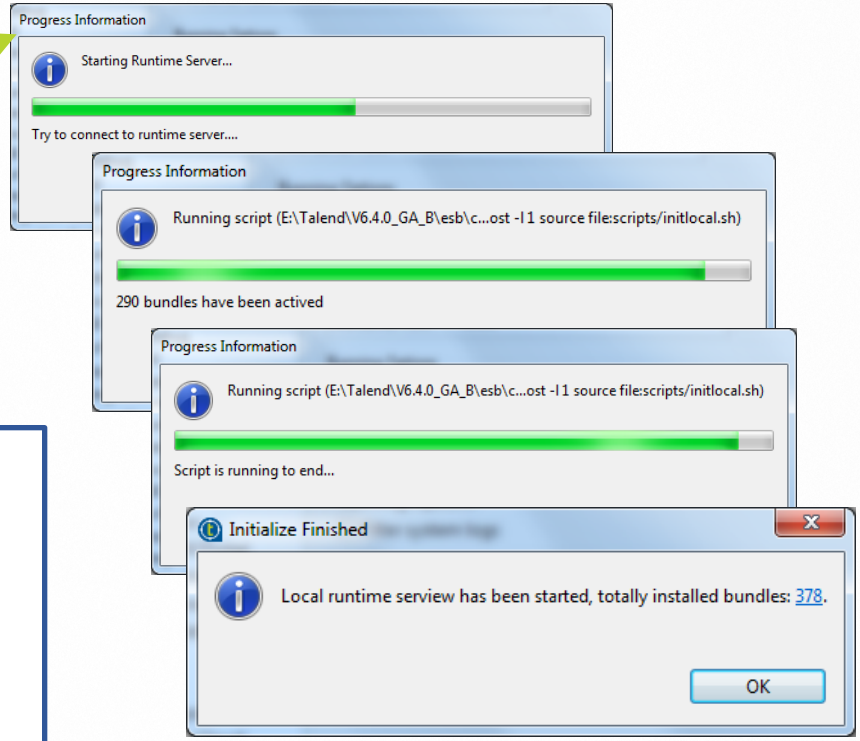
Local Runtime: Setup/Preferences 2

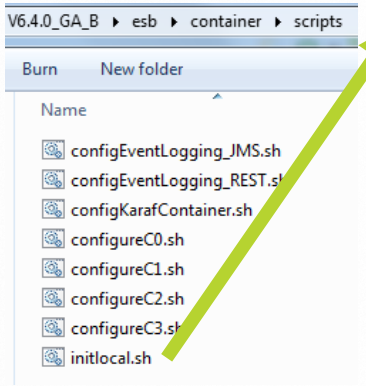
- Add a Runtime either by
  - Using an existing Talend Runtime or
  - Install from Talend Runtime or Talend ESB ZIP file
- Default is STUDIO_HOME/esb/container
- Choose a different folder if Studio is e.g. in C:\Program Files\Talend

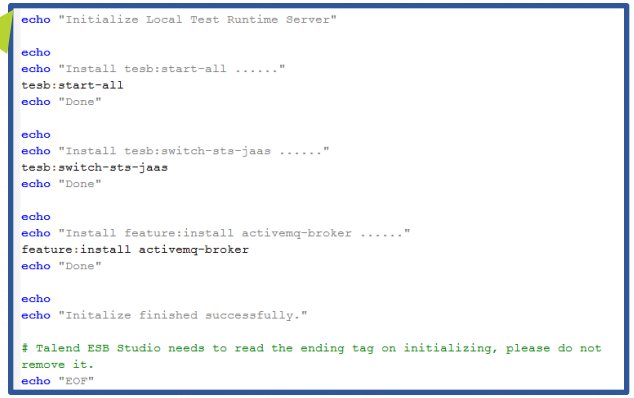


Local Runtime: Setup/Preferences 3

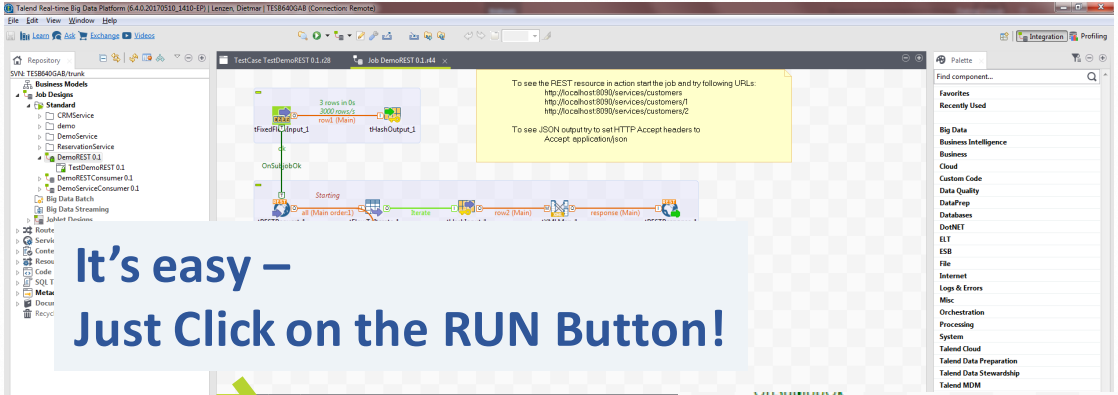




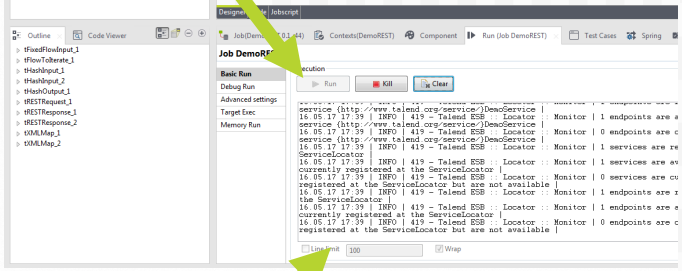




Local Runtime: How to use it?

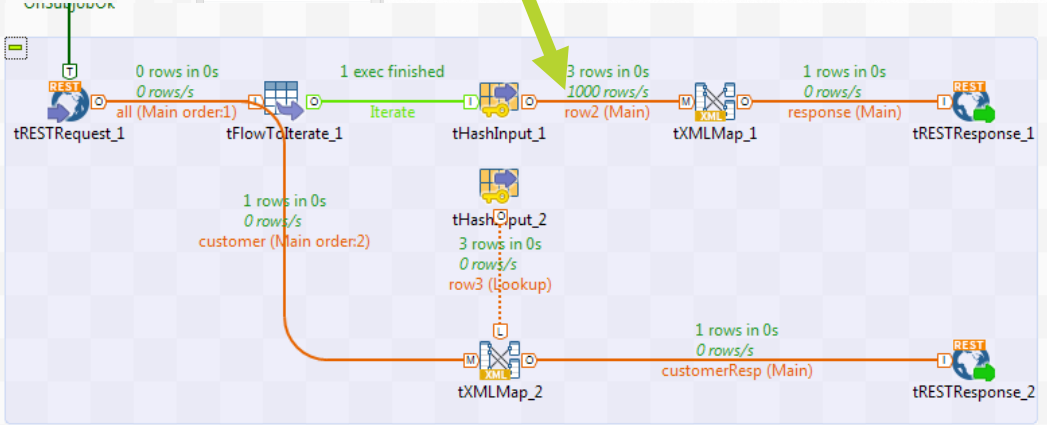


It’s easy –
Just Click on the RUN Button!



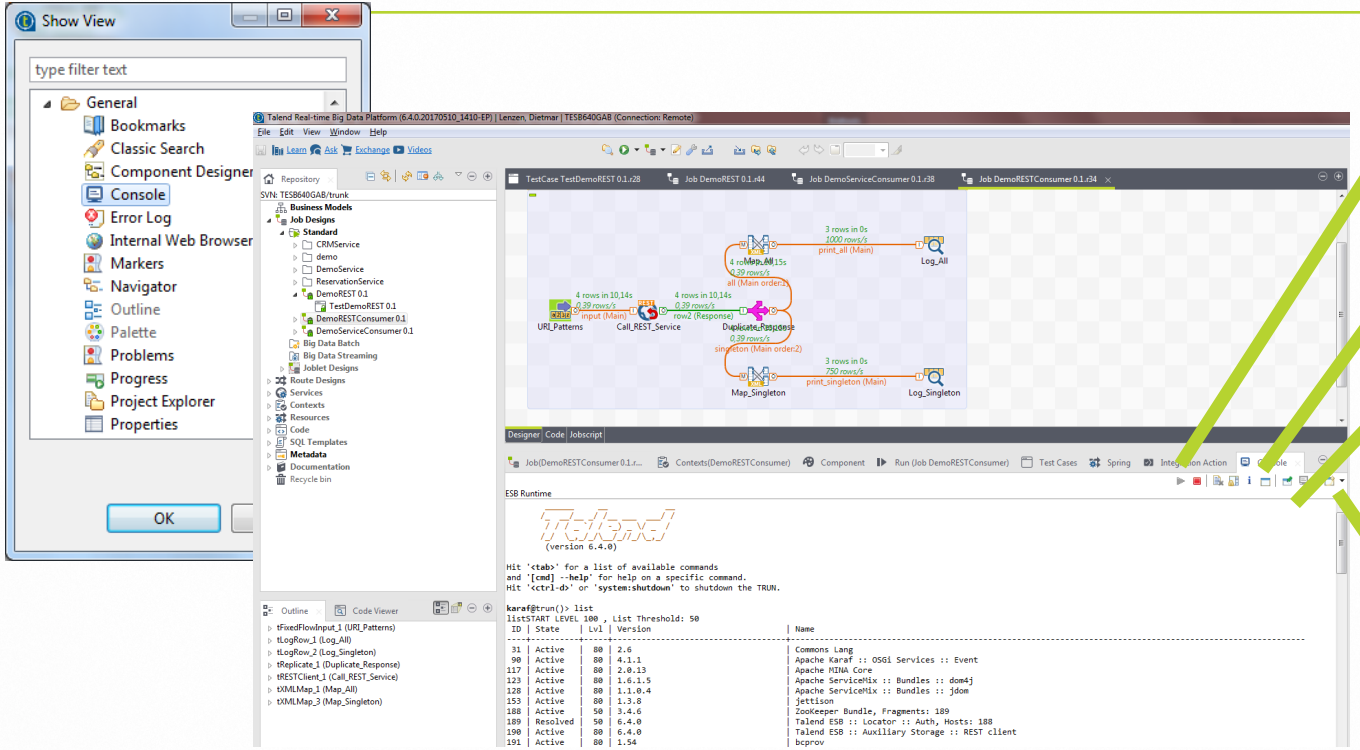
See the Logs just right in the same log view as usual

See the Rows executed directly in the model



talend

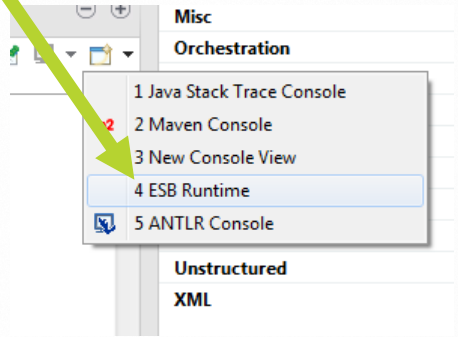
Local Runtime: Console View to the Runtime?



Start / Stop Runtime

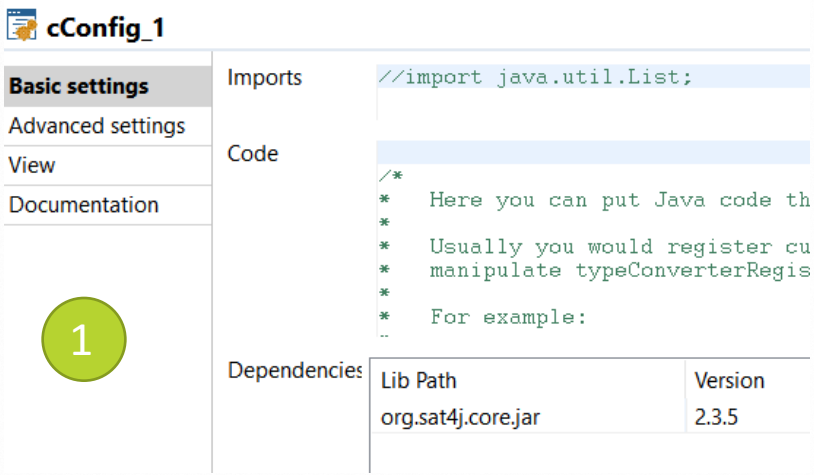
Open Preferences

Run (Karaf) Console Commands

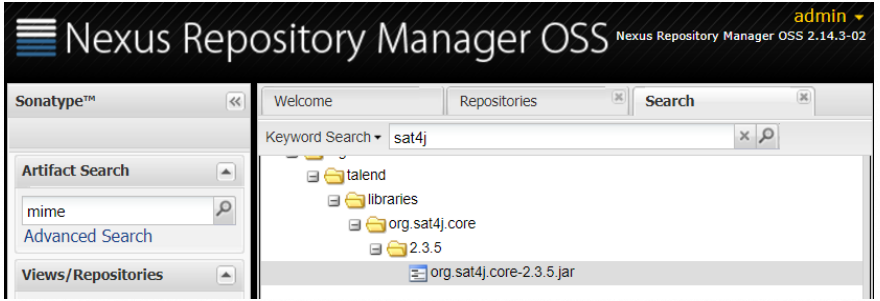


talend

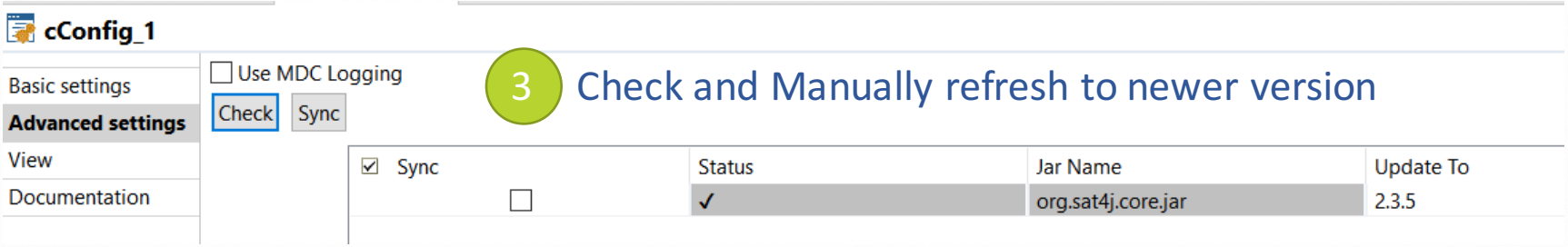
cConfig integrates with Nexus



1



2 Publish from Studio
Publish form Builds



3 Check and Manually refresh to newer version

Route Builder: cREST/cSOAP

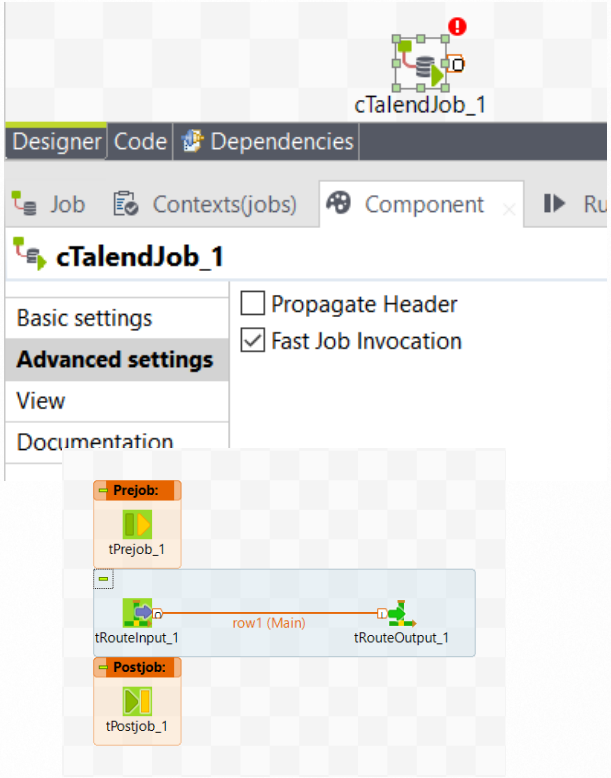
- Renaming
  - cCXFRS → cREST
  - cCXF → cSOAP

- Why?
  - Easier for new customers

Design Patterns

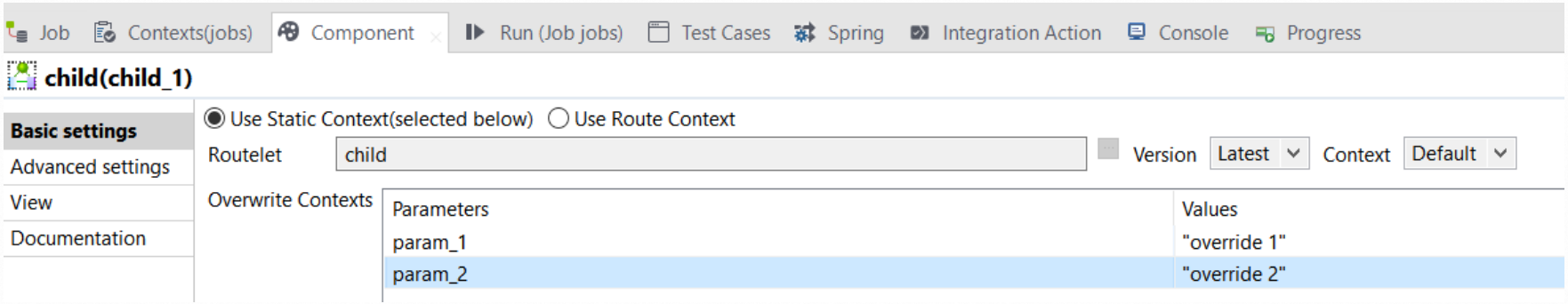
cTalendJob Fast Job Invocation

- Prerequisite:
  - Called Job is thread safe
- Behavior
  - Lifecycle of the embedded Talend Job is bound to the "start" and "stop" state changes of the connected Talend Camel Endpoint
  - Start calls the Job's _begin.javajet
  - Stop calls the Job's _end.javajet
  - Message calls job's _main.javajet
- Recommendation
  - Use tPrejob / tPostjob
  - Test, test, test



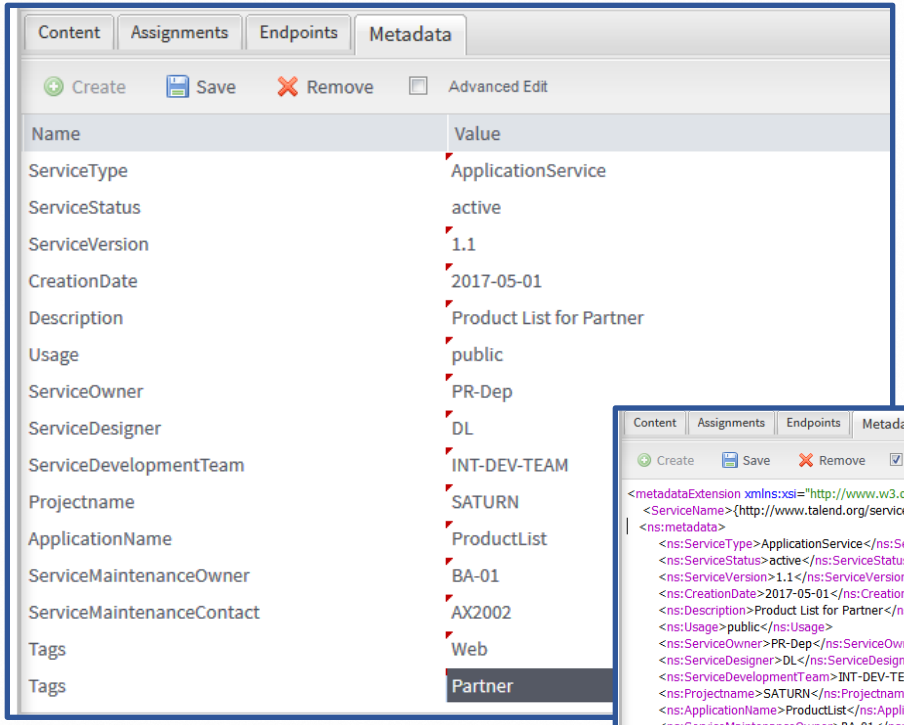
Route Builder: Routelets

- A RouteLet used with multiple routes can be deployed to the same Talend Runtime (& Release Nexus) without conflicts
  - From 6.4 use a unique naming at deployment time (route_routelet)
- More context parameter options and deployment improvements
  - More control how RouteLet context parameter should be used
  - A way to transfer context parameter from the Route to the RouteLet.

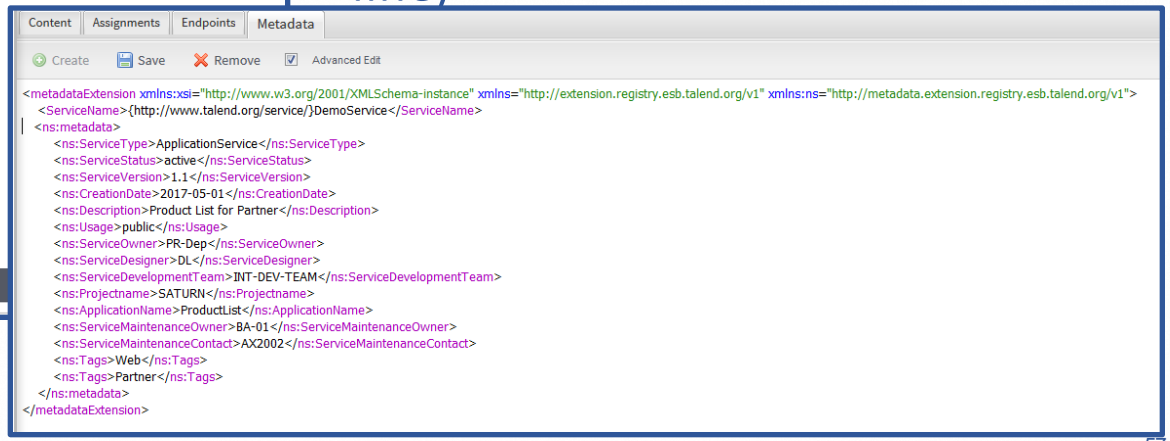


Service Registry – Metadata Support

SR-Metadata: TAC UI – Metadata Tab



- Edit values in the table view
- Or switch to the ‘Advanced Edit’
- Note: Create and Remove create a new Metadata XML and Remove removes the XML (and not an single line)



SR-Metadata: Elasticsearch

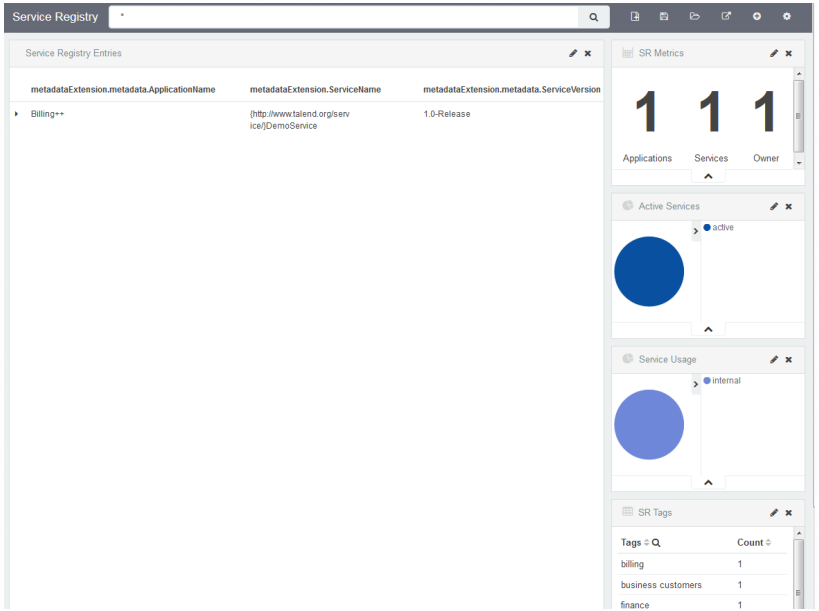
- Search capabilities via ElasticSearch (ES)
- Metadata is synchronized between Registry an ES. Registry is the master.
- Synchronization is done automatically in the background. It is configurable to enable or disable it.
- REST interface of ES is used. HTTPS and Basic Authentication is possible
- XML to JSON transformation is done automatically by Jackson framework.
- The ES index template is generated automatically to configure the ES analyzer correctly.

THIS IS OPTIONAL- The Service Registry Metadata feature
does not require the Talend Log-Server (ES) as a mandatory dependency

SR-Metadata: Logging Dashboard

- Example Dashboard with 6.4.1, which looks like the one. As the dashboard only works with the Schema the customer has defined it is keep it in add-ons/registry/dashboard

- Via the Dashboard the Metadata becomes searchable means beside the TAC ESB Service Registry Screen the User can use the Dashboard to find Services with certain metadata.



SR-Metadata: Metadata Support

- Extension to the Service Registry to maintain custom specific meta data about the service
- Allows custom metadata to be added to a service
- It’s only for information not as a runtime filtering of services
- It can be (optionally) configured to use Talend Log-Server and a dashboard
- TAC allows to edit the data in a UI way
- Talend Runtime: tregistry: commands can be used to CRUD and sync metadata (Means the metadata support is fully integrated in the commandline)

Data Mapper

Data Mapper on Big Data

- See Big Data section

- Spark 2.1 support
- New Spark Streaming component tHMapRecord
- Spark Batch Signature Enhancements
  - E.g. automatically infer COBOL, XML ends-with,

TAC

TAC 6.4 in a nutshell

Security

- Segregation of Admin rights
- Custom Roles

SSO

- Allow non-email login
- SSO w/o Talend custom attributes
- Configurable mappings from SAML token

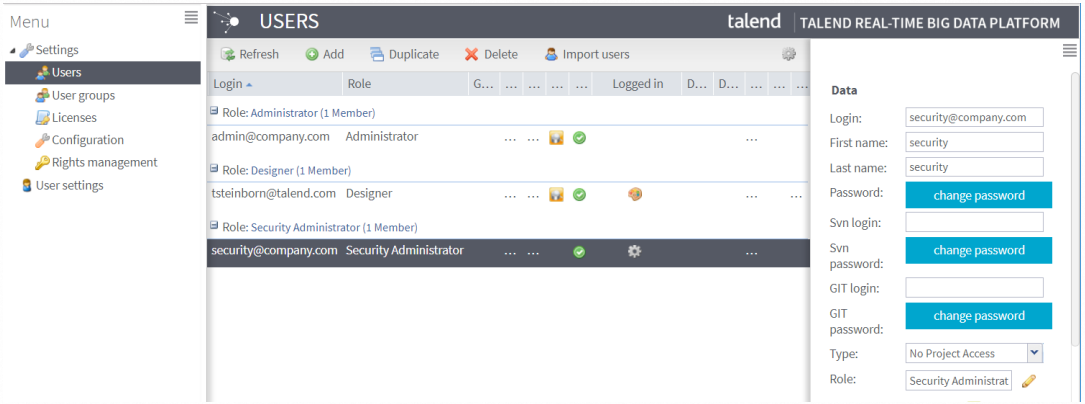
License Management

- Change to Max exceeds rules
- Real Time Big Data license combinations

Security

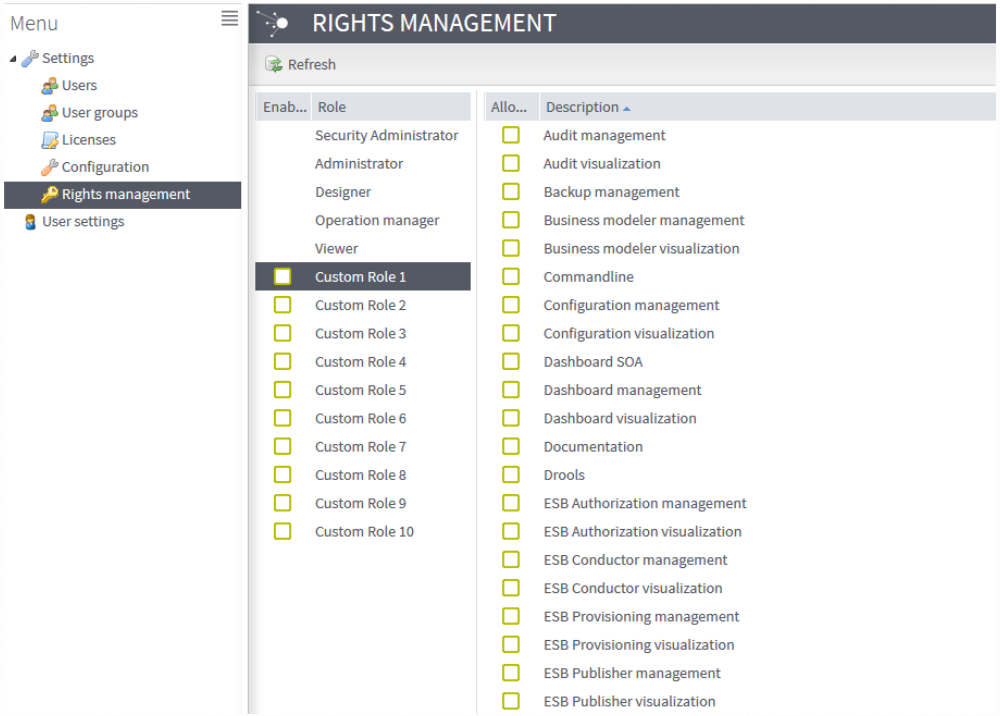
Segregation of Admin Rights

1. **Administrator**
   - Removed: User Management / Visualization, License Management, Security configuration management / Visualization rights
   - Removed: Rights Management.
2. **Security Administrator (security@company.com)**
   - Has now above
   - No project access
   - Is the default for new installations



TAC Custom roles

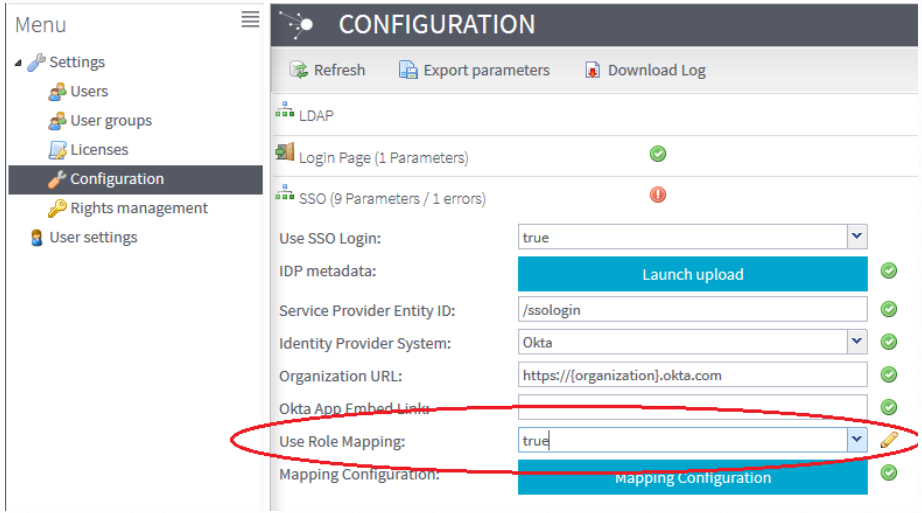
- 10 additional roles with no rights selected
- Scope is TAC / Studio only
  - Data Preparation and Stewardship cannot use them
- And no, they cannot be renamed



sso

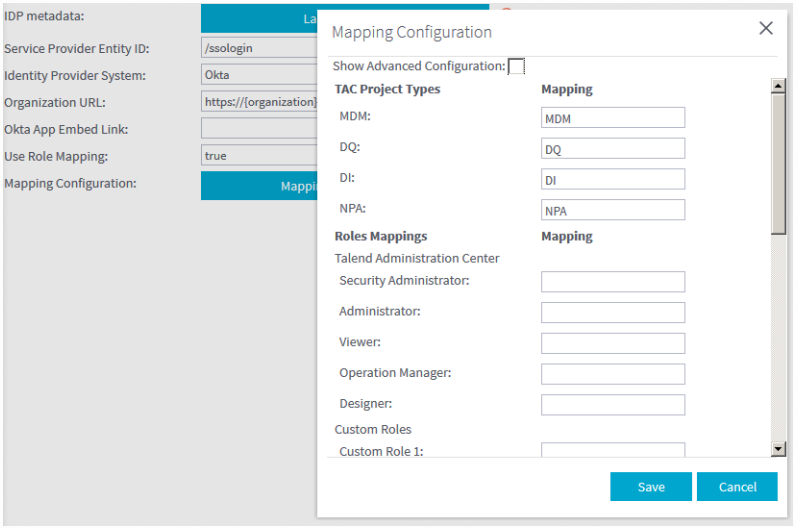
SSO: non email login and no custom attributes

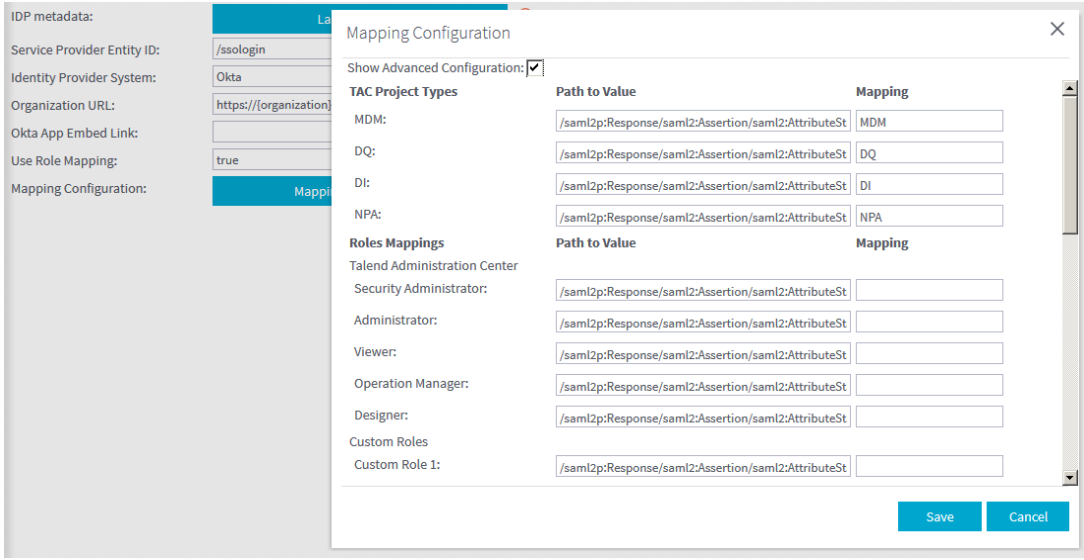
- Requirements
  - Allow non-email login, e.g. SID
  - Need for SSO work without adding Talend custom attributes
- Changes
  - Allow ‘No Project access’ type even if DP/TDS not selected
  - ‘Use role mapping’ flag
  - When user login first time to TAC we create this user with type NPA and without TAC roles



SSO: Make role mapping configurable for SAML

- Big customers have fixed SAMLs and do not want to change it specially for TAC





License Management

Licensing: Change max exceeds rules

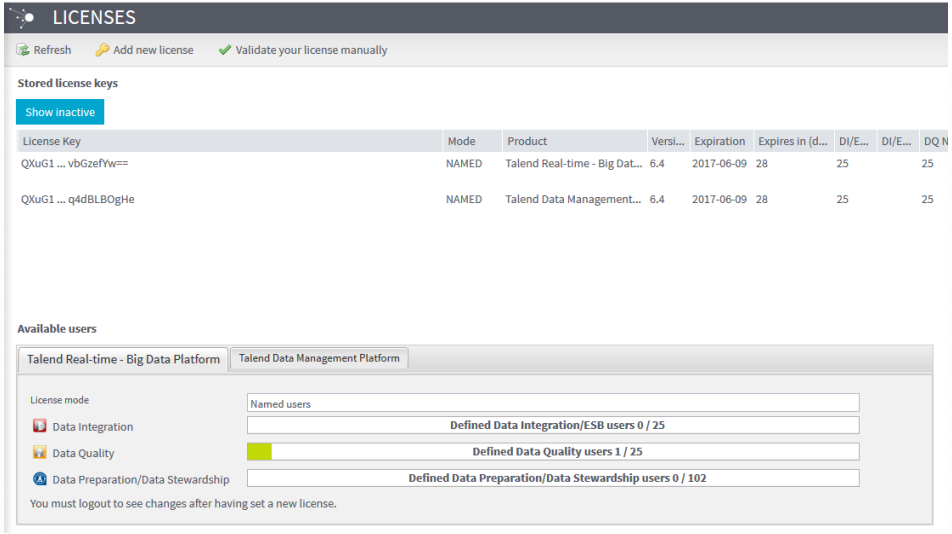
- Max exceeds N for create and login from only being percent based to be the lesser of:
  - +20%, or
  - 5 users
- Prior Versions only had +20%

- **named users**:
  - create N + new limit
  - login N + new limit
- **concurrent users**:
  - create any number
  - login N + new limit

- Additional warning Dialog with text exactly the same as in banner. Dialog shown every time for security admins after login.

Talend Real Time Big Data Platform & One other

Non-Big Data	Big Data
Data Management Platform	Real Time Big Data Platform
Data Services Platform	Real Time Big Data Platform
Master Data Management Platform	Real Time Big Data Platform

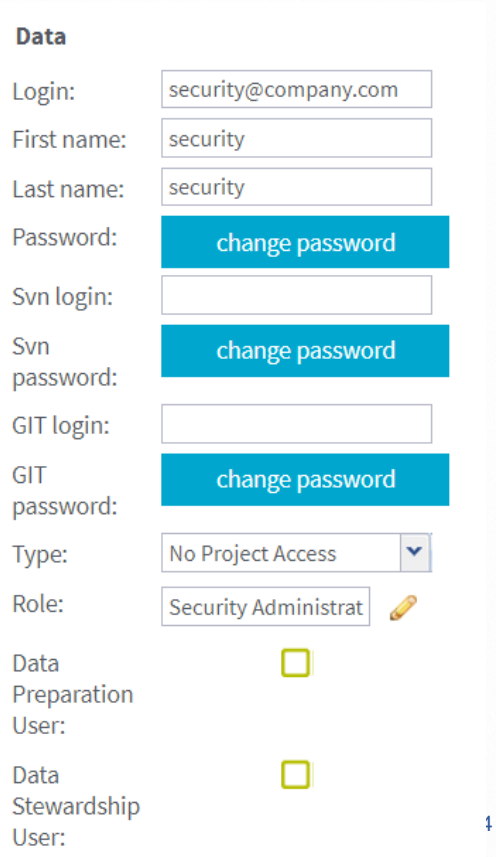


https://cms.talend.com/display/PRIC/Pricebooks

Licensing of the Security Admin

- If a user is configured exactly as:
  - Only one role: Security Administrator
  - No Project Access (aka no Studio)
  - No Data Preparation roles
  - No Data Stewardship roles

- Then the user is not counted against any license quota / free of charge



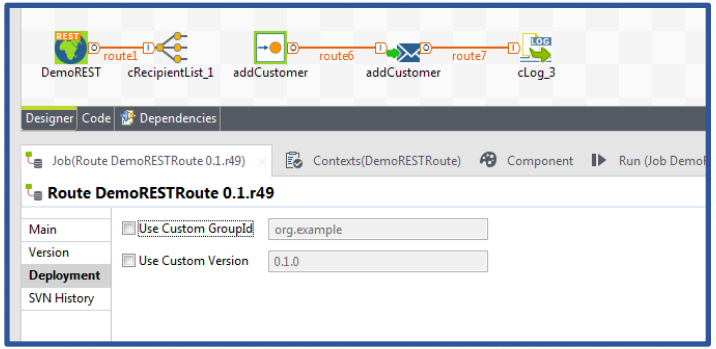
TAC 6.4: Deprecate TAC features

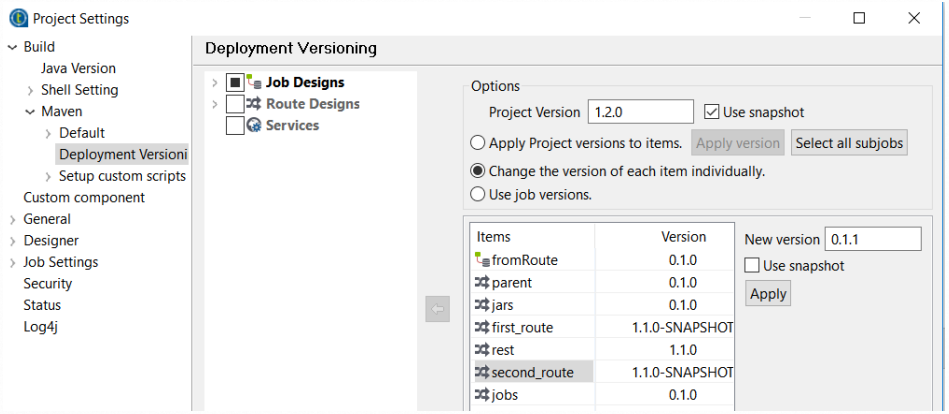
- TAC: Publisher
- TAC: SVN/Git based task in Job Conductor
- TAC: SVN Backup
- TAC: DB Backup
- TAC: Repository Browser
- Repository Manager: Repository Migration

Continuous Integration

Custom GroupId and Custom Deploy Version

- Jobs, Routes and Data Services
- Allow a custom GroupId and Version





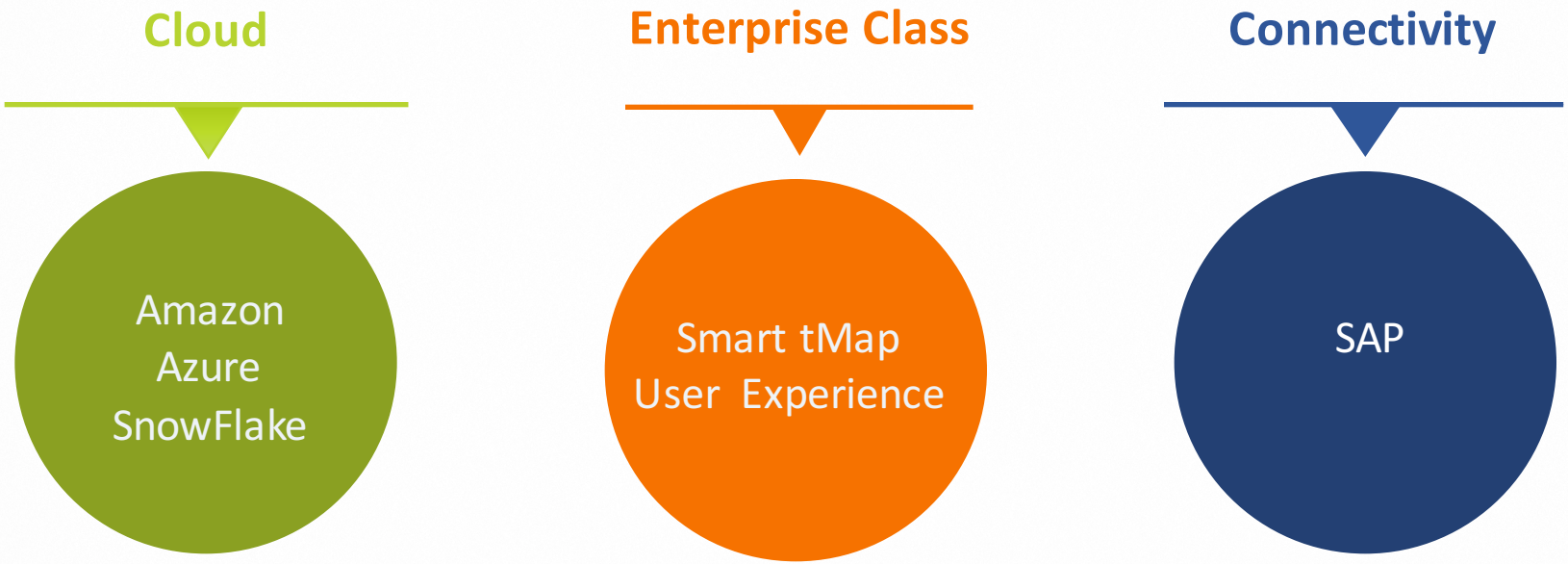
- Bulk changes at Project level

Misc

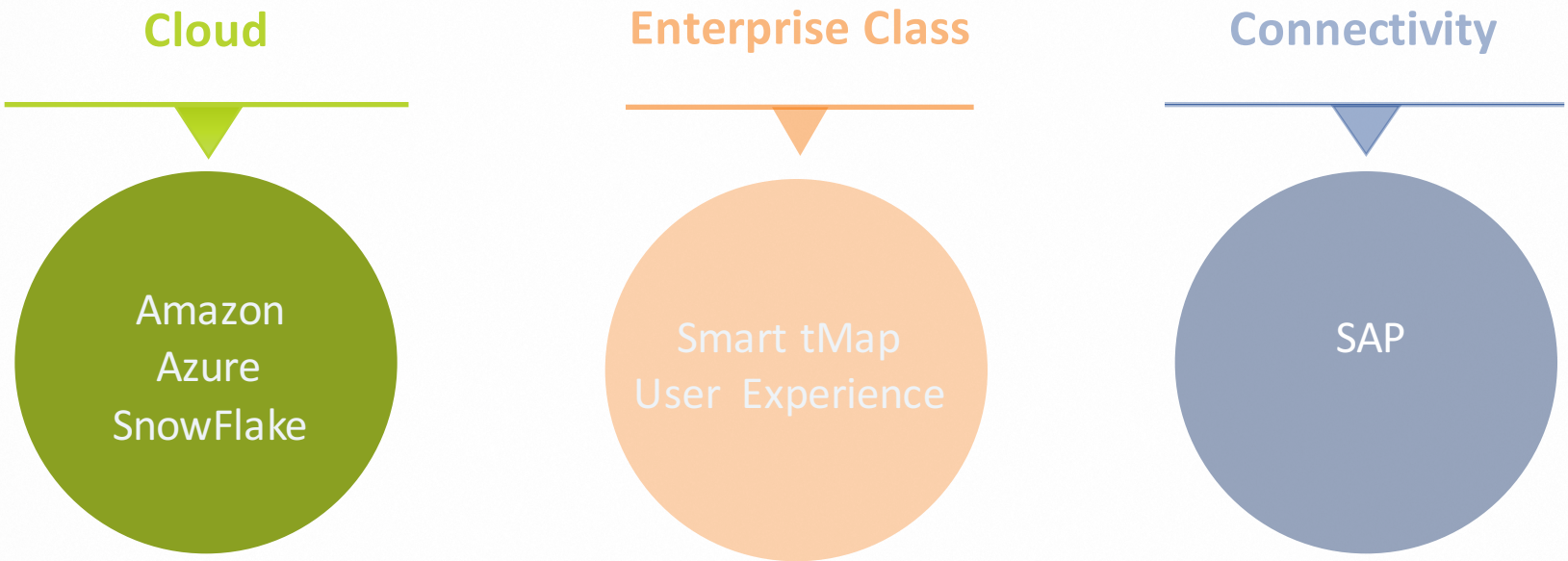
- Documentation how to build a single job using Maven and CI Builder
- Nexus 2.14

Data Integration

Talend Data Integration 6.4



Talend Data Integration 6.4



Cloud connectivity

Loading 20x faster that regular output
Supports dynamic schema

New Guess Query and guess Schema
Retrieve Id used when creating record
Bulk extraction with query mode (PK-Chunking)

Cloud connectivity: Azure

Refactor of Azure Blob

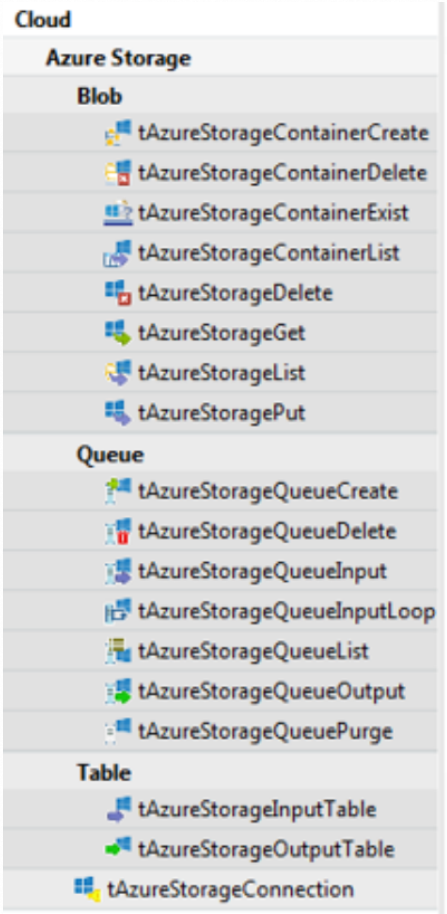
New components

Azure Storage
(Microsoft implementation of a distributed NoSQL big table.)

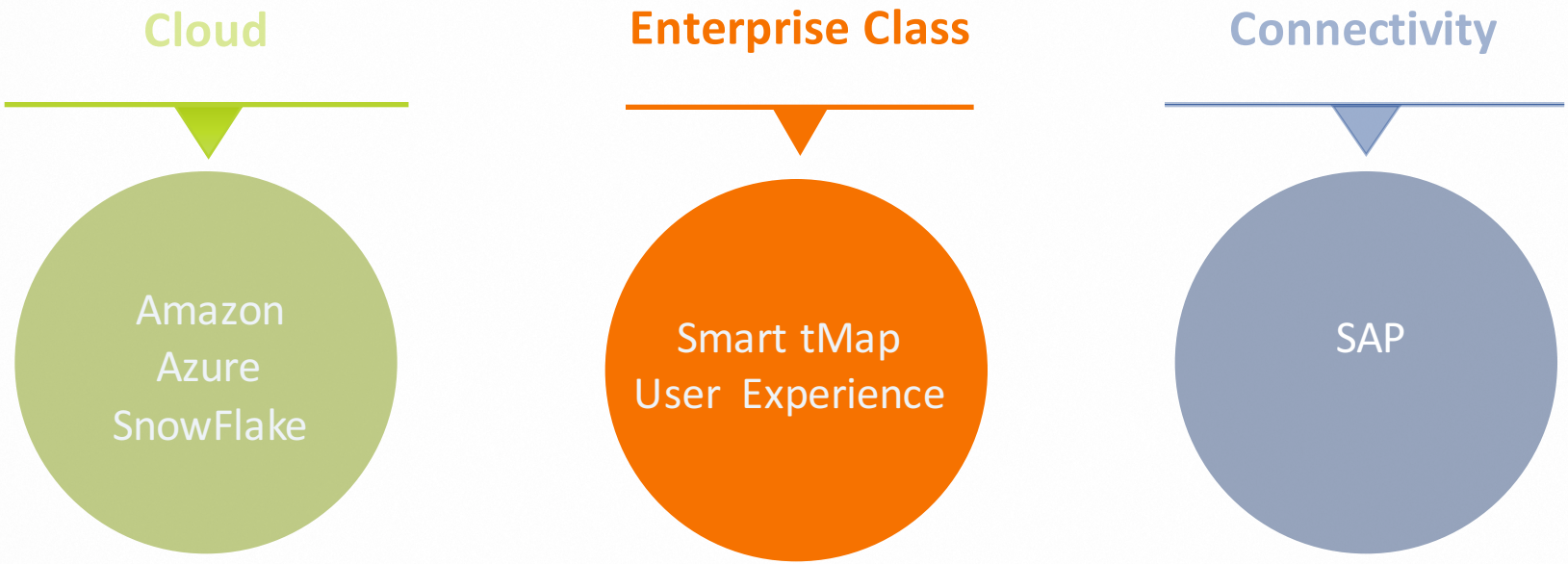
Azure Queue
(Similar to Amazon SQS)

Azure CosmoDB (formerly DocumentDB)
Support Mongo API only

AzureDWH (still working in progress)



Talend Data Integration 6.4

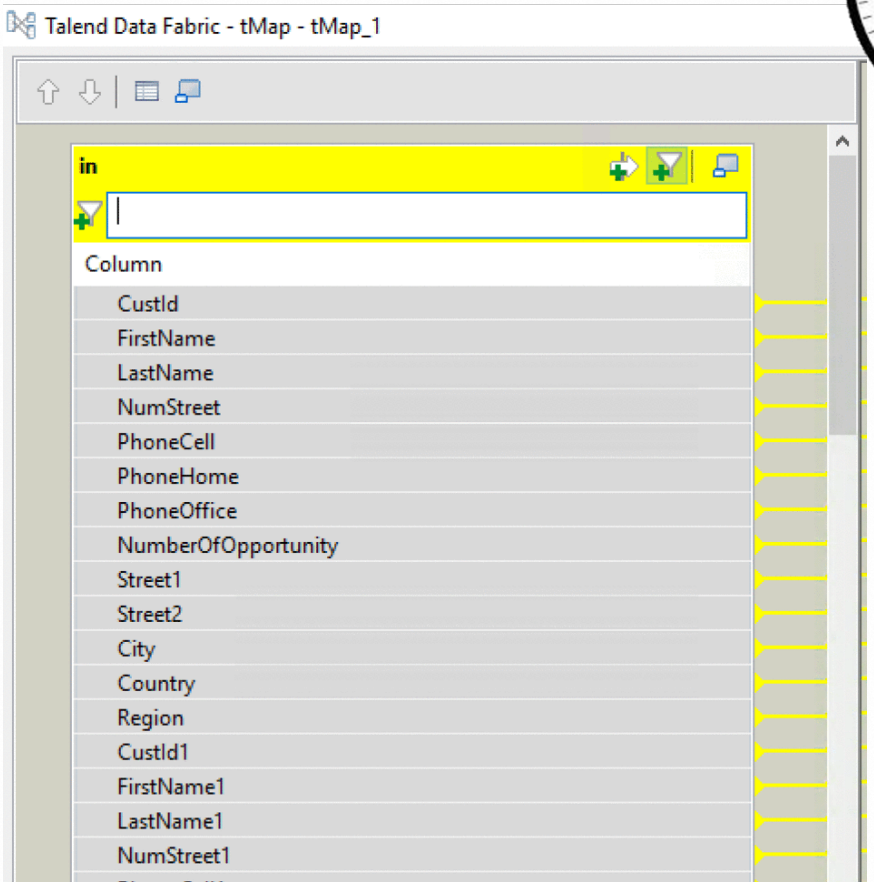


Productivity: Dynamic Column search

Find columns you want to work with in seconds!

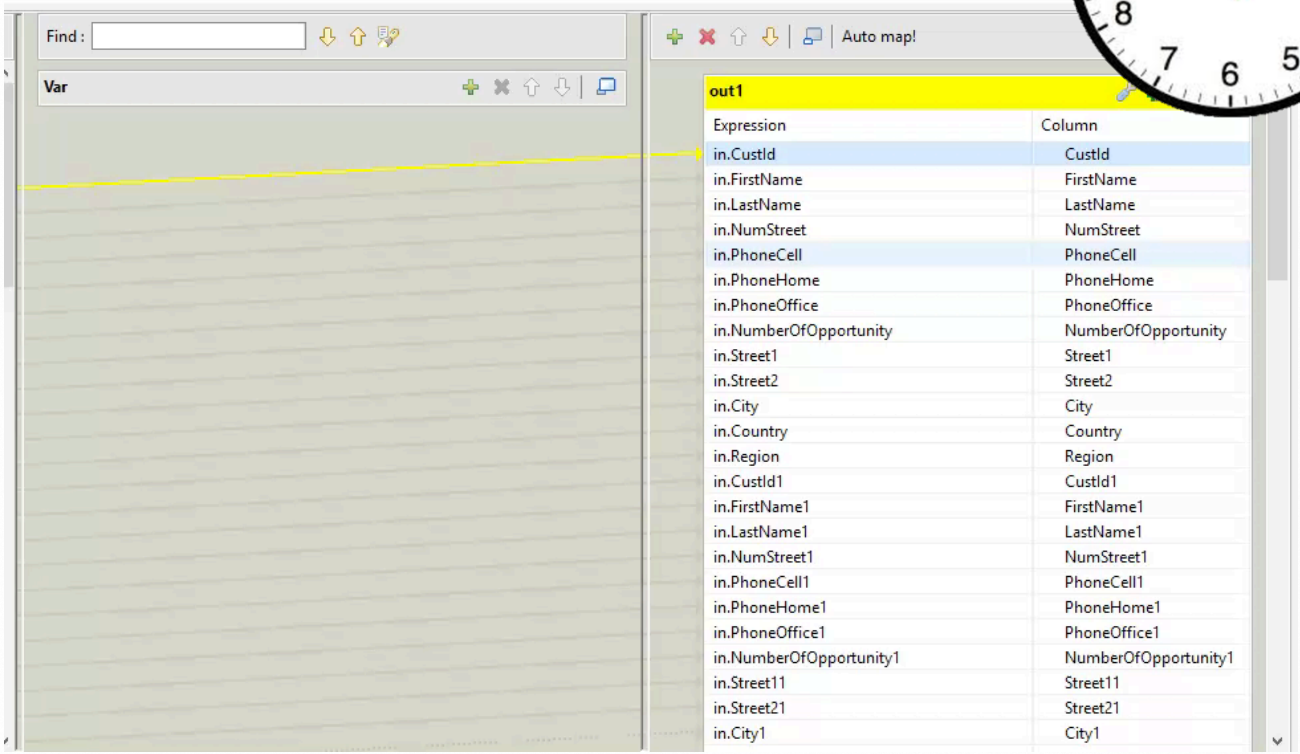
Works for sources &Targets

Full Text Search engine



Productivity: Batch Change expressions

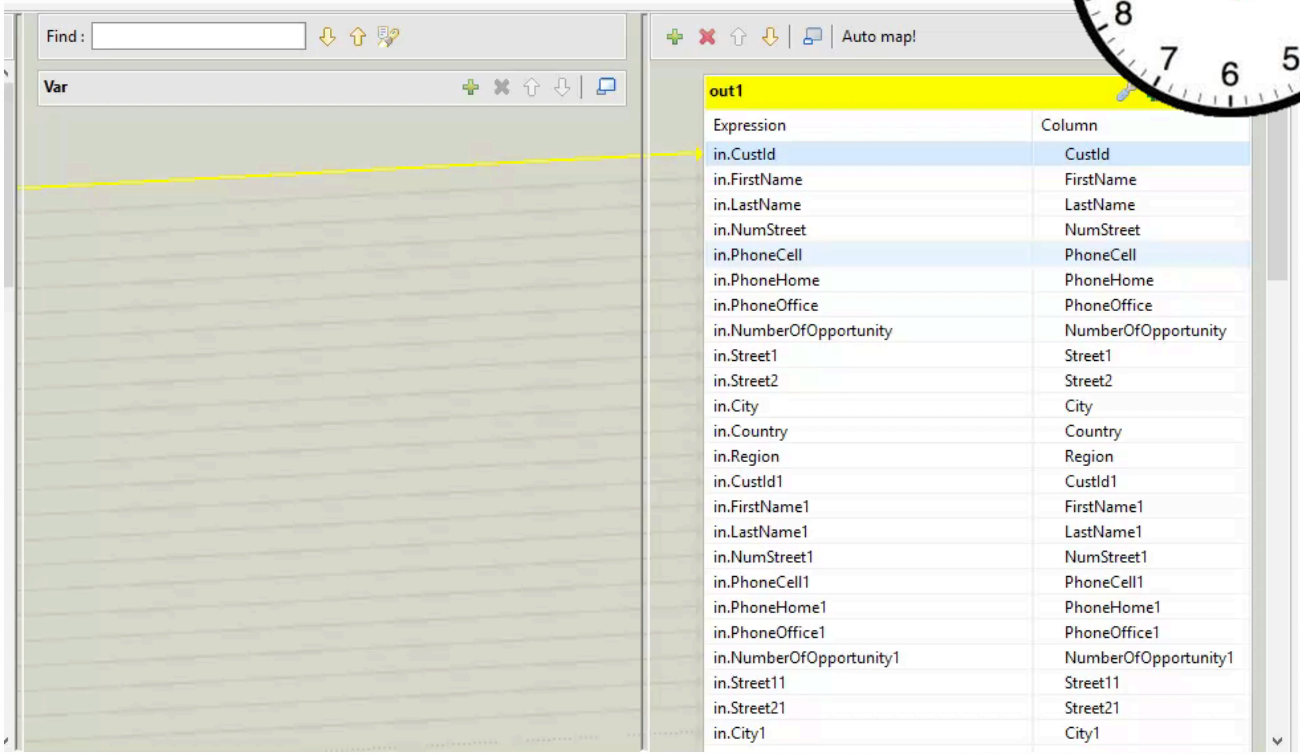
Excel style batch expression change. Massively apply transformation to selected columns



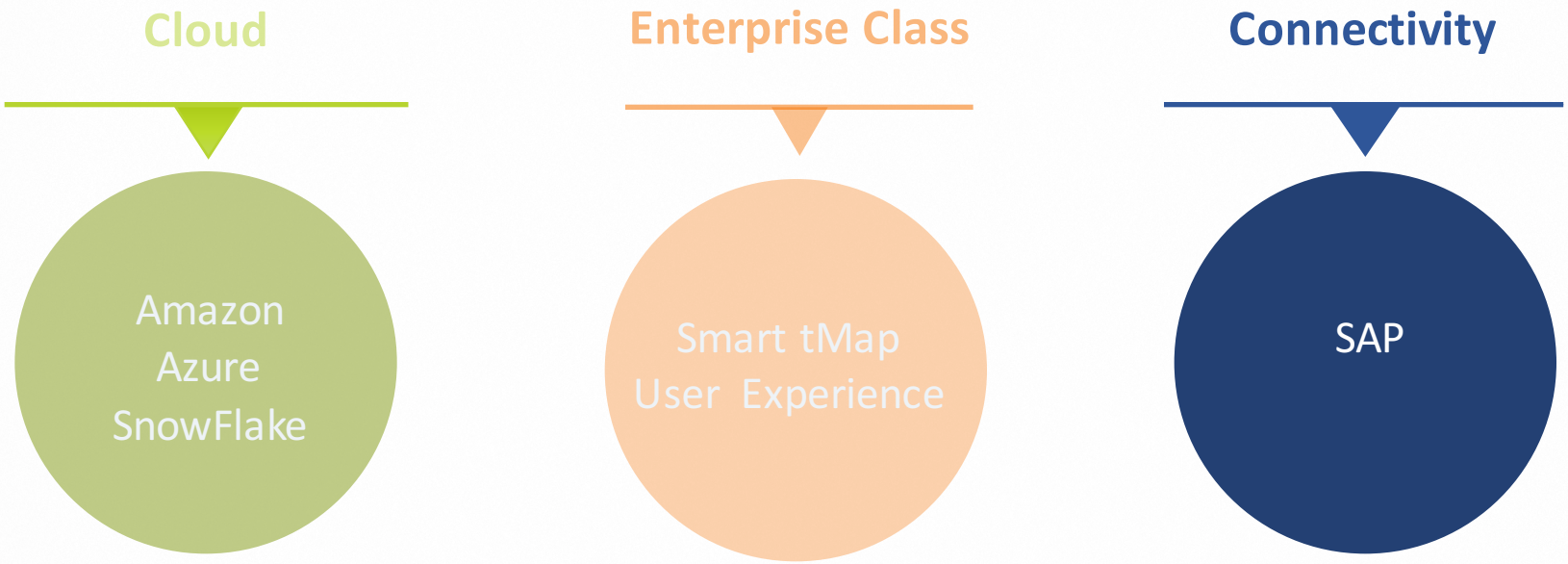
Productivity: Batch Change expressions

Excel style batch expression change.

Massively apply transformation to selected columns



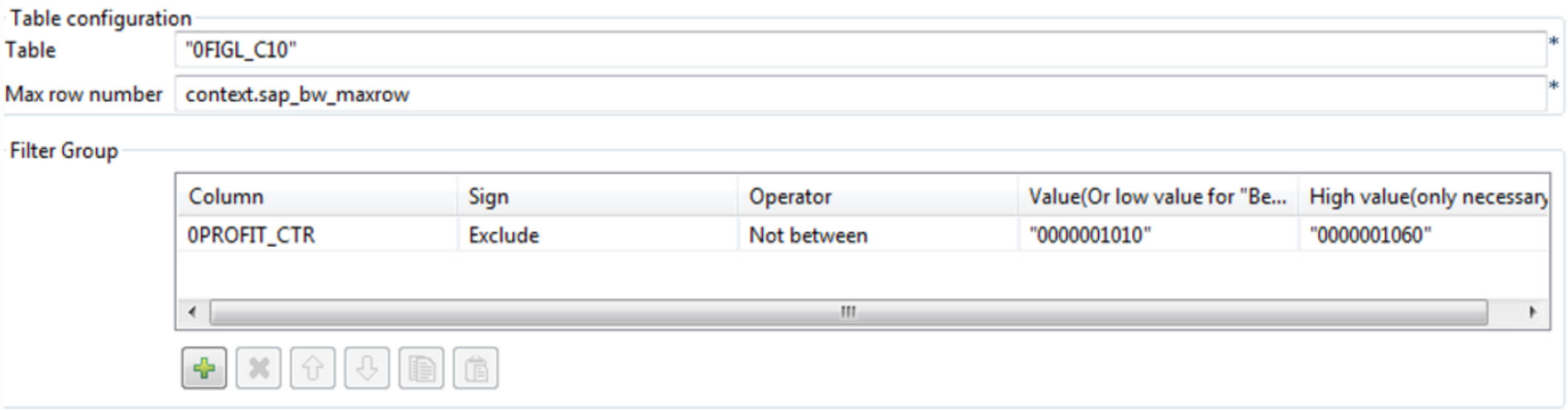
Talend Data Integration 6.4



SAP: BW Filtering and dynamic extraction

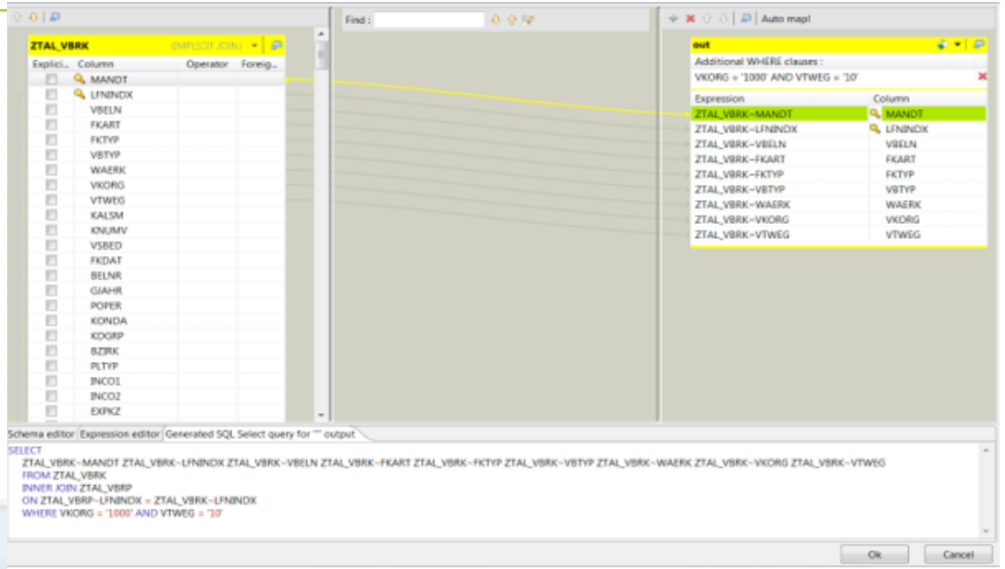
- Filter DataSource Objects and Infocube

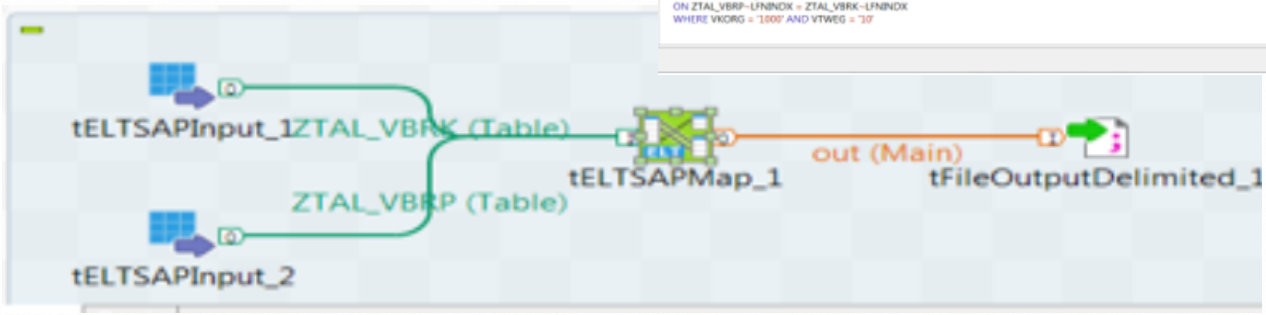
- They also support dynamic schemas



SAP: Bulk Extraction and ELT/Push-down

- End to end FTP transfer integrated.
- Graphical design without ABAP guru
- Server side file extraction





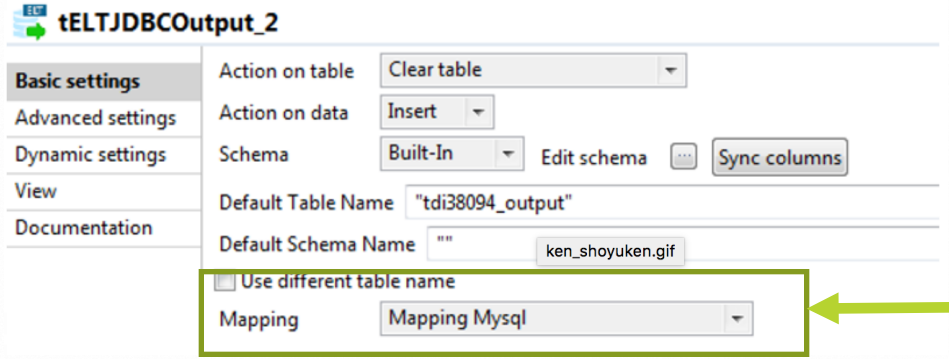
SAP: Misc

- RFC Server can use a Standalone ActiveMQ (High-Availability, Monitoring)

- BW 7.5 support

Industrialization: JDBC

- New Action on Table for ELT JDBC connectors



Do not forget Mapping

- New SCDELTDB connector
  - Supports Netezza, Redshift, Exasol.
  - Should work with others Databases.

Other upgrades

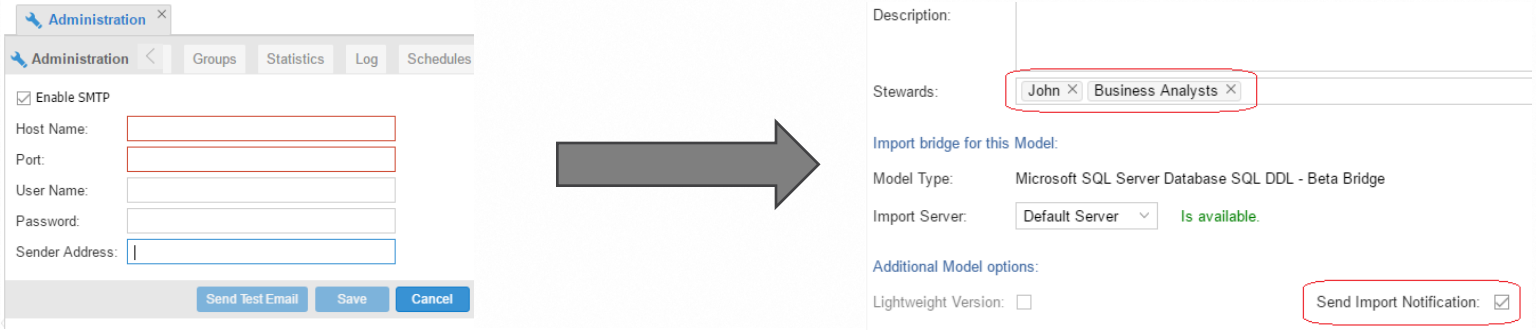
- NetSuite (API upgrade)
- New Microsoft CRM cloud (365) support
- Exasol (support v6 + Dynamic Schemas)
- Sybase (support v16)
- Excel component upgrade
- New AWS SQS components (tSQSDelete, tSQSMessageChangeVisibility)

talend

Metadata Management

Notification

- Subscribe to changes of metadata and get notified when a metadata **you are** interested in gets harvested or changed



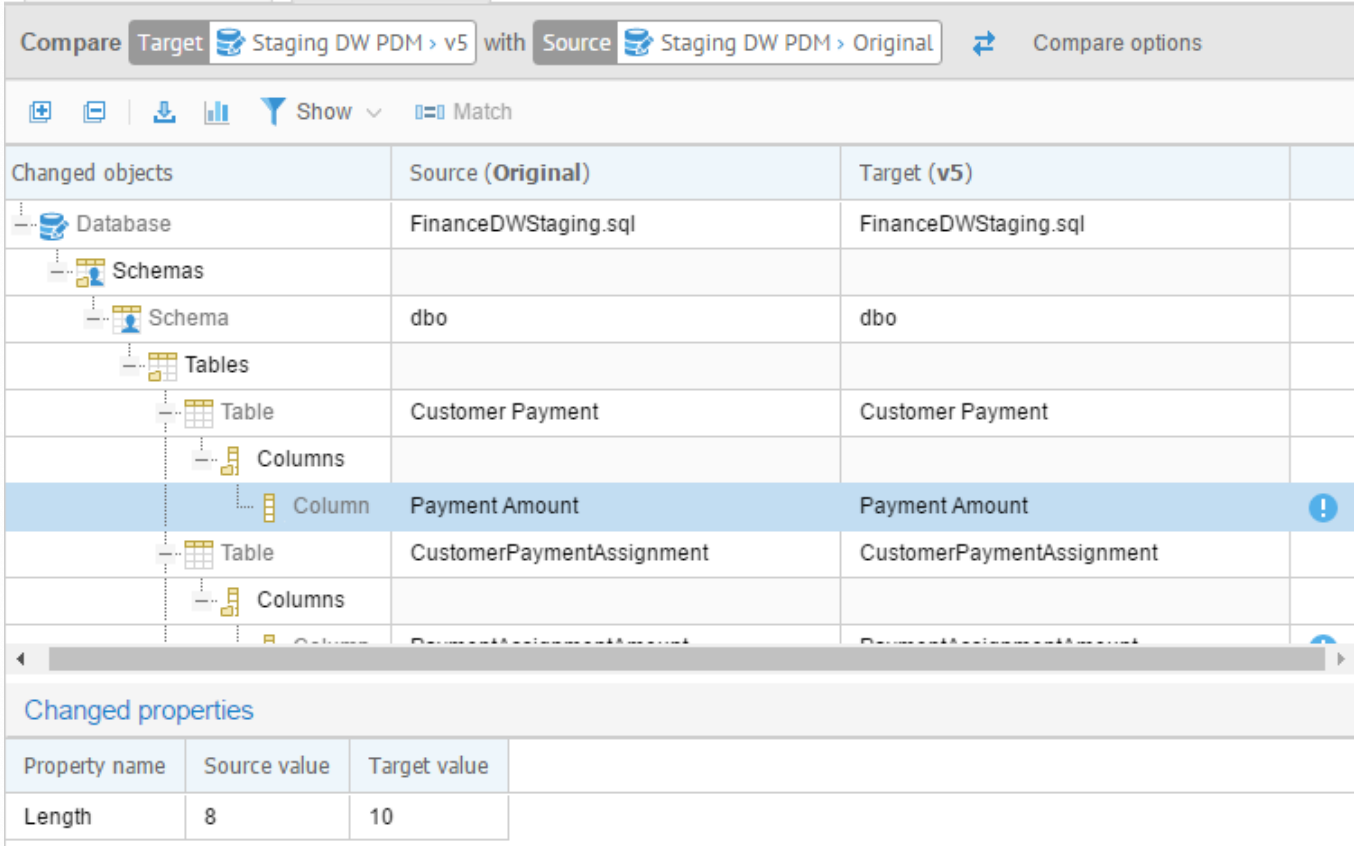
The content /New Folder1/Tutorials/Metadata Management/End of Chapter 6/2 - Data Warehouse/Staging DW PDM was imported at 2017-06-02 09:29:28.

There is no error in the import log.
A new version is created.
Summary of differences from the previous version:

DISPLAY NAME	ID	ADDED	REMOVED	CHANGED
Column	305000003	0	0	9
Relationship	305000016	47	11	3
Primary Key	305000019	29	29	0
Table	305000026	0	3	0

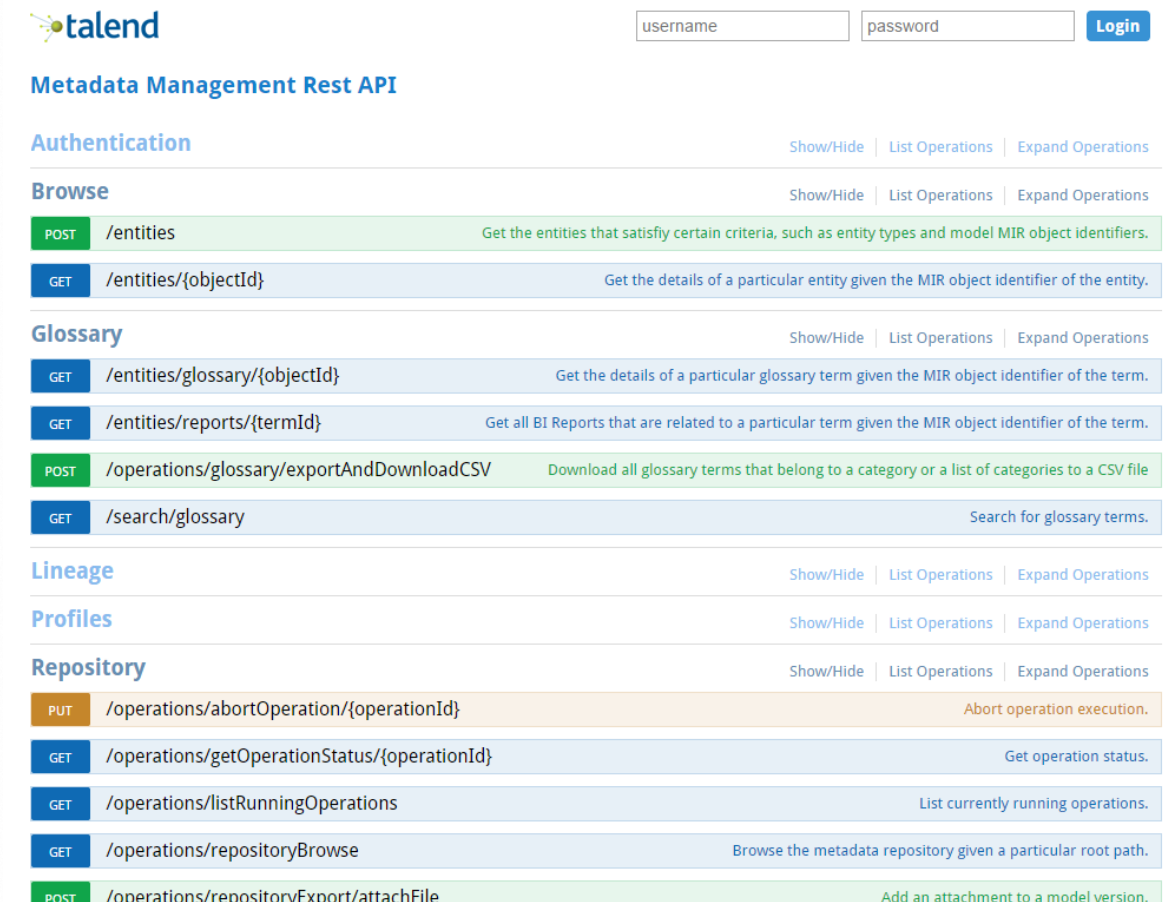
Metadata Change detection

- Only Harvest metadata when required
- Harvest faster
- Save storage $$$
- Model comparison



New Rest API

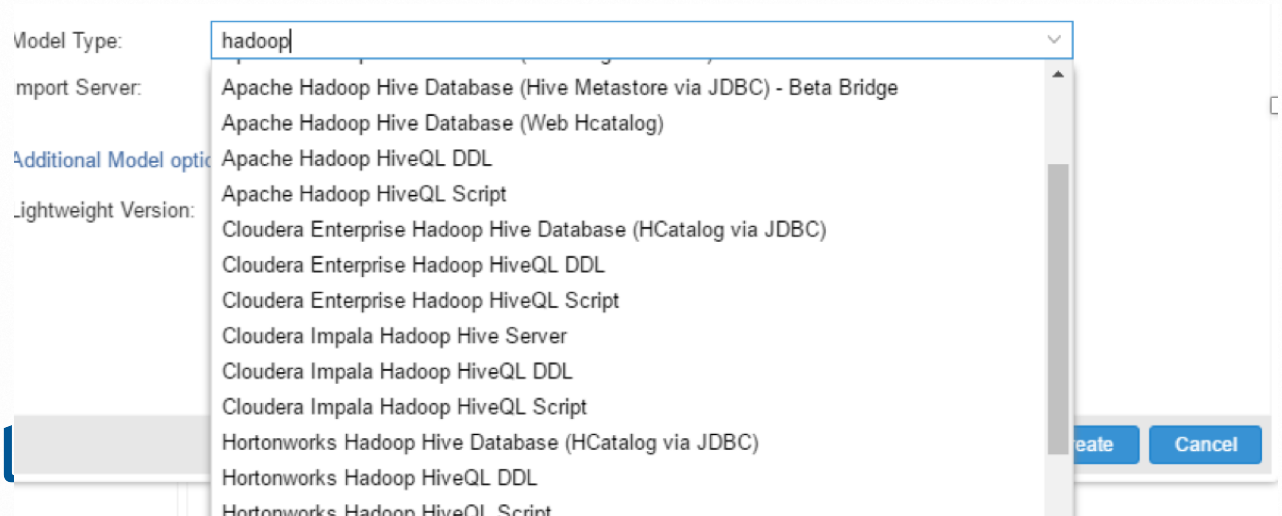
- Take control of TMM from exposed API.
- Load, search and insert metadata
- Easier access from business glossary



talend

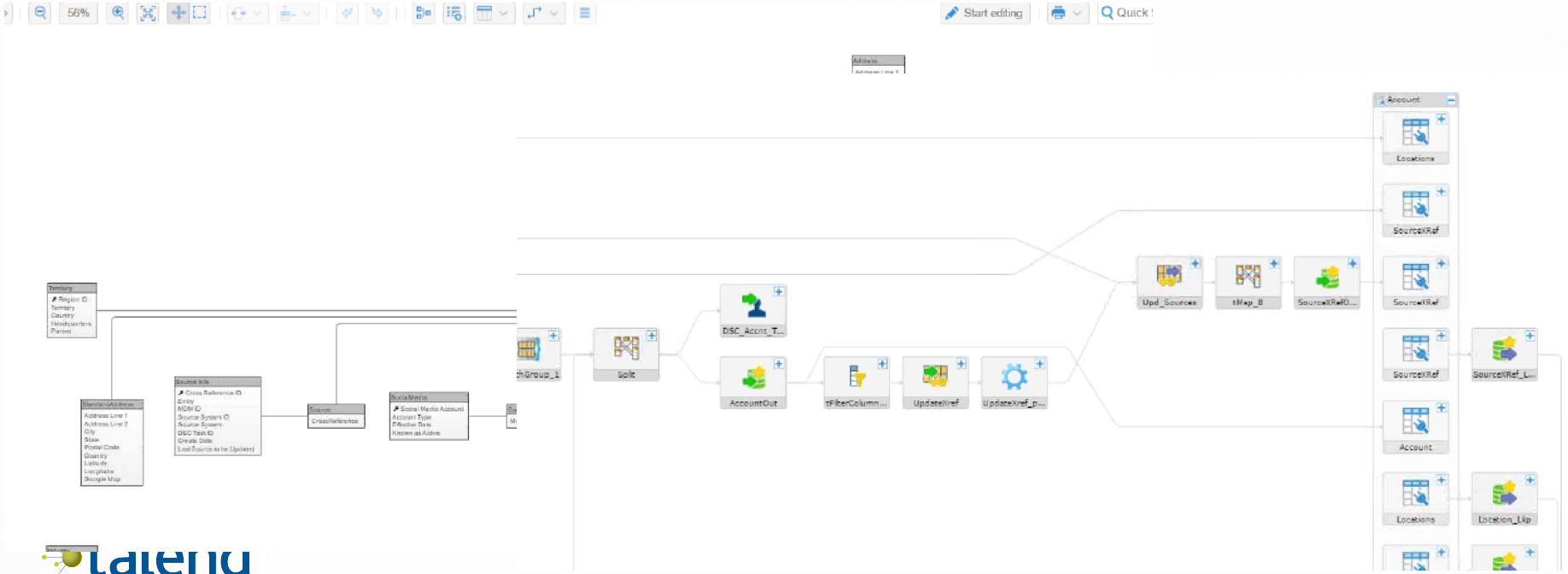
Big Data Inventory

- Harvest Big Data Sources and extract their metadata
- HDFS, Hive, AVRO, PARQUET, WebHDFS, MongoDB , CouchDb, MarkLogic, Cassandra and others.
- Full list at http://www.metaintegration.net/Products/MIMB/MIMBReadme.html



New MDM Integration

- Integration of MDM into TMM
- End to end lineage of metadata



New Bridges and improvements

- SQL Parsing enhanced
- Atlas bridge (import)
- Amazon S3
- JSON Files

Data Quality

Data Quality 6.4 in a nutshell

Dictionary service

- UI in Data Prep & Data Stewardship
- Compound types
- Fine-grain validation

Matching with ML on Spark

- Labeling via Data Stewardship
- Continuous matching

Survivorship enhancements

Natural Language Processing

- Named entity recognition

Dictionary Service

Reminder - Dictionary service in 6.3

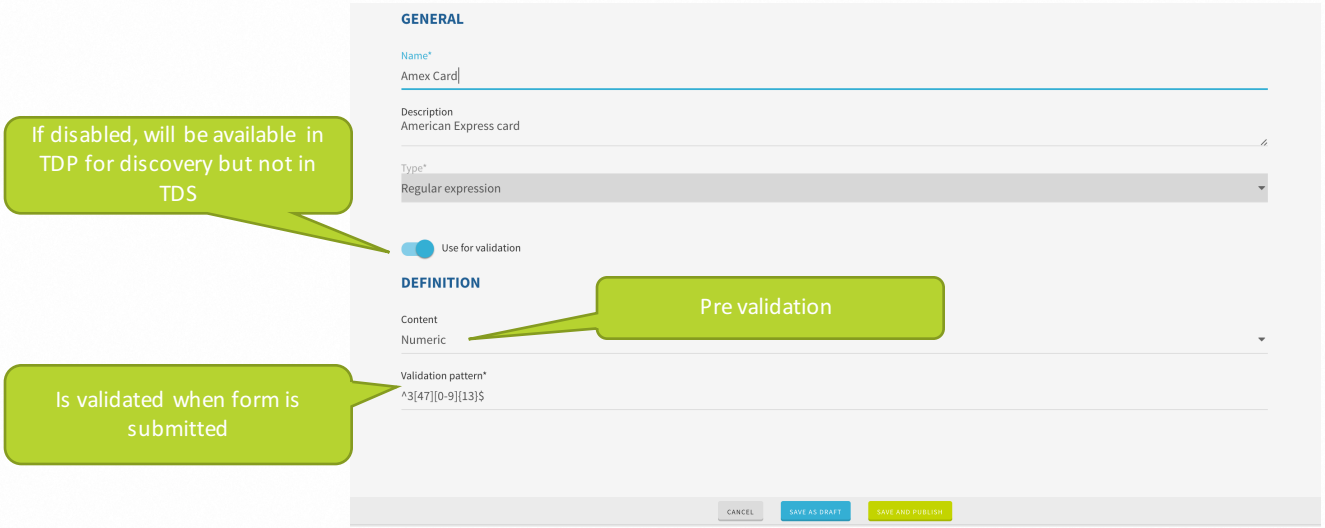
- The Dictionary service was introduced in 6.3
- Manages semantic types definition – custom and predefined ones
- Integrates with Data Prep and Data Stewardship

Type	Data Prepaparation	Data Stewardship
Regular expression	Discovery / validation	Validation
Open dictionary	Discovery	N/A
Closed dictionary	Discovery / validation	Validation
Keywords (*)	Discovery	N/A

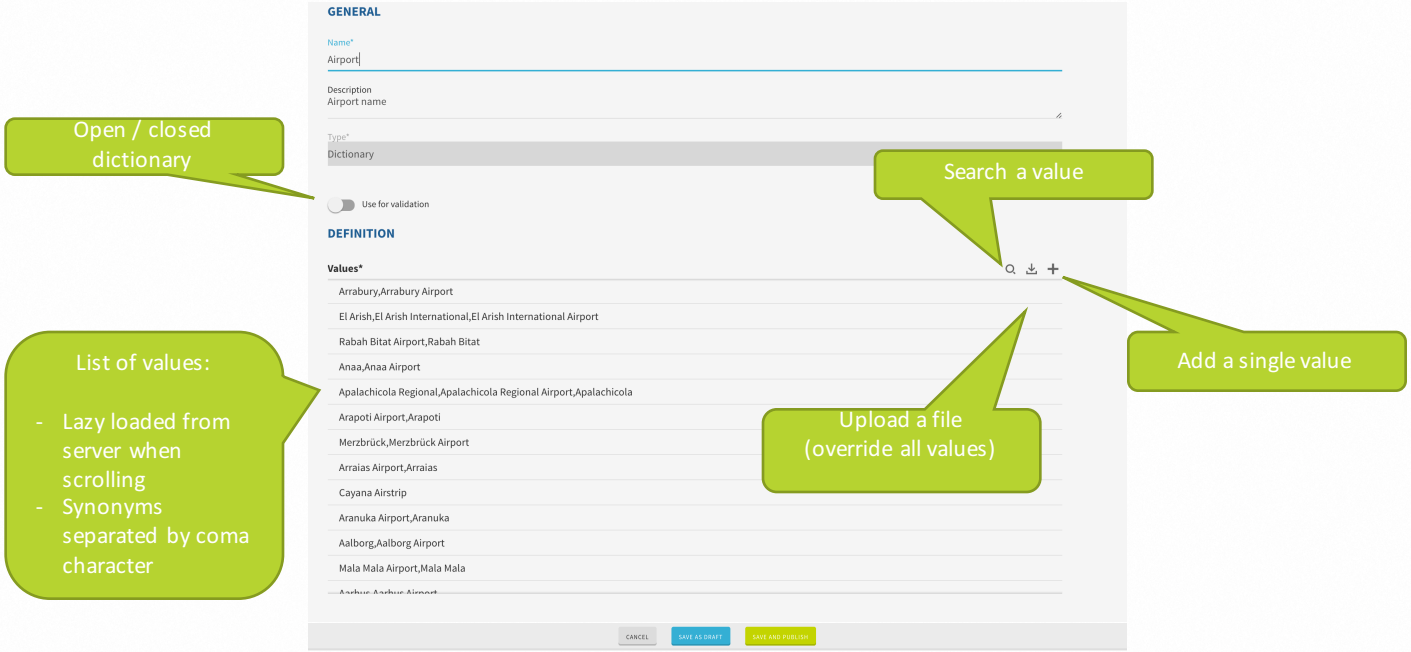
(*) Keywords were not available in dictionary service, but only packaged with Data Prep

Semantic type management in Prep & Stewardship

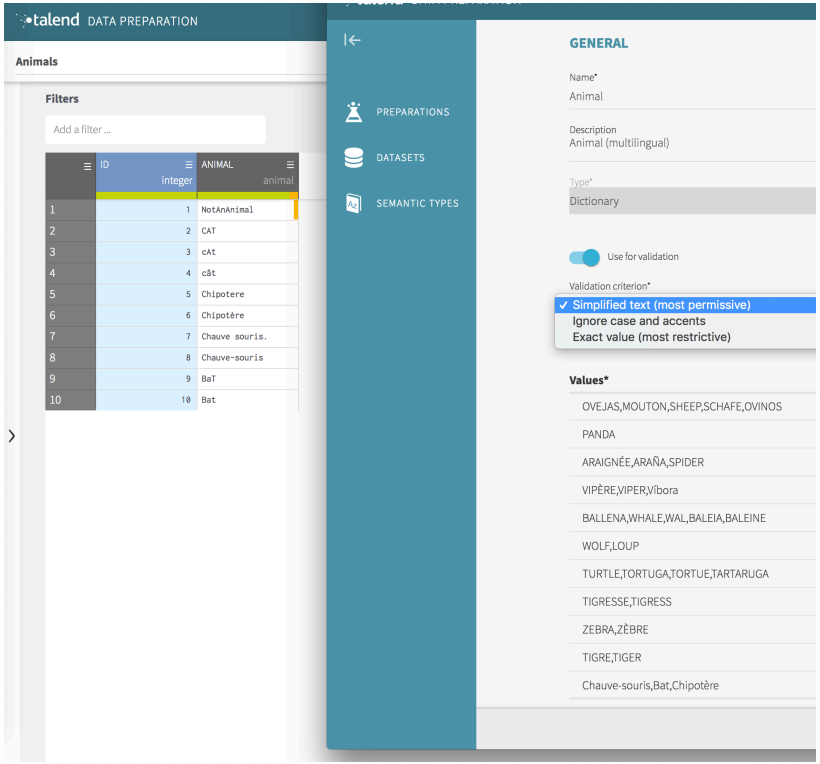
Regular expressions



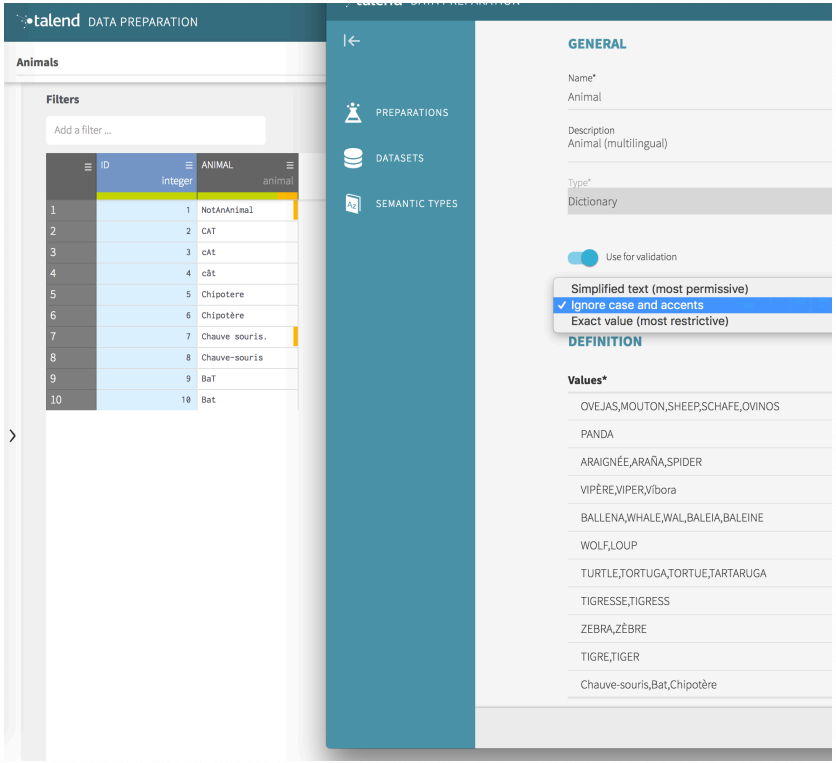
Dictionaries



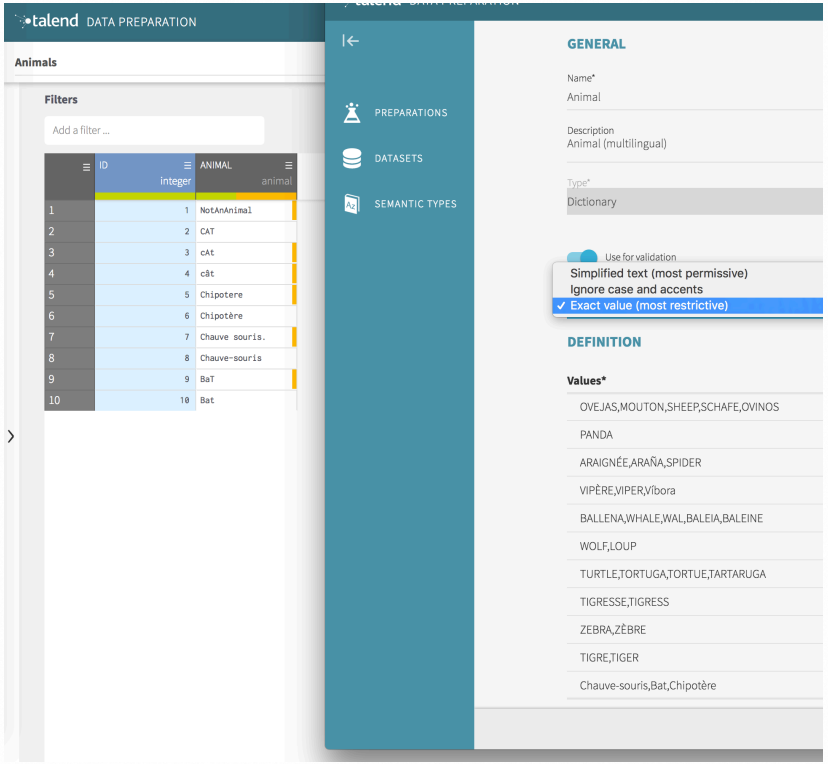
Validation – Simplified text *(6.3 behavior)*



Validation – Case and accent insensitive



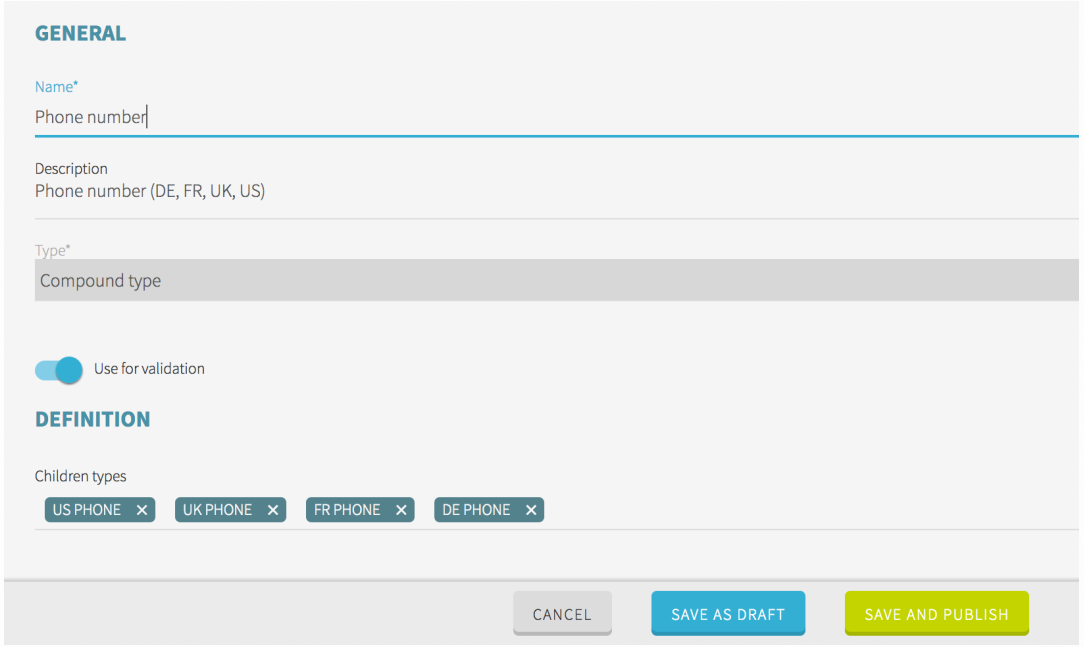
Validation – Exact value



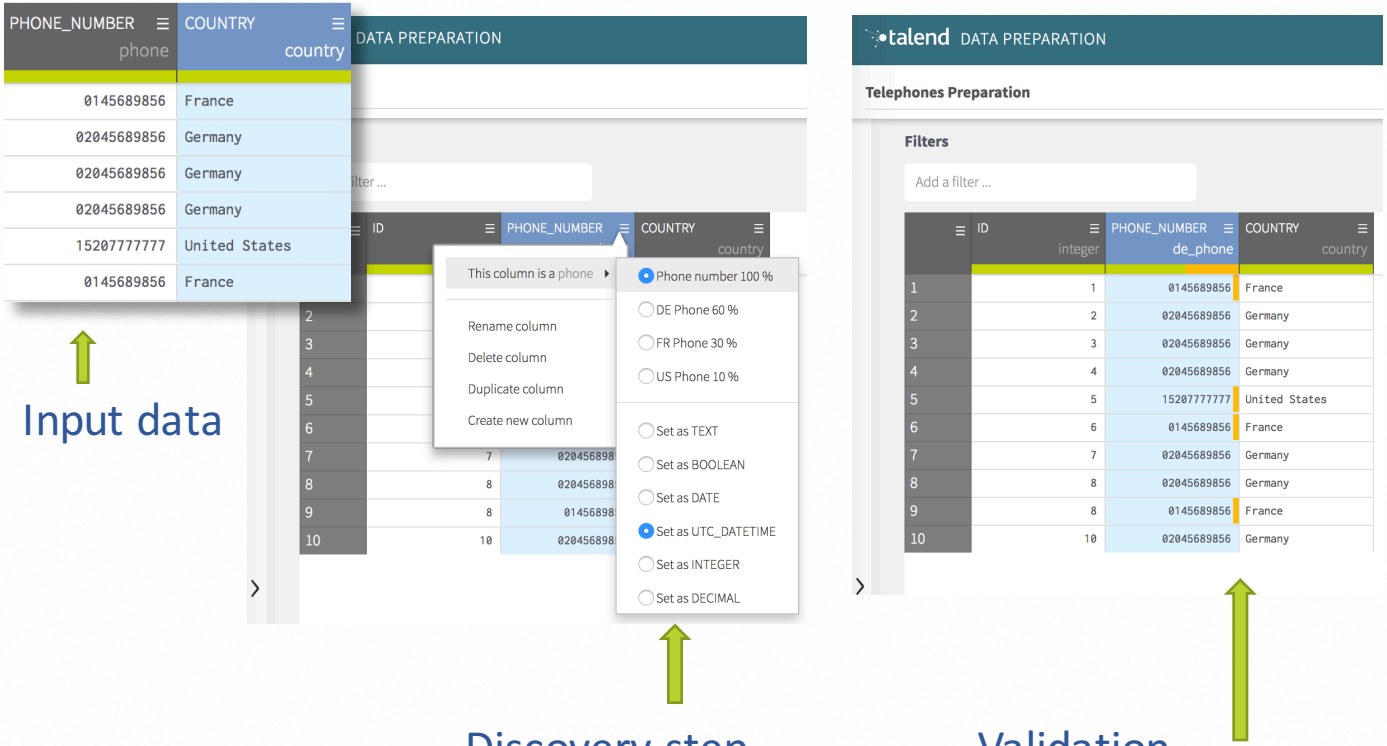
Compound Semantic Types

- A compound semantic type refers to other semantic types
- Compound types can be hierarchical/cascading
- Can be used for validation

- Examples (predefined):
  - North American states
  - Phone numbers

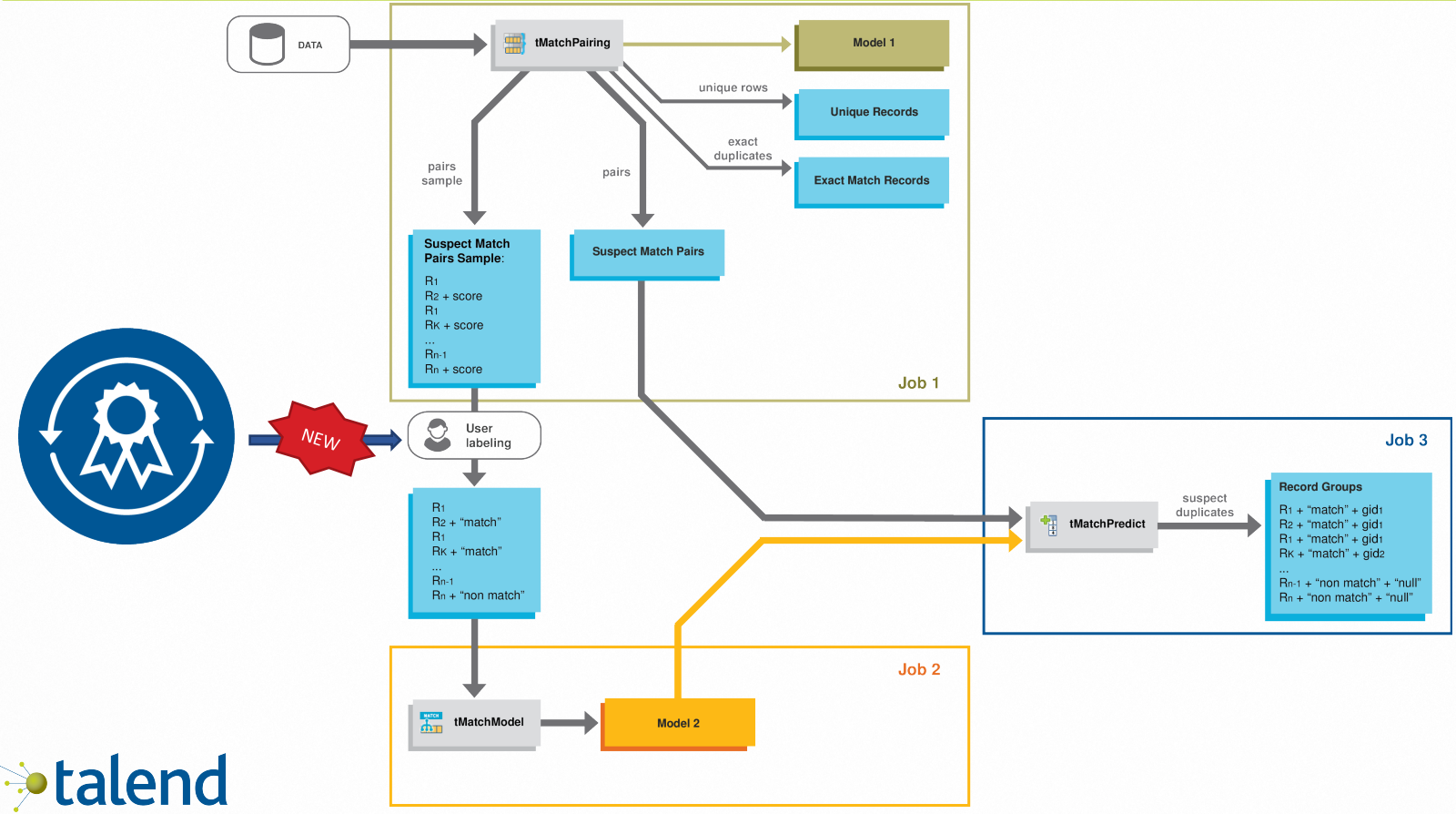


Compound Semantic Types

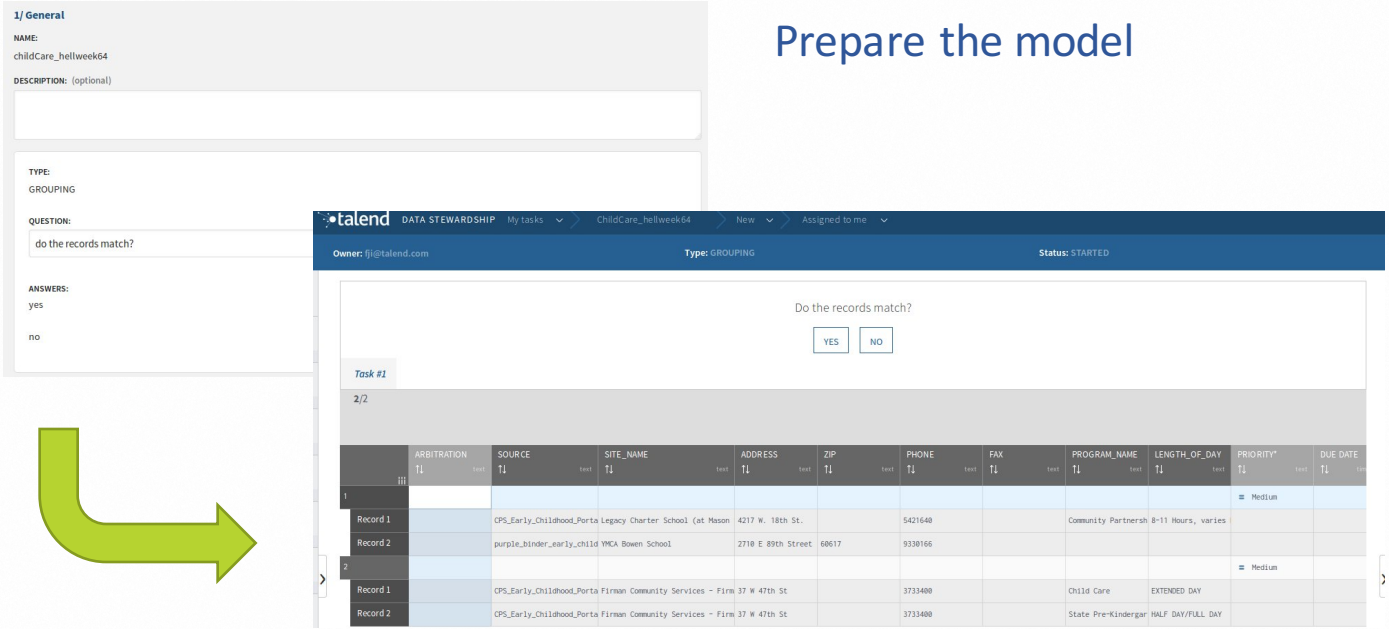


Matching with ML on Spark

Matching with Machine Learning



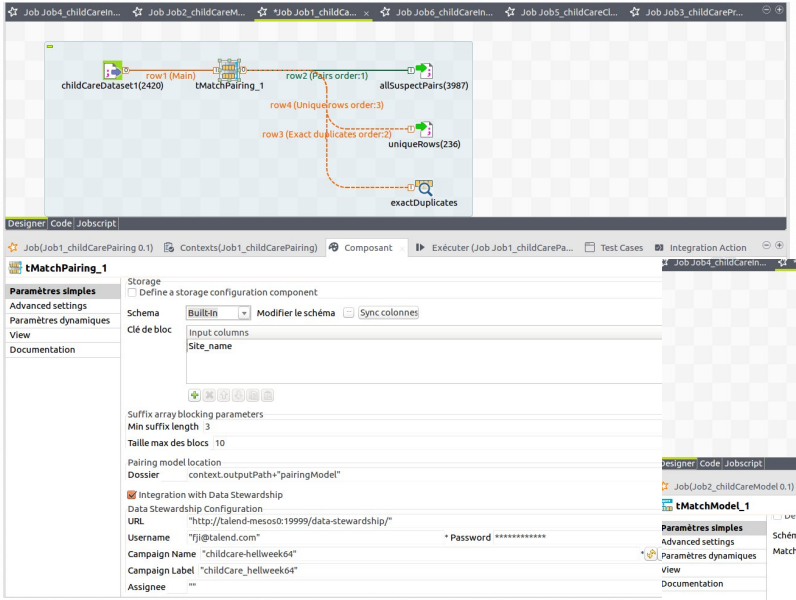
Stewardship grouping campaign



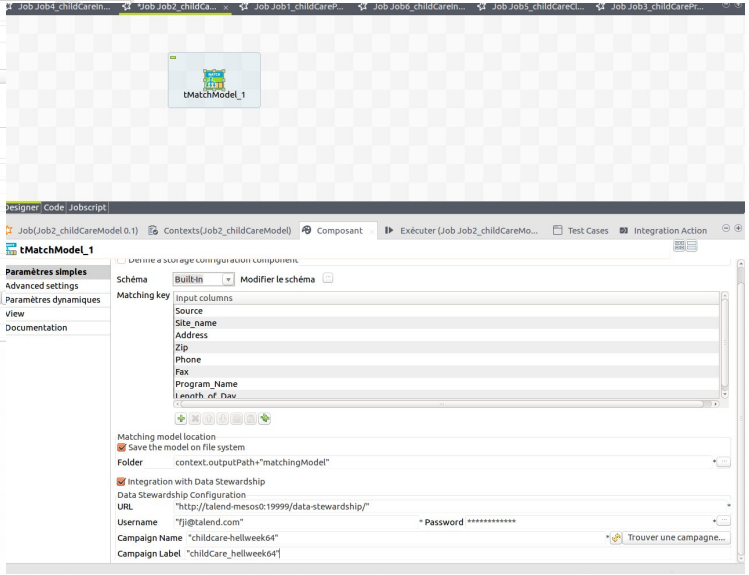
Prepare the model

Label pairs of records

Connect components to Data Stewardship



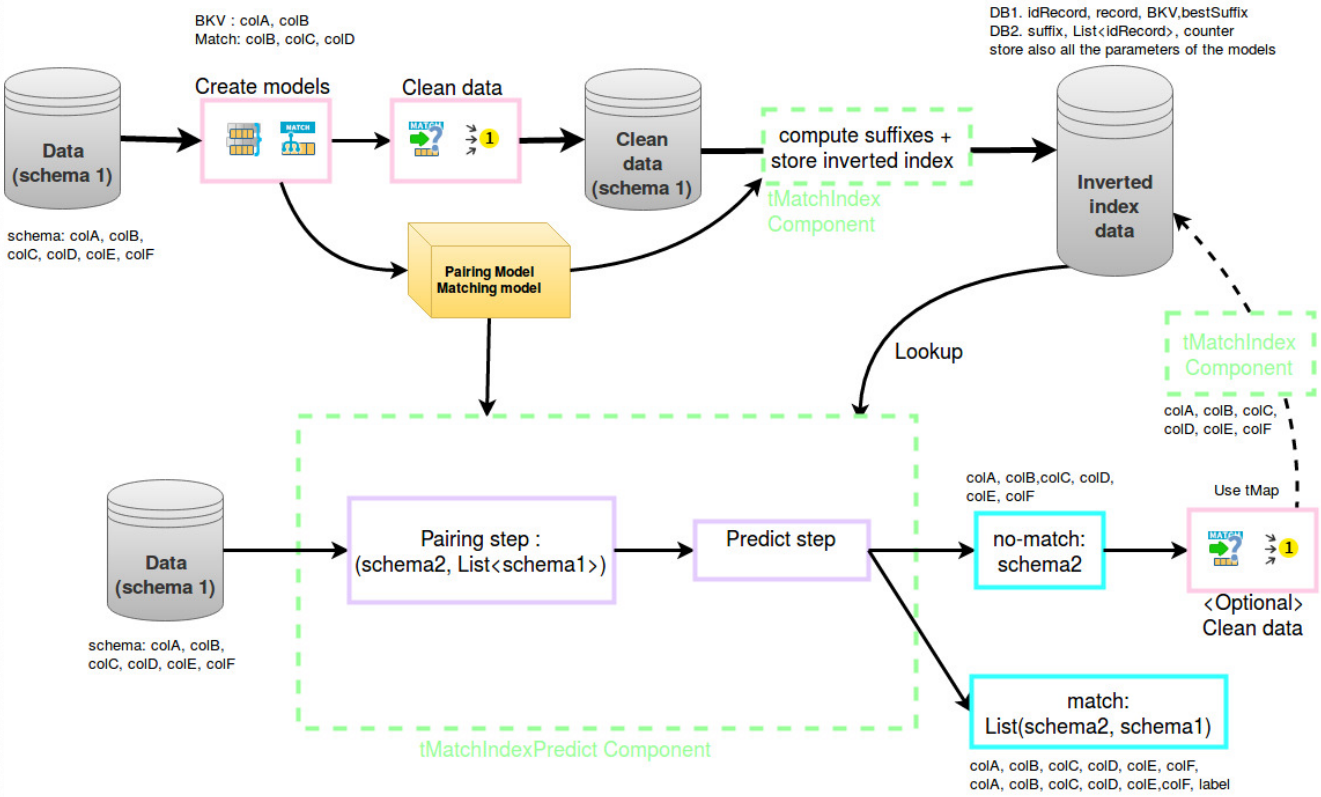
Write into Data Stewardship



Read from Data Stewardship and learn ML model

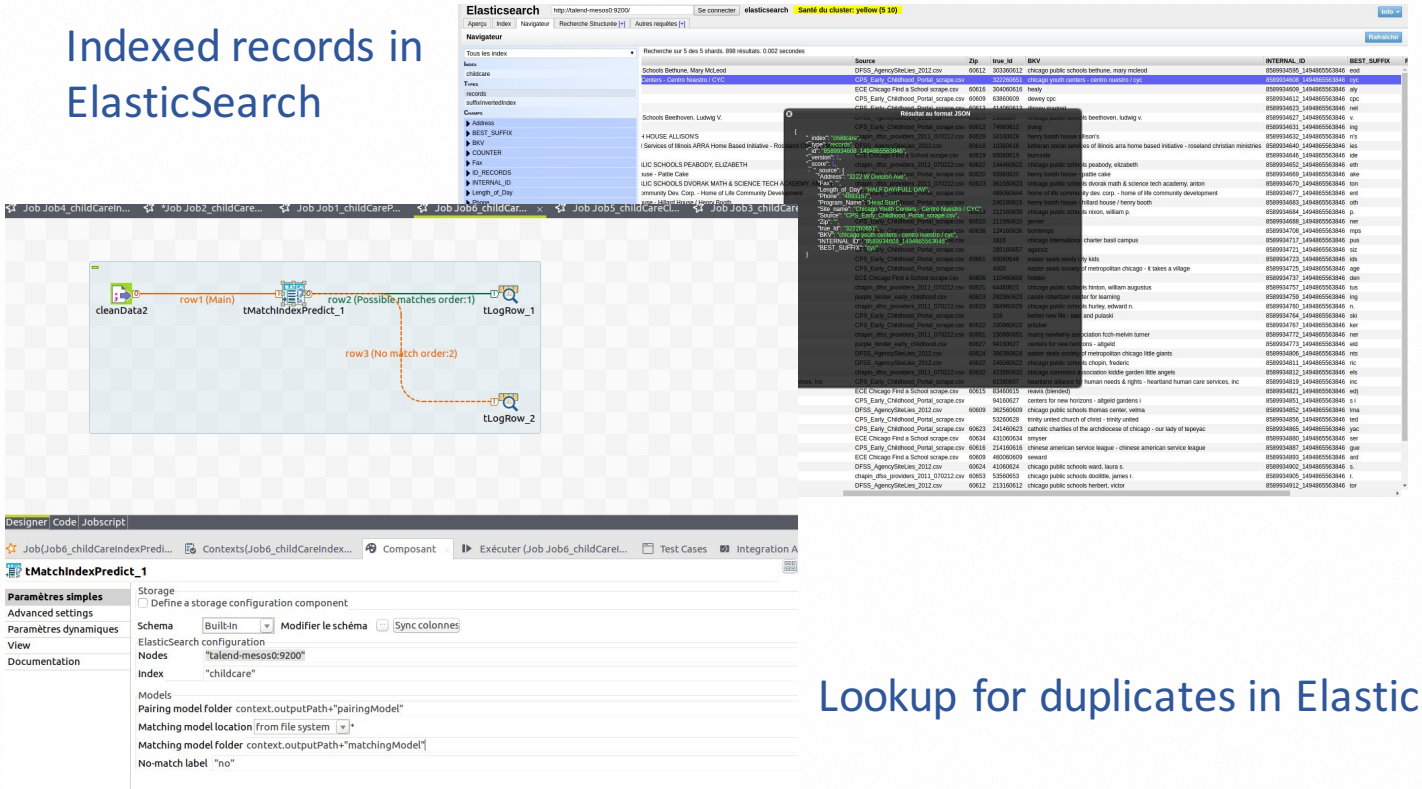
Continuous matching

Aka "rematch a limited number of records against the humongous initial dataset"



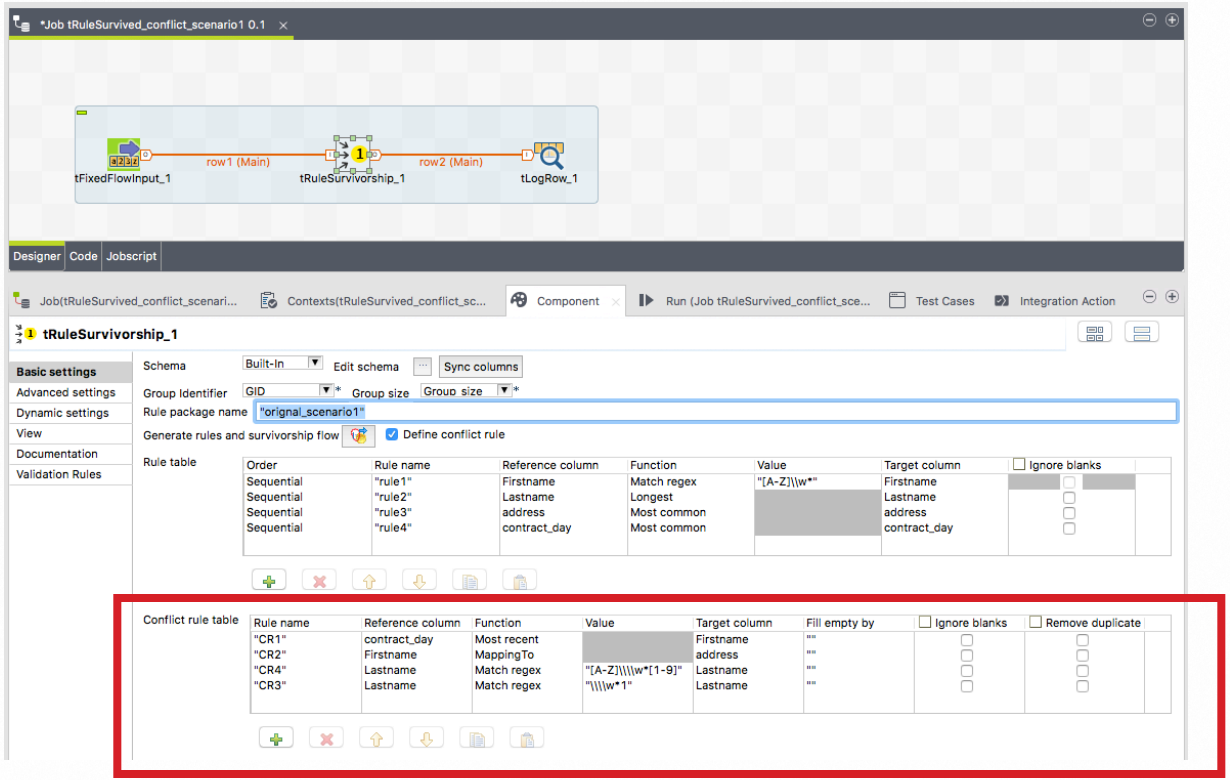
Match new records with deduplicated records

Indexed records in ElasticSearch



Lookup for duplicates in ElasticSearch

Survivorship

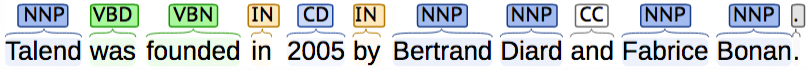


Natural Language Processing

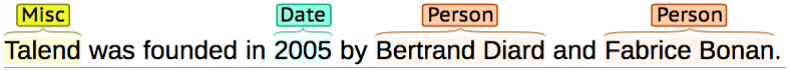
Natural Language Processing

- **What is natural language processing?**
  - Text tokenization
  - Sentence splitting
  - **Part-of-Speech tagging**

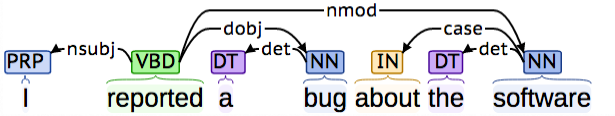
    http://www.clips.ua.ac.be/pages/mbsp-tags



  - Syntactic parsing
  - Shallow parsing (aka chunking)
  - **Named Entity Recognition**



  - Co-reference resolution
  - **Dependency parsing**



  - Sentiment analysis

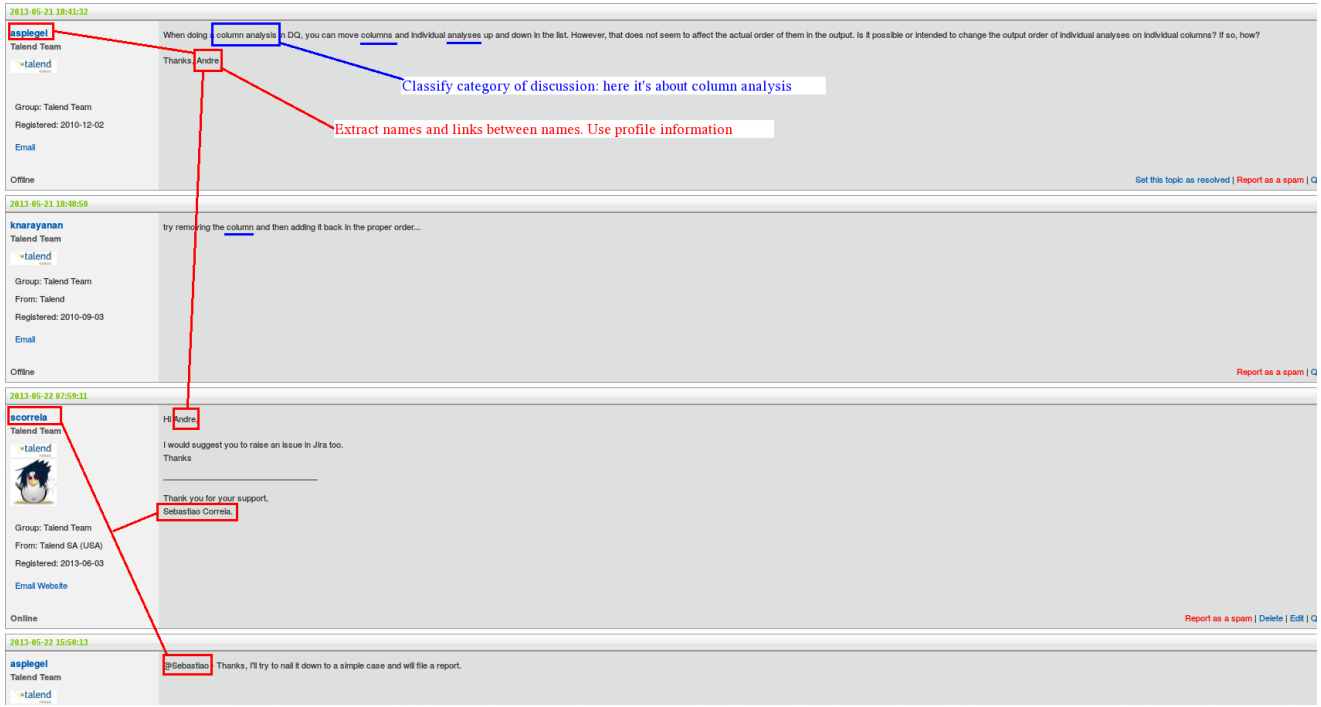
Play at http://nlp.stanford.edu:8080/corenlp/process

talend

Where can this be useful?

- Extract useful information from the textual resources (such as forums, notes in salesforce, etc.)
  - names of persons
  - names of companies (competitors…)
  - names of tools (concurrent tools…)
- Classify discussions by topics
  - Group discussions together
  - Find discussions where people are mentioned but don't participate to the discussion.
- Entity linking
  - links between profiles and mentions in the text
  - links between persons and organizations
  - links between persons and any other information that may be used for re-identification

Where can this be useful?



Relationship with data quality?

- Use textual data to get more information about your structured data
  - Analyze CRM notes
    - Extract contact names
    - Get information about their status (left the company, new phone number, got married and changed name…)
  - Compare them with the current values in your structured data
    - Contact information up-to-date?
    - Name changed?
    - Phone changed?
    - Address changed?
    - …

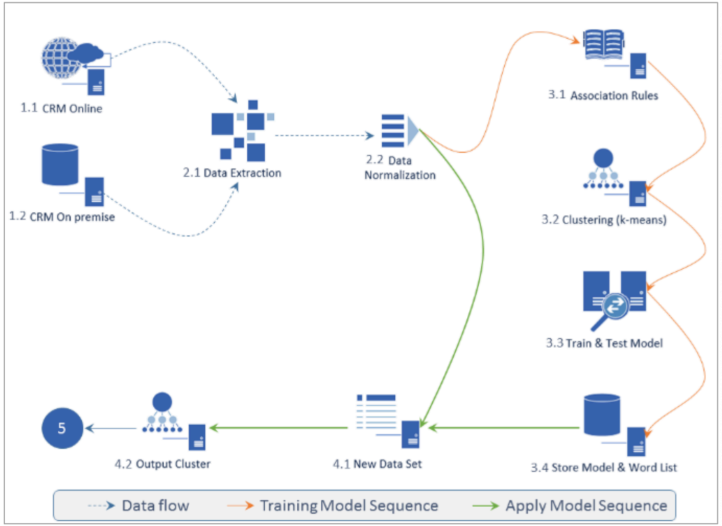
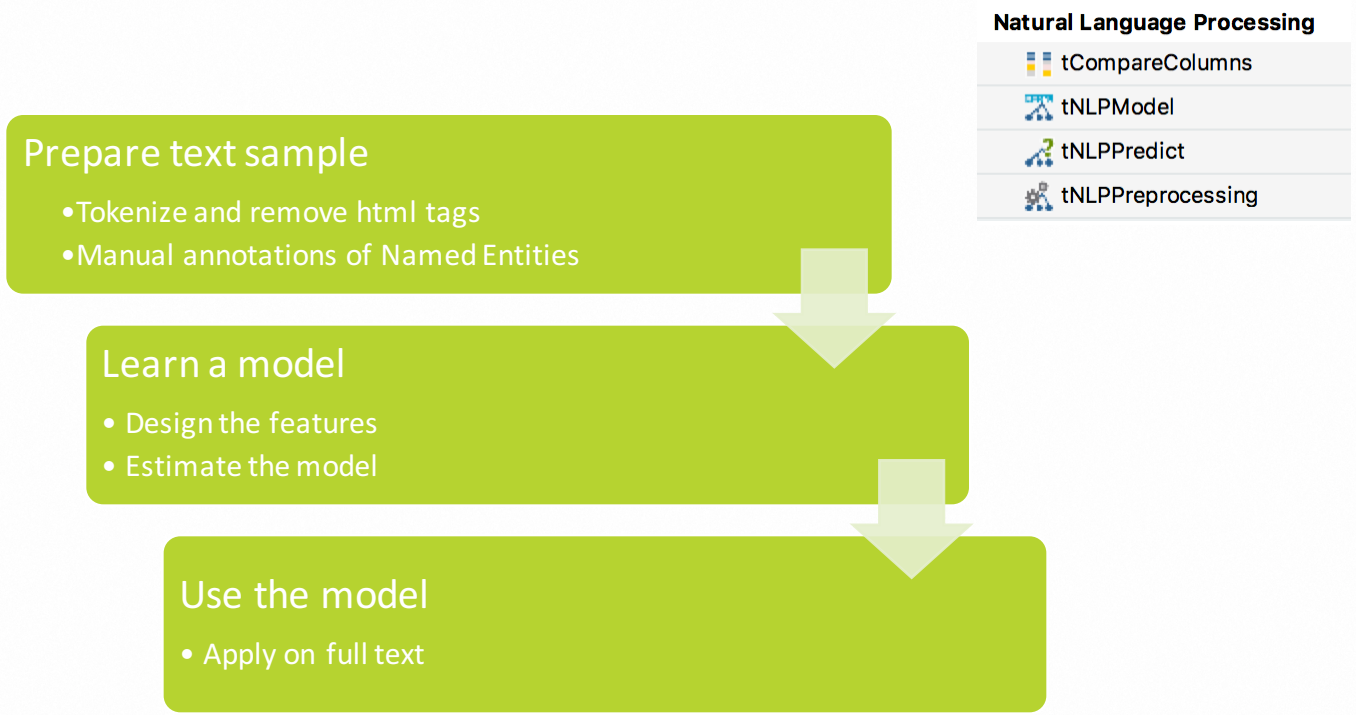


Figure 2: CRM Notes Mining & Machine Learning workflow

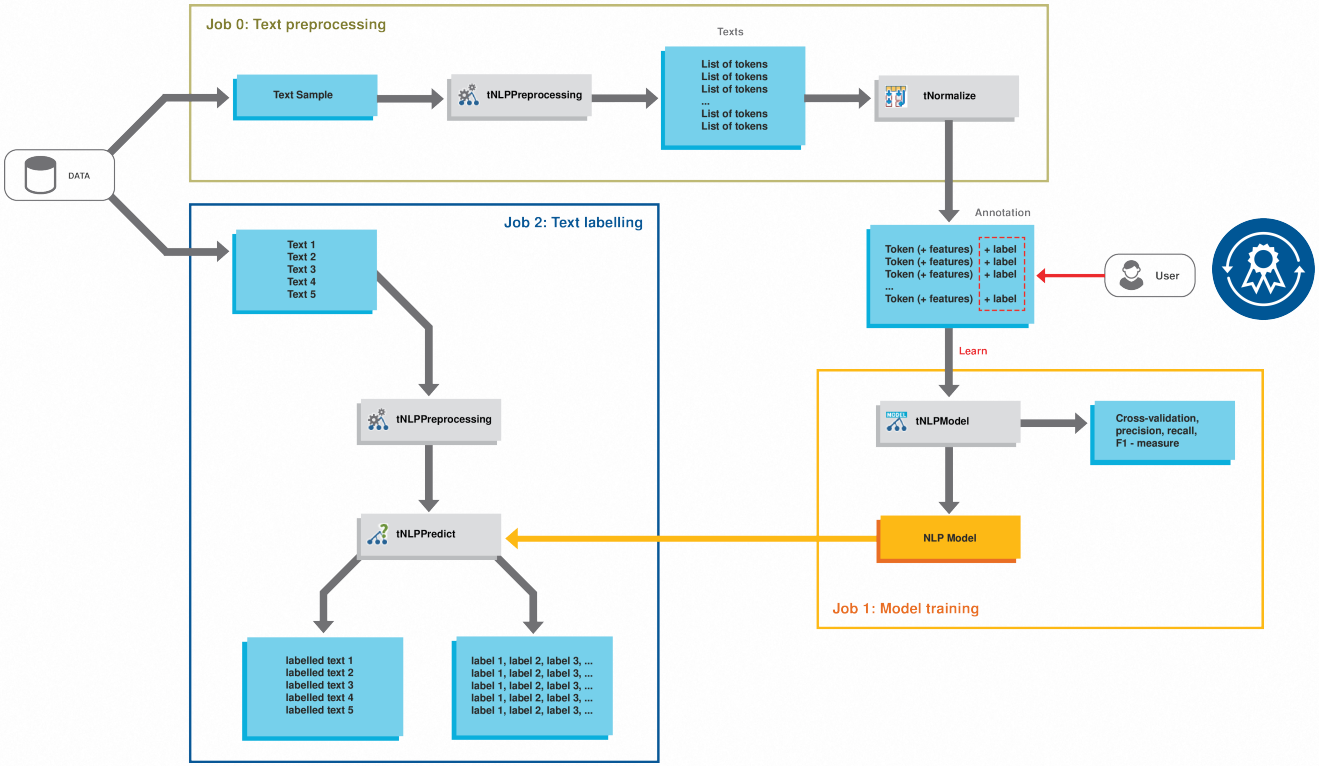
http://ualr.edu/informationquality/iciq-proceedings/iciq-2015/

Self-healing customer data quality issues through interpretation of unstructured data (Chandrasekaran.K, Clement.D)

Great! How does it work?



Component workflow



Text transformations

```
--------newtopic--------
--------newpost--------
Stéphane|Guest|
Hi,

First of all, congratulation for this application.

I had an error when a tried to register as an user in the community Users section.

I fill my email and my Geo localisation but I had this error message.

>**One or both fields are empty!**
>Please click on the following link to get back.

regards,
<PERSON>Stéphane</PERSON>
--------newpost--------
smallet|smallet@talend.com|stephane mallet
Hello,
This problem is now fixed.
Thanks for having declared it.
```

Convert in Conll-2003 format
add optional features
and label tokens

```
using    0.46    F    F    O
Postgres    0.64    F    F    O
with    0.6 F    F    O
Talend    0.56    F    F    O
?    0.0 F    F    O
Best    0.75    F    F    O
regards 0.68    F    T    O
,    0.0 F    F    O
Tim 0.83    F    F    PER

Hi    0.56    F    F    O
,    0.0 F    F    O
ppm 0.5 F    F    O
install 0.44    F    F    O
Text-CSV_XS 0.51    F    F    O
does    0.61    F    F    O
```

Extract named entities with <PER> labels

```
Grif Thanks for the reply. I did for get to mention all 3 computers are running XP. Two have the Home Ver and 1 the Pro Ver. The one
giving me the problem has the Home Ver and is my "main" desktop computer. The weird thing is the other two (a laptop & 2nd desktop) are
still networked. The "main" desktop has my printer connected to it, so until I can get it "networked" with the others, I can't print
from either of them. Can I assume that since none of them have Vista that I do not need to download and run the program you suggested?
Alan
;<PER>Grif</PER> Thanks for the reply. I did for get to mention all 3 computers are running XP. Two have the Home Ver and 1 the Pro
Ver. The one giving me the problem has the Home Ver and is my "main" desktop computer. The weird thing is the other two (a laptop & 2nd
desktop) are still networked. The "main" desktop has my printer connected to it, so until I can get it "networked" with the others, I
can't print from either of them. Can I assume that since none of them have Vista that I do not need to download and run the program you
suggested? <PER>Alan</PER>
;PER    O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O
O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O
O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O
O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  O  PER
```

talend

NLP libraries

	Stanford	ScalaNLP
Web site	https://stanfordnlp.github.io/CoreNLP/	http://www.scalanlp.org/
License	GPL	Apache v2
download	http://nlp.stanford.edu/software/stanford-corenlp-full-2015-12-09.zip	Integrated in Studio

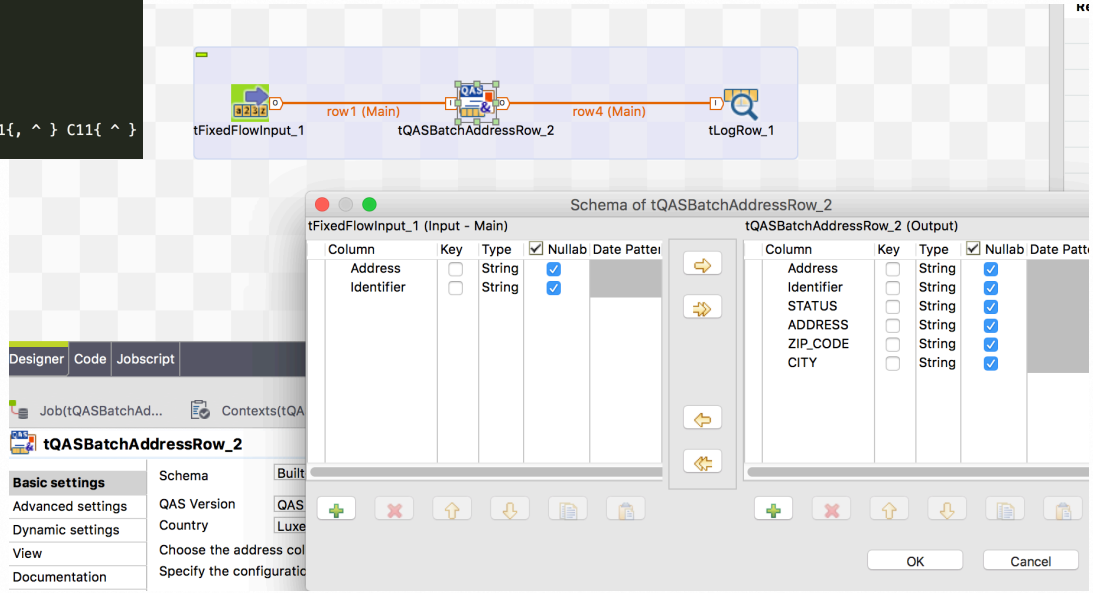
Precomputed models exist.

Customer requests

QAS component – Configurable output schema

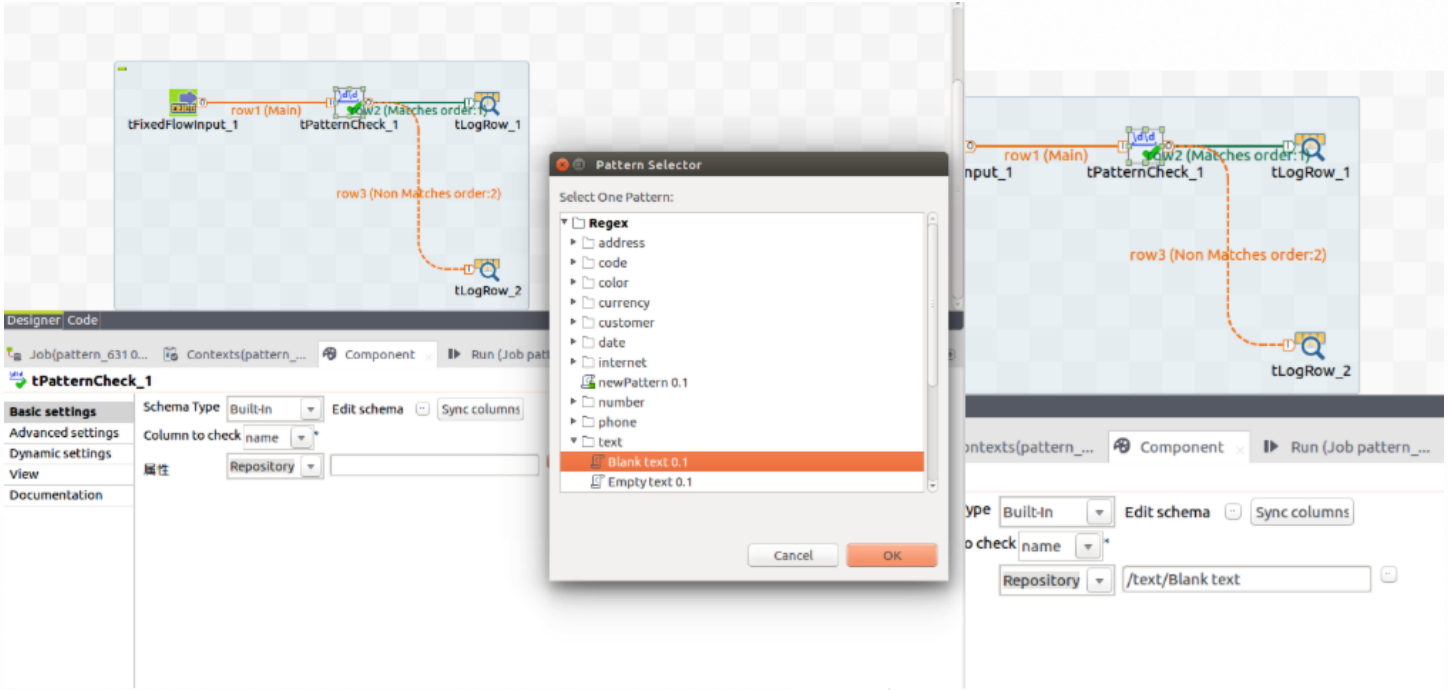
```
qaworld.ini
[FRA]

CountryBase=FRA
CleaningAction=Address

FRAAddressLineCount=6
FRAAddressLine1=W60,O11
FRAAddressLine2=W60,O21
FRAAddressLine3=W60,L41
FRAAddressLine4=W40,P11,S11
FRAAddressLine5=W40,B10,L31
FRAAddressLine6=W40,C11,L21

FRASeparateElements=Yes
FRAElementSeparator={, } P11{, ^ } C11{ ^ }
FRACapitaliseItem=L21
```

Configure required output in qaworld.ini and configure output schema accordingly



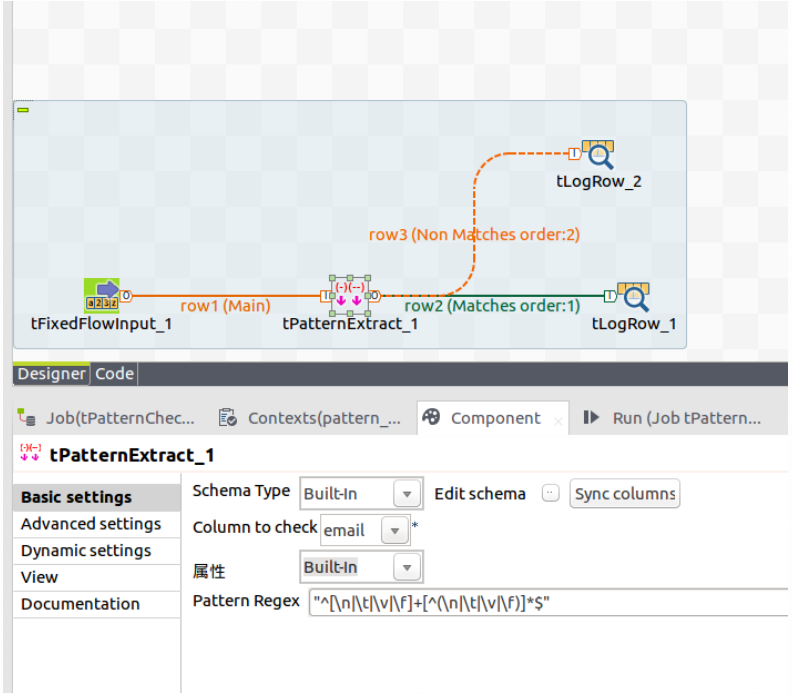
Pattern components – Repository mode

- Get patterns from repository



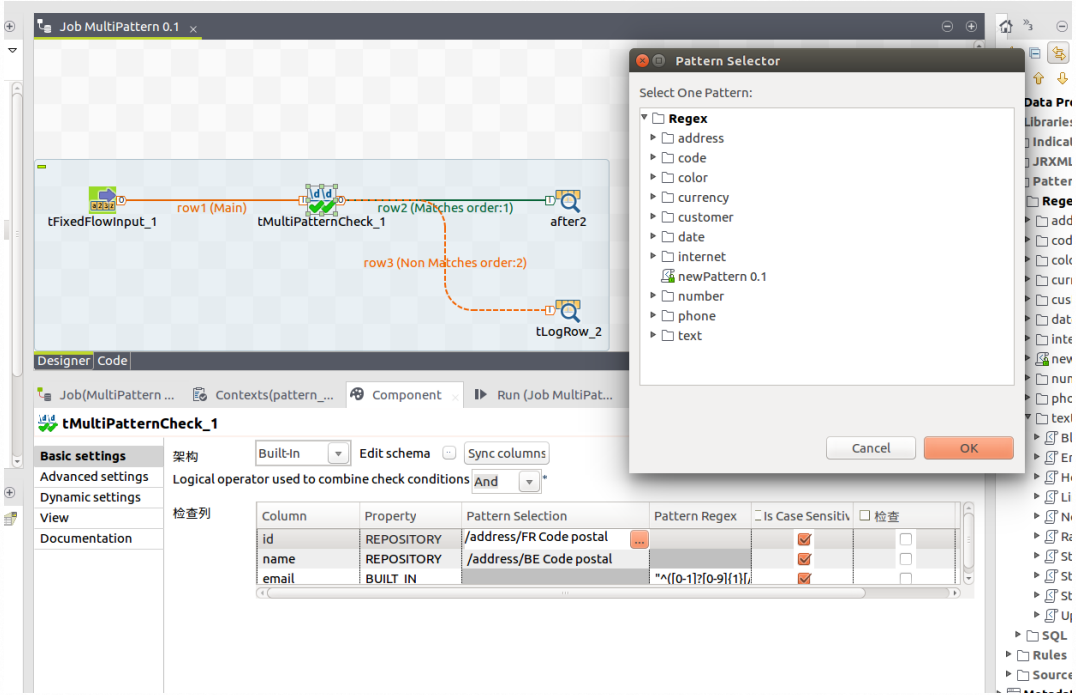
Pattern components – Built-in mode

- The user can write his own regular expression or modify a regular expression copied from the repository.

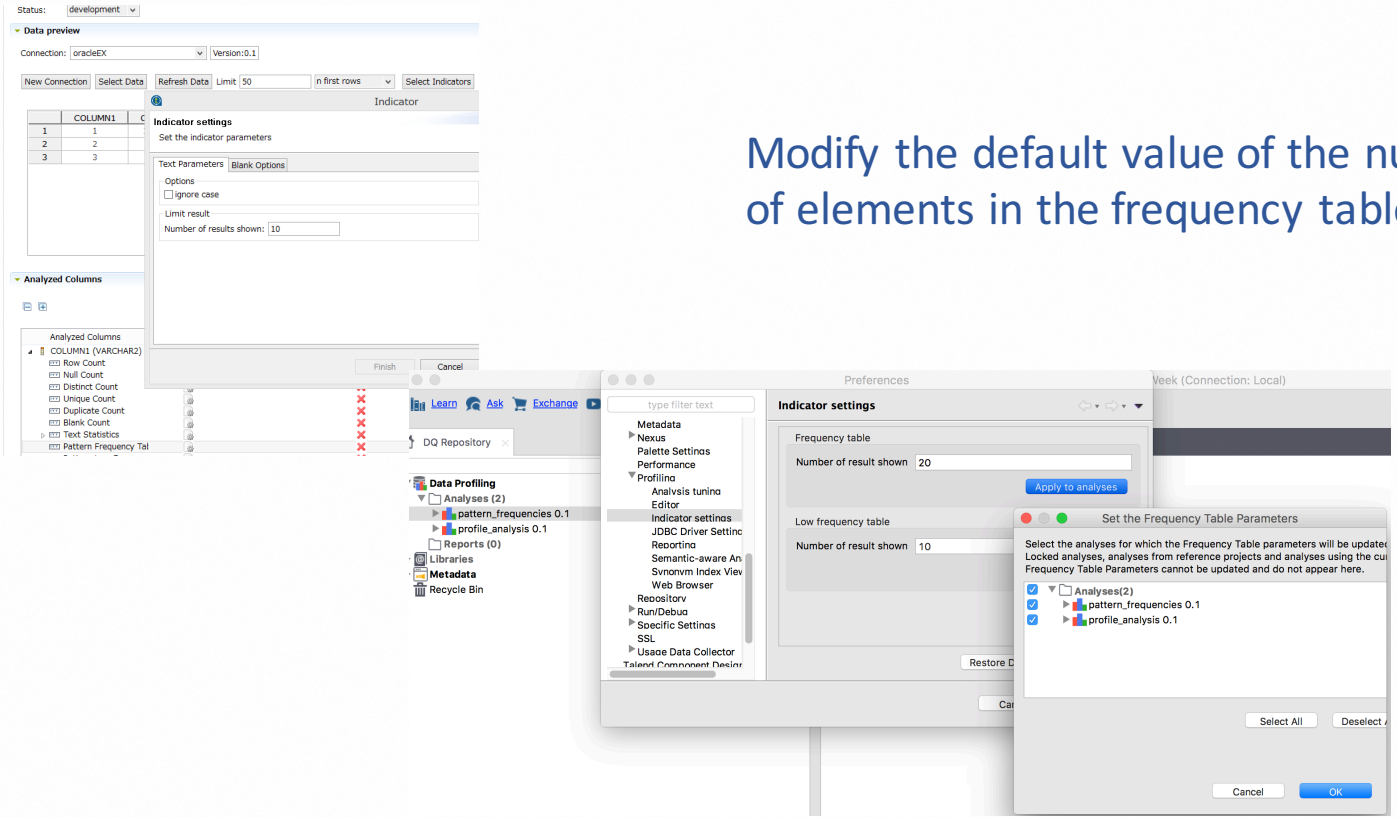


tMultiPattern component

- The built-in vs repository mode can be selected on each column of the schema.



Frequency table parameter



Modify the default value of the number of elements in the frequency tables

Data Preparation

Data Preparation 6.4 in a nutshell

Connectors

- Amazon S3
- Salesforce

Preparation versioning

Preparation promotion across environment

- Via preparation export/import

Operationalize a preparation in real-time Big Data

- AKA “tDataPrepRun in Spark Streaming jobs”

SSO between Data Prep and Data Stewardship

Dictionary service UI

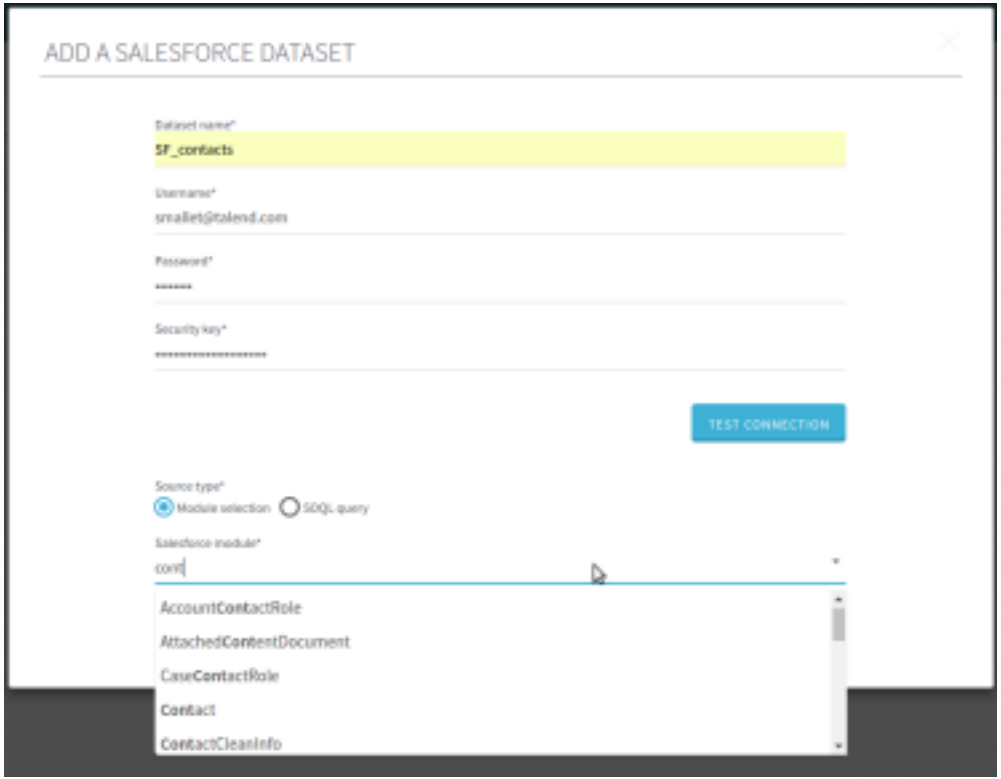
Additional improvements changes

- New functions
- Live dataset improvements

Salesforce Connector

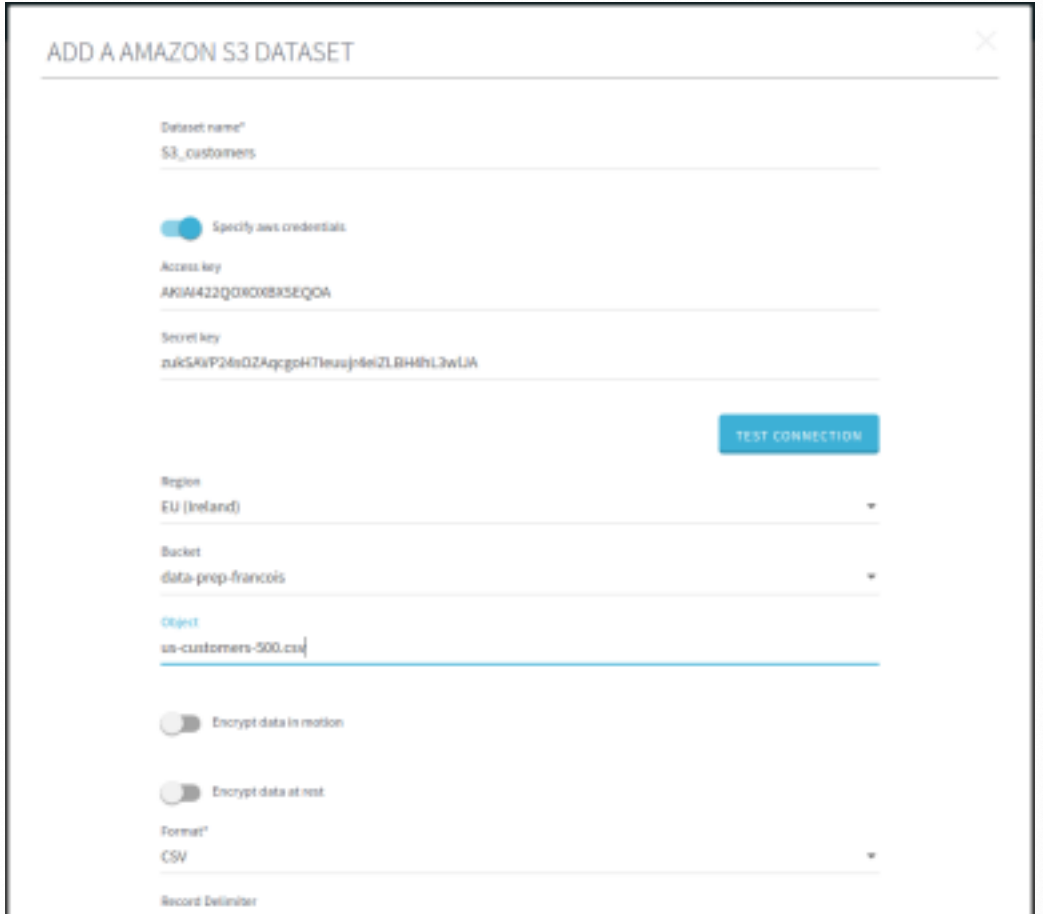
- Input only (for now)
- Module section or SOQL query
- Column selection for module selection

- Technical details
  - Based on TCOMP
  - Processing done on Data Prep server, not Beam runtime



Amazon S3 Connector

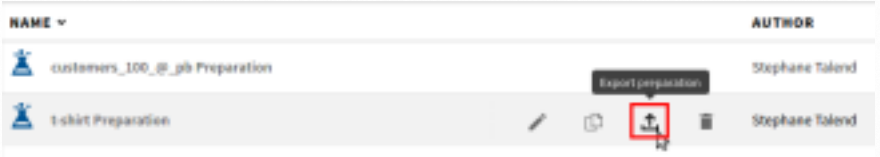
- Input and output
- Technical details
  - Based on TCOMP
  - Processing on Beam runtime if available, otherwise on Data Prep server

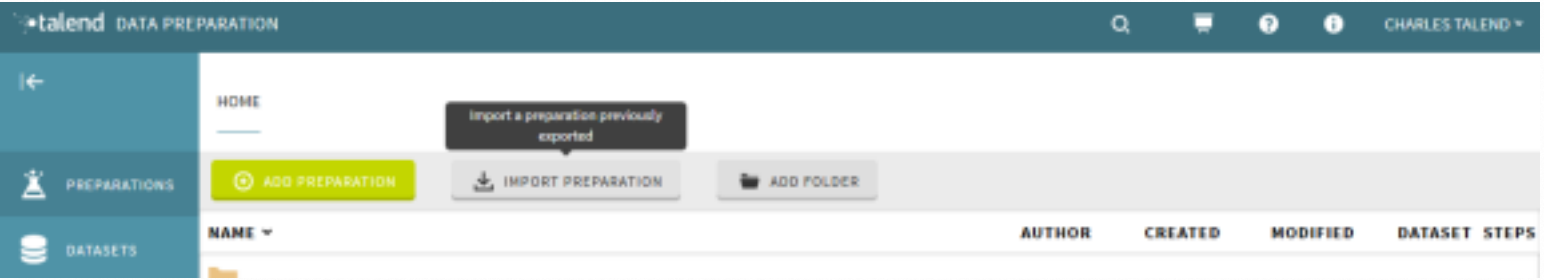


talend

Manage promotion across environments

- Based on preparation import/export
- Export/import of a single preparation at a time
- Only preparation is imported, not the dataset
- Dataset with the same name must exists and be accessible for the user
- All versions of a preparation are exported

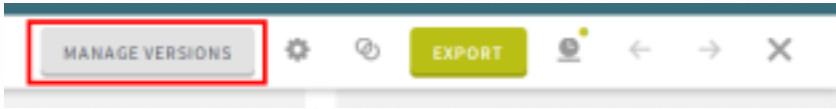


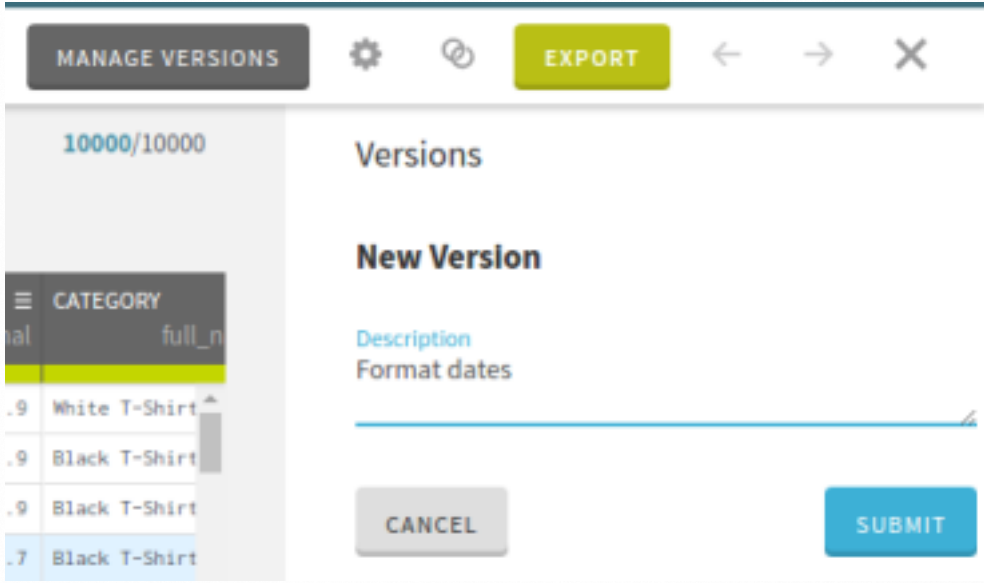


Preparation versioning

Leverage a stable state of a preparation in a job

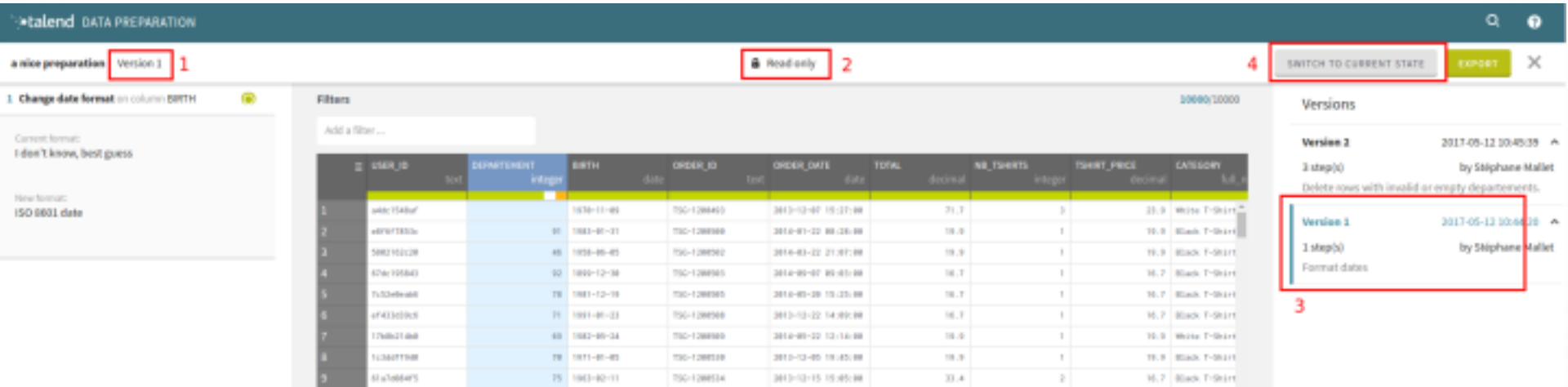
- Versions are managed directly in the Data Preparation UI
- Only Data Preparation Administrators can create versions
- Version number is an auto-increment
- A description is available





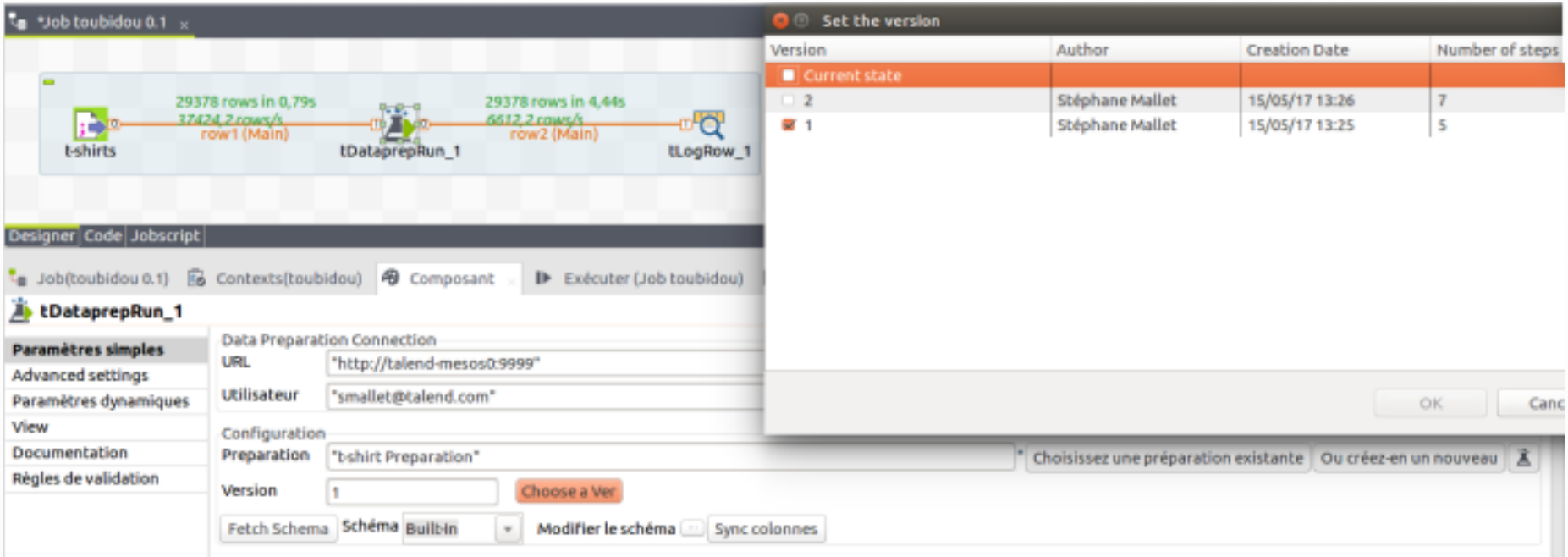
Preparation versioning

- Existing versions are visible by all users, regardless of their role
- Existing versions are read-only
- Versioning is linear and sequential, there is no full-blown SDLC with branching and so on
- Versions cannot be removed (for now)



Preparation versioning

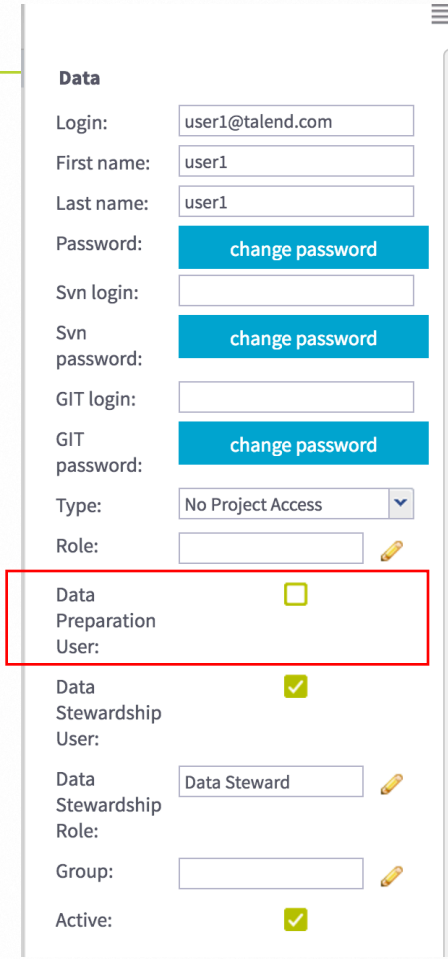
- In tDataPrepRun, after selecting a Preparation, you can now select the desired version
- The default value is “Current state”, which mimics 6.3 behavior



SSO integration

Identity management in 6.3

- In 6.3, TDP delegated identity and access management to TAC
- TDP was communicating directly with TAC for
  - Authentication
  - Current user information
  - List all available TDP users
- A TAC admin account was required in TDP configuration

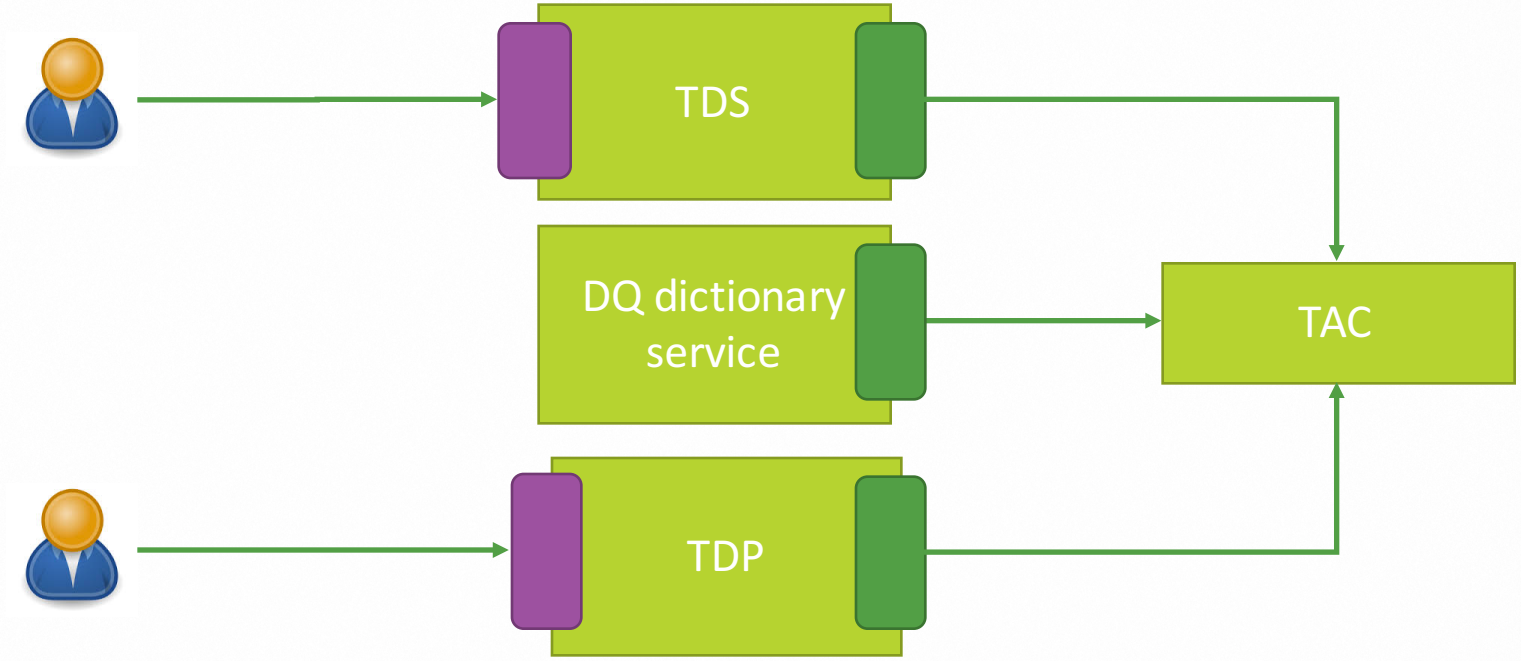


talend

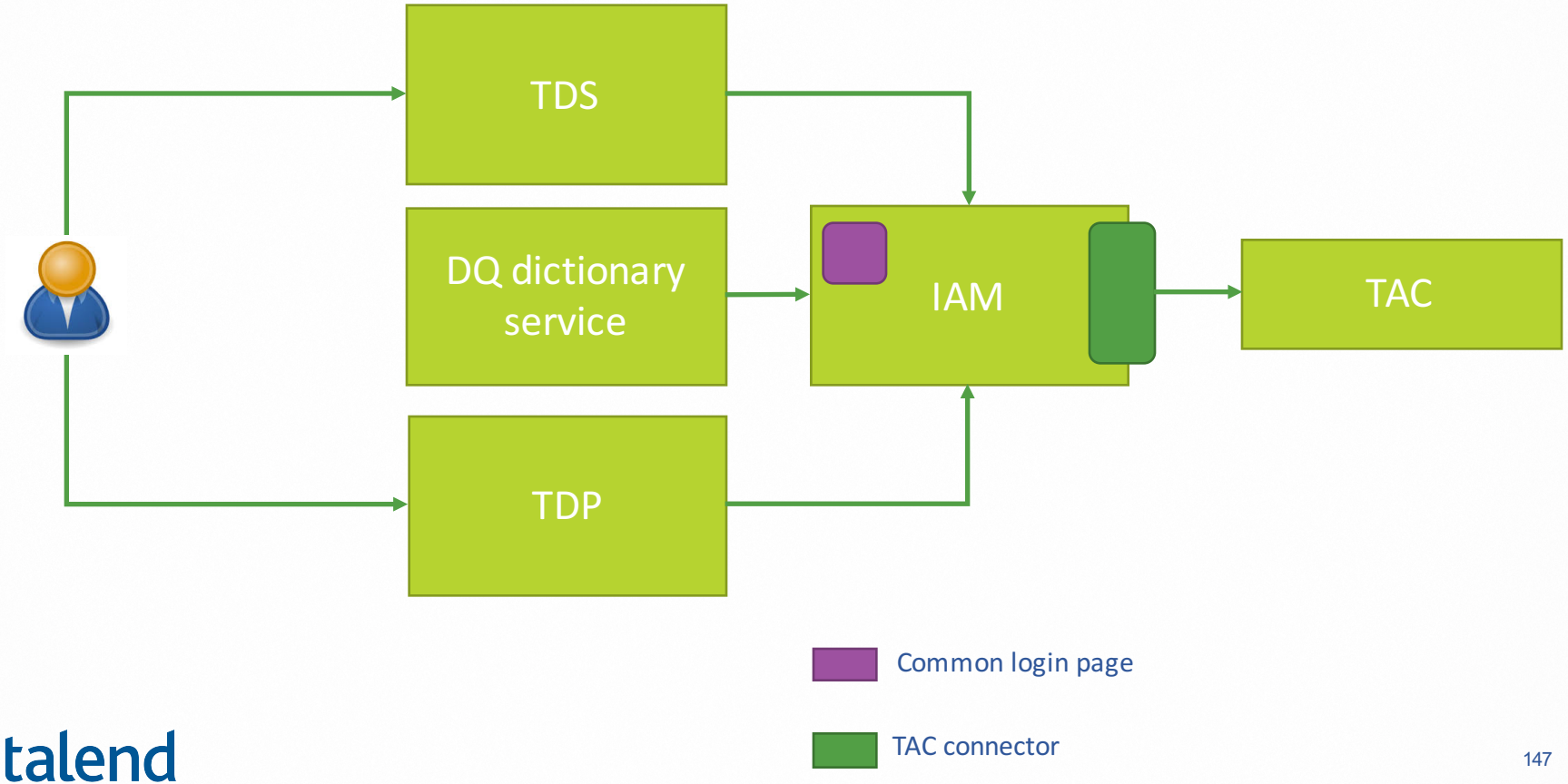
Identity management in 6.4

- TDP users are still managed in TAC
- A new module has been added: “Talend IAM”
  - TDP communicates to Talend IAM only for identity & access management
  - Talend IAM acts as an internal SSO agent for Data Prep and TDS (and all web apps later on)

6.3 users management implementation



6.4 implementation



Talend IAM solution

- Developed by the Platform service team
- Shared across Data Stewardship, Data Preparation and the Dictionary service

Shared presentation

New functions

Delete multiple columns at once

"Contains"

Conversion functions

- Convert date (ex: from **Gregorian calendar** to **Julian days)**
- Convert distance (ex: from **miles** to **kilometers)**
- Convert duration (ex: from **day** to **hour)**

Remove consecutive repeated characters

- change **Hoodiiiie** to **Hoodie**

Trim on a different character (or sequence) than whitespace

- change **00012** to **12** or **TSG-45515** to **45515**

Data Preparation Cloud Beta

Data Preparation Cloud

Timeline

- Beta as part of Summer ‘17
- GA as part of Fall ‘17

Scope

- (Almost) Identical to the on-premises release
- No Beam runtime yet
- No Dictionary service yet, only predefined semantic types
- Seamless integration in Talend Cloud

MDM

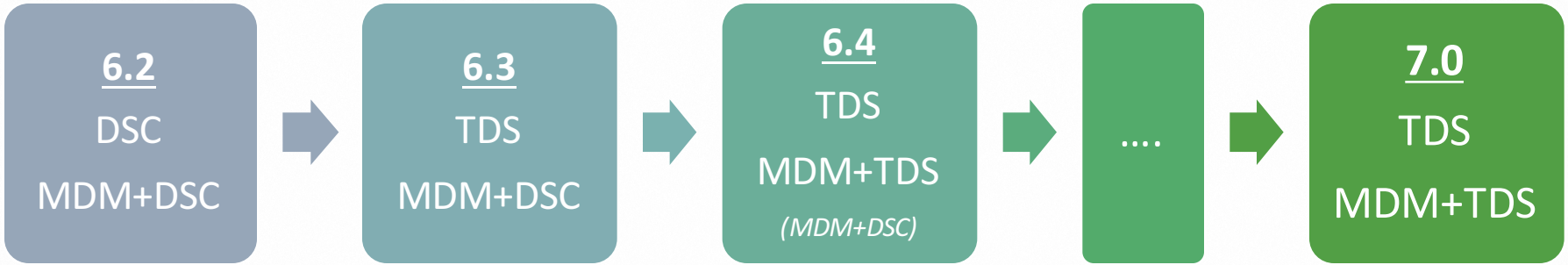
MDM 6.4 in a nutshell

Edition	Epic
EE	Integrating Matching with TDS
SE + EE	Improve read/write performances
EE	Improve cluster performances
EE	Ease customer migration
EE	Bonita UI Form Designer
EE	Customer requests
EE	E/R Editor improvements
SE + EE	Impact analysis improvements

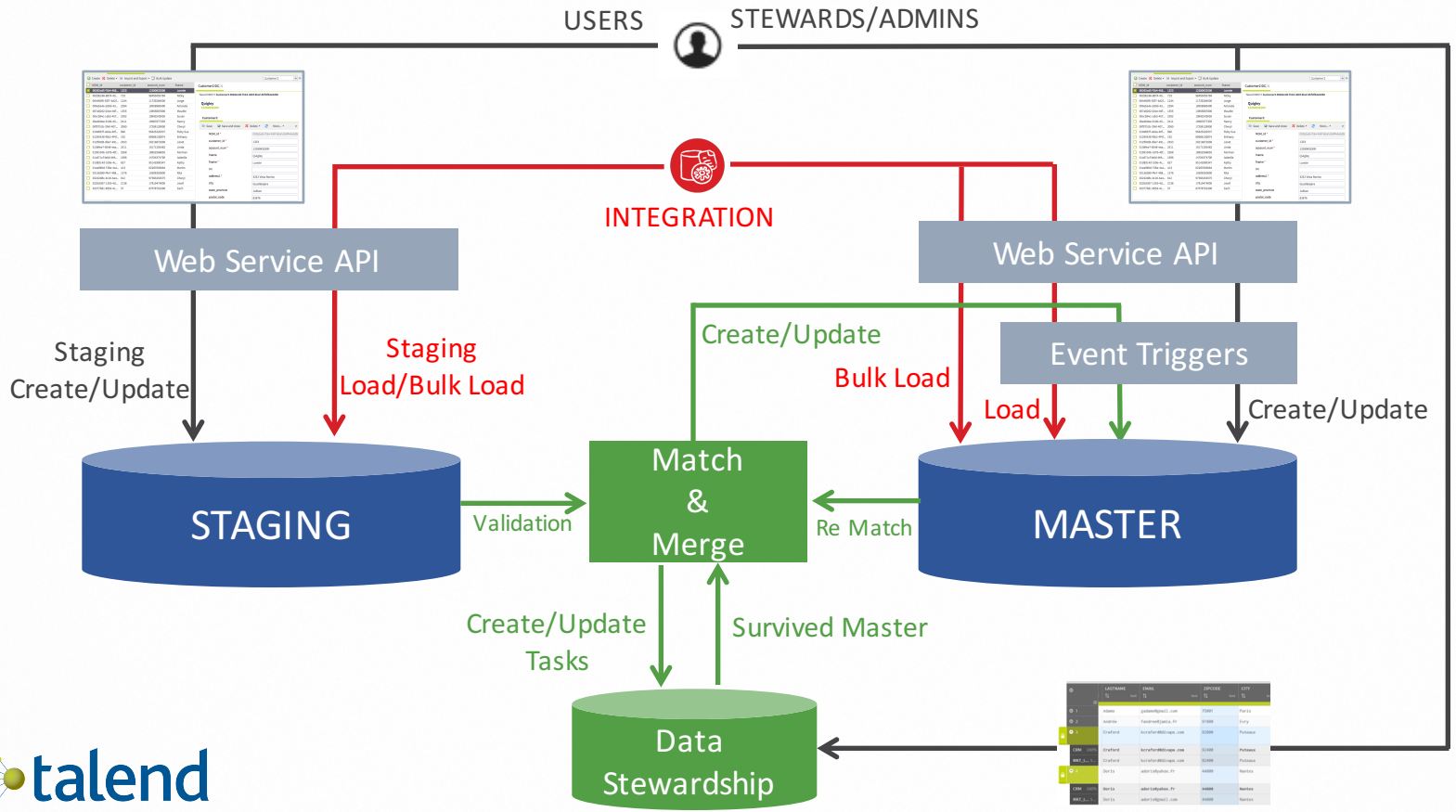
DSC

- Since 6.3, DSC standalone is no more officially delivered (can be still delivered upon request)
- Starting from 6.4 onwards, embedded DSC will be deprecated (still provided with MDM)

  "Deprecated features are no longer recommended for use and may cease to exist in future versions of the product."
- DSC will be retired as of 7.0



MDM Integrated Matching



Integrated Matching

- DSC/TDS enablement is done by commenting/uncommenting configuration properties in mdm.conf
- Authentication configuration
  - No “SSO” between MDM & TDS: each app gets individually authenticated
  - MDM users are authenticated via JAAS (with or without TAC)
  - TDS users are authenticated via SSO (along with TDP, DQ Dictionary)
  - MDM to TDS communication internally uses a dedicated “MDM Campaign Owner (MDM CO)” identity – i.e. a TDS user with the Campaign Owner role

```
#####################################
TDS settings
#####################################
tds.root.url=http://localhost:19999
tds.user=MDMCampaignOwner@talend.com
tds.password=LhAnf3Un7GJURGppg8yC5g==,Encrypt
tds.batchsize=50
```

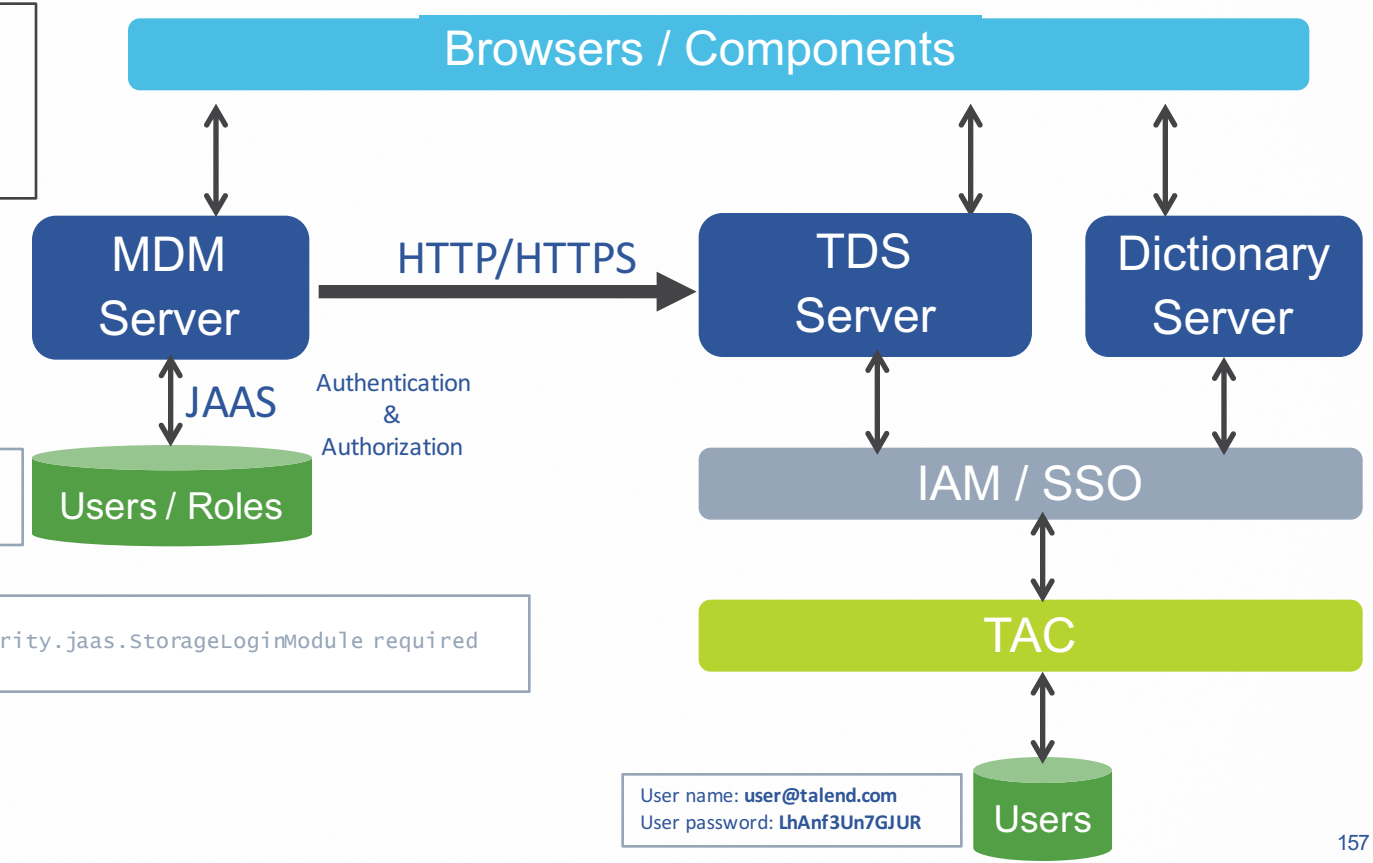
MDM CO credentials

Tasks creation batch size

Integrated Matching

MDM without TAC

Both TAC and MDM must remain synchronized:
User **names, passwords and email addresses** must be consistent

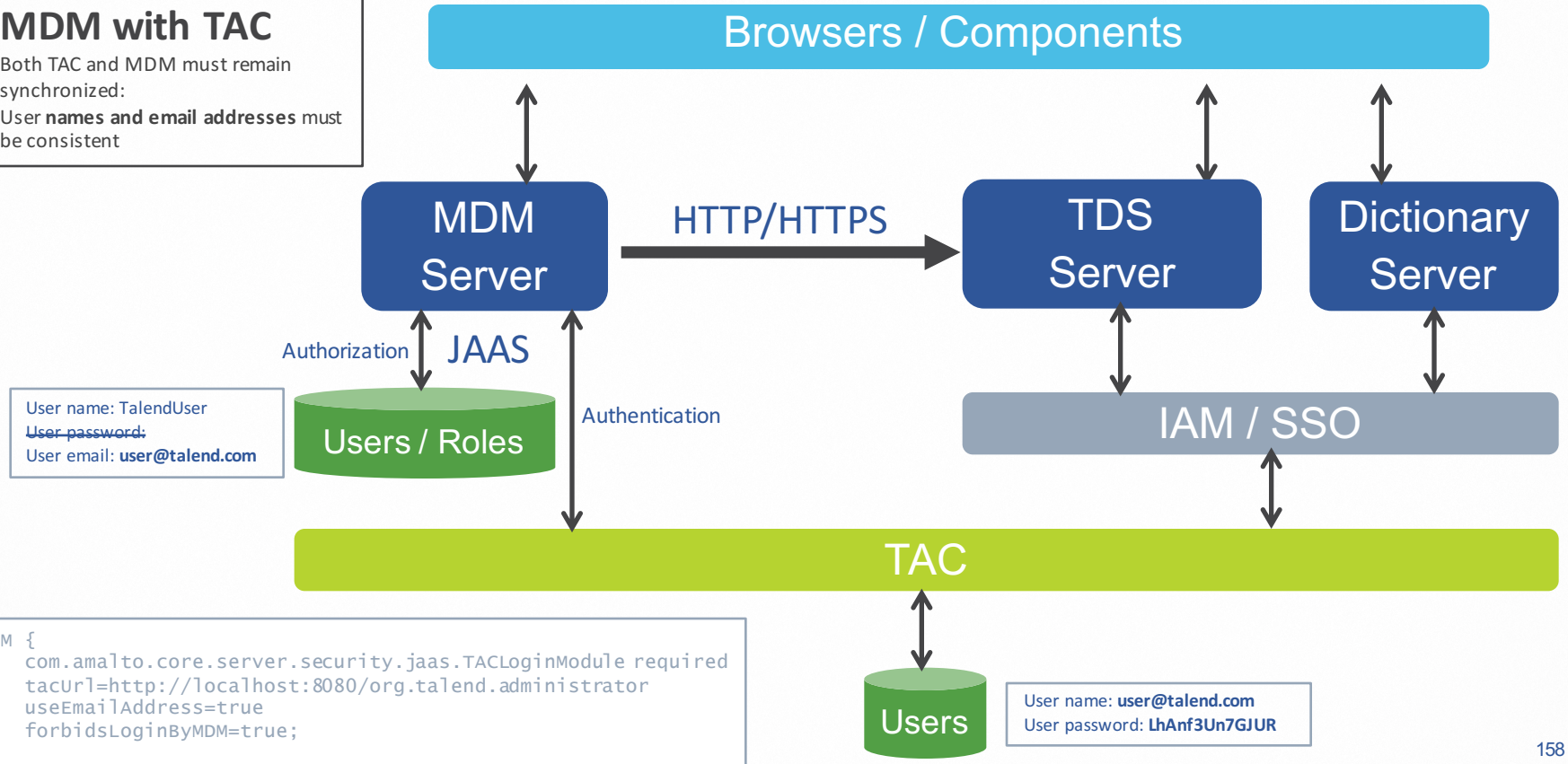



```
MDM {
    com.amalto.core.server.security.jaas.StorageLoginModule required
};
```

Integrated Matching

MDM with TAC

Both TAC and MDM must remain synchronized:
User **names and email addresses** must be consistent




```
MDM {
    com.amalto.core.server.security.jaas.TACLoginModule required
    tacUrl=http://localhost:8080/org.talend.administrator
    useEmailAddress=true
    forbidsLoginByMDM=true;
};
```

Integrated Matching

- Upon MDM Data Model deployment
  - A TDS tuple Campaign/DataModel is created for each MDM Entity having match rule(s)
  - Tuples creation/update are automatically done upon Studio DM deployment
  - Each campaign is created and owned by the dedicated MDM CO

- On Match & Merge run, the MDM CO
  - creates TDS tasks
  - consumes Resolved tasks from TDS

Integrated Matching

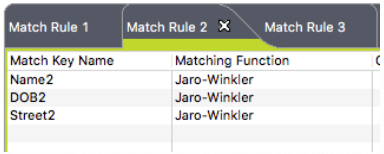
- Naming convention for Campaigns and DataModels
  <DataModel name>_<Entity name>_TMDM
- TDS new task metadata EXTERNAL_ID: corresponds to MDM group ID
- Synchronization is done one way: from MDM to TDS (ie TDS never calls MDM)
- TDS related materials are not deleted when DM is removed from MDM
- Matching process always tries to match a record with both golden record and source records

Integrated Matching (EE only)

Integrated Matching Function	Who?	Needs User Synchro?	Comment
Create Campaigns & Data Models	The TDS "MDM Campaign Owner"	No	• The MDM CO is used to create/update campaigns & data models • The user does not necessary exist in MDM
Match & Merge	The TDS "MDM Campaign Owner"	No	• The MDM CO is used to create/read tasks • The user does not necessary exist in MDM
DataStewardship menu	Any TDS User (CO or DS)	No	• Opens a browser tab on TDS webapp only • User can log into TDS using any account
Open related task	Any TDS Steward of the related campaign	No	• Opens a browser tab on filtered TDS task • User must log into TDS using a correct DS account in order to see the given task
Tasks summary in welcome page	The current MDM user	**Yes** **MDM JAAS** **or** **TAC JAAS**	• Uses MDM user's email as the TDS account name • User password is either from TAC (TAC JAAS) or must be the same in MDM (MDM JAAS) • Must be a DS user in TAC (otherwise 0 tasks) • Counts tasks of all the TDS campaigns user is assigned on

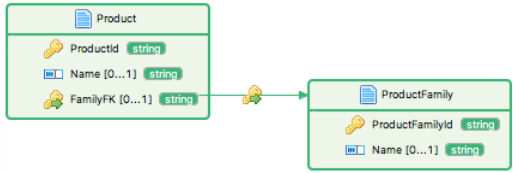
Integrated Matching (EE only)

- Support multiple match rules (TMDM-10501)



- Support a 0..1 FK field in a match rule (TMDM-9173)

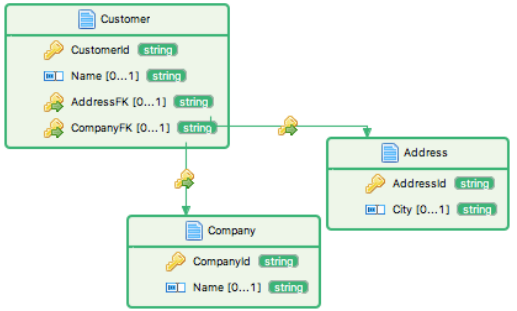
  e.g. „Match Product records with their Name and FamilyFK“



- Support match rules on multiples entities joined by 0..1 FKs (TMDM-9634)

  Still flat entities

  e.g. „Match Customer records with their Name, Address City and Company Name“



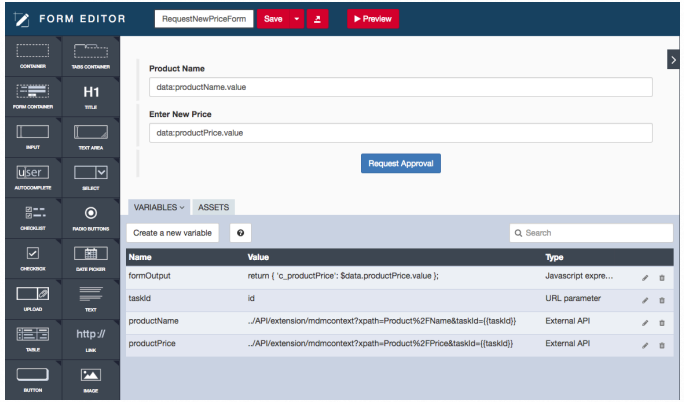
Bonita UI Form Designer

- Full support of Bonita BPM 7 UI Designer (6.x legacy forms still supported)
  http://www.bonitasoft.com/for-you-to-read/videos/bonita-bpm-7-ui-designer-basics
- Use of a REST API Extension (groovy) to retrieve data within the user's context from the mdm_context. Done in UI through endpoints of the form:

Form Variable	Value (External API Endpoint)
ProductName	../API/extension/mdmcontext?xpath=Product%2FName&taskId={ {taskId} }

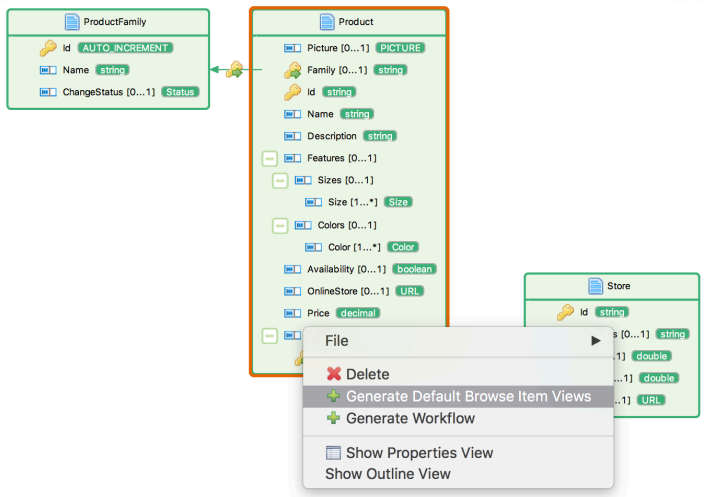
Request parameter "Product%2FName" is the URL encoded XPath for the attribute: "Product/Name"

- REST API extension automatically deployed with the workflow
- Product Demo workflow has been updated to use the form designer



Entity/Relations Diagram Editor (EE only)

- ‘Generate Default View’ and ‘Generate Workflow’ contextual menus
- Display occurrence range on Entity fields
- Create FK link from an existing element
  - Target element type is converted to String
- Highlight Entity/Reusable Type when selected
- Re-order the elements of an Entity
- Reset graphical design



Impact Analysis Improvements (SE+EE)

Allow finer-grain changes:

Data Model Change	6.4 Impact *(6.3 Impact)*	6.4 Results *(6.3 Results)*
Change a simple type element from optional to mandatory 1- without a default value (column contains null values)	**High/Medium**	The NOT NULL constraint is added to the corresponding column in the entity table, the impact level is High. If the database you are using does not add the NOT NULL constraint, the impact level is Medium.
Change a simple type element from optional to mandatory 2- with a default value (column contains null values)	**Medium** *(High/Medium)*	The column in the entity table is updated with the default value.
Change a simple type element from optional to mandatory 3- with a default value (column has no null value)	**Low** *(High/Medium)*	The column definition in the entity table is changed accordingly.
Change a simple type element from mandatory to optional	**Low** *(High)*	The column definition in the entity table is changed accordingly. The NOT NULL constraint is removed.
Change a multi-occurrence element from optional to mandatory or vice versa	**Low** *(High)*	The column definition in the entity table is changed accordingly.
Remove a simple type element (mandatory/optional)	**Medium** *(High/Medium)*	The column in the entity table is removed. *(Column was not deleted)*

Impact Analysis Improvements (SE+EE)

"If you want to change an element from optional to a mandatory, better to make sure a default value is defined for the element"

Simple type Element change

From/ To	0→1	1→1	0→n	1→n
0→1	-	H/M/L	High	High
1→1	Low	-	High	High
0→n	High	High	-	Low
1→n	High	High	Low	-

Complex type Element change

From/ To	0→1	1→1	0→n	1→n
0→1	-	High	High	High
1→1	Low	-	High	High
0→n	High	High	-	Low
1→n	High	High	Low	-

Under the hood:

- Use of Liquibase library (www.liquibase.org) to address Hibernate schema change lacks
- DB Schema changeset saved in *<MDM_HOME>/data/liquibase-changelog/{date}* and change history is kept in a table named *databasechangelog*

Performances Improvements (SE+EE)

- Management of Auto-Increment (cluster)
  - Implementation based on Hazelcast distributed caching support to replace DB-based implementation in cluster mode
  - Settings in mdm.conf
  - Advanced Spring configuration can be done in <MDM_HOME>/apache-tomcat/webapps/talendmdm/WEB-INF/beans.xml (Ref: http://docs.hazelcast.org/docs/latest/manual/html-single/index.html#spring-integration)

```
################################
Hazelcast basic settings
################################
hz.group.name=dev
hz.group.password=password
hz.network.port=5705
hz.network.port-auto-increment=true
hz.multicast.enabled=false
hz.tcp-ip.enabled=true
#Write comma-separated IP addresses
hz.members=127.0.0.1
```

- Bulkload
  - Throughput increased in cluster
  - ”Backpressure” tuning (TMDM-10560) with settings in mdm.conf and JVM parameters
  - However, in some situations, e.g. using FKs, 5.6x has higher perfs (due to Hibernate)

    => Work in progress

```
//Bulkload Component-side tuning
-Dbulkload.concurrent.http.requests=25
```

```
##########################################
Bulkload Server-side tuning
##########################################
#To avoid connection pool and database overload
bulkload.concurrent.database.requests=25
#Control how many milliseconds to wait before retry
bulkload.concurrent.wait.milliseconds=200
```

Excel import/export (SE+EE)

Enhance FKs import (TMDM-10358)

- **Export** (Already available in 6.3 version)
  1. FK value (e.g. [id1])
  2. FK + FKInfo values (e.g. [id1]|Info1|Info2)
  3. FKInfo values (e.g. Info1|Info2)

- **Import** (New in 6.4)
  1. FK value (e.g. [id1] or id1)
  2. FK + FKInfo values (e.g. [id1]|Info1|Info2 or id1|Info1|Info2)
  3. FKInfo values => not supported

  Case of a composite FK:
  => must be always surrounded by brackets (e.g. [id1]|[id2] or [id1]|[id2]|Info1|Info2)

Misc

- Web UI
  - Allow copy of greyed fields (TMDM-10028)
  - Roles list sorted alphabetically (TMDM-9873)
  - Horizontal scroll bar on records list (TMDM-10030)
- Migration Tool
  - Identify potential issues before migration (using –v validation option)
  - Direct migration from [5.5.1,5.6.x] range to 6.x
- Upgrade to Tomcat 8.0.42

Appendix: Integrated Matching

MDM -> TDS: Types mapping

MDM	TDS
boolean	boolean
date	date
time	time
dateTime	timestamp
int	integer
integer	integer
short	integer
long	integer
base64Binary	integer
byte	integer
negativeInteger	integer
nonNegativeInteger	integer
positiveInteger	integer
nonPositiveInteger	integer
unsignedByte	integer
unsignedInt	integer
unsignedLong	integer
unsignedShort	integer

MDM	TDS
decimal	decimal
double	decimal
float	decimal
string	text
anyURI	text
normalizedString	text
tokenstring	text
language	text
hexBinary	text
duration	text
AUTO_INCREMENT	text
MULTILINGUAL	text
PICTURE	text
URL	text
UUID	text

MDM -> TDS: Constraints mapping

Boolean: no constraints

String

MDM	TDS
minLength	minLengthText
maxLength	maxLengthText
length	minLengthText & maxLengthText
pattern	patternText
enumeration	allowedValues

Decimal

MDM	TDS
minInclusive	minDecimal
maxInclusive	maxDecimal
fractionDigits	scaleDecimal

- Integer

MDM	TDS
minInclusive	minInteger
maxInclusive	maxInteger

Date

MDM	TDS
minInclusive	minTime
maxInclusive	maxTime

Time

MDM	TDS
minInclusive	minDate
maxInclusive	maxDate

DateTime

MDM	TDS
minInclusive	minDatetime
maxInclusive	maxDatetime

MDM -> TDS: Permissions mapping

- If no permission declared on the MDM entity
  => define a campaign role "Steward" with READ_ONLY on all fields.
- If permissions are defined on some fields
  => define campaigns roles for the defined fields with the MDM permissions
  => all the rest is READ_ONLY

Data Stewardship

TDS 6.4 in a Nutshell

- Grouping campaigns
- SLA on tasks
- Dictionary service UI
- SSO integration
- MDM integrated matching with TDS
- More flexible impact analysis on Data Model changes
- Miscellaneous changes
  - Bulk functions
  - TDS components enhancements

Grouping campaigns

Implemented campaign types in 6.3

- Arbitration campaign
  - Answer a question on data
  - 1 task = 1 read-only record

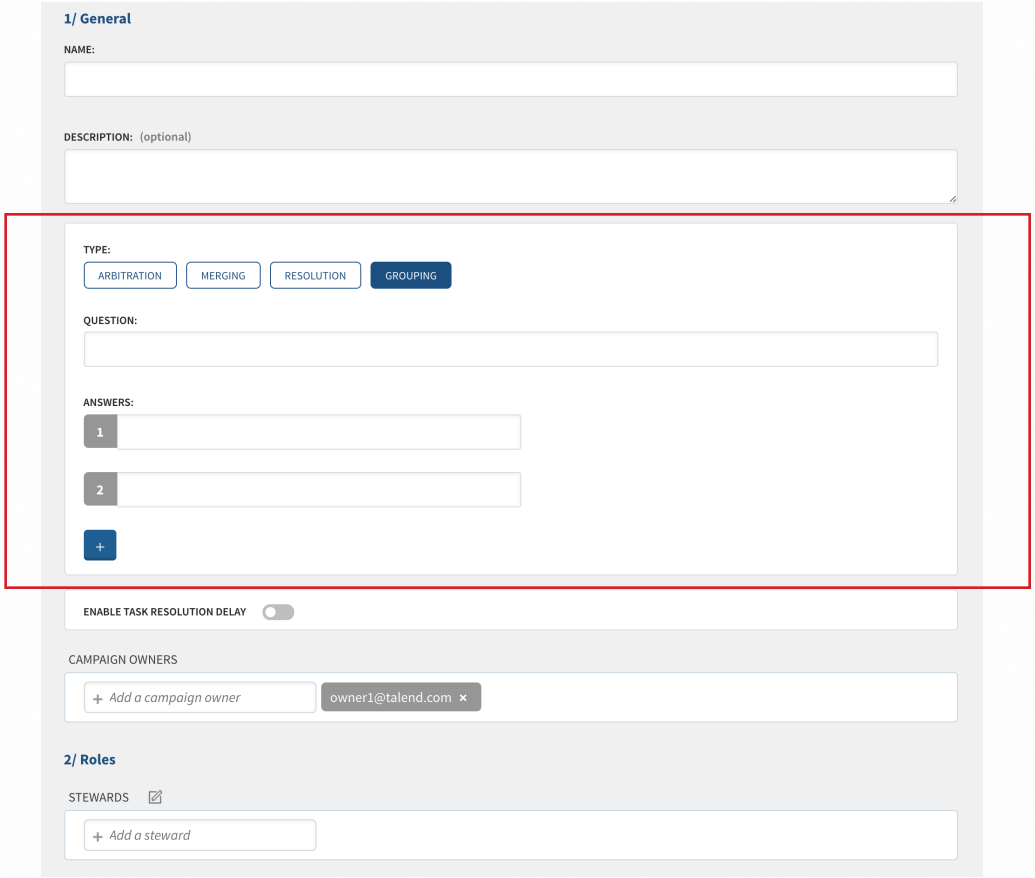
- Resolution campaign
  - Data cleansing
  - 1 task = 1 modifiable record

- Merging campaign
  - Data reconciliation / deduplication
  - 1 task = 1 modifiable record (golden record)
    + N read-only records (sources)

Grouping campaigns

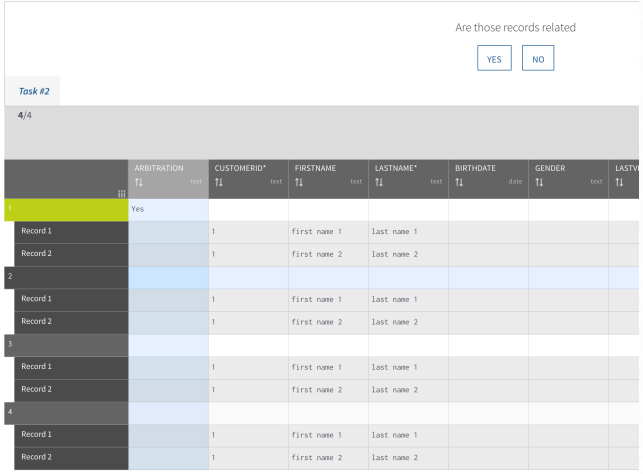
- Answer a question on a group of records
- 1 task = N read-only records (not limited to 2 records)
- Use cases examples:
  - Potential duplicates labeling: “Are those records duplicates?”
  - Potential relationship labeling: “Are those records related?”

Grouping campaigns creation



Grouping tasks creation

- Via tMatchPairing for Spark matching
- Standard DI jobs using tDataStewardshipTaskOutput
  - One record per row
  - “TDS_GID” to group records together
    - Same as for merging tasks
  - “TDS_SCORE”
    - new read-only metadata (decimal)
    - Also added to merging tasks
  - ”TDS_SOURCE”



Grouping tasks retrieval

- Via tMatchModel for Spark matching
- Standard DI jobs using tDataStewardshipTaskInput
  - “TDS_ID”: to group records together
  - “TDS_ARBITRATION”: selected arbitration choice index
  - “TDS_ARBITRATION_LABEL”: selected arbitration choice label (as in the UI)
  - All other metadata as usual

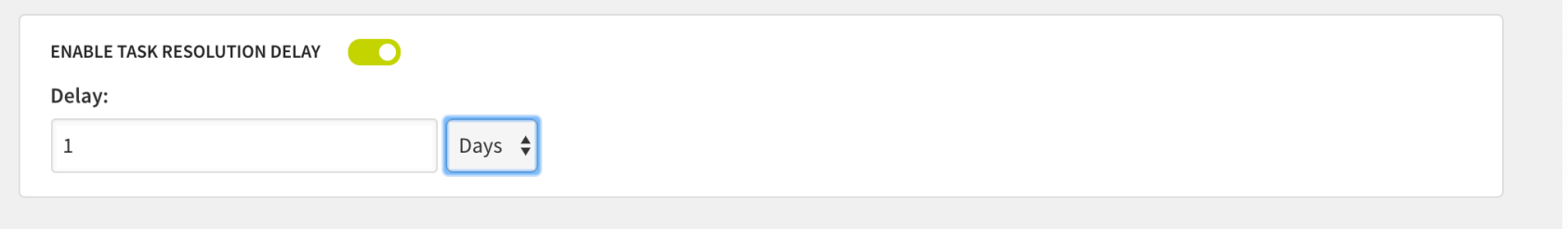
SLA on tasks

Use case

- As a campaign owner, I want to define due dates for my campaigns' tasks so that data stewards can prioritize their work according to my clients requirements

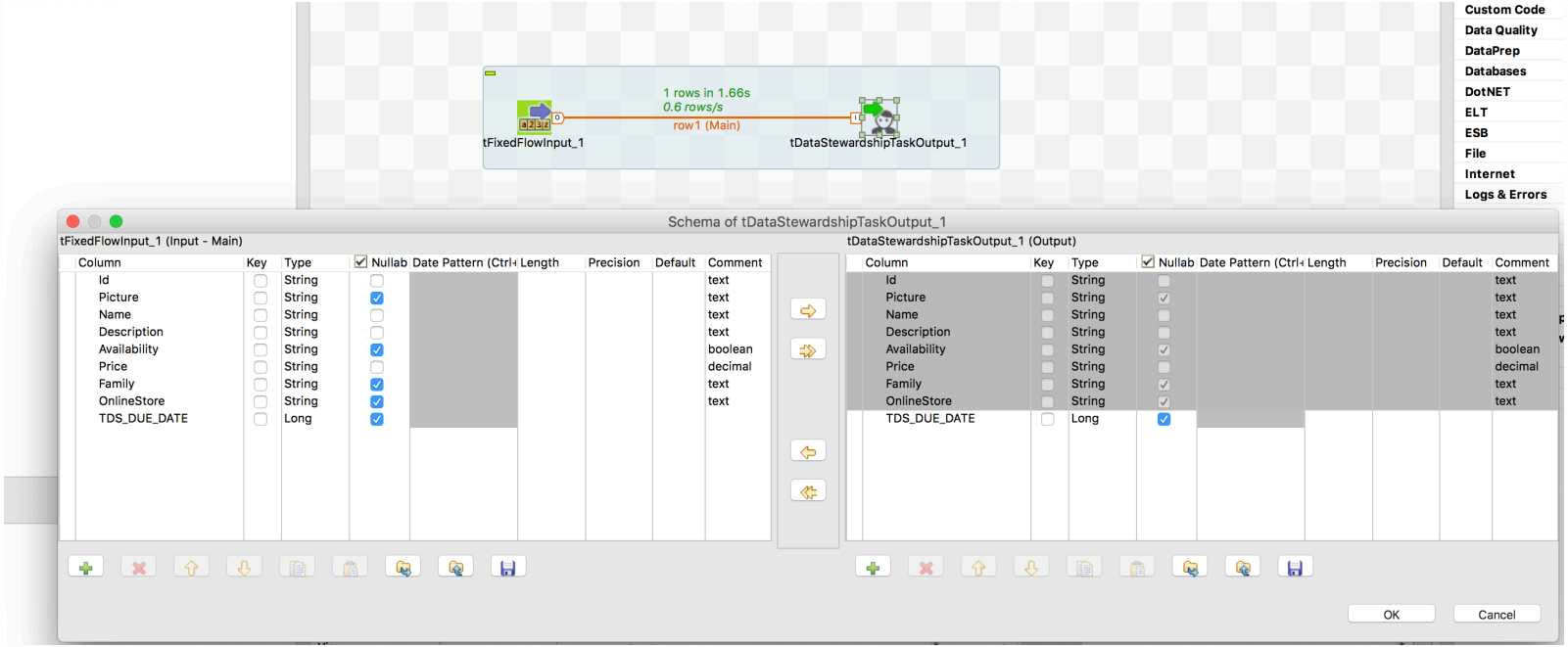
Configure default due dates at campaign level

- A default resolution delay can be configured at campaign level



- When configured, a due date will be computed on each new created task unless it is already provided

Due date definition at task level (studio)



Due date definition at task level (studio)

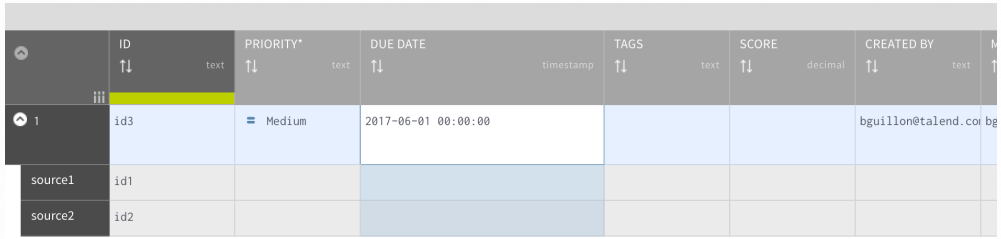
- New optional metadata column TDS_DUE_DATE
- Java timestamp (number of milliseconds since 1970-01-01 00:00:00 UTC)
- On merging tasks, the due date must be provided on golden record only

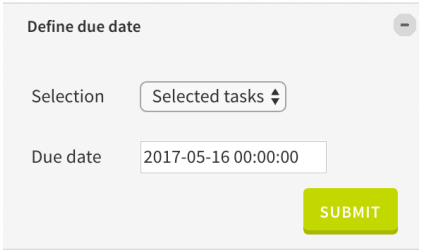
Email	TDS_GID	TDS_MASTER	TDS_SOURCE	TDS_SCORE	TDS_CREATION	TDS_LAST_UPDAT	TDS_DUE_DATE
"email1@talend.com"	"12345"	false	"source1"				
"email2@talend.com"	"12345"	false	"source2"				
"golden@talend.com"	"12345"	true					9999999L

- On grouping tasks, the last provided due date is taken

Due date definition at task level (user interface)

- New metadata column in TDS grid
- Read-only for Data Stewards / Editable by Campaign Owners
- Format is YYYY-MM-DD HH:mm:ss
- Stored in UTC and displayed in user's local timezone
- Mass update function available on right panel for campaign owners

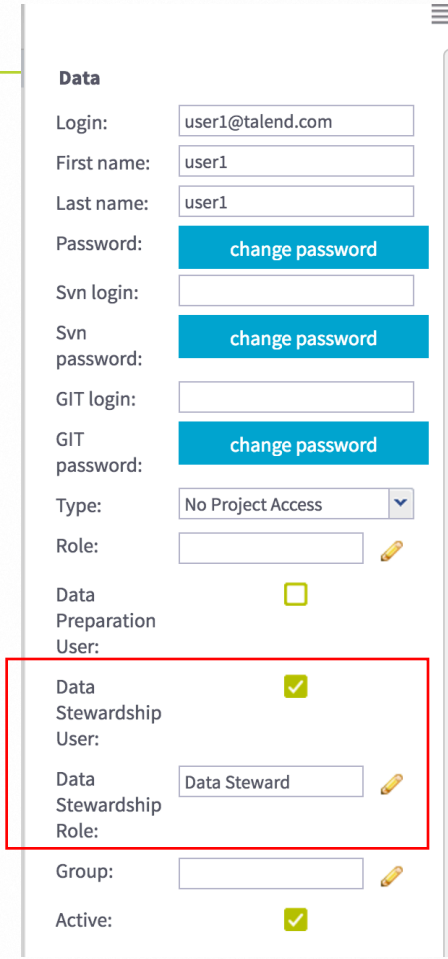




SSO integration

Identity management in 6.3

- TDS delegated identity and access management to TAC
  - New user type “Data Stewardship”
  - Built-in roles “Campaign Owner” and “Data Steward”
- TDS was communicating directly with TAC for
  - Authentication
  - Current user information
  - List all available TDS users
- A TAC admin account was required in TDS configuration

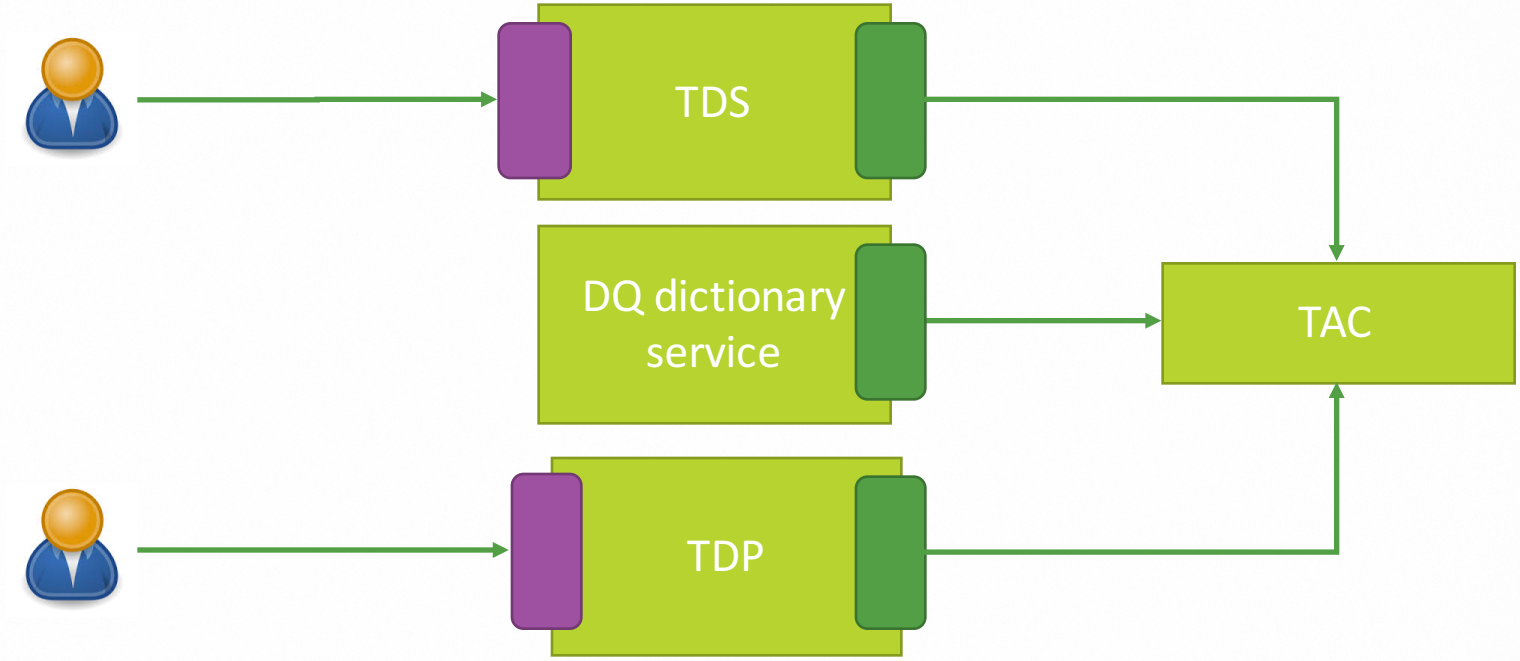


talend

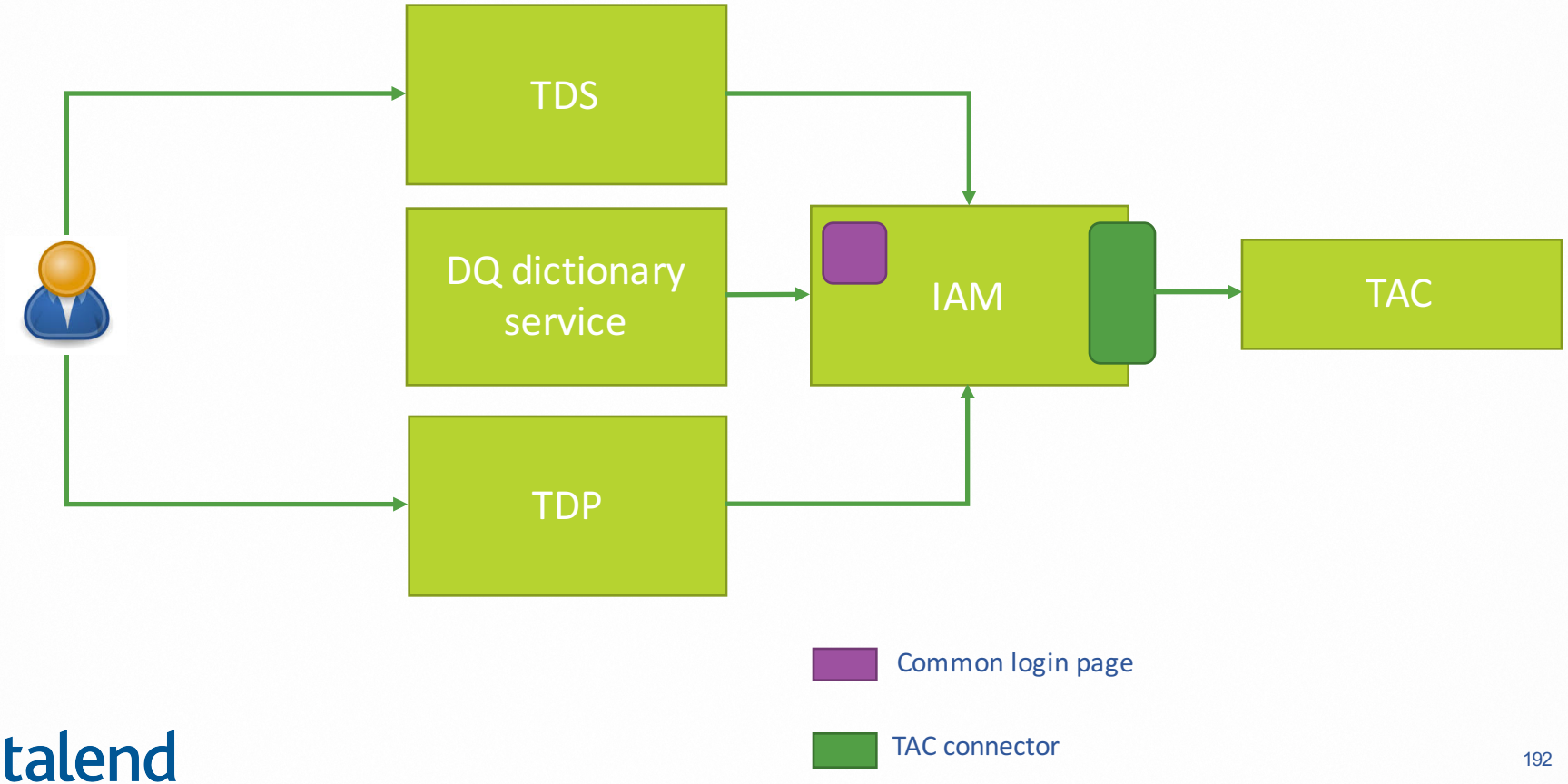
Identity management in 6.4

- TDS users are still managed in TAC
- A new module has been added: “Talend IAM”
  - TDS communicates to Talend IAM only
  - Talend IAM acts as an internal SSO agent for Data Prep and TDS (and all web apps later on)

6.3 users management implementation



6.4 implementation



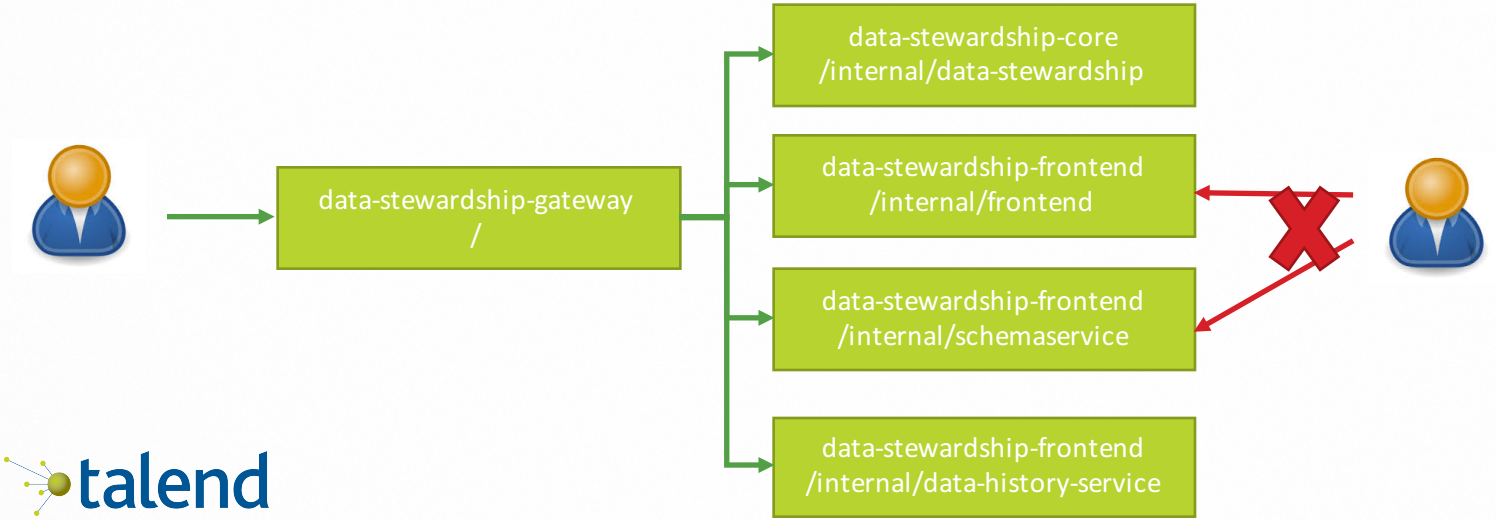
Talend IAM solution

- Developed by the Platform service team
- Shared across Data Stewardship, Data Preparation and the Dictionary service

Shared presentation

Impacts on TDS

- All client-server communication must go through a gateway (Web-UI, components)
  - data-stewardship-gateway.war is a new service provided by TDS
  - Tomcat deployment was reworked to make it transparent
  - The gateway is responsible to check authentication



Impacts on TDS configuration file

New configuration properties in data-stewardship.properties:

tds.security=iam

oidc.url=http://tal-qa151:9080/oidc

oidc.userauth.url=http://tal-qa151:9080/oidc

scim.url=http://tal-qa151:9080/scim

oidc.gateway.id=tl6K6ac7tSE-LQ

oidc.gateway.secret=sLbyFKTzM8F0dTL10mHd3A

Data model updates & impact analysis

Data model updates in 6.3

- When a data model was used by at least one campaign, it was only possible to modify
  - Attributes label (column headers in the grid)
  - Attributes constraints (min & max values, …)
  - Data model description

Data model updates in 6.4

- Everything in a data model can be updated, even when used by a campaign

- All tasks are smartly recomputed

- Technical considerations
  - Campaigns are temporarily “disabled”, time to update the tasks
  - Changing an attribute identifier might require to adapt DI jobs

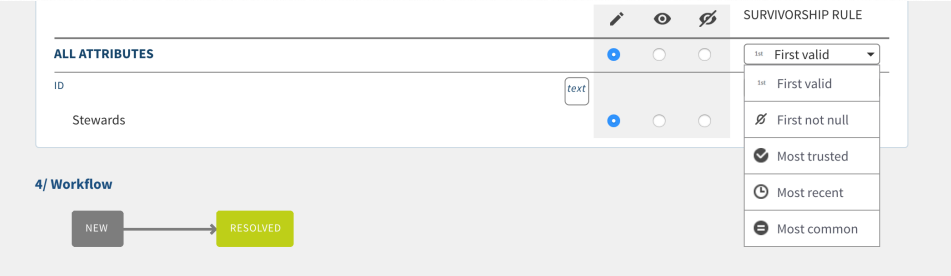
Miscellaneous changes

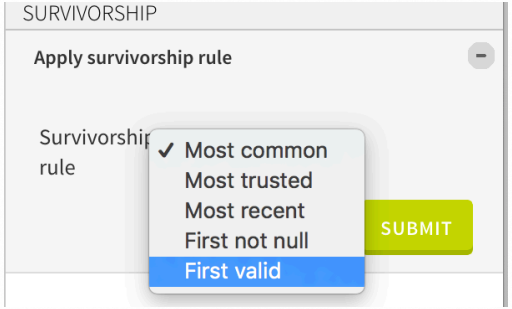
New bulk functions on tasks

Function name	Available to	Description
Assign tasks	• Campaign owners and Data Stewards • Any kind of campaigns	Same as the assignment panel but can work on a filter or all tasks
Define due date	• Campaign owner • Any kind of campaigns	See SLA support
Arbitrate tasks	• Data Stewards • Arbitration / Grouping campaigns	Same as arbitration panel but can work on a filter or all tasks
Transition tasks	• Data Stewards • Campaigns with validation step only during the validation step (Accept / Reject)	Same as transition panel but can work on a filter or all tasks

“First valid value” survivorship rule

- When creating a merging tasks, automatically creates the golden record with the first valid value found in sources with respect to the semantic type
- Available in
  - Campaign creation / edition form (applies to all subsequently created tasks)
  - Apply survivorship rule function (applies to selection / filtered / all tasks)





TDS Components

- Grouping campaign support in DI components (Input/Output/Delete)
- Drag and drop support for TDS metadata
- Support a due date per task (overrides default tasks resolution delay)
- Matching on Spark using TDS
- Arbitration label as a new meta-data in TDS Input component