



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 parallel 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>



Big Data

Talend Summer'17 - Big Data: In a nutshell



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 ?*

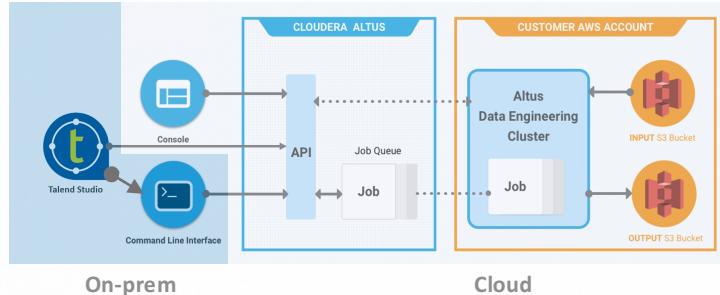
- Cloudera Altus is a PaaS offering, managed Big Data service
- Cloudera 5.11 on AWS (for the first release)
- “*Bring your own AWS account*”, and Altus does the rest: provisioning, setup & management
- “*Pay per use*” for Cloudera, eventually with transient clusters

The screenshot displays two windows of the Cloudera Altus interface. The top window is the 'Jobs' page, which lists completed Spark jobs. The table includes columns for Name, Group, Status, Type, Submitter, Start Time, Cluster, and Actions. Two entries are shown:

Name	Group	Status	Type	Submitter	Start Time	Cluster	Actions
talend_spark_ml_pipeline_process_0_1	talend_spark_ml_pipeline_process_0_1	Completed	Spark	Cyril Sonnefraud	05/19/2017 12:00 PM CEST	talend-pipeline	[Actions]
talend_spark_ml_pipeline_process_0_1	talend_spark_ml_pipeline_process_0_1	Completed	Spark	Cyril Sonnefraud	05/19/2017 11:59 AM CEST	talend	[Actions]

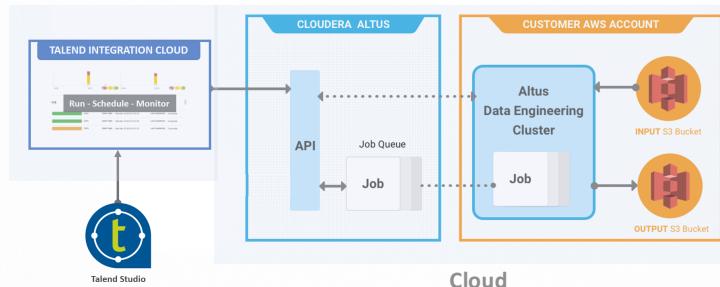
The bottom window is the 'Designer/Code Advisor' tool, specifically the 'Job demarshal' tab. It shows configuration details for a job, including the cluster name ('spain_elastic'), environment ('spain_elastic'), and AWS provider ('AWS'). Other settings include instance type ('t2large'), worker nodes ('10'), and various security and log-related parameters.

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

Job demoAltus 0.1

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

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

Access Key

Cluster Name

Cloudera Altus 1.0

Advanced settings

Target Exec

Memory Path to Altus CLI

/opt/altus-cli/altus

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.

The screenshot displays the Talend Data Fabric interface. At the top, there's a navigation bar with links like File, Edit, View, Window, Help, Learn, Ask, Exchange, Integration, Profiling, MDM, and Mapping. Below the navigation is a toolbar with various icons. The main workspace shows a job named 'Job demoAltus 0.1' with a flowchart of components: tS3Configuration_1, tFileInputParquet_1, tMap_1, tTopBy_1, tAggregateRow_1, rev_sum (Main), and tFileOutputParquet_1. To the left is a sidebar with a tree view of local resources under 'demos'. The bottom half of the screen shows the 'Job demoAltus' configuration panel. It includes sections for Basic Run, Spark Configuration (with fields for Access Key and Cluster Name), Advanced settings, Target Exec, and Memory. A large circular callout highlights the 'Altus Cluster' section, which contains checkboxes for 'Use an existing Altus cluster' (checked) and 'Delete cluster after execution (transient)' (checked), along with a 'Cluster Name' field set to 'talend-altus-cluter'. A tooltip box with the text 'Running a Talend big data (Spark) job has never been easier with Altus: just set the cluster name, Altus does the rest.' is overlaid on the configuration area.

Repository

Palette

Job demoAltus 0.1

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

[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

 Use local mode

Spark Configuration

Property Type

Distribution

Version Cloudera Altus 1.0

Spark Mode

Altus Configuration

Altus Access Key

Altus Secret Key

Path to Altus CLI

Or let Altus dynamically spin up
& down a new cluster for this job
= easy cluster creation + only pay
for the computation time

Altus Cluster

 Use an existing Altus cluster

Cluster Name

Environment

Cloud Provider

AWS

 Delete cluster after execution (transient mode)

AWS Configuration

 Override with a custom JSON configuration

Instance Type

Worker nodes

10

SSH Private Key

Cloudera Manager Username

Cloudera Manager Password

 Enable log

s3://altus_logs_bucket/spark/talend

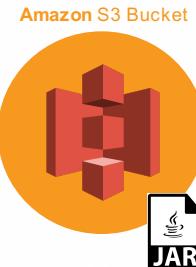
More Cloud ! Talend and Altus: Under the hood



Generate a Java program from a Graphical Job in Talend Studio

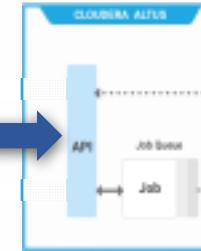


Talend builds the Job as a JAR file

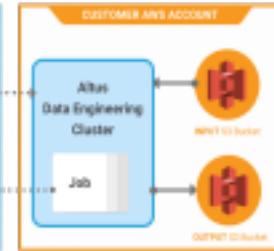


Talend uploads the Job on Amazon S3 Bucket

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



Cloudera Altus™



Monitor the execution in the Cloudera Altus Console.

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

The screenshot shows the Cloudera Altus interface. At the top, there are filters for Environment (All), Cluster (All), Submitter (Cyril Sonne), Status (Completed), and Type (All). Below this is a table titled 'Jobs' with two entries:

Name	Group	Status	Type	Submitter	Start Time	Cluster
talend_spark_ml_pipeline_process_0_1	talend_spark_ml_pipeline_process_0_1	Completed	Spark	Cyril Sonne	05/19/2017 12:00 PM CEST	talend pipeline
talend_spark_ml_pipeline_process_0_1	talend_spark_ml_pipeline_process_0_1	Completed	Spark	Cyril Sonne	05/19/2017 11:59 AM CEST	talend

Below the table are sections for 'cloudera MANAGER' (Clusters, Metrics, Diagnostics, Environments) and 'Accueil' (Jobs, Tables, Clusters). On the right side, there are two graphs: 'Total Events vs. Time' and 'Total Events vs. Delta Time'.

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

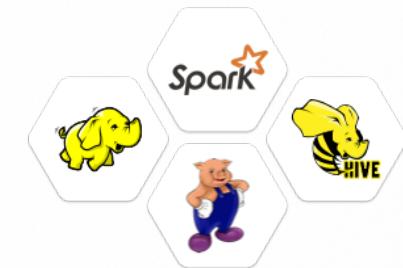
Supported as a new distribution in the Studio

The screenshot shows the 'Spark Configuration' tab of the Talend Studio interface. The configuration includes:

- Basic Run
- Spark Configuration (selected)
- Debug Run
- Advanced settings
- Target Exec
- Memory Run

Configuration details:

- Use local mode:
- Cluster Version: Built-In
- Property Type: Built-In
- Distribution: Google Cloud Dataproc
- Version: Dataproc 1.1 (Apache 2.7.3)
- Spark Version: 2.0
- Spark Mode: YARN Client
- Configuration:
 - Project identifier: "my-google-project"
 - Cluster identifier: "my-cluster-id"
 - Region: "global"
 - Google Storage staging bucket: "gs://my-bucket/talend/jars"
- Authentication:
 - Provide Google Credentials in file
 - Path to Google Credentials file: "/opt/secure/mycredentials"



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

tGoogleDataprocManage_1

Basic settings

Project identifier "my-google-project" *

Cluster identifier "my-cluster-id"

Provide Google Credentials in file

Action Start *
Version Latest *
Zone US Central1 (Iowa) - A *

Instance Configuration

Number of master instances (1 or 3 for HA) instances type n1-standard-2 *
Number of worker instances 2 * Instances type n1-standard-2 *
Number of secondary worker instances 0 *



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

	tBigQueryConfiguration_1	tBigQueryInput_1	tBigQueryOutput_1	tBigQueryOutputBulk_1	tBigQueryBulkExec_1
DI					
Spark Batch					
Spark Stream.					



Google PubSub
components

	tPubSubInput_1	tPubSubInputAvro_1	tPubSubOutput_1
DI			
Spark Batch			
Spark Stream.			



Google Cloud
Storage
components

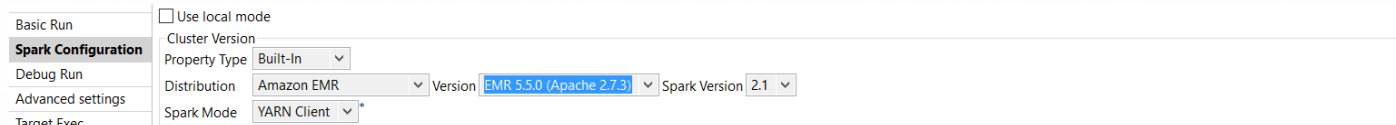
	tGSConfiguration_1	tGSConnection_1	tGSPut_1	tGSGet_1	tGSCopy_1	tGSList_1	tGSBucketCreate_1	tGSBucketList_1	tGSBucketExist_1	tGSClose_1
DI										
Spark Batch										
Spark Stream.										

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)

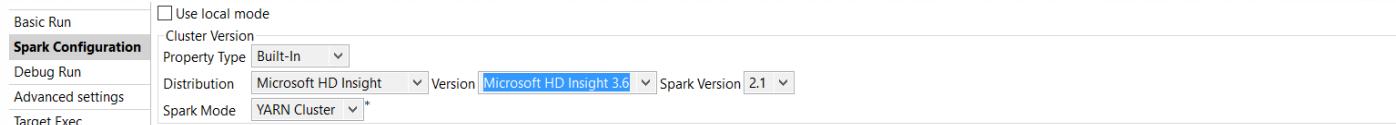


A screenshot of the Studio interface showing the "Spark Configuration" tab selected. The configuration includes:

- Basic Run
- Spark Configuration (selected)
- Debug Run
- Advanced settings
- Target Exec

Configuration fields:
Use local mode:
Cluster Version: Built-In
Property Type: Built-In
Distribution: Amazon EMR
Version: EMR 5.5.0 (Apache 2.7.3)
Spark Version: 2.1
Spark Mode: YARN Client

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



A screenshot of the Studio interface showing the "Spark Configuration" tab selected. The configuration includes:

- Basic Run
- Spark Configuration (selected)
- Debug Run
- Advanced settings
- Target Exec

Configuration fields:
Use local mode:
Cluster Version: Built-In
Property Type: Built-In
Distribution: Microsoft HD Insight
Version: Microsoft HD Insight 3.6
Spark Version: 2.1
Spark Mode: YARN Cluster

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



The screenshot shows the Apache Studio interface for managing cluster configurations. A red oval highlights the 'Spark Version' dropdown menu, which lists '2.1', '1.6', and '2.1'. Other visible dropdown menus include 'Property Type' (set to 'Built-In'), 'Distribution' (set to 'Cloudera'), 'Version' (set to 'Cloudera CDH5.10(YARN mode)'), and 'Spark Mode' (set to 'YARN Client').

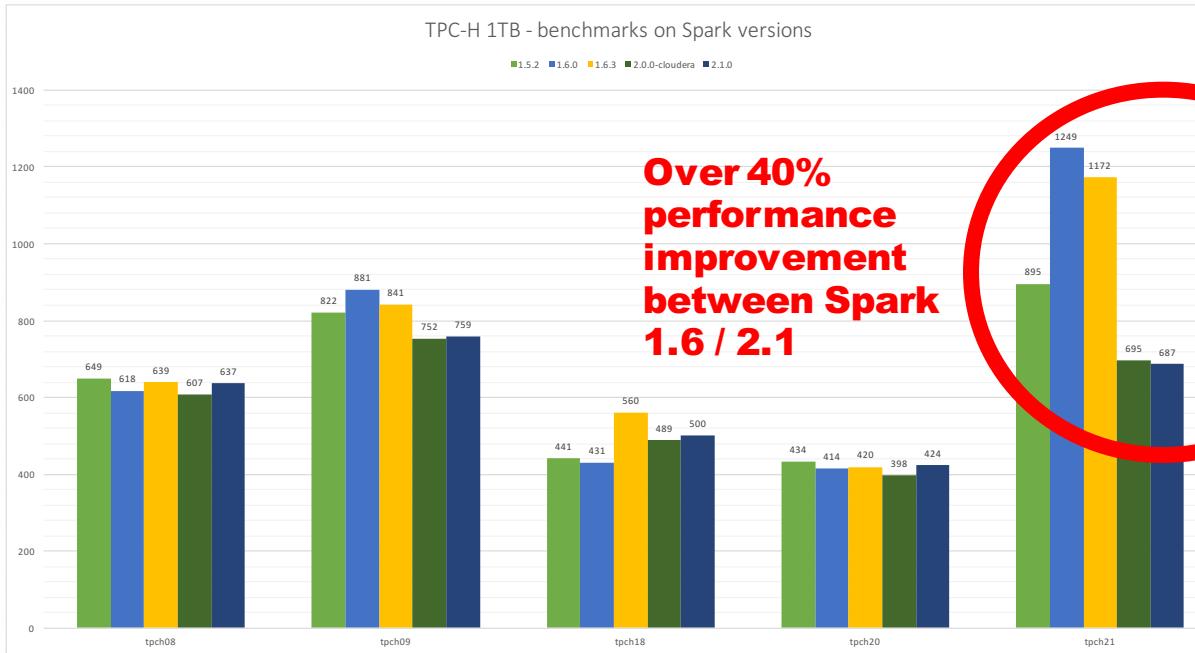


Supported with:

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

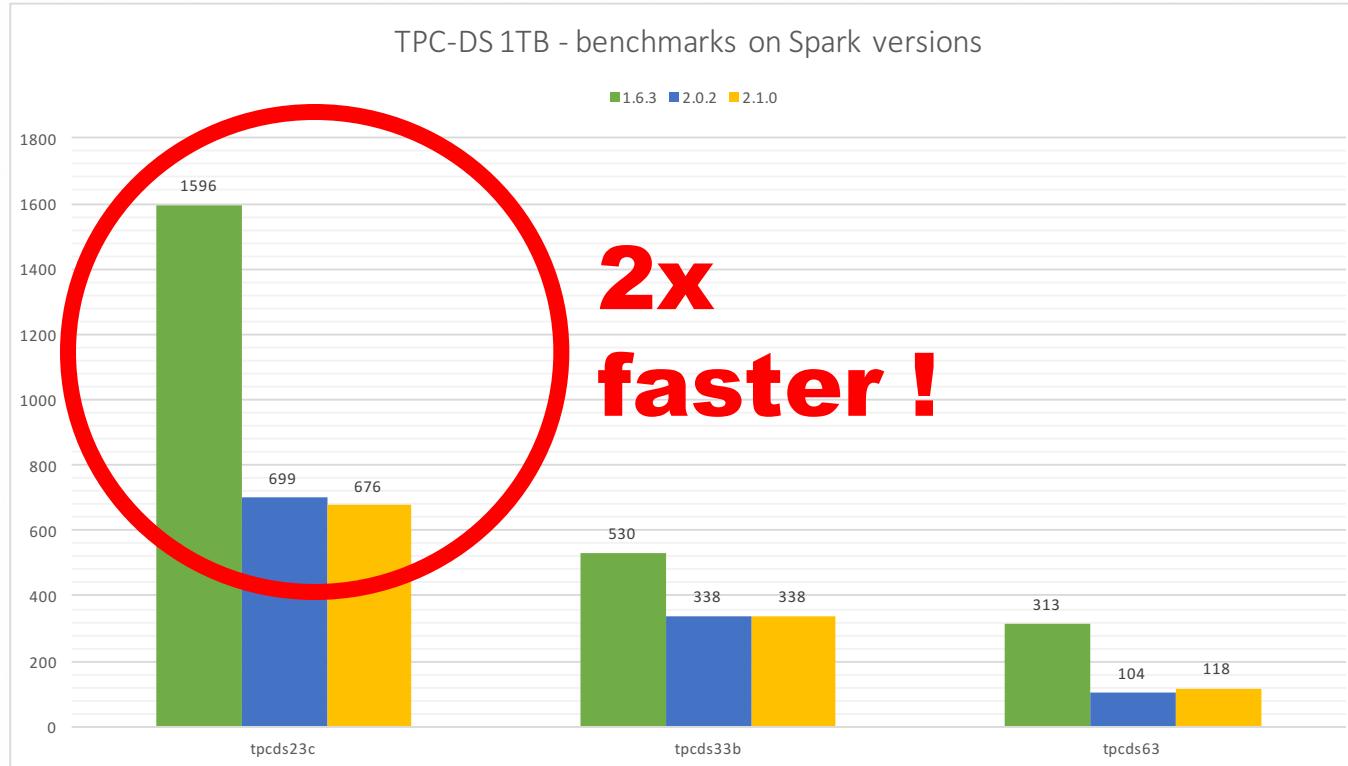
Benefits: Production-ready with Spark 2.1. Enterprise-class robustness.

More Spark ! – Better performance with Spark 2.1



**Over 40%
performance
improvement
between Spark
1.6 / 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

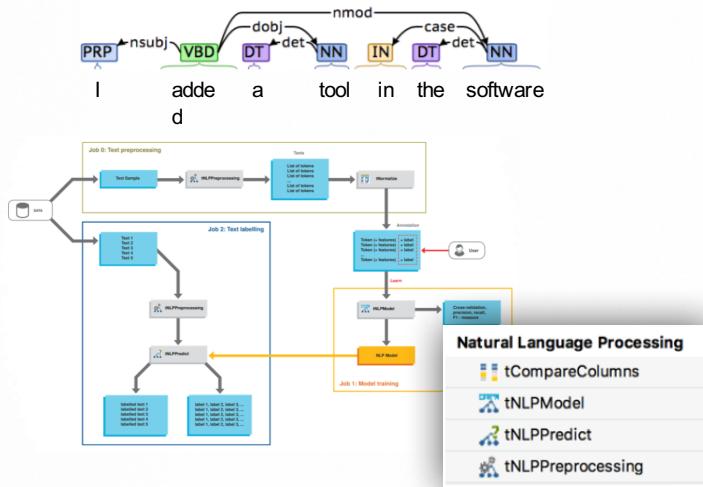
Benefits: Leverage governed self-service within your Real-Time Big Data pipelines.

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



NNP VBD VBN IN CD IN NNP NNP CC NNP
Talend was founded in 2005 by Bertrand Diard and Fabrice Bonan.

Misc Date Person Person
Talend was founded in 2005 by Bertrand Diard and Fabrice Bonan.



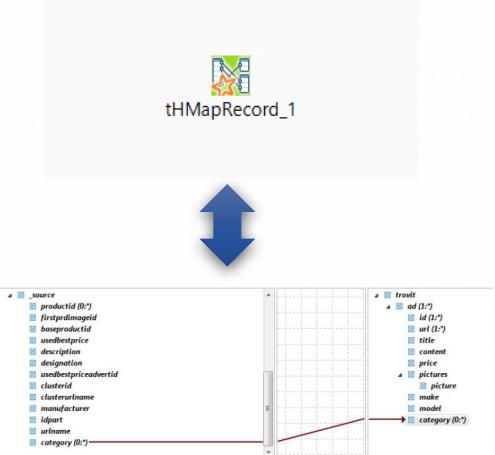
- New Natural Language Processing (NLP) components are available in Spark Batch
- What does it do exactly ?
 - Extract useful information from textual resources (people names, companies, tools...)
 - Classify discussions by topics (group discussions together, find discussions where people are mentioned)
 - 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

The screenshot shows the configuration panel for a tKafkaConfiguration component named "tKafkaConfiguration_1". The left sidebar lists "Basic settings", "Advanced settings", "Dynamic settings", "View", and "Documentation". The main panel has sections for "Connection" (Broker list set to "kafka.weave.local:9092"), "Security" (checkboxes for "Use SSL/TLS" and "Use kerberos authentication" are present, with the latter checked), and "JAAS configuration path" set to "/etc/kafka/kafka_client_jaas.conf". Other fields include "Kafka brokers principal name" set to "kafka" and checkboxes for "Set kinit command path" and "Set kerberos configuration path" both set to "/etc/kafka krb5.conf".

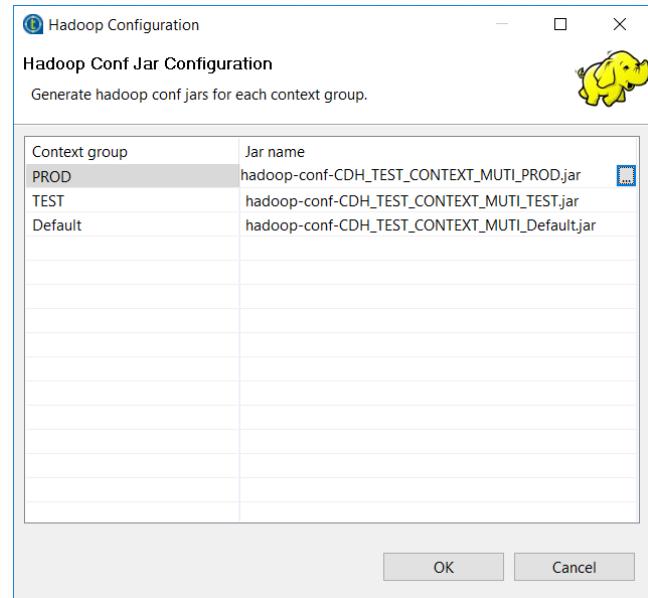
- Hive partitioning for Spark Batch & Streaming

The screenshot shows the configuration panel for a tHiveOutput component named "tHiveOutput_1". The left sidebar lists "Basic settings", "Advanced settings", "Dynamic settings", "View", and "Documentation". The main panel includes sections for "Property Type" (set to "Built-In"), "Storage" (Hive Storage Configuration set to "thiveConfiguration_1" and HDFS Storage Configuration set to "tHDFSConfiguration_1"), "Schema" (set to "Built-In"), "Output source" (set to "Hive table"), "Database" (set to "default"), "Table format" (set to "orc"), "Save mode" (set to "Append"), and "Enable Hive partitions" (checkbox checked). Under "Partition keys", there is a "Column" input field containing "age".

(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

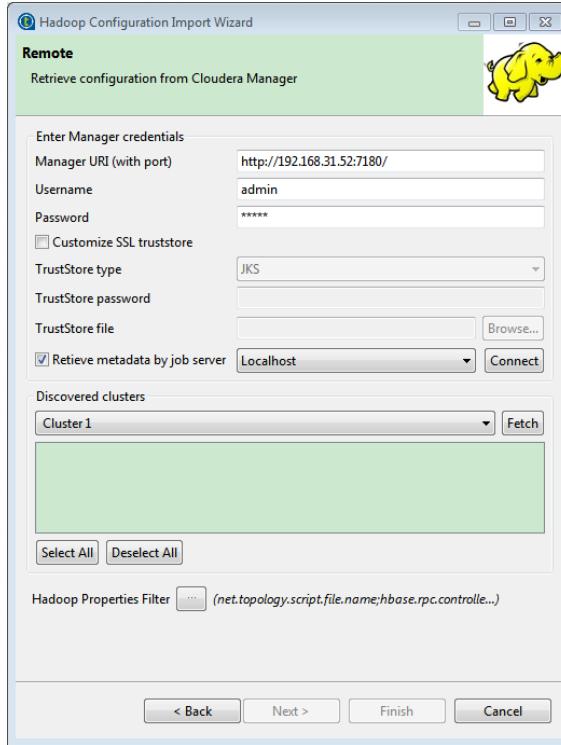
The screenshot shows two windows from the Talend interface:

- Hadoop Cluster Connection**: A window titled "New Hadoop Cluster Connection on repository - Step 2/2". It defines connection parameters for an "Amazon EMR" distribution (Version: EMR 5.5.0 (Apache 2.7.3)).
 - Namenode URI: hdfs://localhost:8020
 - Resource Manager: localhost:8032
 - Resource Manager Scheduler: localhost:8030
 - Job History: 0.0.0:10020
 - Staging directory: /user
 - Use datanode hostname
 - Enable kerberos security
 - User name: hdfs
- Spark Properties**: A window titled "Spark Properties" showing "Spark properties configurations".

Property	Value
spark.memory.offHeap.enabled	true
spark.memory.offHeap.size	1073741824

Buttons at the bottom include OK, Cancel, and a blue "Use" button.

(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

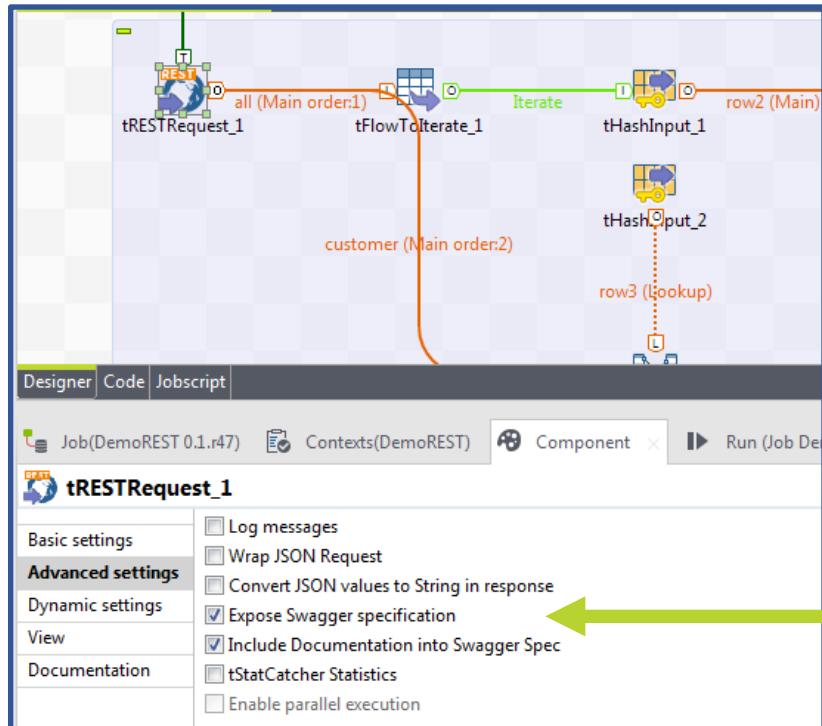
OSS Project	6.3.1	6.4.1
Apache CXF	3.1.7	3.1.12
Apache Camel	2.17.3	2.17.6
Apache Karaf	4.0.7	4.1.1
Apache ActiveMQ	5.14.1 (5.14.2 WebConsole)	5.14.5
Jetty	9.2.1	9.3.14
Syncope	1.2.9	2.0.2 (via Talend IAM)
Spring-Boot	1.3.7	1.3.7



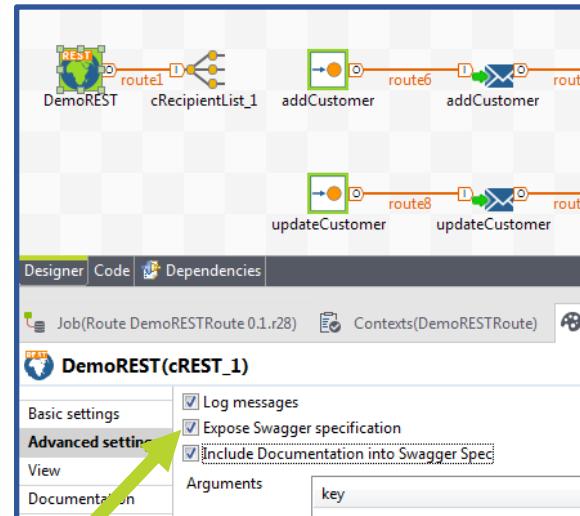
REST API Design

Swagger: How to enable it for REST Provider

REST DS - tRESTRequest



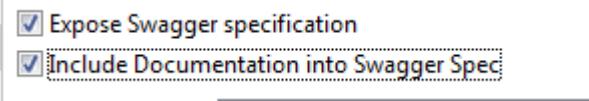
REST Routes – cREST (Provider)



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:



A screenshot of the Talend Studio interface showing a component configuration page for "DemoREST(cREST_1)". The top navigation bar includes tabs for "Job(Route DemoRESTRoute 0.1)", "Contexts(DemoRESTRoute)", "Component" (which is active), "Run (Job DemoRESTRoute)", "Test Cases", "Spring", and "Integration Action".

The main area shows a sidebar with tabs: "Basic settings", "Advanced settings", "View", and "Documentation". The "Documentation" tab is highlighted with a red oval. To its right, there's a "Comment" section containing the following text:

Contents of this tab will appear in Swagger doc and also on Swagger UI page
Please note, that Swagger supports formatting of this section with **Markdown** <https://en.wikipedia.org/wiki/Markdownformatting>.
This is an example of 2nd level header.
Not ordered list
* Apples
* Oranges
* Pears

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=/service>

```

"swagger": "2.0",
"info":
  {
    "version": "1.0.0",
    "title": "DemoRESTHouse REST Application"
  },
"basePath": "/services/customers",
"paths": {
  "/":
    {
      "get": {
        "summary": "Operation getAllCustomers",
        "description": "Produces: ( application/xml, text/xml, application/json )",
        "operationsId": "getAllCustomers",
        "produces": [
          "application/xml",
          "text/xml",
          "application/json"
        ],
        "responses": {
          "200": {
            "description": "Successful operation",
            "schema": {
              "type": "object"
            }
          }
        }
      }
    }
},
"responses": {
  "200": {
    "description": "Successful operation"
  }
}
}

```

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 used 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

The screenshot shows the Talend Platform Swagger UI interface. At the top, the URL is `localhost:8040/services/customers/api-docs?url=/services/customers/swagger.json#/default/getAllCustomers`. The title bar says "talend Talend Platform". Below the title, it says "DemoRESTRoute REST Application". The main content area is for the "default" operation, specifically the "GET /" endpoint. It includes "Implementation Notes" (produces application/xml, text/xml, application/json), a "Response Class (Status 200)" section (successful operation), and a "Model Example Value" section with XML code. A "Try it out!" button is present. Below this, there are sections for other operations: "POST /" (Operation addCustomer), "DELETE /{id}" (Operation deleteCustomer), "GET /{id}" (Operation getCustomer), and "PUT /{id}" (Operation updateCustomer). The footer indicates a base URL of "/services/customers" and API version 1.0.0.

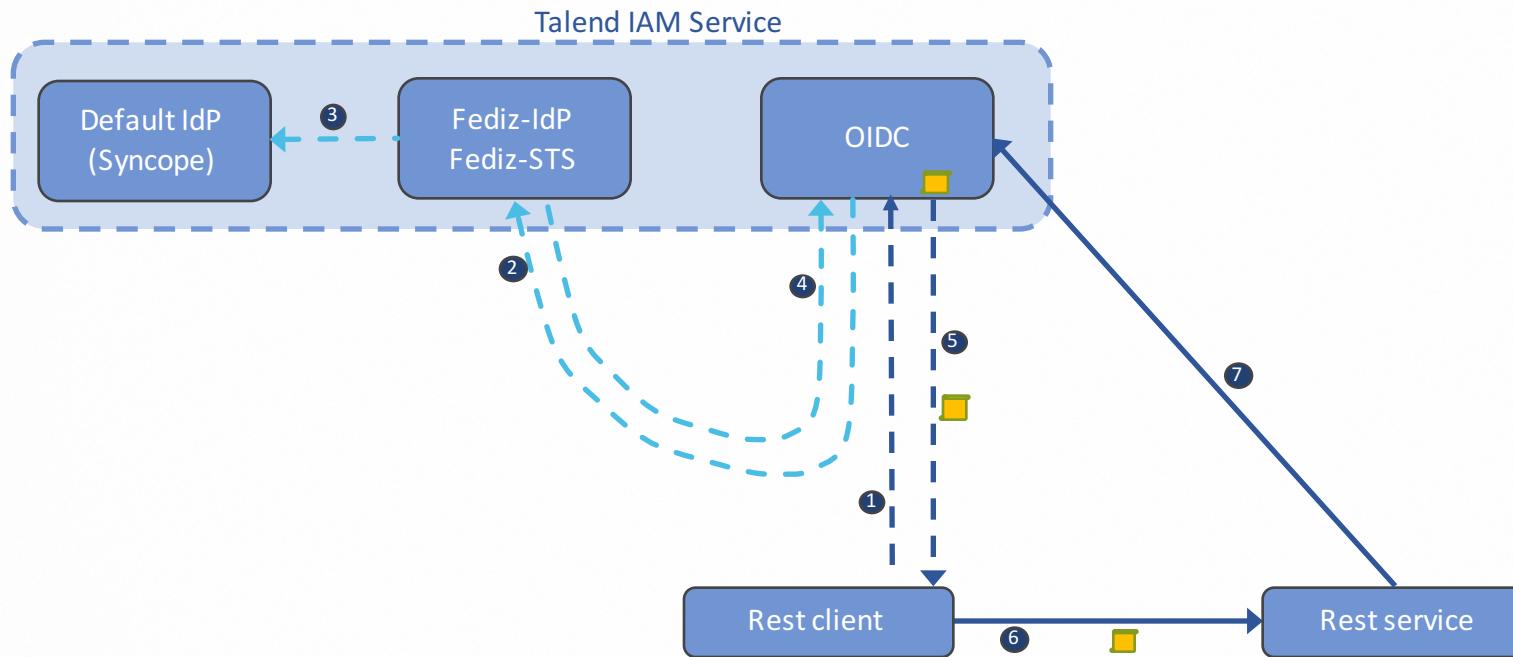
Talend IAM - Syncope, Fediz, STS OIDC

- 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

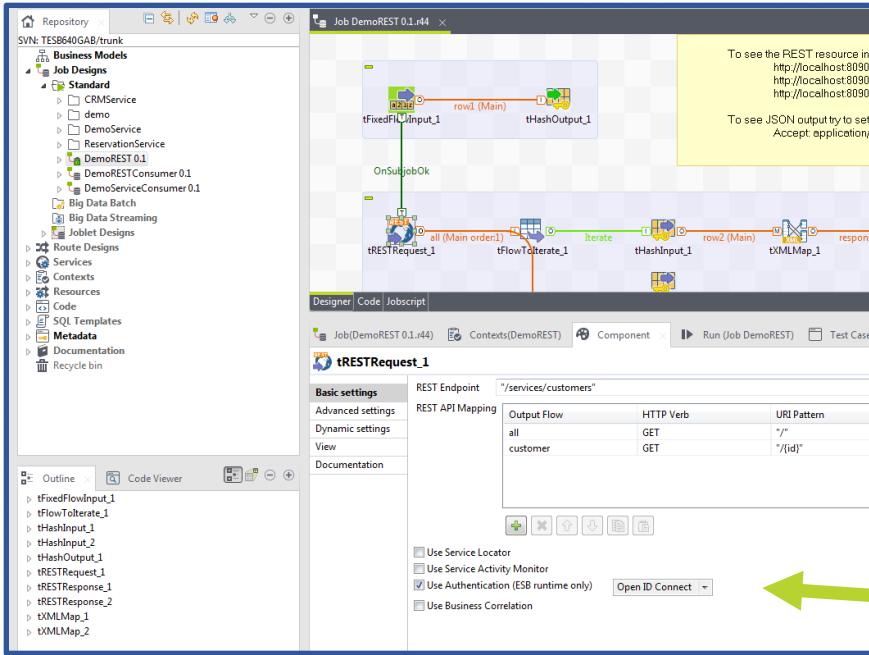
- With 6.4 platform, we are introducing a new IAM service. This IAM service is based on standard specifications and protocols like OpenID Connect (OIDC), OAuth2, WS-Federation and WS-Trust. In 6.4 it provides
 - SSO between TDP and TDS based on OIDC & WS-Federation standards.
 - Authentication to ESB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

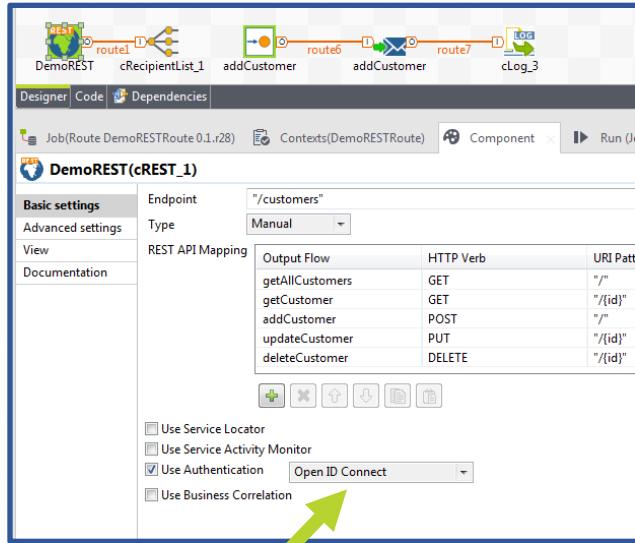


OIDC: ESB Rest – Provider

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

The screenshot shows the Talend Studio interface with a job editor. At the top, there is a flow diagram with components like 'URI_Patterns', 'input (Main)', 'Call_REST_Service', 'row2 (Response)', and 'Duplicate_Response'. Below the diagram, the 'Call_REST_Service' component is selected, and its configuration panel is displayed.

Call_REST_Service(tRESTClient_1) Configuration:

- Basic settings:**
 - URL: "http://127.0.0.1:8040"
 - Relative Path: "/services/customers/" + input.number
 - HTTP Method: GET
 - Accept Type: XML
- Query parameters:** A table with columns 'name' and 'value'.
- Input Schema:** Built-In, Edit schema, Sync columns.
- Response Schema:** Built-In, Edit schema, Sync columns.
- Error Schema:** Built-In, Edit schema, Sync columns.
- Authentication:** Use Authentication is checked, and the Authentication Type is set to Open ID Connect. The 'Username' field contains 'context.username_rest' and the 'Password' field contains 'context.password_rest'.
- Business Correlation:** Use Business Correlation is unchecked.

REST Routes – cREST (Client)

The screenshot shows the Talend Studio interface with a job editor. At the top, there is a flow diagram with components like 'ResourceAsServiceCXF', 'route1', 'Request', 'route2', and 'DemoRESTClient'. Below the diagram, the 'DemoRESTClient' component is selected, and its configuration panel is displayed.

DemoRESTClient(cREST_1) Configuration:

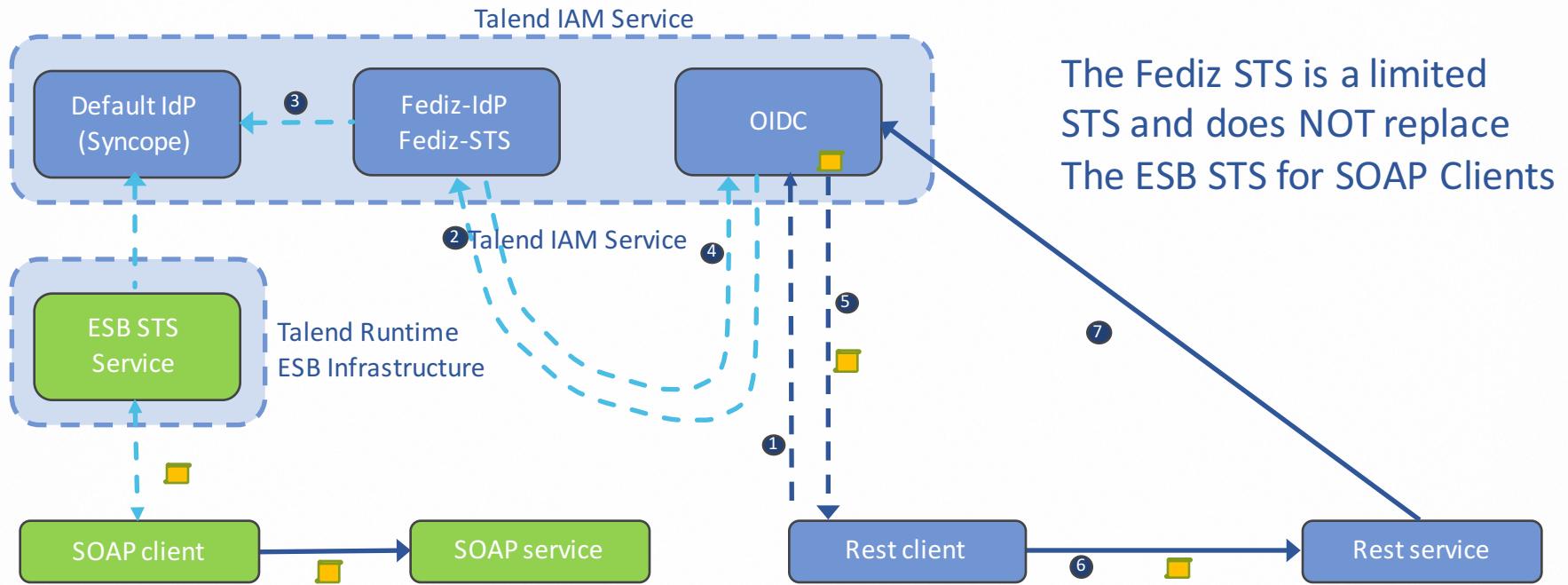
- Basic settings:**
 - Endpoint: "http://localhost:8040/services/customers"
 - Type: Manual
 - Relative Path: simple("\${body}")
 - HTTP Method: GET
 - Accept Type: XML
 - Response class: (empty)
- Authentication:** Use Authentication is checked, and the Authentication Type is set to Open ID Connect. The 'Username' field contains 'context.username_rest' and the 'Password' field contains 'context.password_rest'.
- Business Correlation:** Use Business Correlation is unchecked.

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 technically possible to register own clients but we did 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

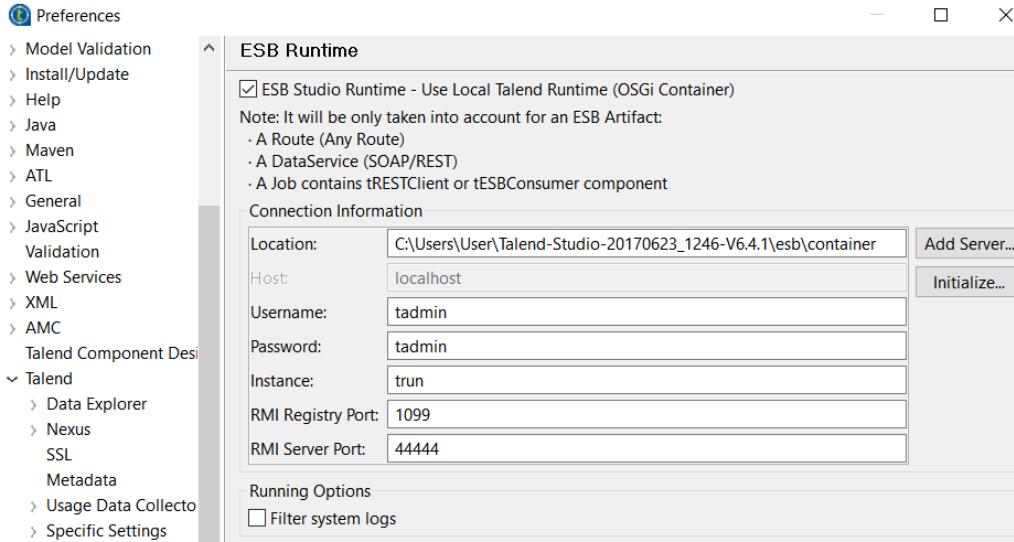
- ESB Capabilities testable from Studio now
 - Data Sources (connection pooling and alias)
 - Service Locator
 - Service Registry
 - Service Activity Monitoring
 - Authentication and Authorization
 - Transport Security (TLS)
 - Concurrent routes (e.g. cVM calls)
 - Concurrency of WebServices

Developer Desktop



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

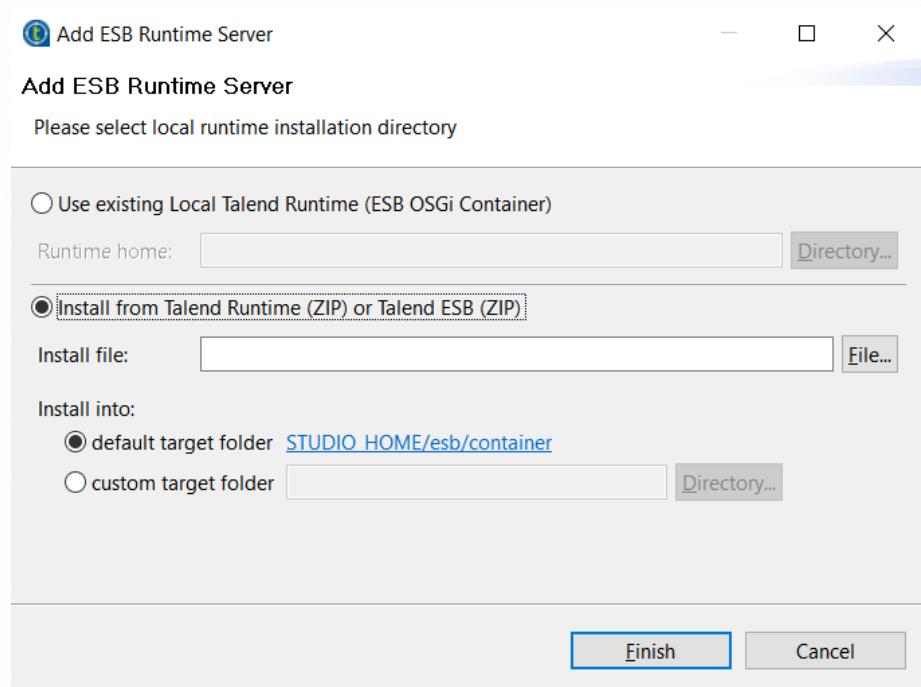
Local Runtime: Setup/Preferences 1



- 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

CONNECTION INFORMATION

Location:	E:\Talend\V6.4.0_GA_B\studio\esb\container
Host:	localhost
Username:	tadmin
Password:	tadmin
Instance:	trun
RMI Registry Port:	1099
RMI Server Port:	44444

Add Server... Initialize...

The screenshot shows the Talend Studio interface. On the left, there's a 'CONNECTION INFORMATION' panel with fields for Location (E:\Talend\V6.4.0_GA_B\studio\esb\container), Host (localhost), Username (tadmin), Password (tadmin), Instance (trun), RMI Registry Port (1099), and RMI Server Port (44444). Below it is an 'Add Server...' and an 'Initialize...' button. A green arrow points from the 'Initialize...' button to a stack of four progress dialog boxes. The top dialog says 'Starting Runtime Server...'. The second dialog says 'Running script (E:\Talend\V6.4.0_GA_B\esb\c...ost -l1 source file:scripts/initlocal.sh)' and '290 bundles have been activated'. The third dialog says 'Running script (E:\Talend\V6.4.0_GA_B\esb\c...ost -l1 source file:scripts/initlocal.sh)' and 'Script is running to end...'. The bottom dialog is titled 'Initialize Finished' and says 'Local runtime serview has been started, totally installed bundles: 378.' with an 'OK' button. On the right, there's a file browser window showing a directory structure: V6.4.0_GA_B > esb > container > scripts. Inside the scripts folder, there are several files: configEventLogging_JMS.sh, configEventLogging_REST.sh, configKarafContainer.sh, configureC0.sh, configureC1.sh, configureC2.sh, configureC3.sh, and initlocal.sh. A blue box highlights the 'initlocal.sh' file, and a green arrow points from its name in the browser to the content of the file in the code editor below. The code in 'initlocal.sh' is as follows:

```
echo "Initialize Local Test Runtime Server"
echo
echo "Install tesb:start-all ....."
tesb:start-all
echo "Done"

echo
echo "Install tesb:switch-sts-jaaas ....."
tesb:switch-sts-jaaas
echo "Done"

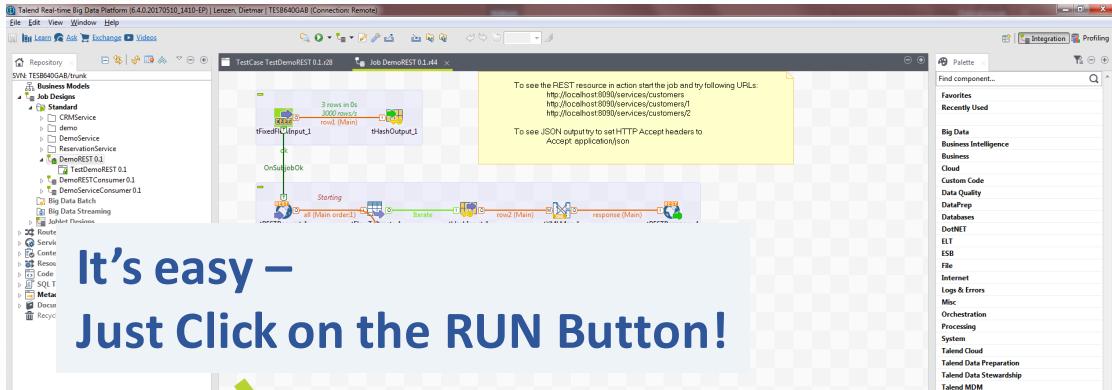
echo
echo "Install feature:install activemq-broker ....."
feature:install activemq-broker
echo "Done"

echo
echo "Initialize finished successfully."

# Talend ESB Studio needs to read the ending tag on initializing, please do not
remove it.
echo "EOF"
```

Local Runtime: How to use it?

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

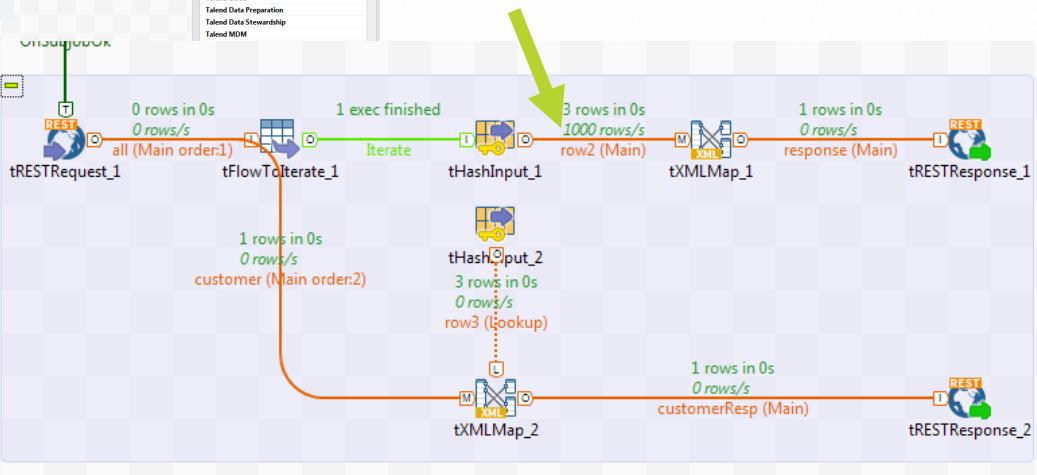


The screenshot shows the JBoss Seam Test Runner interface with the following details:

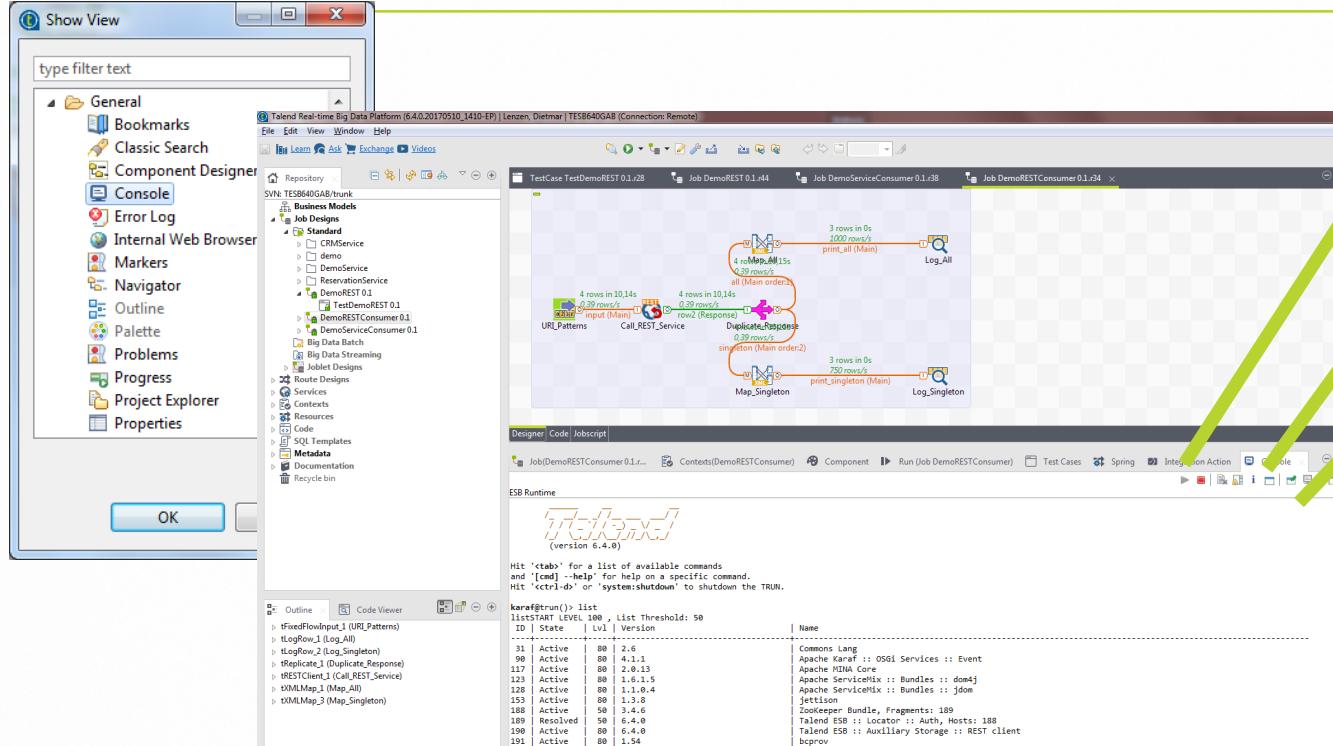
- Outline** and **Code Viewer** tabs are visible at the top.
- The main title is **Job Demo REST**.
- Job Demo REST** is selected in the dropdown menu.
- Job Demo REST** is also listed in the **Component** tab.
- Execution** tab is active.
- Basic Run** is selected in the dropdown.
- Run** button is highlighted in blue.
- Advanced settings** and **Target Exec** dropdown are present.
- Memory Run** is also listed in the dropdown.
- Logs** tab is open, displaying log entries for the REST service. The logs show several INFO messages from the `Talend ESB - Locator` and `ServiceLocator` components, indicating the registration of endpoints and services.

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

See the Rows executed directly in the model



Local Runtime: Console View to the Runtime?



Start / Stop Runtime
Open Preferences
Run (Karaf) Console Commands

- 1 Java Stack Trace Console
- 2 Maven Console
- 3 New Console View
- 4 ESB Runtime
- 5 ANTLR Console

cConfig integrates with Nexus

 cConfig_1

Basic settings

Imports `//import java.util.List;`

Code
`/*
 * Here you can put Java code th
 *
 * Usually you would register cu
 * manipulate typeConverterRegis
 *
 * For example:
 ..`

Dependencies Lib Path: org.sat4j.core.jar Version: 2.3.5

1

 Nexus Repository Manager OSS admin

Welcome | Repositories | Search

Artifact Search: mime | Advanced Search | Views/Repositories

Keyword Search: sat4j

sonatype™

talend / libraries / org.sat4j.core / 2.3.5 / org.sat4j.core-2.3.5.jar

2

2 Publish from Studio
Publish from Builds

 cConfig_1

Basic settings

Advanced settings

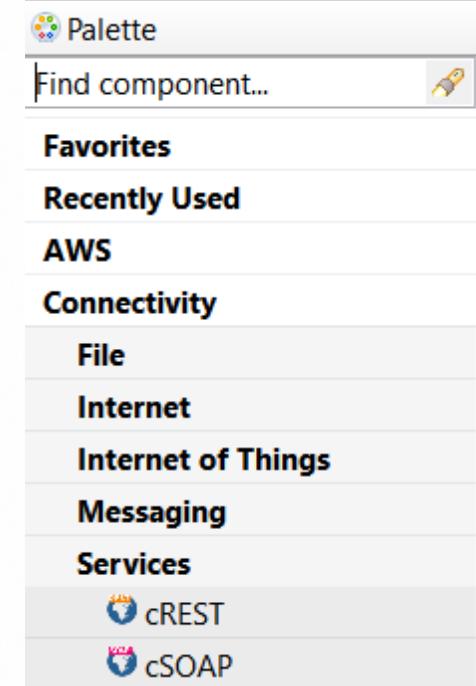
Use MDC Logging **Check** Sync

3 Check and Manually refresh to newer version

Sync	Status	Jar Name	Update To	
<input checked="" type="checkbox"/>	<input type="checkbox"/>	✓	org.sat4j.core.jar	2.3.5

Route Builder: cREST/cSOAP

- Renaming
 - cCXFRS → cREST
 - cCXF → cSOAP
- Why?
 - Easier for new customers

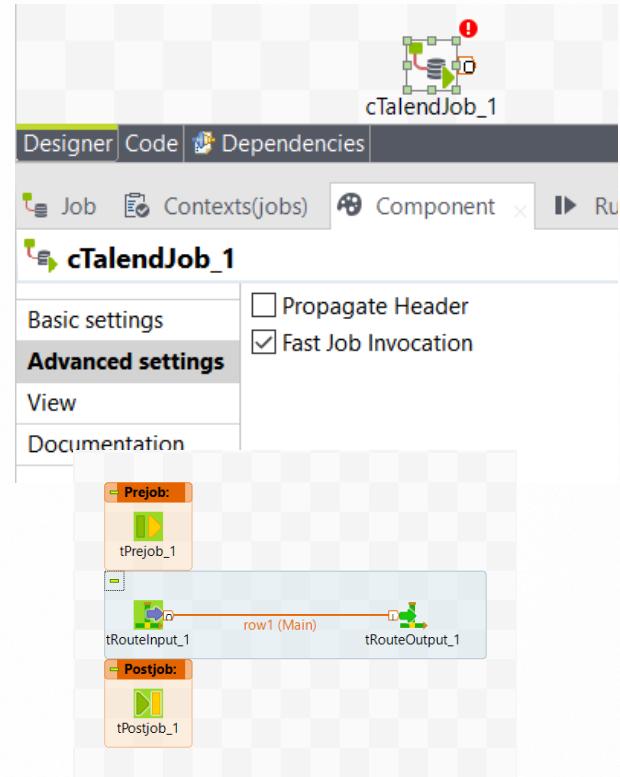




Design Patterns

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.

The screenshot shows the Talend Component view with the following details:

- Toolbar:** Job, Contexts(jobs), Component (selected), Run (Job jobs), Test Cases, Spring, Integration Action, Console, Progress.
- Panel:** child(child_1)
- Basic settings:**
 - Use Static Context(selected below) Use Route Context
 - Routelet: child
 - Version: Latest
 - Context: Default
- Advanced settings:** Advanced settings, View, Documentation.
- Overwrite Contexts:** Parameters: param_1, param_2. Values: "override 1", "override 2".



Service Registry – Metadata Support

SR-Metadata: TAC UI – Metadata Tab

The screenshot shows the Talend Administration Center (TAC) interface with the 'Metadata' tab selected. The top navigation bar includes tabs for Content, Assignments, Endpoints, and Metadata. Below the tabs are buttons for Create, Save, Remove, and Advanced Edit. The main area displays a table with columns for Name and Value. The data entries are:

Name	Value
ServiceType	ApplicationService
ServiceStatus	active
ServiceVersion	1.1
CreationDate	2017-05-01
Description	Product List for Partner
Usage	public
ServiceOwner	PR-Dep
ServiceDesigner	DL
ServiceDevelopmentTeam	INT-DEV-TEAM
Projectname	SATURN
ApplicationName	ProductList
ServiceMaintenanceOwner	BA-01
ServiceMaintenanceContact	AX2002
Tags	Web
Tags	Partner

- 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 a single line)

The screenshot shows the Talend Administration Center (TAC) interface with the 'Metadata' tab selected. The top navigation bar includes tabs for Content, Assignments, Endpoints, and Metadata. Below the tabs are buttons for Create, Save, Remove, and Advanced Edit. The main area displays the XML representation of the service metadata. The XML code is as follows:

```
<metadataExtension xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://extension.registry.esb.talend.org/v1" xmlns:ns="http://metadata.extension.registry.esb.talend.org/v1">
  <ServiceName>(http://www.talend.org/service)/DemoService</ServiceName>
  | <ns:metadata>
    <ns:ServiceType>ApplicationService</ns:ServiceType>
    <ns:ServiceStatus>active</ns:ServiceStatus>
    <ns:ServiceVersion>1.1</ns:ServiceVersion>
    <ns:CreationDate>2017-05-01</ns:CreationDate>
    <ns:Description>Product List for Partner</ns:Description>
    <ns:Usage>public</ns:Usage>
    <ns:ServiceOwner>PR-Dep</ns:ServiceOwner>
    <ns:ServiceDesigner>DL</ns:ServiceDesigner>
    <ns:ServiceDevelopmentTeam>INT-DEV-TEAM</ns:ServiceDevelopmentTeam>
    <ns:Projectname>SATURN</ns:Projectname>
    <ns:ApplicationName>ProductList</ns:ApplicationName>
    <ns:ServiceMaintenanceOwner>BA-01</ns:ServiceMaintenanceOwner>
    <ns:ServiceMaintenanceContact>AX2002</ns:ServiceMaintenanceContact>
    <ns:Tags>Web</ns:Tags>
    <ns:Tags>Partner</ns:Tags>
  </ns:metadata>
</metadataExtension>
```

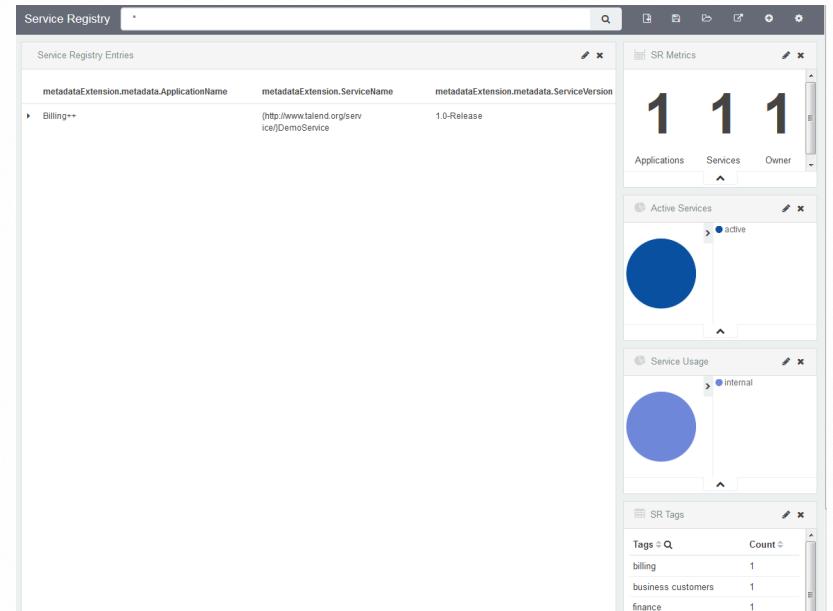
SR-Metadata: Elasticsearch

- Search capabilities via ElasticSearch (ES)
- Metadata is synchronized between Registry and 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**

The screenshot shows the Talend Real-Time Big Data Platform's user management interface. The left sidebar has a 'Menu' with options: Settings, Users (selected), User groups, Licenses, Configuration, Rights management, and User settings. The main area is titled 'USERS' and lists three users:

Login	Role	Actions
admin@company.com	Administrator	... ⚙️ 🗑️ 🔍
tsteinborn@talend.com	Designer	... ⚙️ 🗑️ 🔍
security@company.com	Security Administrator	... 🔍

The 'Data' panel on the right contains the following fields:

- Login: security@company.com
- First name: security
- Last name: security
- Password: [change password](#)
- Svn login:
- Svn password: [change password](#)
- GIT login:
- GIT password: [change password](#)
- Type: [No Project Access](#)
- Role: [Security Administrat](#)

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

Enab...	Role	Allo...	Description ▾
	Security Administrator		<input type="checkbox"/> Audit management
	Administrator		<input type="checkbox"/> Audit visualization
	Designer		<input type="checkbox"/> Backup management
	Operation manager		<input type="checkbox"/> Business modeler management
	Viewer		<input type="checkbox"/> Business modeler visualization
	Custom Role 1		<input type="checkbox"/> Commandline
	Custom Role 2		<input type="checkbox"/> Configuration management
	Custom Role 3		<input type="checkbox"/> Configuration visualization
	Custom Role 4		<input type="checkbox"/> Dashboard SOA
	Custom Role 5		<input type="checkbox"/> Dashboard management
	Custom Role 6		<input type="checkbox"/> Dashboard visualization
	Custom Role 7		<input type="checkbox"/> Documentation
	Custom Role 8		<input type="checkbox"/> Drools
	Custom Role 9		<input type="checkbox"/> ESB Authorization management
	Custom Role 10		<input type="checkbox"/> ESB Authorization visualization



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

The screenshot shows the Talend Configuration interface. On the left, a sidebar menu includes 'Settings', 'Users', 'User groups', 'Licenses', 'Configuration' (which is selected), and 'Rights management' and 'User settings'. The main area is titled 'CONFIGURATION' and shows 'LDAP' and 'SSO' sections. Under 'SSO', there are fields for 'Use SSO Login' (set to 'true'), 'IDP metadata' (containing '/ssologin'), 'Service Provider Entity ID' (set to 'Okta'), 'Identity Provider System' (set to 'Okta'), 'Organization URL' (set to 'https://[organization].okta.com'), 'Okta App Embed Link' (empty), 'Use Role Mapping' (set to 'true'), and 'Mapping Configuration' (with a red circle around it). A red oval highlights the 'Use Role Mapping' and 'Mapping Configuration' fields.

SSO: Make role mapping configurable for SAML

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

The image displays two 'Mapping Configuration' dialog boxes side-by-side, illustrating the configuration of role mappings for SAML.

Left Dialog (Basic Configuration):

- IDP metadata:
 - Service Provider Entity ID: /sslogin
 - Identity Provider System: Okta
 - Organization URL: https://[organization]
 - Okta App Embed Link: [redacted]
 - Use Role Mapping: true
- Mapping Configuration:
 - TAC Project Types: MDM, DQ, DI, NPA
 - Roles Mappings:
 - Talend Administration Center:
 - Security Administrator: [redacted]
 - Administrator: [redacted]
 - Viewer: [redacted]
 - Operation Manager: [redacted]
 - Designer: [redacted]
 - Custom Roles:
 - Custom Role 1: [redacted]
- Buttons: Save, Cancel

Right Dialog (Advanced Configuration):

- IDP metadata:
 - Service Provider Entity ID: /sslogin
 - Identity Provider System: Okta
 - Organization URL: https://[organization]
 - Okta App Embed Link: [redacted]
 - Use Role Mapping: true
- Mapping Configuration:
 - Show Advanced Configuration: checked
 - TAC Project Types: MDM, DQ, DI, NPA
 - Path to Value: MDM, DQ, DI, NPA
 - Roles Mappings:
 - Talend Administration Center:
 - Security Administrator: [/saml2p:Response/saml2:Assertion/saml2:AttributeSet]
 - Administrator: [/saml2p:Response/saml2:Assertion/saml2:AttributeSet]
 - Viewer: [/saml2p:Response/saml2:Assertion/saml2:AttributeSet]
 - Operation Manager: [/saml2p:Response/saml2:Assertion/saml2:AttributeSet]
 - Designer: [/saml2p:Response/saml2:Assertion/saml2:AttributeSet]
 - Custom Roles:
 - Custom Role 1: [/saml2p:Response/saml2:Assertion/saml2:AttributeSet]
- Buttons: Save, Cancel



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

LICENSES

Refresh Add new license Validate your license manually

Stored license keys

Show inactive

License Key	Mode	Product	Versi...	Expiration	Expires in (d...)	DI/E...	DI/E...	DQN
QXuG1 ... vbGzeFyw==	NAMED	Talend Real-time - Big Dat...	6.4	2017-06-09	28	25	25	
QXuG1 ... q4dBLoBgHe	NAMED	Talend Data Management...	6.4	2017-06-09	28	25	25	

Available users

Talend Real-time - Big Data Platform Talend Data Management Platform

License mode

Data Integration
 Data Quality
 Data Preparation/Data Stewardship

Defined Data Integration/ESB users 0 / 25
Defined Data Quality users 1 / 25
Defined Data Preparation/Data Stewardship users 0 / 102

You must logout to see changes after having set a new license.

[https://cms.talend.com/
display/PRIC/Pricebooks](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

Data	
Login:	security@company.com
First name:	security
Last name:	security
Password:	change password
Svn login:	
Svn password:	change password
GIT login:	
GIT password:	change password
Type:	No Project Access
Role:	Security Administrat 
Data Preparation User:	<input type="checkbox"/>
Data Stewardship User:	<input type="checkbox"/>

TAC 6.4: Deprecate TAC features

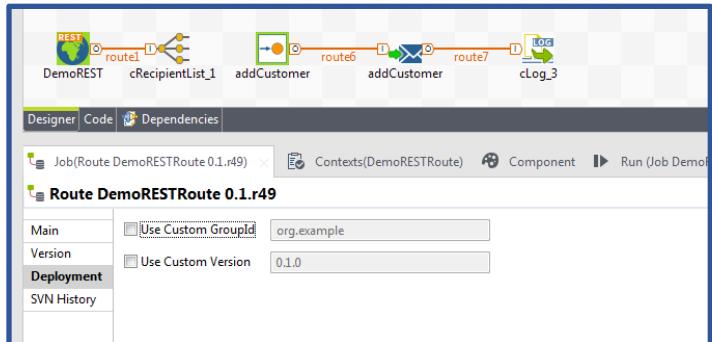
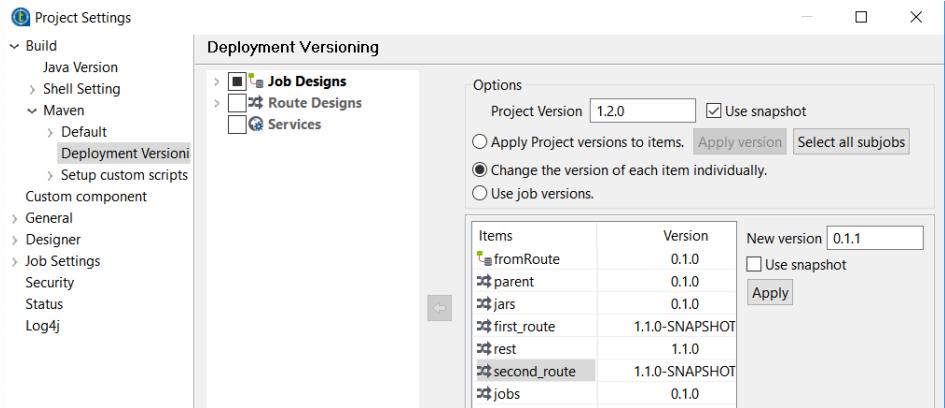
- TAC: Publisher
- TAC: SVN/Git based task in Job Conductor
- TAC: SVN Backup
- TAC: DB Backup
- TAC: Repository Browser
- Repository Manager: Repository Migration



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

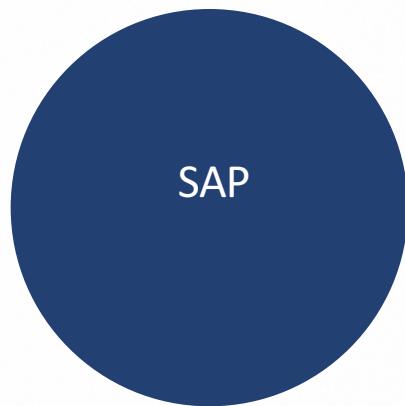
Cloud



Enterprise Class



Connectivity



Talend Data Integration 6.4

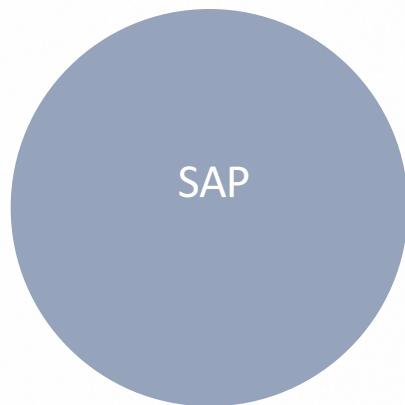
Cloud



Enterprise Class



Connectivity



Cloud connectivity



Loading 20x faster than 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)

Cloud
Azure Storage
Blob
tAzureStorageContainerCreate
tAzureStorageContainerDelete
tAzureStorageContainerExist
tAzureStorageContainerList
tAzureStorageDelete
tAzureStorageGet
tAzureStorageList
tAzureStoragePut
Queue
tAzureStorageQueueCreate
tAzureStorageQueueDelete
tAzureStorageQueueInput
tAzureStorageQueueInputLoop
tAzureStorageQueueList
tAzureStorageQueueOutput
tAzureStorageQueuePurge
Table
tAzureStorageInputTable
tAzureStorageOutputTable
tAzureStorageConnection

Talend Data Integration 6.4

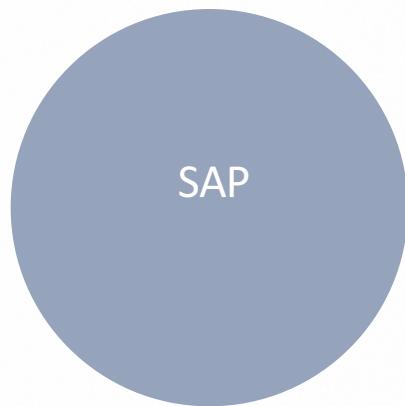
Cloud



Enterprise Class



Connectivity



Productivity: Dynamic Column search

Find columns you want to work with in seconds!

Works for sources & Targets

Full Text Search engine

Talend Data Fabric - tMap - tMap_1

in

Column

- CustId
- FirstName
- LastName
- NumStreet
- PhoneCell
- PhoneHome
- PhoneOffice
- NumberOfOpportunity
- Street1
- Street2
- City
- Country
- Region
- CustId1
- FirstName1
- LastName1
- NumStreet1

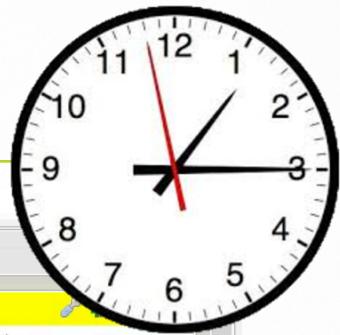


Productivity: Batch Change expressions

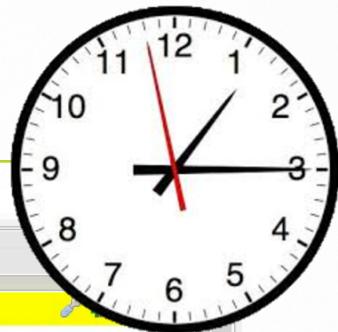
Excel style batch expression change.
Massively apply transformation to selected columns

The screenshot shows the Talend Data Integration interface. On the left, there is a 'Var' panel with a search bar at the top. On the right, there is a mapping table titled 'out1' with two columns: 'Expression' and 'Column'. A yellow arrow points from the 'Var' panel towards the mapping table.

Expression	Column
in.CustId	CustId
in.FirstName	FirstName
in.LastName	LastName
in.NumStreet	NumStreet
in.PhoneCell	PhoneCell
in.PhoneHome	PhoneHome
in.PhoneOffice	PhoneOffice
in.NumberOfOpportunity	NumberOfOpportunity
in.Street1	Street1
in.Street2	Street2
in.City	City
in.Country	Country
in.Region	Region
in.CustId1	CustId1
in.FirstName1	FirstName1
in.LastName1	LastName1
in.NumStreet1	NumStreet1
in.PhoneCell1	PhoneCell1
in.PhoneHome1	PhoneHome1
in.PhoneOffice1	PhoneOffice1
in.NumberOfOpportunity1	NumberOfOpportunity1
in.Street11	Street11
in.Street21	Street21
in.City1	City1



Productivity: Batch Change expressions



Excel style batch expression change.

Massively apply transformation to selected columns

The screenshot shows two windows from a data integration tool:

- Var View:** A large window on the left labeled "Var" containing a grid of rows. A yellow arrow points from the "in.CustId" entry in the Var view to its corresponding row in the "out1" mapping table.
- Auto map! View:** A smaller window on the right titled "Auto map!" showing a mapping table titled "out1". The table lists input expressions (e.g., "in.CustId", "in.FirstName") and their corresponding output columns (e.g., "CustomerId", "FirstName").

Expression	Column
in.CustId	CustomerId
in.FirstName	FirstName
in.LastName	LastName
in.NumStreet	NumStreet
in.PhoneCell	PhoneCell
in.PhoneHome	PhoneHome
in.PhoneOffice	PhoneOffice
in.NumberOfOpportunity	NumberOfOpportunity
in.Street1	Street1
in.Street2	Street2
in.City	City
in.Country	Country
in.Region	Region
in.CustId1	CustomerId1
in.FirstName1	FirstName1
in.LastName1	LastName1
in.NumStreet1	NumStreet1
in.PhoneCell1	PhoneCell1
in.PhoneHome1	PhoneHome1
in.PhoneOffice1	PhoneOffice1
in.NumberOfOpportunity1	NumberOfOpportunity1
in.Street11	Street11
in.Street21	Street21
in.City1	City1

Talend Data Integration 6.4

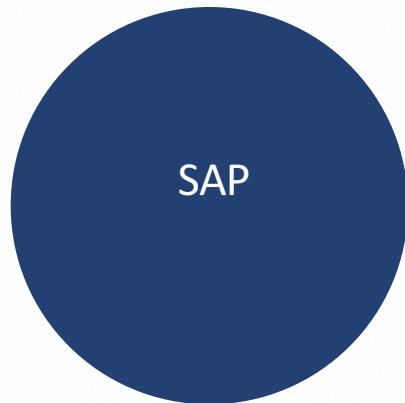
Cloud



Enterprise Class



Connectivity



SAP: BW Filtering and dynamic extraction

- Filter DataSource Objects and Infocube
- They also support dynamic schemas

Table configuration

Table	"0FIGL_C10"
Max row number	context.sap_bw_maxrow

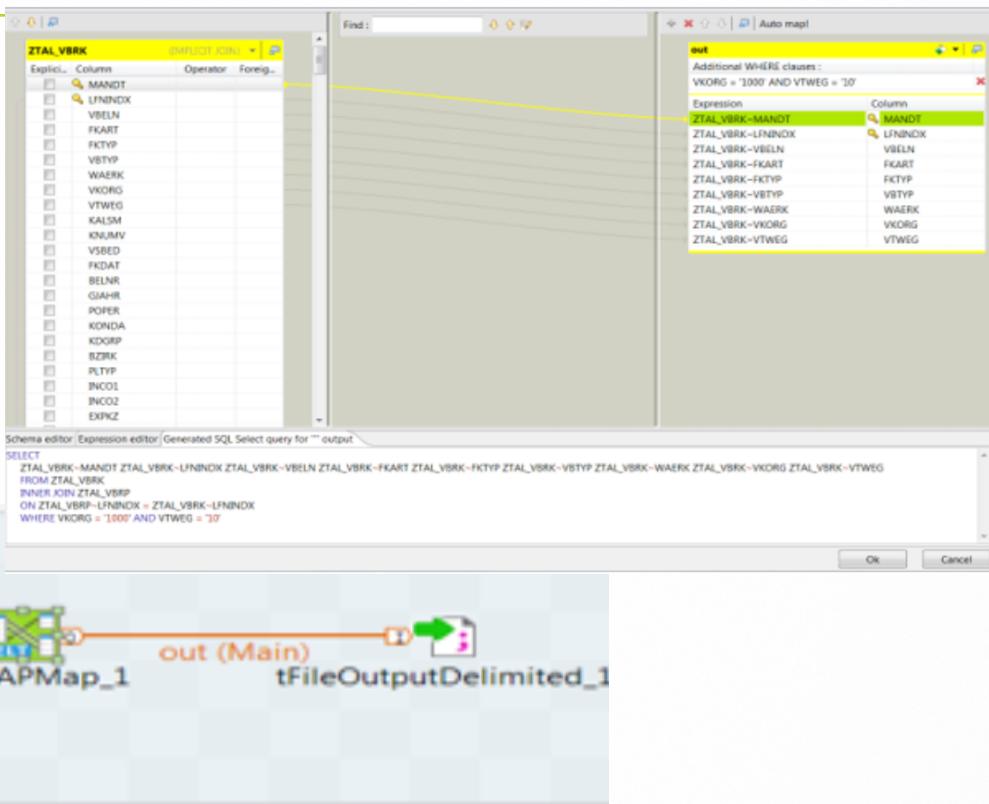
Filter Group

Column	Sign	Operator	Value(Or low value for "Be...")	High value(only necessary)
OPROFIT_CTR	Exclude	Not between	"0000001010"	"0000001060"



SAP: Bulk Extraction and ELT/Push-down

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

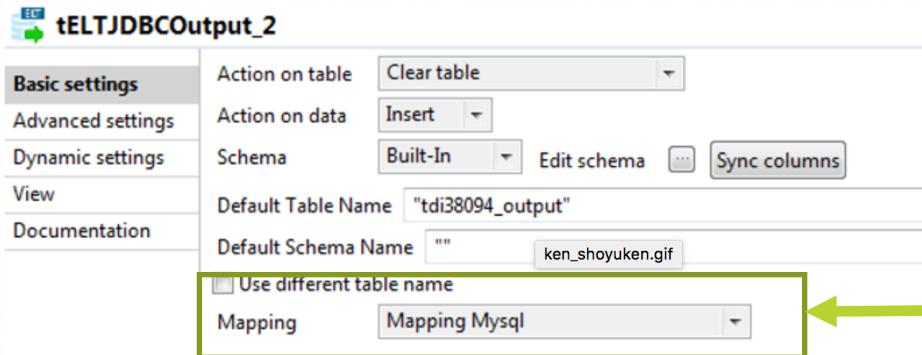


SAP: Misc

- RFC Server can use a Standalone ActiveMQ (High-Availability, Monitoring)
- BW 7.5 support

Industrialization: JDBC

- New Action on Table for ELT JDBC connectors



tELTJDBCOutput_2

Basic settings	Action on table	Clear table
Advanced settings	Action on data	Insert
Dynamic settings	Schema	Built-In
View	Edit schema ... Sync columns	
Documentation	Default Table Name "tdi38094_output"	
	Default Schema Name "" ken_shoyukan.gif	
<input type="checkbox"/> Use different table name		
Mapping		Mapping Mysql

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)



Metadata Management

Notification

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

The screenshot shows two panels. On the left, the 'Administration' panel has tabs for Administration, Groups, Statistics, Log, and Schedules. It includes fields for Enable SMTP (checked), Host Name, Port, User Name, Password, and Sender Address. Buttons at the bottom are Send Test Email, Save, and Cancel. A large grey arrow points from this panel to the right panel. On the right, the 'Import bridge for this Model' panel shows a 'Description' field, a 'Stewards' field containing 'John' and 'Business Analysts' (both highlighted with a red box), and an 'Import Server' dropdown set to 'Default Server' (highlighted with a red box) with the status 'Is available'. Under 'Additional Model options', there is a 'Lightweight Version' checkbox and a 'Send Import Notification' checkbox (highlighted with a red box). The 'Model Type' is listed as 'Microsoft SQL Server Database SQL DDL - Beta Bridge'.

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

Compare Target Staging DW PDM > v5 with Source Staging DW PDM > Original

Show Match

Changed objects	Source (Original)	Target (v5)
Database	FinanceDWStaging.sql	FinanceDWStaging.sql
Schemas		
Schema	dbo	dbo
Tables		
Table	Customer Payment	Customer Payment
Columns		
Column	Payment Amount	Payment Amount
Table	CustomerPaymentAssignment	CustomerPaymentAssignment
Columns		

Changed properties

Property name	Source value	Target value
Length	8	10

New Rest API

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

 talend

username password [Login](#)

Metadata Management Rest API

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

Authentication

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

Browse

[POST](#) /entities Get the entities that satisfy certain criteria, such as entity types and model MIR object identifiers.

[GET](#) /entities/{objectId} Get the details of a particular entity given the MIR object identifier of the entity.

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

Glossary

[GET](#) /entities/glossary/{objectId} Get the details of a particular glossary term given the MIR object identifier of the term.

[GET](#) /entities/reports/{termId} Get all BI Reports that are related to a particular term given the MIR object identifier of the term.

[POST](#) /operations/glossary/exportAndDownloadCSV Download all glossary terms that belong to a category or a list of categories to a CSV file

[GET](#) /search/glossary Search for glossary terms.

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

Lineage

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

Profiles

[Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

Repository

[PUT](#) /operations/abortOperation/{operationId} Abort operation execution.

[GET](#) /operations/getOperationStatus/{operationId} Get operation status.

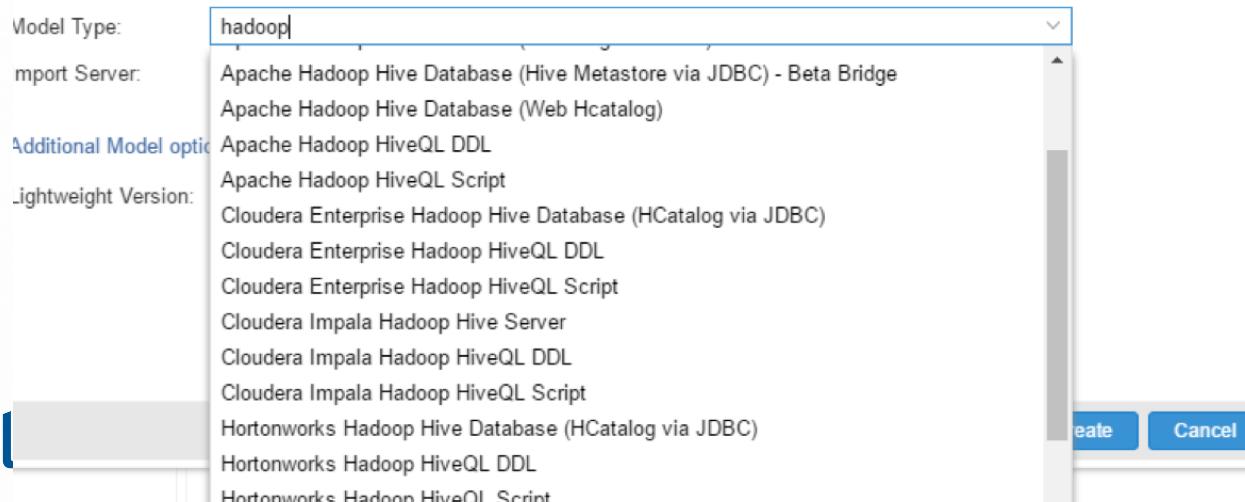
[GET](#) /operations/listRunningOperations List currently running operations.

[GET](#) /operations/repositoryBrowse Browse the metadata repository given a particular root path.

[POST](#) /operations/repositoryExport/attachFile Add an attachment to a model version.

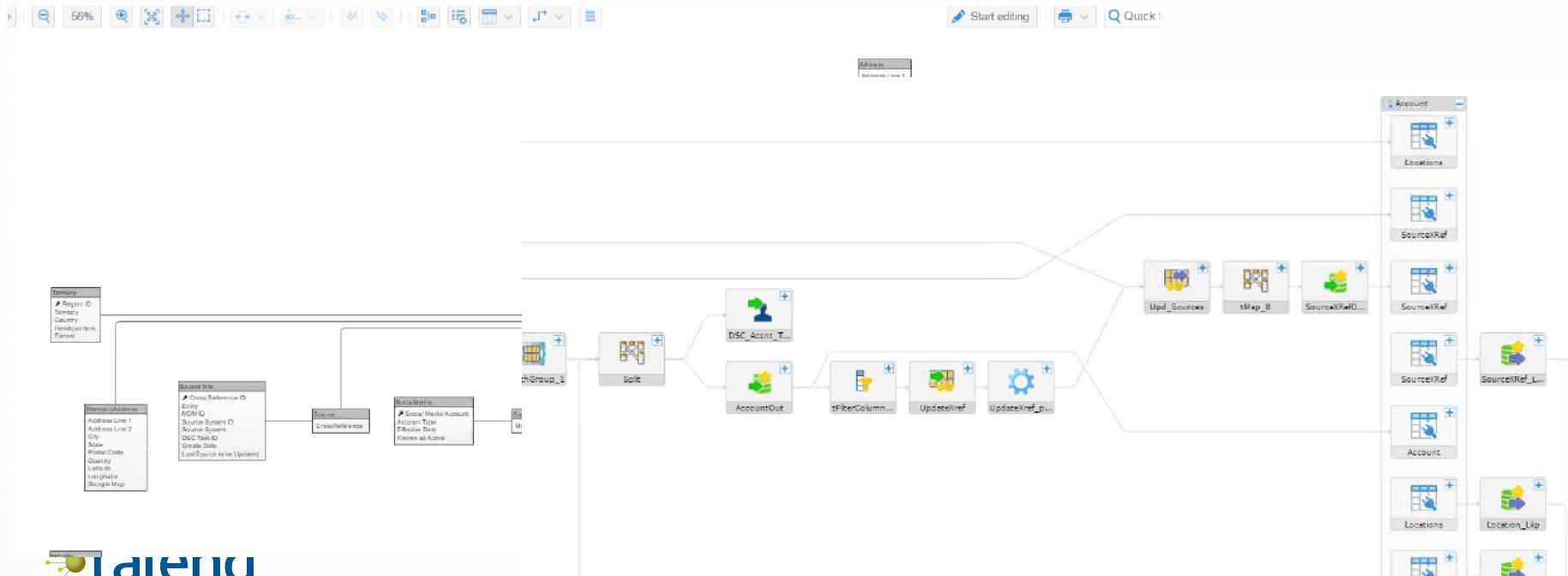
Big Data Inventory

- 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 Preparation	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

The screenshot shows the Talend Data Stewardship interface. On the left, there is a sidebar with menu items: MY CAMPAIGNS, MY DATA MODELS, and MY SEMANTIC TYPES. The MY SEMANTIC TYPES item is highlighted with a blue box and has a green callout bubble pointing to it labeled "New menu entry in TDS and TDP". The main content area displays a list of semantic types. Each entry includes a thumbnail, the semantic type name, a description, creation and modification details, and a status. An "Edit semantic type" button is located at the bottom of the list, with a green callout bubble pointing to it. A "Remove semantic type" link is visible on the right side of the list. A green callout bubble points to the top of the list with the text "Add new semantic type". Another green callout bubble points to the top of the list with the text "List all categories (draft and published)". A green callout bubble at the bottom right points to the "Delete semantic type" link.

Category	Description	Created On	Modified On	Status
Airport	• DESCRIPTION: Airport name • TYPE: Dictionary	• CREATED ON: 2016-12-08 15:38:49 • CREATED BY: Talend	• MODIFIED ON: N/A • MODIFIED BY: N/A	• PUBLISHED ON: 2016-12-08 15:38:49 • STATUS: Published
Airport Code	• DESCRIPTION: Airport name • TYPE: Dictionary	• CREATED ON: 2016-12-08 15:38:49 • CREATED BY: Talend	• MODIFIED ON: N/A • MODIFIED BY: N/A	• PUBLISHED ON: 2016-12-08 15:38:49 • STATUS: Published
Amex Card	• DESCRIPTION: American Express card • TYPE: Regular expression	• CREATED ON: 2016-12-08 15:38:49 • CREATED BY: Talend	• MODIFIED ON: N/A • MODIFIED BY: N/A	• PUBLISHED ON: 2016-12-08 15:38:49 • STATUS: Published
Animal	• DESCRIPTION: Animal (multilingual) • TYPE: Dictionary	• CREATED ON: 2016-12-08 15:38:49 • CREATED BY: Talend	• MODIFIED ON: N/A • MODIFIED BY: N/A	• PUBLISHED ON: 2016-12-08 15:38:49 • STATUS: Published
Answer	• DESCRIPTION: Yes/No (in EN, FR, DE and ES) • TYPE: Dictionary	• CREATED ON: 2016-12-08 15:38:49 • CREATED BY: Talend	• MODIFIED ON: N/A • MODIFIED BY: N/A	• PUBLISHED ON: 2016-12-08 15:38:49 • STATUS: Published

Regular expressions

The screenshot shows a configuration dialog for a regular expression. The dialog is divided into two main sections: **GENERAL** and **DEFINITION**.

GENERAL section:

- Name*: Amex Card
- Description: American Express card
- Type*: Regular expression
- Use for validation

DEFINITION section:

- Content: Numeric
- Validation pattern*: `\d{4}[\d]{0-9}\d{13}`

Annotations with callouts:

- A green speech bubble points to the "Name*" field in the General section with the text: "If disabled, will be available in TDP for discovery but not in TDS".
- A green speech bubble points to the "Content" field in the Definition section with the text: "Is validated when form is submitted".
- A green speech bubble points to the "Validation pattern*" field in the Definition section with the text: "Pre validation".

Dictionaries

The screenshot shows a Talend dictionary creation interface. The 'GENERAL' section includes fields for Name* (containing 'Airport'), Description (containing 'Airport name'), and Type* (containing 'Dictionary'). A 'Use for validation' checkbox is also present. The 'DEFINITION' section lists values separated by commas. A search bar and a '+' button are available for adding more values. A file upload input field is also shown. At the bottom are 'CANCEL', 'SAVE AS DRAFT', and 'SAVE AND PUBLISH' buttons.

GENERAL

Name*
Airport

Description
Airport name

Type*
Dictionary

Use for validation

DEFINITION

Values*

- Arrabury,Arrabury Airport
- El Arish,El Arish International,El Arish International Airport
- Rabah Bitat Airport,Rabah Bitat
- Anaa,Anaa Airport
- Apalachicola Regional,Apalachicola Regional Airport,Apalachicola
- Arapoti Airport,Arapoti
- Merzbrück,Merzbrück Airport
- Arraias Airport,Arraias
- Cayana Airstrip
- Aranuka Airport,Aranuka
- Aalborg,Aalborg Airport
- Mala Mala Airport,Mala Mala
- Anarua,Anarua Airport

Open / closed dictionary

Search a value

List of values:

- Lazy loaded from server when scrolling
- Synonyms separated by coma character

Add a single value

Upload a file (override all values)

Validation – Simplified text (6.3 behavior)

The screenshot shows the Talend Data Preparation interface. On the left, there's a preview of a table titled "Animals" with columns ID, ANIMAL, and animal. The table contains 10 rows, with the first row having a value "NotAnAnimal" in the ANIMAL column. On the right, a configuration panel for a "PREPARATIONS" step is open. The "GENERAL" tab is selected, showing fields for Name* (Animal), Description (Animal (multilingual)), and Type* (Dictionary). A section for "Validation criterion*" has three options: "Simplified text (most permissive)" (selected with a checked checkbox), "Ignore case and accents", and "Exact value (most restrictive)". Below this, a "Values*" section lists various animal names in multiple languages. The Talend logo is visible at the bottom left.

talend DATA PREPARATION

Animals

Filters

Add a filter ...

ID	ANIMAL	animal
1	NotAnAnimal	
2	CAT	
3	cAT	
4	cAt	
5	Chipottere	
6	Chipotère	
7	Chauve souris.	
8	Chauve-souris	
9	Bat	
10	Bat	

PREPARATIONS

DATASETS

SEMANTIC TYPES

GENERAL

Name*
Animal

Description
Animal (multilingual)

Type*
Dictionary

Use for validation

Validation criterion*

Simplified text (most permissive)
Ignore case and accents
Exact value (most restrictive)

Values*

OVEJAS,MOUTON,SHEEP,SCHAF,E,OVINOS

PANDA

ARAIÑE,ARAÑA,SPIDER

VIPÈRE,VIPER,Vibora

BALLENA,WHALE,WAL,BALEIA,BALEINE

WOLF,LOUP

TURTLE,TORTUGA,TURTUE,TARTARUGA

TIGRESSE,TIGRESS

ZEBRA,ZÈBRE

TIGRE,TIGER

Chauve-souris,Bat,Chipottere

Validation – Case and accent insensitive

The screenshot shows the Talend Data Preparation interface. On the left, there's a preview of a table titled 'Animals' with columns 'ID' (integer), 'ANIMAL' (animal), and 'NAME' (string). The data includes rows for various animals like CAT, cát, Chipotere, etc. On the right, the 'GENERAL' tab of a preparation configuration window is open. It shows the preparation name is 'Animal' and its type is 'Dictionary'. A dropdown menu under 'Type' has 'Ignore case and accents' selected. The 'Values' section lists animal names in both English and Spanish, demonstrating how the system handles case and accent insensitivity.

ID	ANIMAL	NAME
1	NotAnAnimal	
2	CAT	CAT
3	cát	cát
4	cât	cât
5	Chipotere	Chipotere
6	Chipotère	Chipotère
7	Chauve souris.	Chauve souris.
8	Chauve-souris	Chauve-souris
9	Bat	Bat
10	Bat	Bat

GENERAL

Name*: Animal

Description: Animal (multilingual)

Type*: Dictionary

Use for validation

Simplified text (most permissive)
 Ignore case and accents
 Exact value (most restrictive)

DEFINITION

Values*

OVEJAS,MOUTON,SHEEP,SCHAF,OVINOS
PANDA
ARAIÑE,ARAÑA,SPIDER
VIPER,VIPER,Vibora
BALLENA,WHALE,WAL,BALEIA,BALEINE
WOLF,LOUP
TURTLE,TORTUGA,TORTUE,TARTARUGA
TIGRESSE,TIGRESS
ZEBRA,ZÉBRE
TIGRE,TIGER
Chauve-souris,Bat,Chipotère

Validation – Exact value

The screenshot shows the Talend Data Preparation interface. On the left, there's a preview of a dataset named "Animals" with columns "ID" (integer) and "ANIMAL" (animal). The data includes rows from 1 to 10, with row 1 labeled "NotAnAnimal". On the right, a detailed configuration panel is open for the "GENERAL" tab of a preparation step. Under "Type", "Dictionary" is selected. In the "Validation" section, the "Exact value (most restrictive)" option is checked. Below this, a list of values is provided: OVEJAS,MOUTON,SHEEP,SCHAF,OVINOS; PANDA; ARAÑE,ARAÑA,SPIDER; VIPER,VIPER,Víbora; BALLENA,WHALE,WAL,BALEIA,BALEINE; WOLF,LOUP; TURTLE,TORTUGA,TORTUE,TARTARUGA; TIGRESSE,TIGRESS; ZEBRA,ZÉBRE; TIGRE,TIGER; Chauve-souris,Bat,Chipotère.

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

GENERAL

Name*
Phone number|

Description
Phone number (DE, FR, UK, US)

Type*
Compound type

Use for validation

DEFINITION

Children types

US PHONE X UK PHONE X FR PHONE X DE PHONE X

CANCEL **SAVE AS DRAFT** **SAVE AND PUBLISH**

Compound Semantic Types

The screenshot shows a Talend Data Preparation interface. On the left, there is a table labeled "Input data" containing two columns: "PHONE_NUMBER" and "COUNTRY". The data includes rows for France, Germany, and the United States. A green arrow points from this table up to the main preparation area. In the center, there is a table titled "DATA PREPARATION" with columns "ID", "PHONE_NUMBER", and "COUNTRY". A context menu is open over the "PHONE_NUMBER" column, listing options like "This column is a phone", "Phone number 100 %", "Rename column", "Delete column", "Duplicate column", "Create new column", "Set as TEXT", "Set as BOOLEAN", "Set as DATE", "Set as UTC_DATETIME" (which is selected), "Set as INTEGER", and "Set as DECIMAL". A green arrow points from this menu up to the validation step.

PHONE_NUMBER	≡	COUNTRY	≡
phone		country	
0145689856	France		
02045689856	Germany		
02045689856	Germany		
02045689856	Germany		
15207777777	United States		
0145689856	France		

Input data

Discovery step

The screenshot shows a Talend Data Preparation interface. On the left, there is a table titled "Telephones Preparation" with a "Filters" section. The table has columns "ID", "PHONE_NUMBER", and "COUNTRY". The data is identical to the input table. A green arrow points from this table up to the validation step.

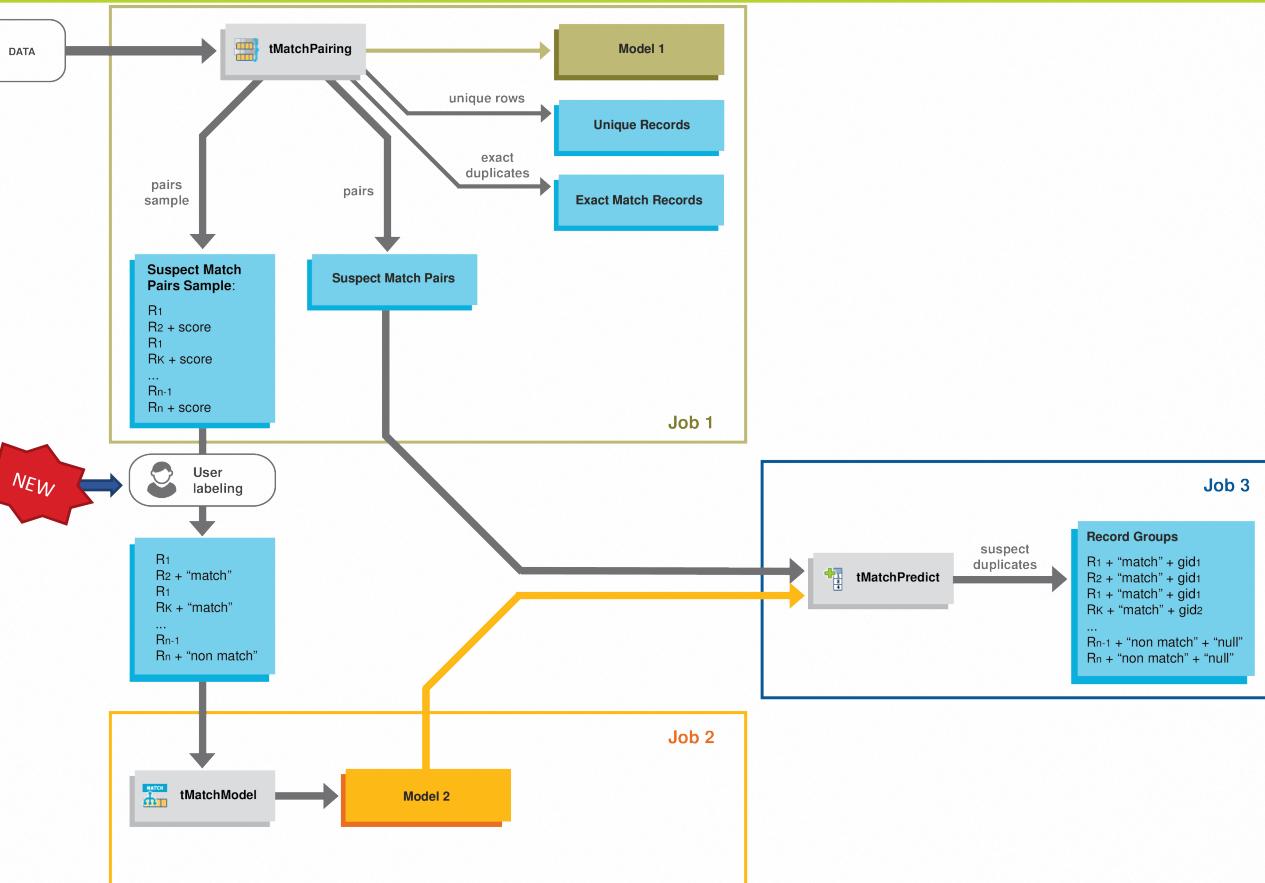
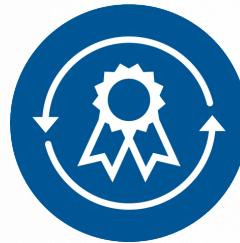
ID	PHONE_NUMBER	COUNTRY
1	0145689856	France
2	02045689856	Germany
3	02045689856	Germany
4	02045689856	Germany
5	15207777777	United States
6	0145689856	France
7	02045689856	Germany
8	02045689856	Germany
9	0145689856	France
10	02045689856	Germany

Validation step



Matching with ML on Spark

Matching with Machine Learning



Stewardship grouping campaign

1/ General

NAME:
childCare_hellweek64

DESCRIPTION: (optional)

TYPE:
GROUPING

QUESTION:
do the records match?

ANSWERS:
yes

no

Prepare the model

talend DATA STEWARDSHIP My tasks > ChildCare_hellweek64 > New > Assigned to me >

Owner: jj@talend.com Type: GROUPING Status: STARTED

Do the records match?

YES NO

Task #1 2/2

	ARBITRATION	SOURCE	SITE_NAME	ADDRESS	ZIP	PHONE	FAX	PROGRAM_NAME	LENGTH_OF_DAY	PRIORITY*	DUUE DATE
1										Medium	
Record 1		CPS_Early_Childhood_Porta Legacy Charter School (at Mason	4217 W. 18th St.		5421640			Community Partnersh	8-11 Hours, varies		
Record 2		purple_binder_early_child YMCA Bowen School	2710 E 89th Street	60617	9330166						
2										Medium	
Record 1		CPS_Early_Childhood_Porta Firman Community Services - Firm	37 W 47th St		3733400			Child Care	EXTENDED DAY		
Record 2		CPS_Early_Childhood_Porta Firman Community Services - Firm	37 W 47th St		3733400			State Pre-Kindergar	HALF DAY/FULL DAY		

Label pairs of records

Connect components to Data Stewardship

The screenshot shows the Talend Data Stewardship interface with two open jobs:

- Job Job1_childCarePairing 0.1**: This job contains a **tMatchPairing_1** component. The component's configuration panel is visible, showing settings for storage, schema, and matching model location. It also includes suffix array blocking parameters and pairing model location.
- Job Job2_childCareModel 0.1**: This job contains a **tMatchModel_1** component. The component's configuration panel is visible, showing settings for storage, schema, and matching key.

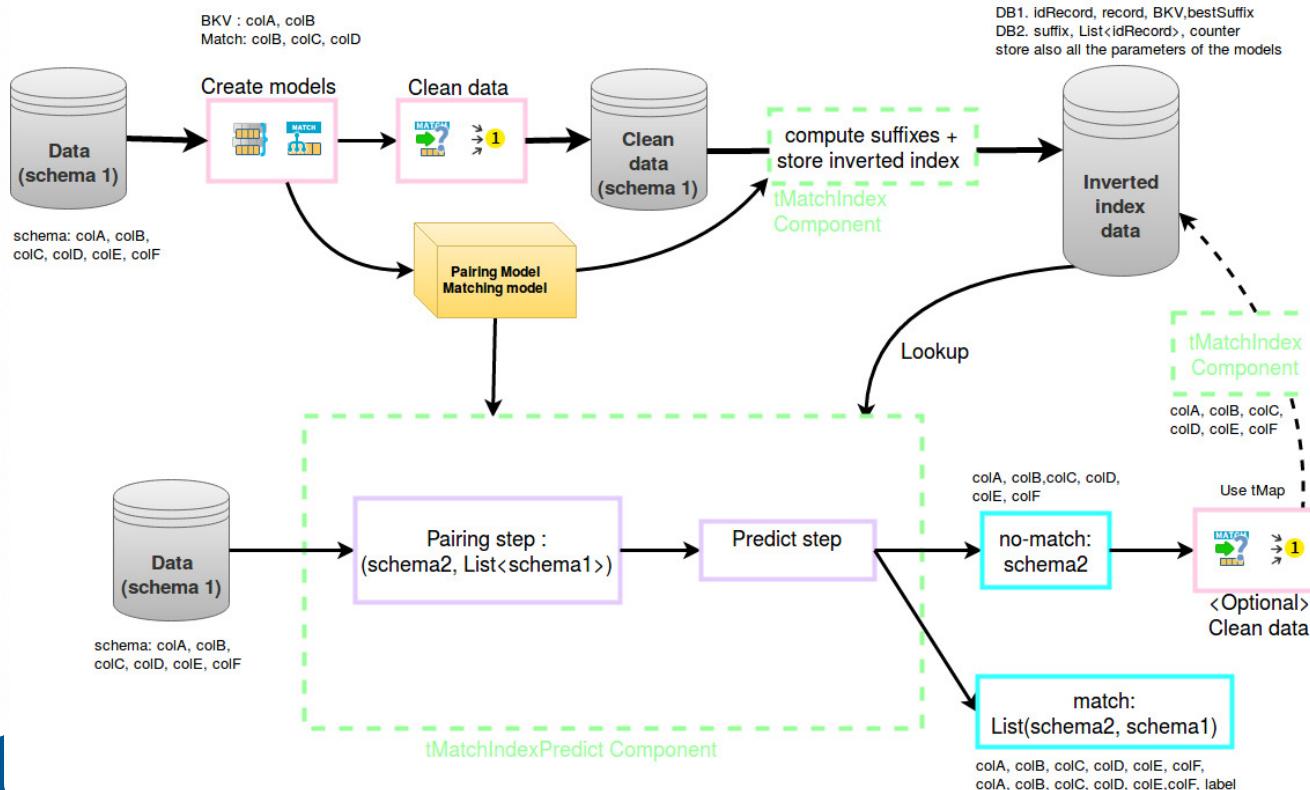
Both components are integrated with Data Stewardship, as indicated by the checked "Integration with Data Stewardship" checkbox in their respective configuration panels.

Read from Data Stewardship and learn ML model

Write into Data Stewardship

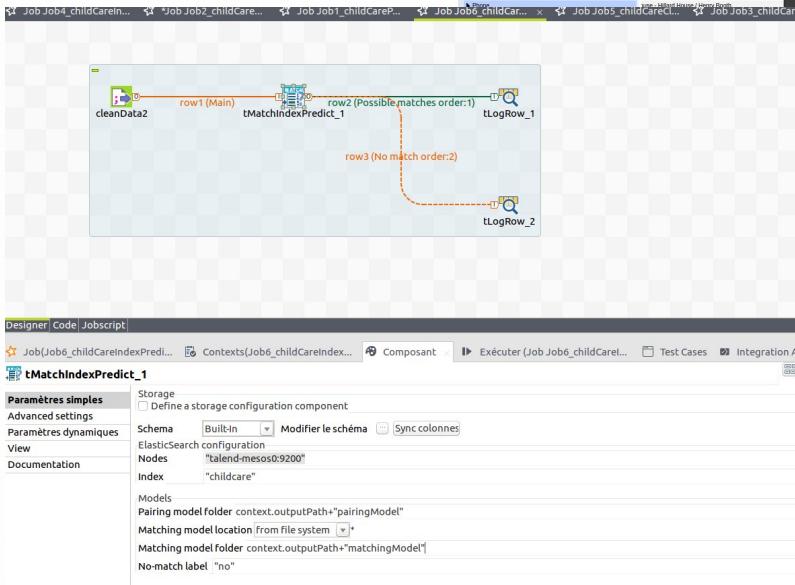
Continuous matching

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

*Job tRuleSurvived_conflict_scenario1 0.1

The screenshot shows the Talend Data Integration Designer interface. At the top, there's a toolbar with tabs like Designer, Code, and Jobscrip. Below the toolbar, the main workspace displays a job flow with components: tFixedFlowInput_1, tRuleSurvivorship_1, and tLogRow_1. A context menu is open over the tRuleSurvivorship_1 component. The context menu has several options: 'Edit schema', 'Sync columns', 'Group identifier', 'Rule package name' (set to 'original_scenario1'), 'Generate rules and survivorship flow' (with a warning icon), 'Define conflict rule' (checked), 'Rule table', 'Conflict rule table', 'Advanced settings', 'Dynamic settings', 'View', 'Documentation', and 'Validation Rules'. The 'Conflict rule table' section is highlighted with a red border. It contains a table with four rows:

Rule name	Reference column	Function	Value	Target column	Fill empty by	Ignore blanks	Remove duplicate
"CR1"	contract_day	Most recent		Firname	""	<input type="checkbox"/>	<input type="checkbox"/>
"CR2"	Firname	MappingTo		address	""	<input type="checkbox"/>	<input type="checkbox"/>
"CR4"	Lastname	Match regex	"[A-Z]{1}[w]{1-9}"	Lastname	""	<input type="checkbox"/>	<input type="checkbox"/>
"CR3"	Lastname	Match regex	"\\w{1}"	Lastname	""	<input type="checkbox"/>	<input type="checkbox"/>



Natural Language Processing

Natural Language Processing

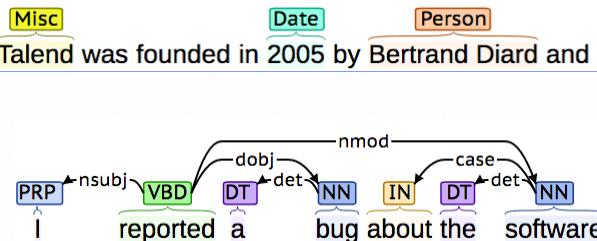
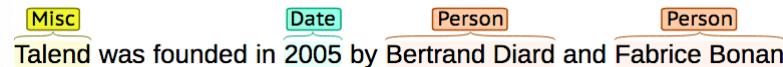
- What is natural language processing?

- 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>

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?

2013-05-21 18:41:32
asplegel
Talend Team
+talend
Group: Talend Team
Registered: 2010-12-02
Email
Offline

When doing column analysis in DQ, you can move columns and individual analyses up and down in the list. However, that does not seem to affect the actual order of them in the output. Is it possible or intended to change the output order of individual analyses on individual columns? If so, how?

2013-05-21 18:48:50
knarayanan
Talend Team
+talend
Group: Talend Team
From: Talend
Registered: 2010-09-03
Email
Offline

try removing the column and then adding it back in the proper order...

2013-05-22 07:59:11
scorrela
Talend Team
+talend
Group: Talend Team
From: Talend SA (USA)
Registered: 2013-06-03
Email Website
Online

Hi Andre
I would suggest you to raise an issue in Jira too.
Thanks

Thank you for your support,
Sebastiao Correia

2013-05-22 15:56:13
asplegel
Talend Team
+talend

@Sebastiao Thanks, I'll try to nail it down to a simple case and will file a report.

Set this topic as resolved | Report as a spam | Q

Report as a spam | Q

Report as a spam | Delete | Edit | Q

Report as a spam | Q

Extract names and links between names. Use profile information

Classify category of discussion: here it's about column analysis

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?
 - ...

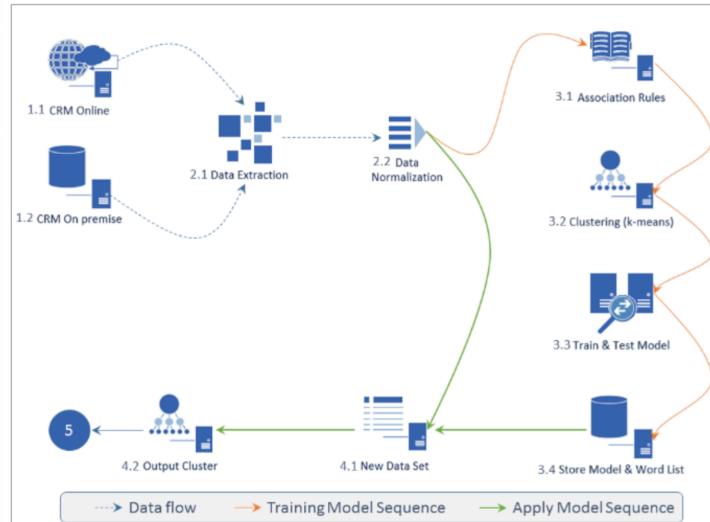


Figure 2: CRM Notes Mining & Machine Learning workflow

<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?

Prepare text sample

- Tokenize and remove html tags
- Manual annotations of Named Entities

Learn a model

- Design the features
- Estimate the model

Use the model

- Apply on full text

Natural Language Processing

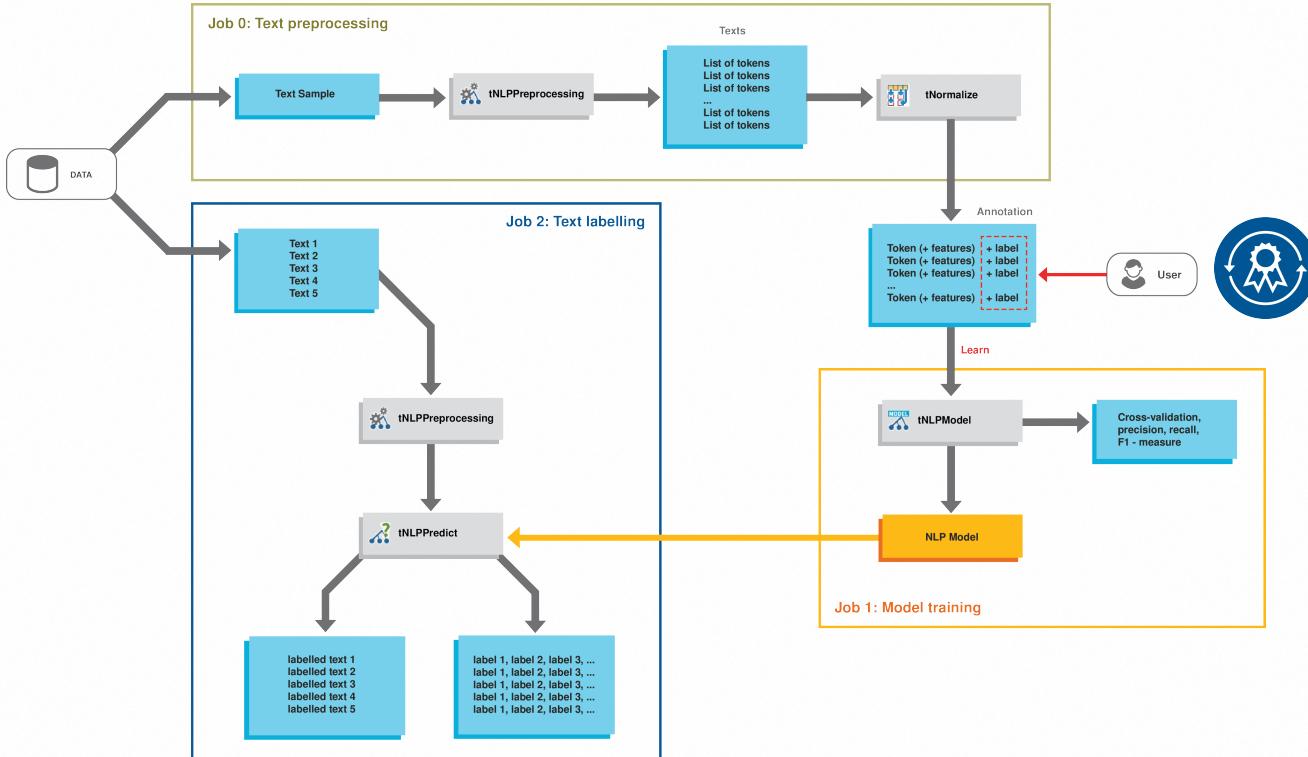
tCompareColumns

tNLPModel

tNLPPredict

tNLPPreprocessing

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.  
using    0.46   F  
Postgres  0.64   F  
with     0.6 F   F  
Talend   0.56   F  
? 0.0 F   F 0
```

```
using    0.46   F   F   0
Postgres 0.64   F   F   0
with     0.6   F   F   0
Talend   0.56   F   F   0
? 0.0 F   F   0
Best    0.78   F   F   0
regards 0.68   F   T   0
,      0.0 F   F   0
Tim    0.83   F   F   PER

Hi   0.56   F   F   0
,      0.0 F   F   0
ppm 0.5 F   F   0
install 0.44   F   F   0
Text-CSV_XS 0.51   F   F   0
does   0.61   F   F   0
```

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

Extract named entities with <PER> labels

NLP libraries

	Stanford	ScalaNLP
Web site	https://stanfordnlp.github.io/CoreNLP/	http://www.scala-nlp.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

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

The screenshot shows the Talend Data Integration interface. On the left, a configuration file 'qaworld.ini' is displayed with the following content:

```
[FRA]
CountryBase=FRA
CleaningAction=Address

FRAAddressLineCount=6
FRAAddressLine1=W60,011
FRAAddressLine2=W60,021
FRAAddressLine3=W60,L41
FRAAddressLine4=W40,P11,S11
FRAAddressLine5=W40,B10,L31
FRAAddressLine6=W40,C11,L21

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

In the center, a job flow is shown with components: tFixedFlowInput_1, tQASBatchAddressRow_2, and tLogRow_1. The tQASBatchAddressRow_2 component is highlighted.

A modal dialog titled "Schema of tQASBatchAddressRow_2" is open, comparing the "tFixedFlowInput_1 (Input - Main)" schema with the "tQASBatchAddressRow_2 (Output)" schema. Both schemas have columns: Address Identifier, STATUS, ADDRESS, ZIP_CODE, CITY. The "Key" column is checked for all columns in both schemas. The "Type" column shows String for all columns. The "Nullab" column has checked checkboxes for Address Identifier, STATUS, ADDRESS, ZIP_CODE, and CITY. The "Date Pattern" column is empty for all columns.

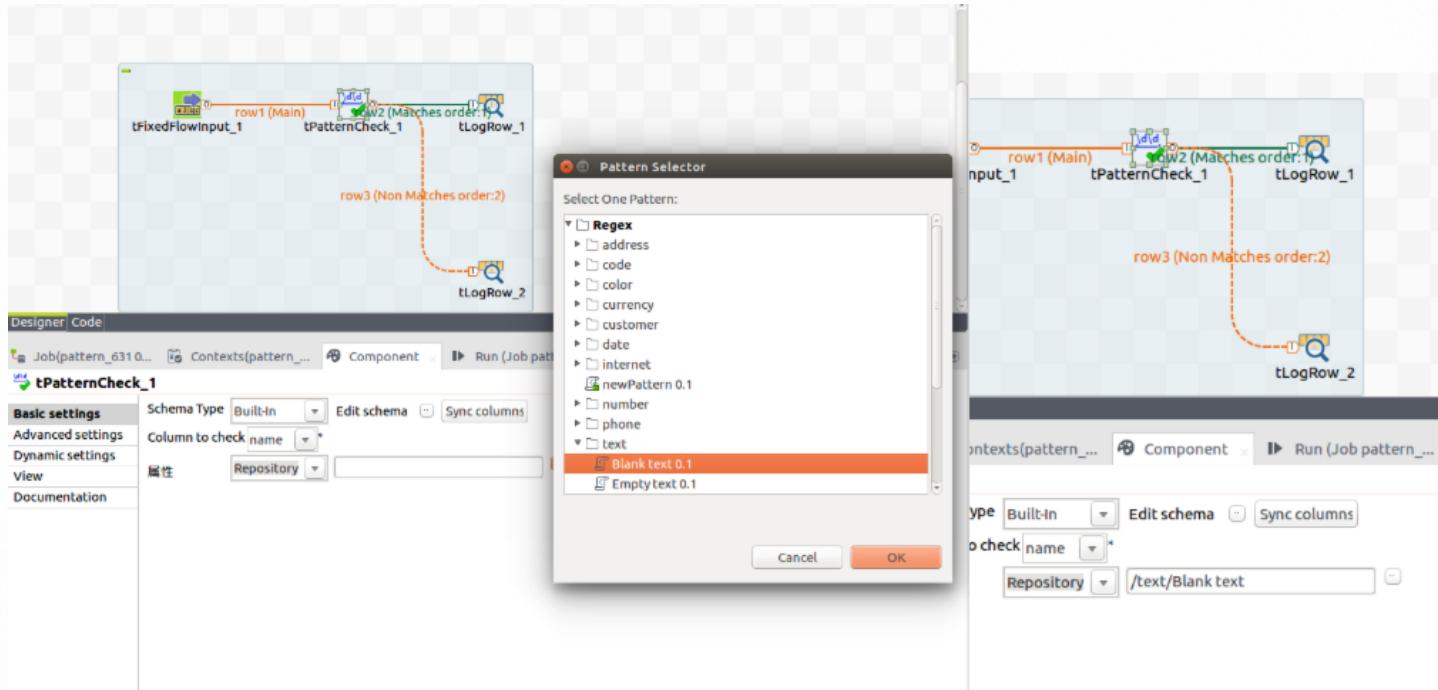
Column	Key	Type	Nullab	Date Pattern
Address Identifier	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>	
	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>	

Column	Key	Type	Nullab	Date Pattern
Address Identifier	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>	
STATUS	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>	
ADDRESS	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>	
ZIP_CODE	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>	
CITY	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>	

At the bottom of the dialog, there are "OK" and "Cancel" buttons.

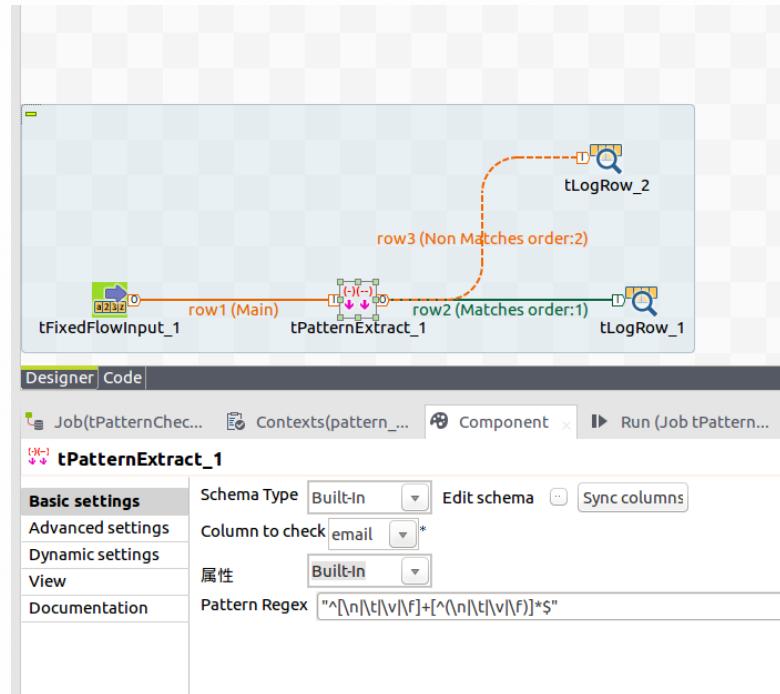
Pattern components – Repository mode

- Get patterns from repository



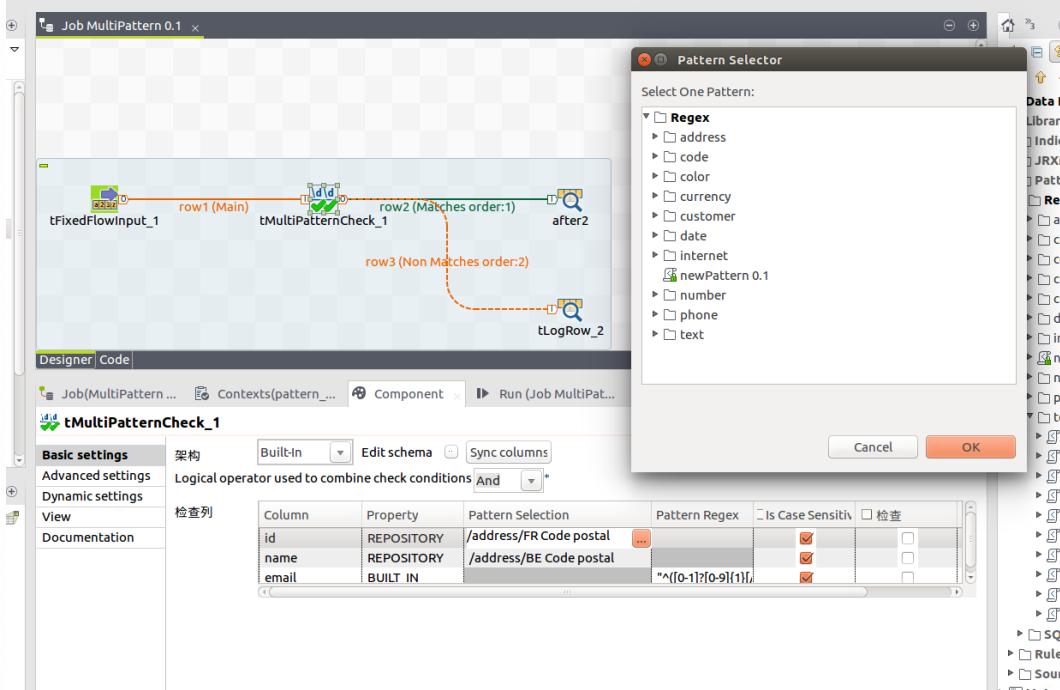
Pattern components – Built-in mode

- 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

The screenshot shows the Talend Data Profiling interface. On the left, there's a 'Data preview' section with a table showing three rows of data: 1, 2, and 3. Below it is an 'Analyzed Columns' tree view with 'COLUMN1 (VARCHAR)' selected, showing various statistics like Row Count, Null Count, Unique Count, etc. In the center, there's a 'Indicator settings' dialog with tabs for 'Text Parameters' and 'Blank Options'. A 'Limit result' section is also present. On the right, there's a 'Preferences' window titled 'Indicator settings' with sections for 'Frequency table' and 'Low frequency table', both set to show 20 results. A modal dialog titled 'Set the Frequency Table Parameters' is open, listing 'Analyses(2)' and 'pattern_frequencies 0.1' with checkboxes checked. Buttons for 'Select All' and 'Deselect' are at the bottom of the modal.

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

ADD A SALESFORCE DATASET

Dataset name*
SF_contacts

Username*
smalleit@talend.com

Password*

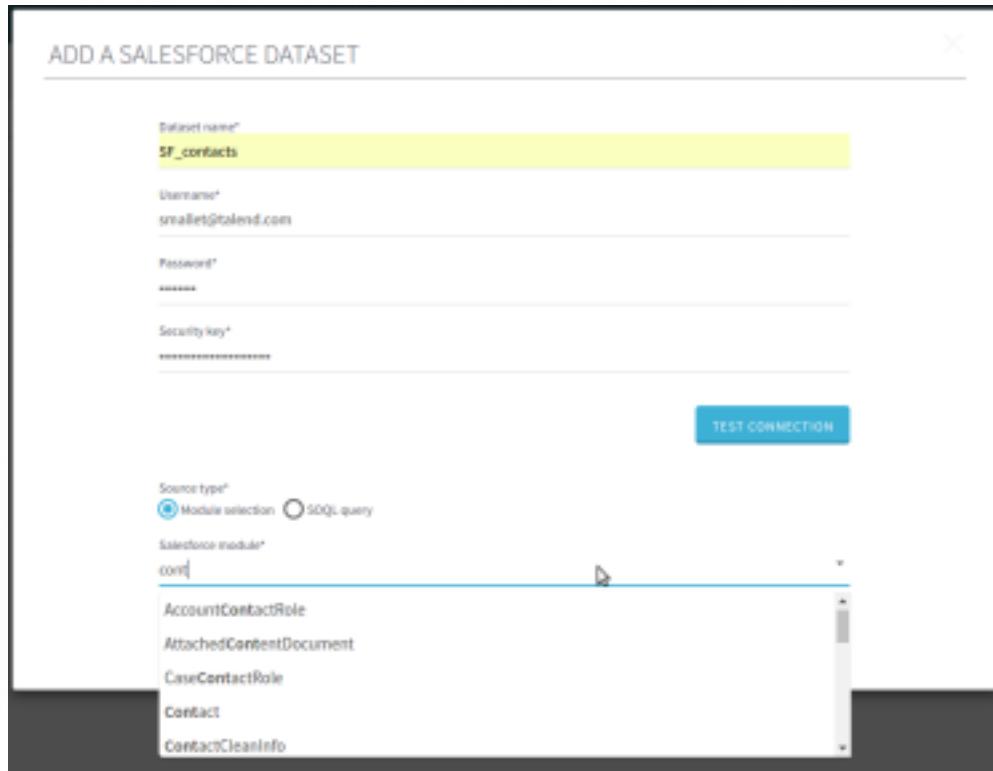
Security key*

Source type*
 Module selection SOQL query

Salesforce module*
com|

AccountContactRole
AttachedContentDocument
CaseContactRole
Contact
ContactCleanInfo

TEST CONNECTION



Amazon S3 Connector

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

ADD A AMAZON S3 DATASET

Dataset name*
S3_customers

Specify aws credentials.

Access key
AKIAI422QOROXBX5EQOA

Secret key
zukSGKxP24nD2AzcgoHTleusj4eiZLBHJhL3wIA

TEST CONNECTION

Region
EU (Ireland)

Bucket
data-prep-francois

Object
us-customers-500.csv

Encrypt data in motion

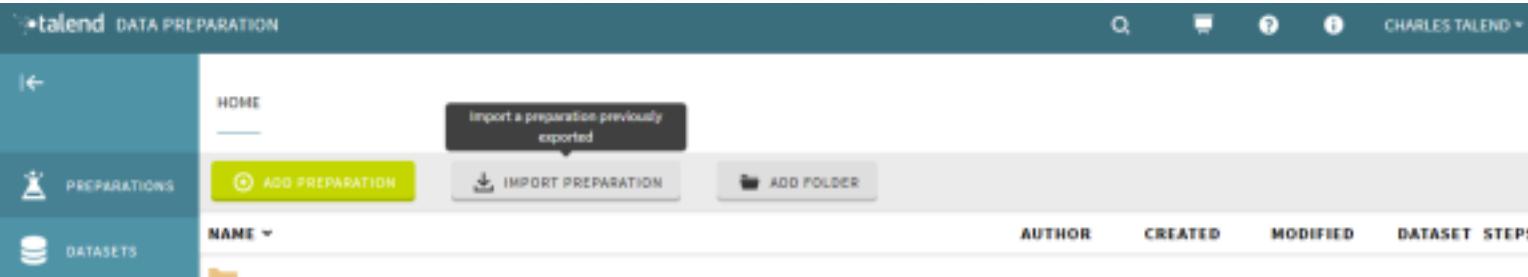
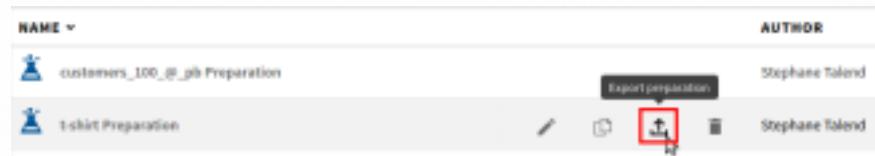
Encrypt data at rest

Format*
CSV

Record Delimiter

Manage promotion across environments

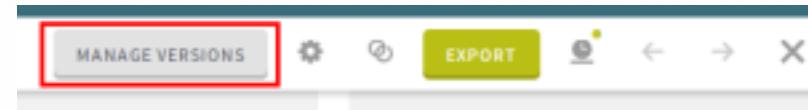
- Based on preparation import/export
- Export/import of a single preparation at a time
- Only preparation is imported, not the dataset
- Dataset with the same name must exists and be accessible for the user
- All versions of a preparation are exported



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

A screenshot of the 'Manage Versions' dialog box. It shows a summary at the top: '10000/10000'. Below is a table with columns 'VERSION' and 'CATEGORY'. The table contains several rows, with the first row being highlighted in yellow. To the right of the table is a 'Versions' section and a 'New Version' form. The 'New Version' form includes fields for 'Description' (set to 'Format dates') and 'Format dates'. At the bottom are 'CANCEL' and 'SUBMIT' buttons.

VERSION	CATEGORY
.9	full_n
.9	White T-Shirt
.9	Black T-Shirt
.9	Black T-Shirt
.7	Black T-Shirt

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)

The screenshot shows the Talend Data Preparation interface with several annotations:

- 1**: A red box highlights the "Version 3" button in the top navigation bar.
- 2**: A red box highlights the "Read-only" button in the top right corner.
- 3**: A red box highlights the "Format dates" note in the "Versions" section.
- 4**: A red box highlights the "SWITCH TO CURRENT STATE" button in the top right corner.

The interface includes:

- Top Bar:** talend DATA PREPARATION, Version 3, Read-only, SWITCH TO CURRENT STATE, EXPORT, X.
- Left Sidebar:** "a nice preparation", Change date format in column BIRTH, Current format: I don't know, best guess, New format: ISO 8601 date.
- Filters:** Add a filter ...
- Data Table:** Shows 10 rows of data with columns: ID, USER_ID, DEPARTMENT, BIRTH, ORDER_ID, ORDER_DATE, TOTAL, NB_TSHIRTS, TSHIRT_PRICE, CATEGORY. The data includes various department names like "Marketing", "Sales", "R&D", etc.
- Versions:** A list of versions:
 - Version 2 (3 step(s) by Shipshape Mallet, 2017-05-12 10:45:39): Delete rows with invalid or empty departments.
 - Version 1 (1 step(s) by Shipshape Mallet, 2017-05-12 10:44:38): Format dates.

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

The screenshot shows the Talend Studio interface with a job named "Job toubidou 0.1". The job consists of three main components: "t-shirts", "tDataprepRun_1", and "tLogRow_1". The "tDataprepRun_1" component has performance metrics: "29378 rows in 0,79s" and "37424,2 rows/s" for "row1 (Main)", and "29378 rows in 4,44s" and "6612,2 rows/s" for "row2 (Main)".

The "tDataprepRun_1" component is selected in the center panel, and its configuration dialog is open. The "Paramètres simples" tab is active. The "Version" dropdown shows "1" is selected. A button labeled "Choose a Ver..." is visible.

A modal dialog titled "Set the version" is displayed on the right. It lists three versions:

Version	Author	Creation Date	Number of steps
Current state	Stéphane Mallet	15/05/17 13:26	7
2	Stéphane Mallet	15/05/17 13:25	5
1	Stéphane Mallet	15/05/17 13:25	5

Buttons for "OK" and "Cancel" are at the bottom right of the modal.



SSO integration

Identity management in 6.3

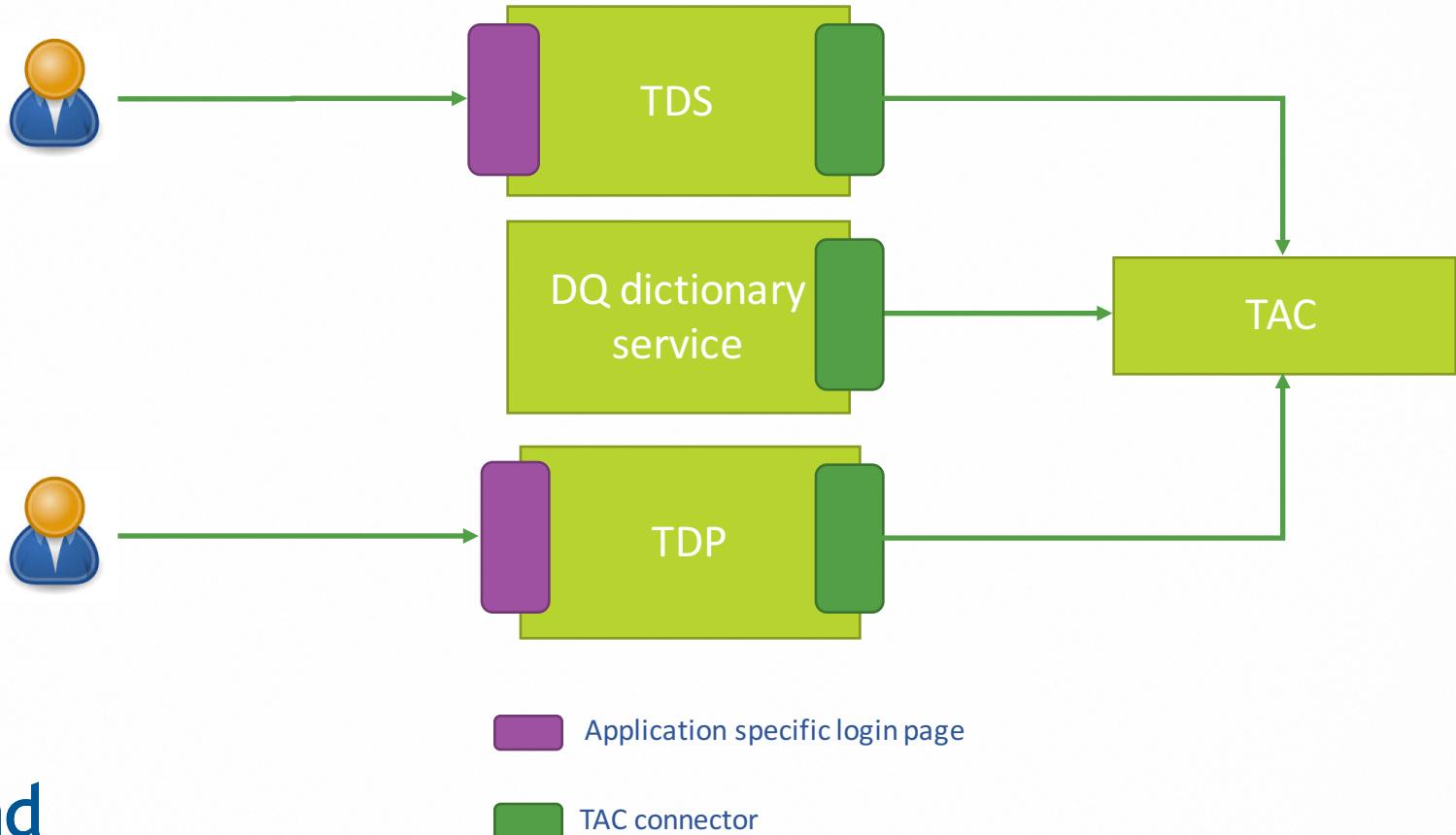
- In 6.3, TDP delegated identity and access management to TAC
- TDP was communicating directly with TAC for
 - Authentication
 - Current user information
 - List all available TDP users
- A TAC admin account was required in TDP configuration

Data	
Login:	<input type="text" value="user1@talend.com"/>
First name:	<input type="text" value="user1"/>
Last name:	<input type="text" value="user1"/>
Password:	<input type="password"/> change password
Svn login:	<input type="text"/>
Svn password:	<input type="password"/> change password
GIT login:	<input type="text"/>
GIT password:	<input type="password"/> change password
Type:	<input type="button" value="No Project Access"/>
Role:	<input type="text"/> 
Data Preparation User:	<input type="checkbox"/>
Data Stewardship User:	<input checked="" type="checkbox"/>
Data Stewardship Role:	<input type="text" value="Data Steward"/> 
Group:	<input type="text"/> 
Active:	<input checked="" type="checkbox"/>

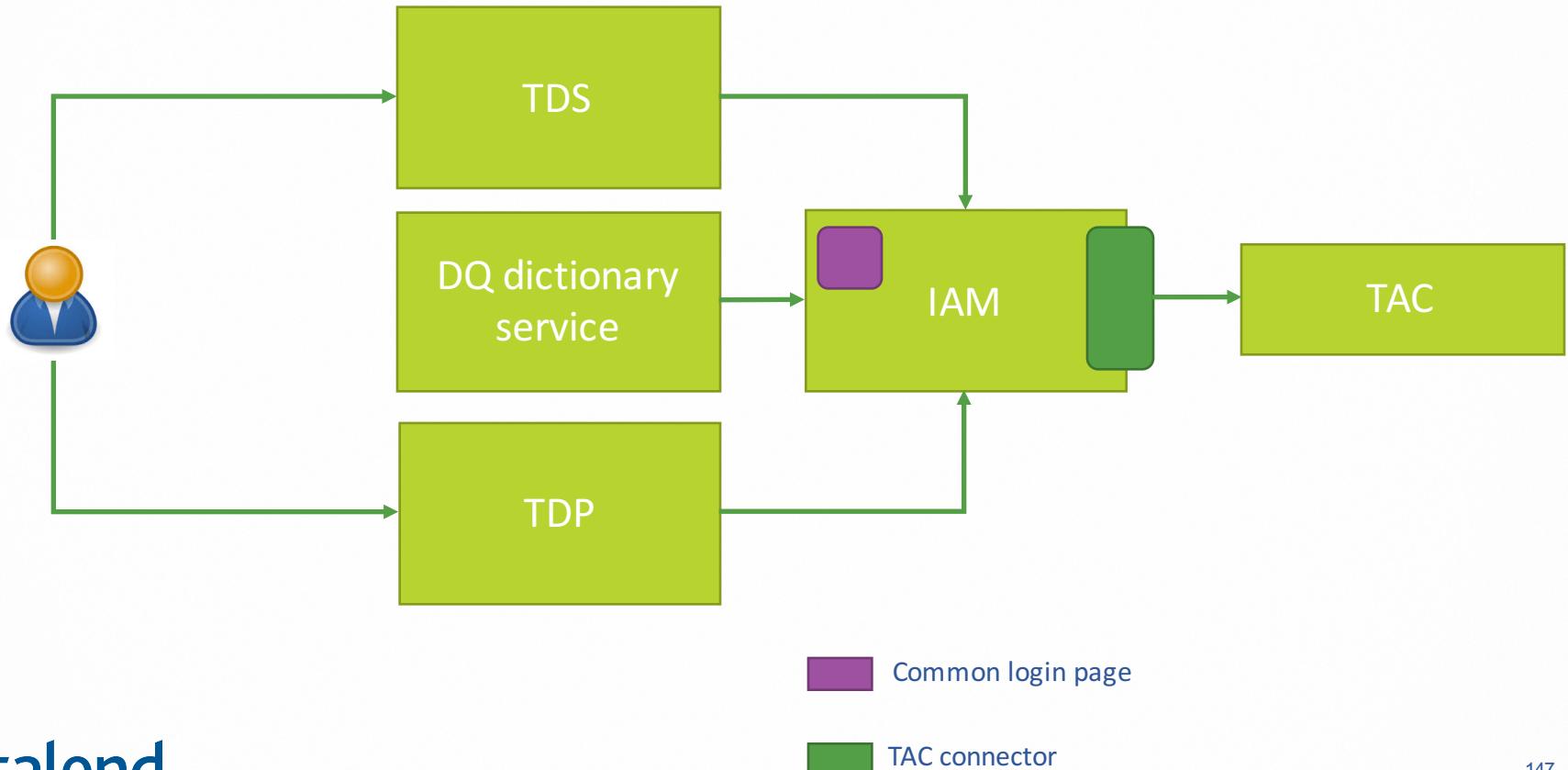
Identity management in 6.4

- 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 **Hoodiiie** 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

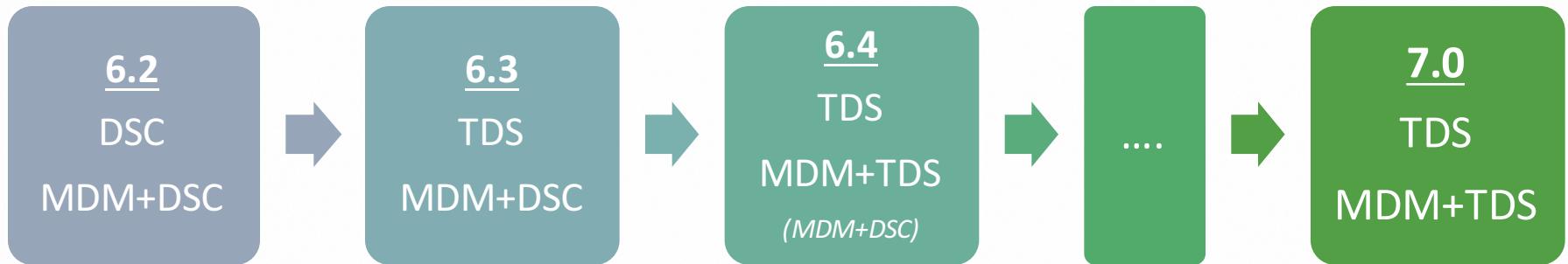


MDM

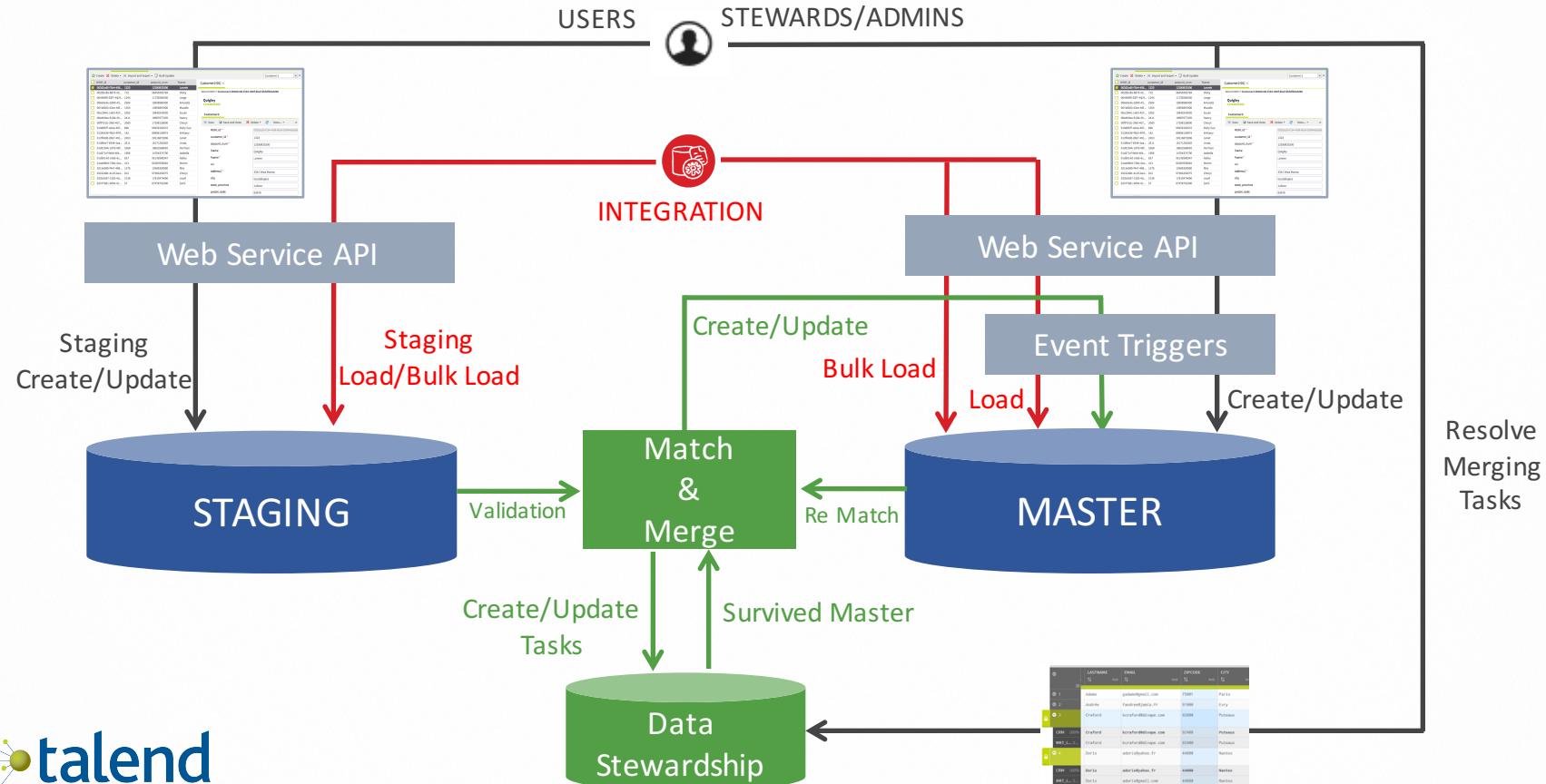
MDM 6.4 in a nutshell

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

- 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
```



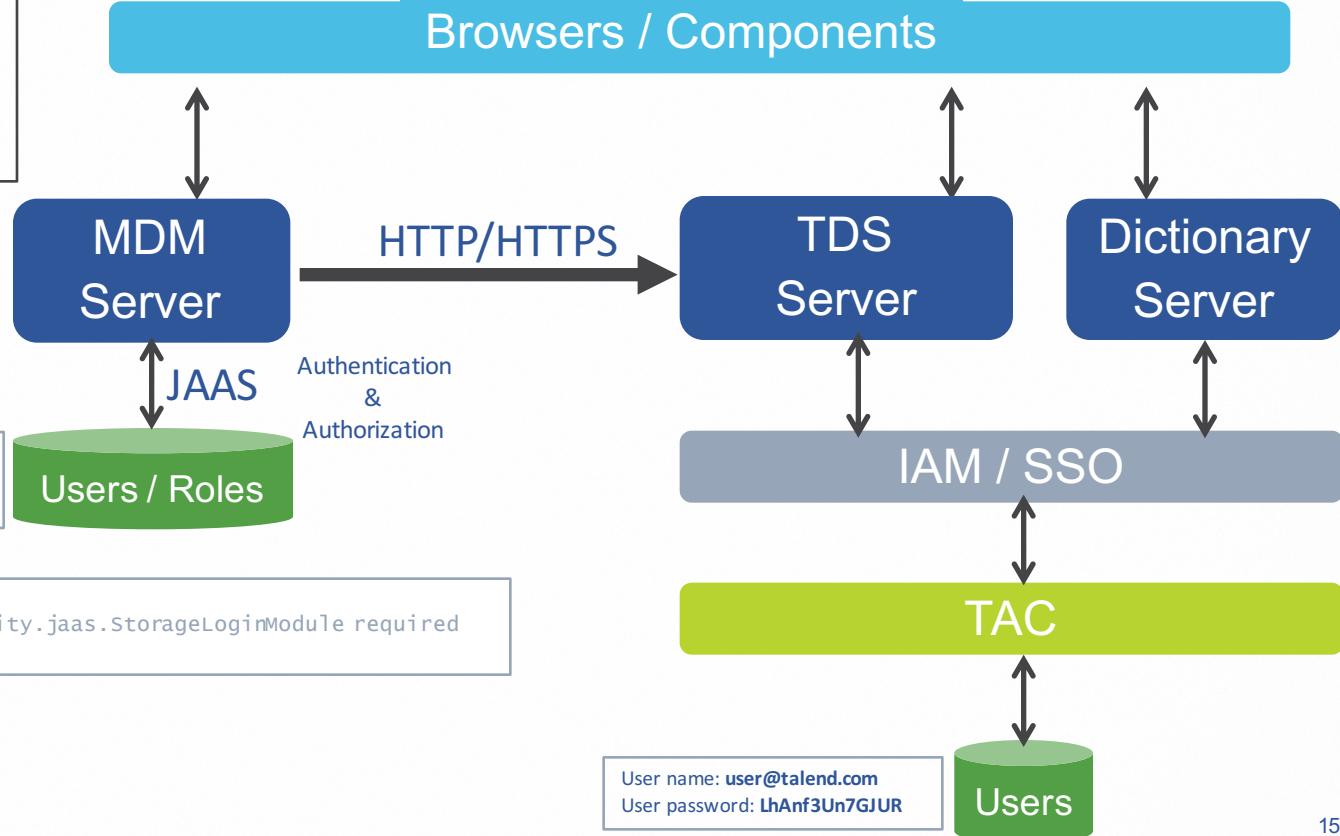
MDM CO credentials

Tasks creation batch size

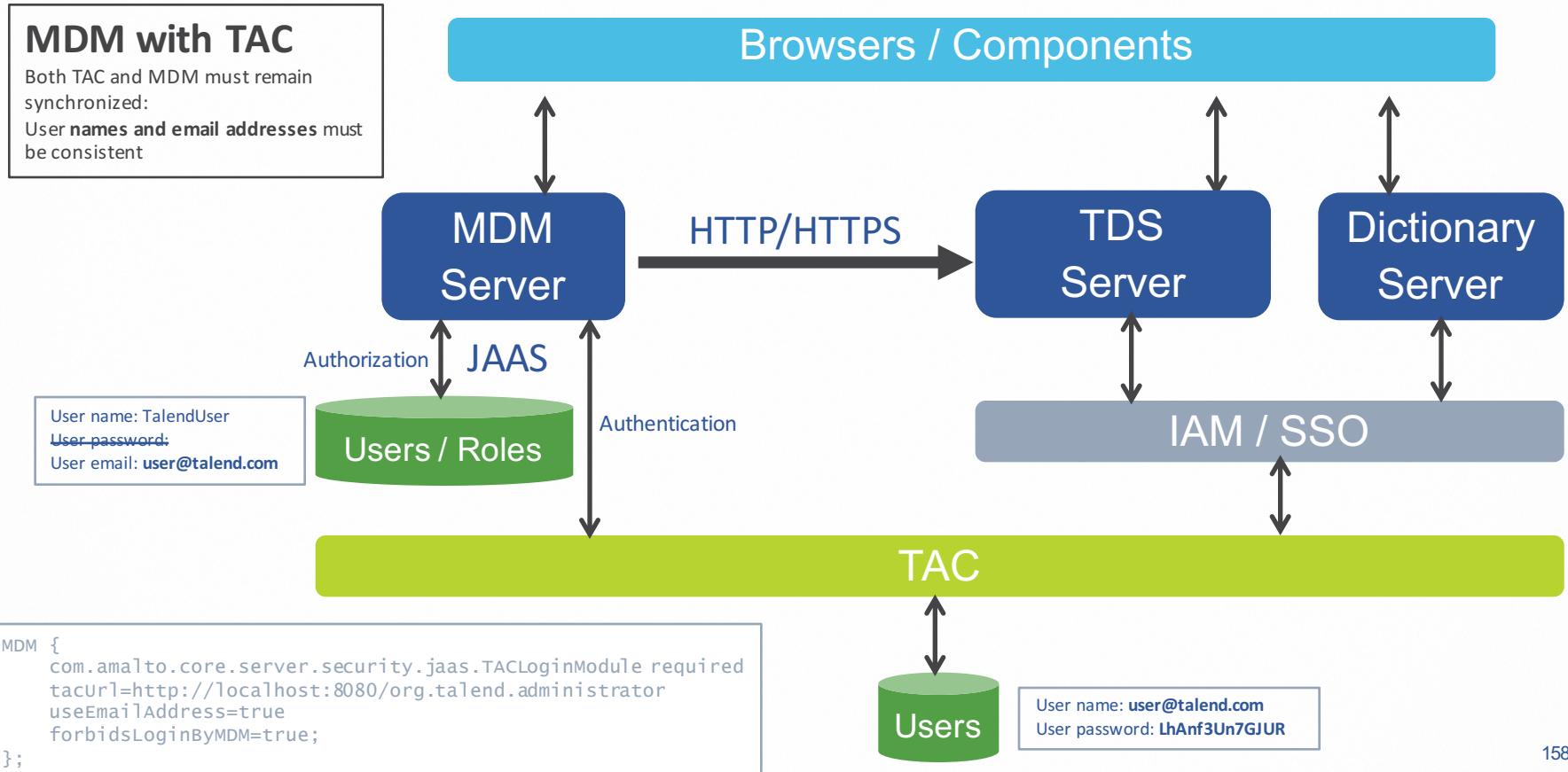
Integrated Matching

MDM without TAC

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



Integrated Matching



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	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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

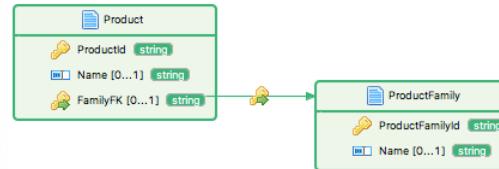
Integrated Matching (EE only)

- Support multiple match rules (TMDM-10501)

Match Rule 1	Match Rule 2	Match Rule 3
Match Key Name	Matching Function	
Name2	Jaro-Winkler	
DOB2	Jaro-Winkler	
Street2	Jaro-Winkler	

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

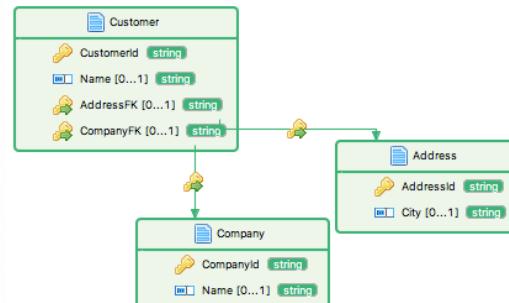
e.g. „Match Product records with their Name and FamilyFK“



- 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“



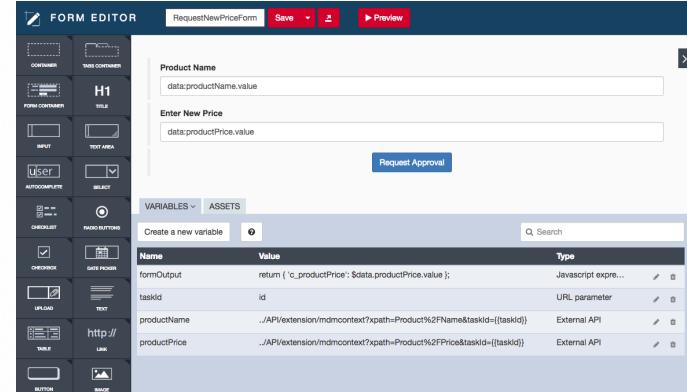
Bonita UI Form Designer

- 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} }

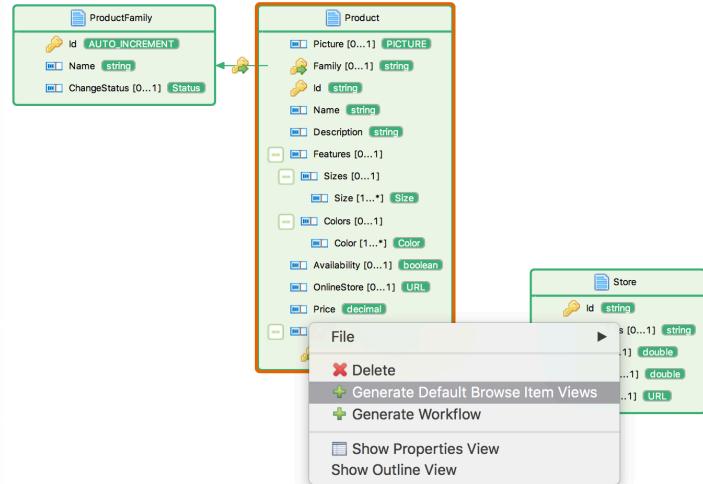
Request parameter "Product%2FName" is the URL encoded XPath for the attribute: "Product/Name"

- 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. <i>(Column was not deleted)</i>

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

1/ General

NAME:

DESCRIPTION: (optional)

TYPE:

ARBITRATION MERGING RESOLUTION GROUPING

QUESTION:

ANSWERS:

1
2

ENABLE TASK RESOLUTION DELAY

CAMPAIGN OWNERS

+ Add a campaign owner owner1@talend.com

2/ Roles

STEWARDS

+ Add a steward

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”

Are those records related

Task #2	4/4
	ARBITRATION <input type="checkbox"/> test
1	CUSTOMERID* <input type="checkbox"/> test
	FIRSTNAME <input type="checkbox"/> test
	LASTNAME* <input type="checkbox"/> test
	BIRTHDATE <input type="checkbox"/> date
	GENDER <input type="checkbox"/> test
	LASTV <input type="checkbox"/> test
1	Yes
Record 1	1 first name 1 last name 1
Record 2	1 first name 2 last name 2
2	
Record 1	1 first name 1 last name 1
Record 2	1 first name 2 last name 2
3	
Record 1	1 first name 1 last name 1
Record 2	1 first name 2 last name 2
4	
Record 1	1 first name 1 last name 1
Record 2	1 first name 2 last name 2

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



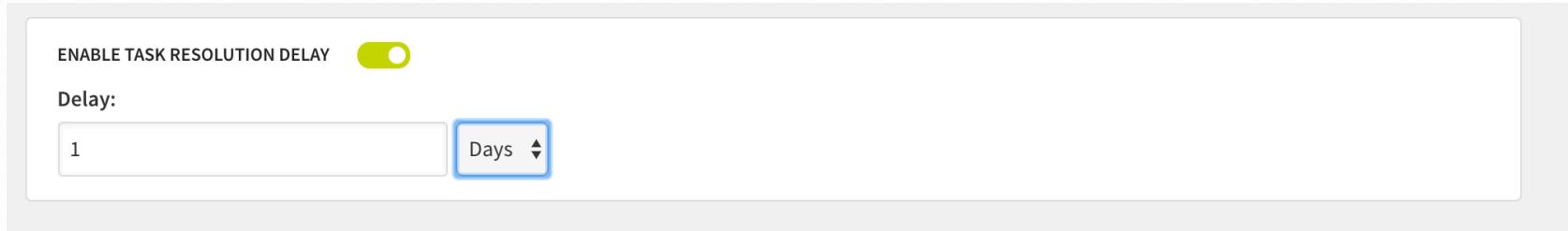
SLA on tasks

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

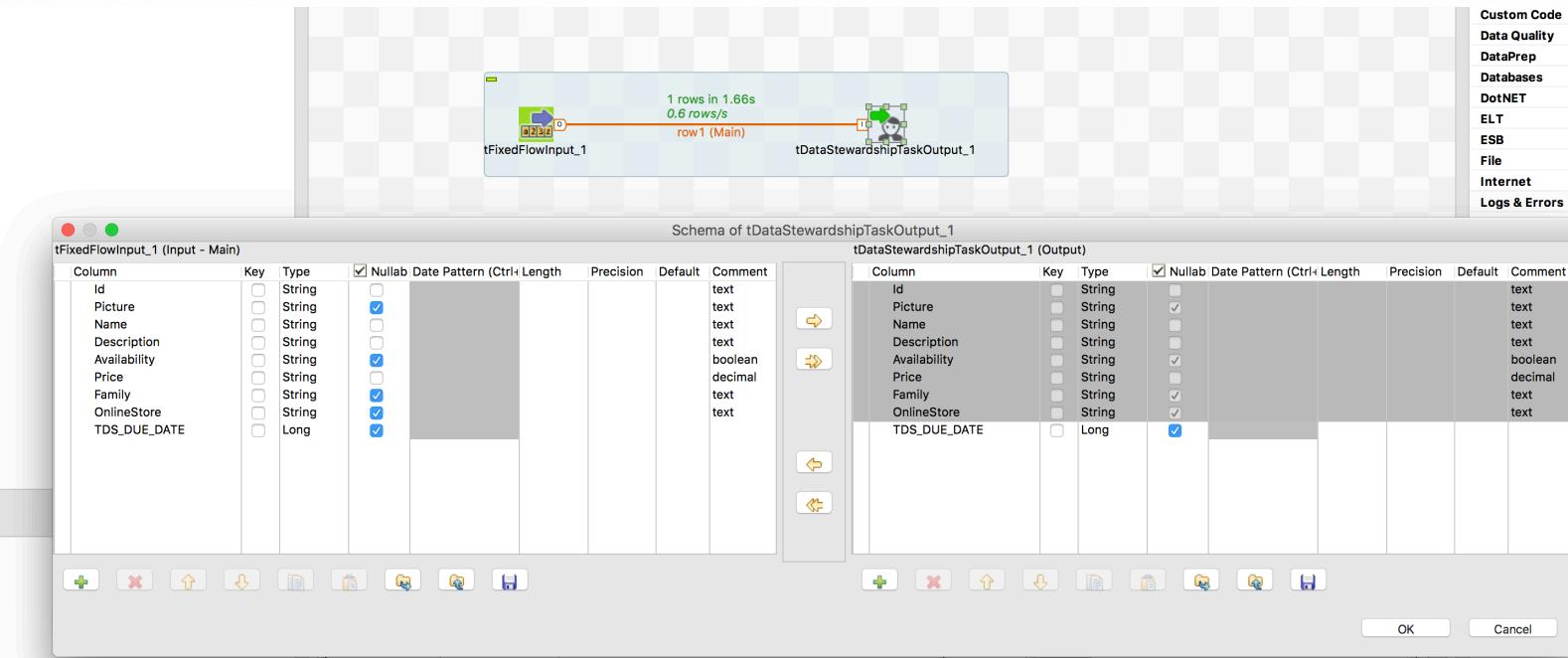
Configure default due dates at campaign level

- 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

	ID	PRIORITY*	DUE DATE	TAGS	SCORE	CREATED BY	M
1	id3	= Medium	2017-06-01 00:00:00			bguillon@talend.co	bg
source1	id1						
source2	id2						

Define due date

Selection

Due date



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

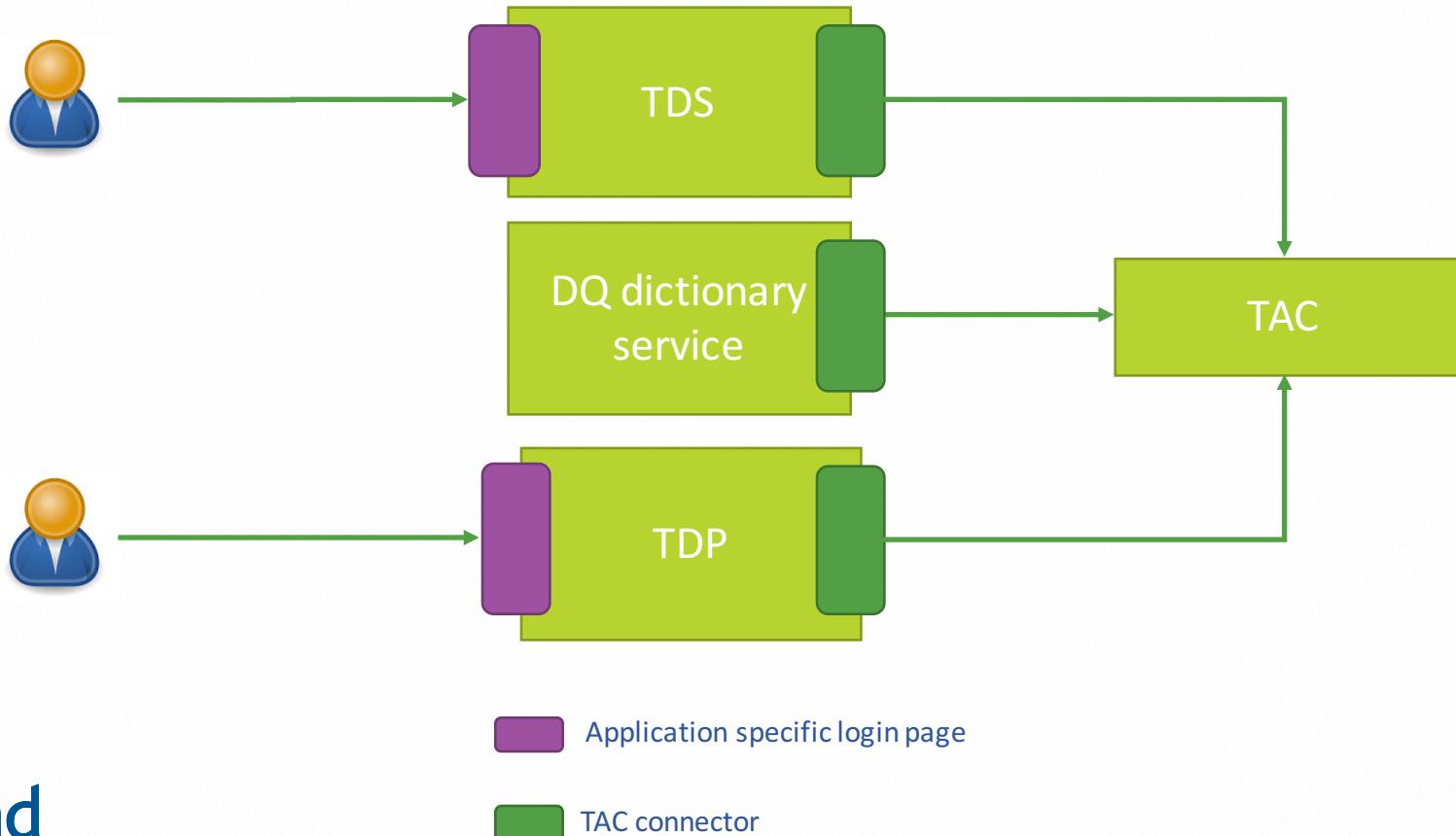
Data

Login:	user1@talend.com
First name:	user1
Last name:	user1
Password:	change password
Svn login:	
Svn password:	change password
GIT login:	
GIT password:	change password
Type:	No Project Access
Role:	
Data Preparation User:	<input type="checkbox"/>
Data Stewardship User:	<input checked="" type="checkbox"/>
Data Stewardship Role:	Data Steward
Group:	
Active:	<input checked="" type="checkbox"/>

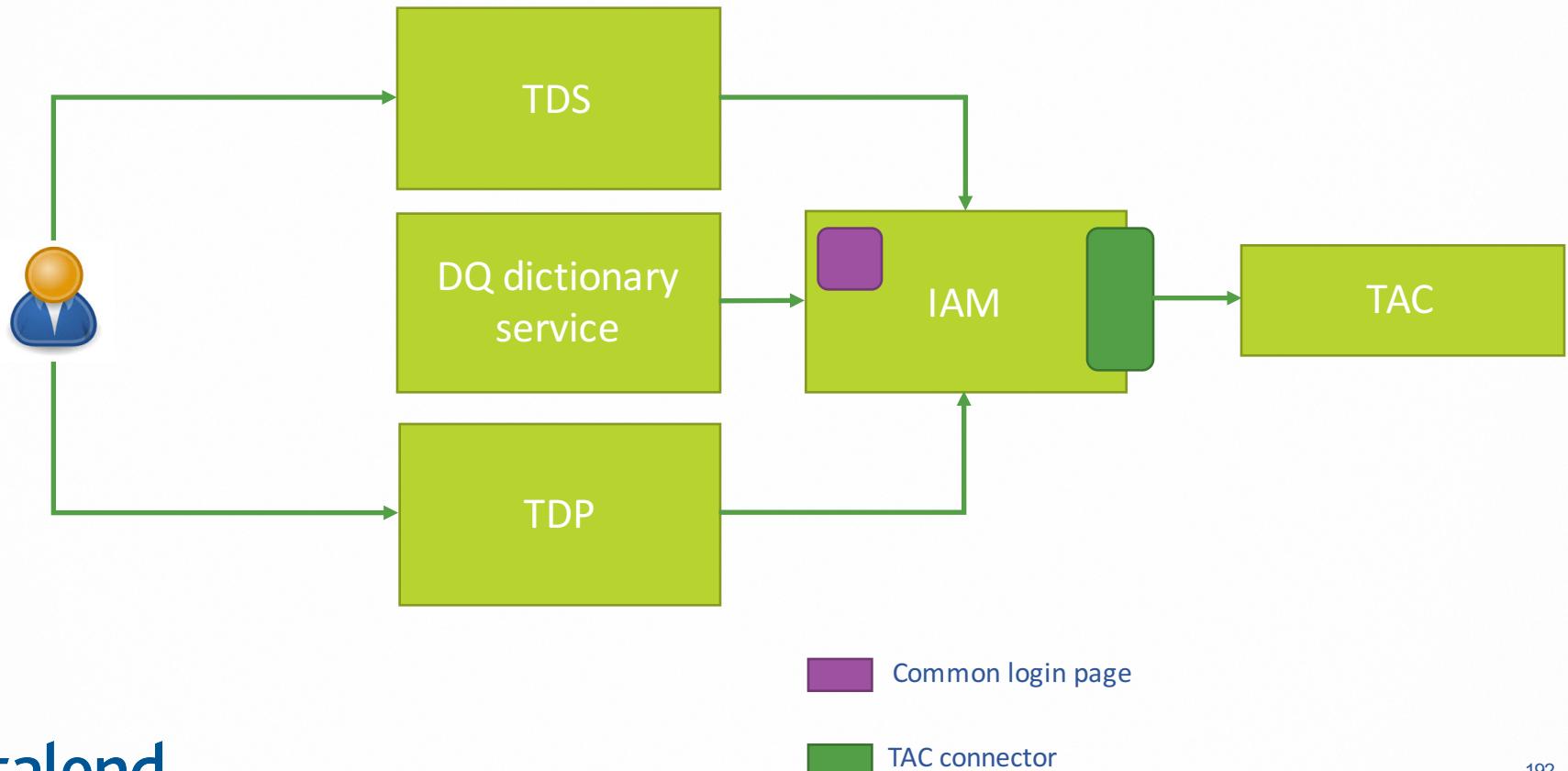
Identity management in 6.4

- 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



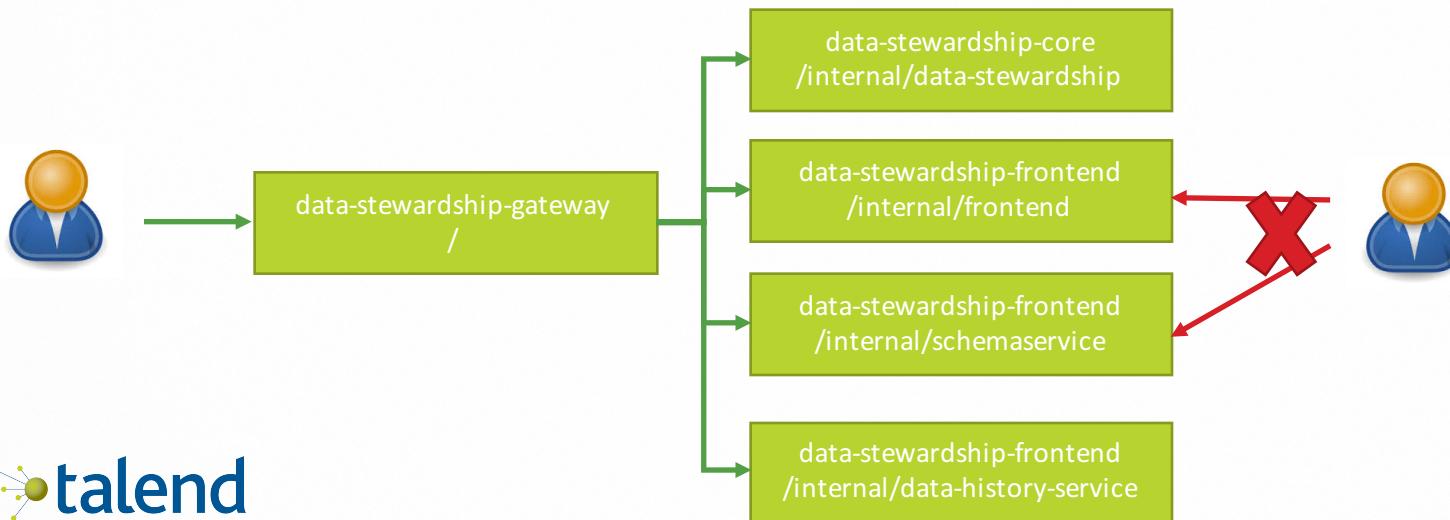
Talend IAM solution

- 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=tI6K6ac7tSE-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



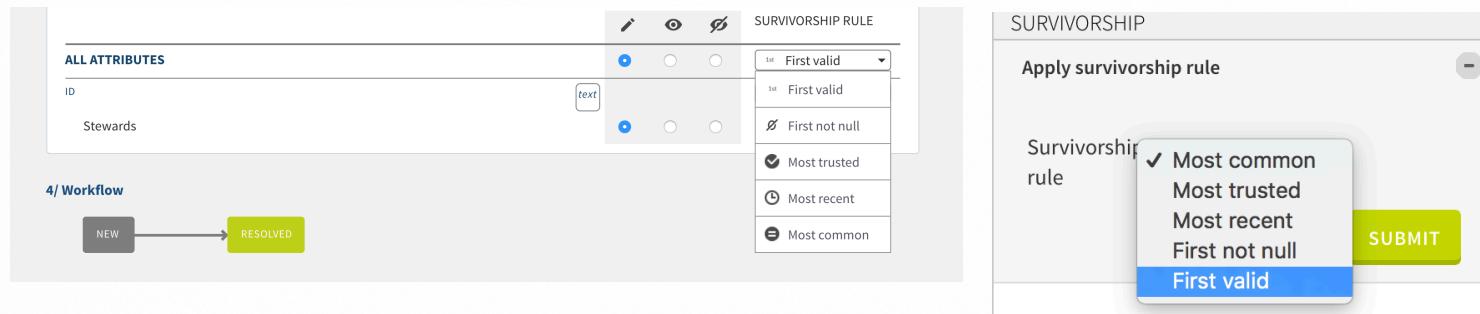
Miscellaneous changes

New bulk functions on tasks

Function name	Available to	Description
Assign tasks	<ul style="list-style-type: none">Campaign owners and Data StewardsAny kind of campaigns	Same as the assignment panel but can work on a filter or all tasks
Define due date	<ul style="list-style-type: none">Campaign ownerAny kind of campaigns	See SLA support
Arbitrate tasks	<ul style="list-style-type: none">Data StewardsArbitration/Grouping campaigns	Same as arbitration panel but can work on a filter or all tasks
Transition tasks	<ul style="list-style-type: none">Data StewardsCampaigns 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