talend | Technical Note >

New Features & Product Changes for Talend Summer '16

- Talend 6.2

July 2016

Table of Contents

Highlights

This technical note highlights the important new features and capabilities of Talend Summer ’16, including our new data preparation capability build into the Talend Data Fabric and of course many new features for big data integration, data integration, application integration, master data management, data quality and cloud integration. Supported features vary between Talend Open Studio and subscription products. Please refer to the talend.com product pages for more detail.

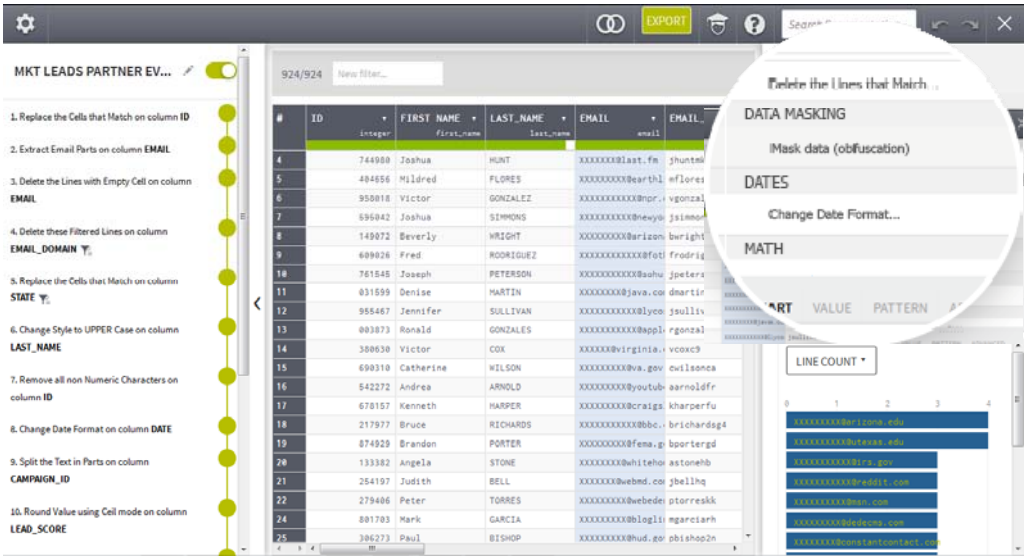
A few of the highlights of this release include:

- Talend combines data preparation and data integration into a single, unified platform that transforms how IT and business turn data into insight. Empower any decision maker to catalog, cleanse, and shape data from any source for use anywhere. Your data experts design the integration rules, while IT provides data governance and facilitates collaboration across batch, bulk, and master data management scenarios. Talend Data Preparation is available with every subscription product, so you can deliver self-service data prep at enterprise scale.
- Talend Data Mapper streamlines complex data processing on Spark and Hadoop so you can increase productivity and performance for Big Data, Real-Time Big Data, and Data Fabric integrations. You can now parse, validate, and transform complex message formats on Spark without hand-coding in XML, CSV, IDoc, Avro, JSON, EDI, COBOL and more.
- You can now maximize Amazon Web Services Redshift elasticity, getting the most out of your cloud resources. Dynamic cluster resizing for AWS EMR and Redshift lets you control the cost for the workload you need to process. Use Talend data profiling on AWS Redshift to analyze metadata and optimize data warehousing jobs more quickly.

1. Data Preparation

Talend Data Preparation combines data preparation and data integration to transform how IT and business can turn data into insight. While IT delivers governed self-service data access

and cleansing without putting data at risk or undermining compliance, business users using graphical tools can then find, visualize, clean, transform, enrich, catalog and consolidate data.



In particular:

- Teams can collaborate better by sharing datasets and data preparation recipes.
- IT ensures governance with appropriate role-based access to published, certified data.
- Data integration is accelerated by incorporating any data preparation recipe back into enterprise data integration scenarios including batch, bulk, and master data management.
- Data preparation is delivered at enterprise scale with support for hundreds of data sources and targets.

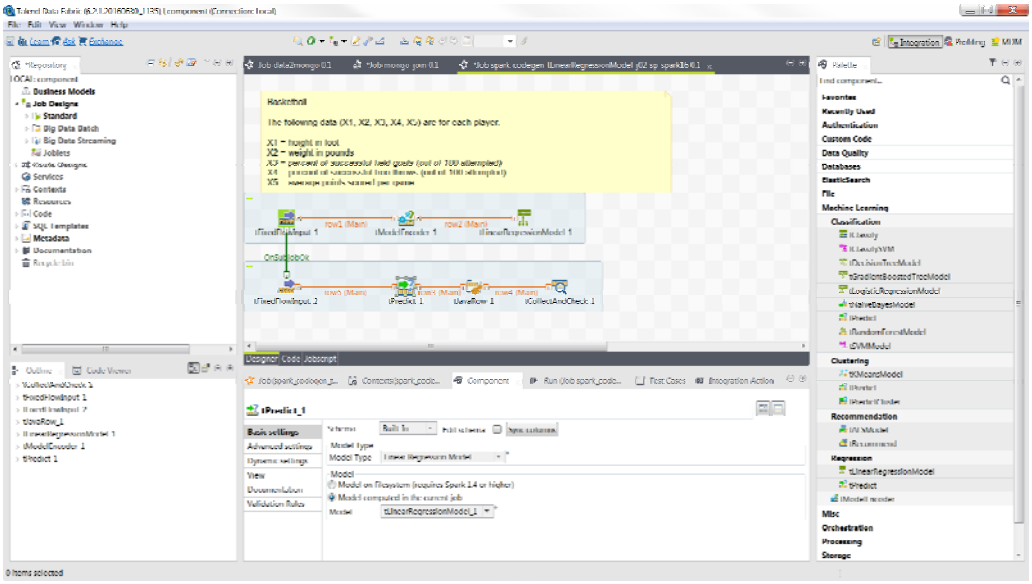
When you upgrade to Talend 6.2, you will receive two free Talend Data Preparation named-user licenses.

2. Big Data

Talend 6.2 introduces complex data mapping processing on Spark and Hadoop. See *Section 4. Data Mapper* for more details.

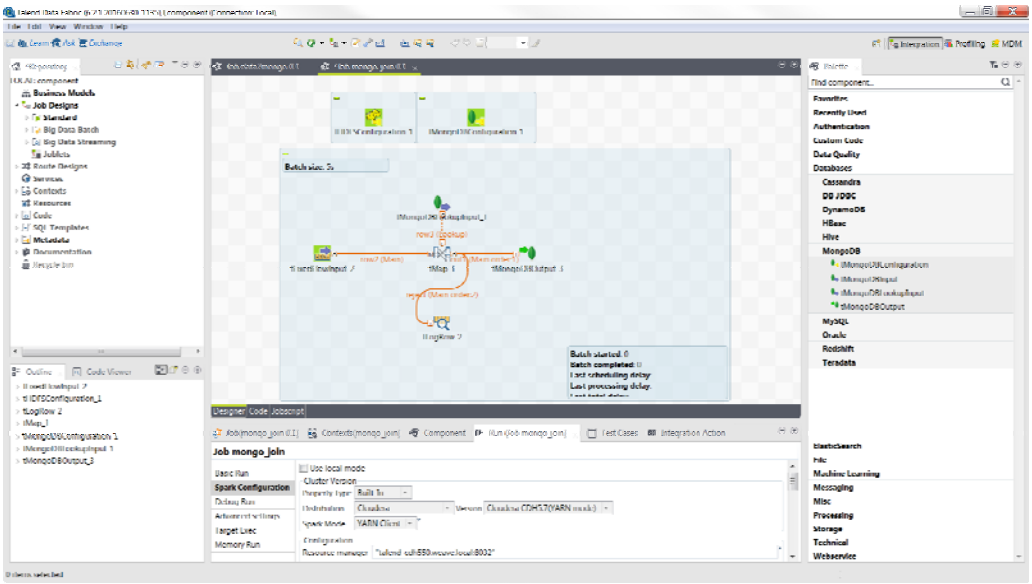
Talend 6.2 leverages Spark MLlib (machine learning library) to expand its machine learning capabilities and provides smarter and faster data-quality processing with support for intelligent

matching (as a Technical Preview), row standardization, reservoir sampling and transliterate functions.



V6.2 improves big data integration support, including:

- Higher performance for analytics and big data applications through support for distributed (parallel) processing between Spark and AWS RedShift, MongoDB and AWS DynamoDB.



- Native Kafka support in each Hadoop cluster provides better interoperability and easier to optimize performance.

Amazon Web Services support is also extended with:

- Expanded NoSQL ingestion capabilities by adding connectivity for AWS DynamoDB, with the unique ability to execute high performance reads and writes from a Spark job.
- Support for the latest AWS EMR and Redshift APIs, delivering high performance, distributed extract and load operations.
- Support for cluster resizing for AWS EMR and Redshift, so you can optimize the use of computing and storage resources (with the associated cost savings) by dynamically changing the number of nodes.

This release also supports the latest Hadoop distributions and versions of NoSQL databases, offering increased functionality and performance:

- Cloudera 5.7
- Hortonworks 2.4
- MapR 5.1
- Spark 1.6.2
- Amazon EMR 4.5, 4.6
- Microsoft Azure HDInsight 3.4
- Cassandra 3.4
- MongoDB 3.2
- AWS DynamoDB

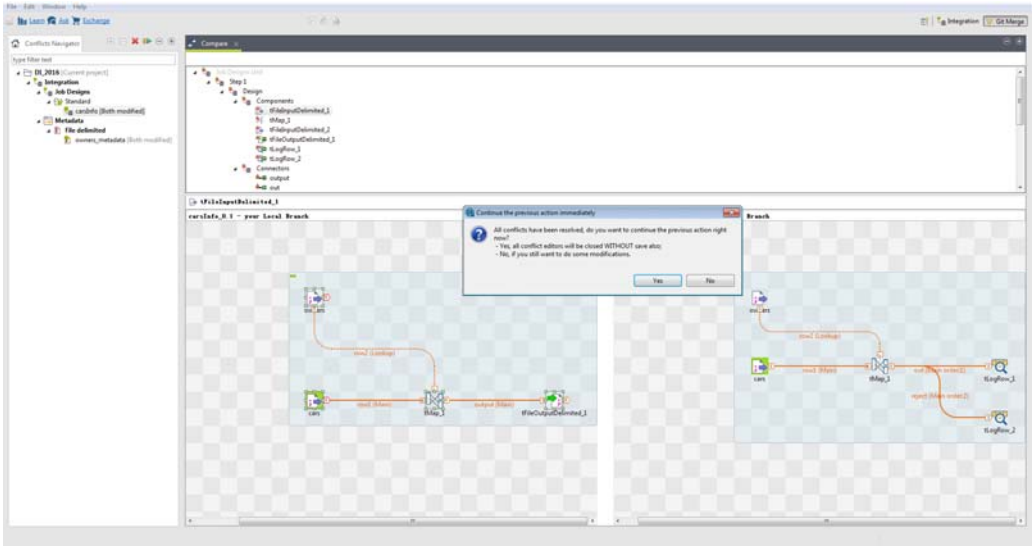
It also ships with four additional machine learning classification and regression components (Linear SVM, Decision Tree, Gradient-boosted Tree, and Linear Regression) to automate actionable insight in data pipelines.

It is now possible to update Hadoop distributions without having to reinstall Talend Studio.

3. Data Integration

Talend 6.2 provides productivity improvements and continuous delivery enhancements with Git support for graphical component level diff and merge, the ability to create feature and bug

branches, and to merge to and from branches; and support for Bitbucket.

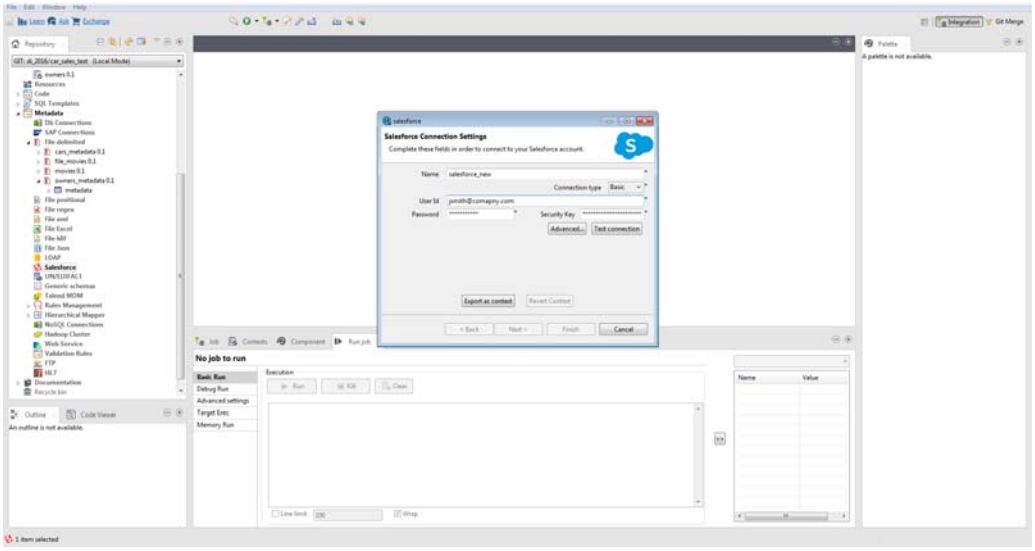


Support for Amazon Web Services is extended, including:

- The ability to perform AWS Redshift data-quality profiling with the collection of metadata used to analyze and optimize data warehousing.
- Support for the latest AWS Redshift APIs, delivering high-performance, distributed extract and load operations.
- Support for cluster resizing for AWS Redshift, so you can optimize the use of computing and storage resources (with the associated cost savings) by dynamically changing the number of nodes.
- Enterprise-grade SSL communication between Talend Jobs and AWS Redshift, as well as support for AWS S3 server and client-side encryption.
- Role-based access to AWS services and resources by inheriting credentials from AWS Identity and Access Management (IAM).

Enterprise connectivity updates include:

- Support for SAP Business Warehouse
- Support for JIRA
- Support for the latest Salesforce.com and Salesforce Wave Spring ’16 APIs



- Support for the latest Marketo REST API
- SAP recertification
- Support for Splunk Event Collector
- Updates for ExaSol ELT
- Updates for Vertica

In Talend Administration Center, users can now be grouped by user type: an MDM user can be part of an MDM, Data Quality or Data Integration group; a Data Quality user can be part of a Data Quality or Data Integration group, but not an MDM group, and a Data Integration user can only be part of a Data Integration group.

Talend Administration Center now has two repositories to store custom libraries: snapshot and release.

The Data Integration and ESB Studio interface perspectives have been merged, which helps improve productivity on data services projects.

A new components framework has been added, making it easier to incorporate your own components into Talend.

4. Data Mapper

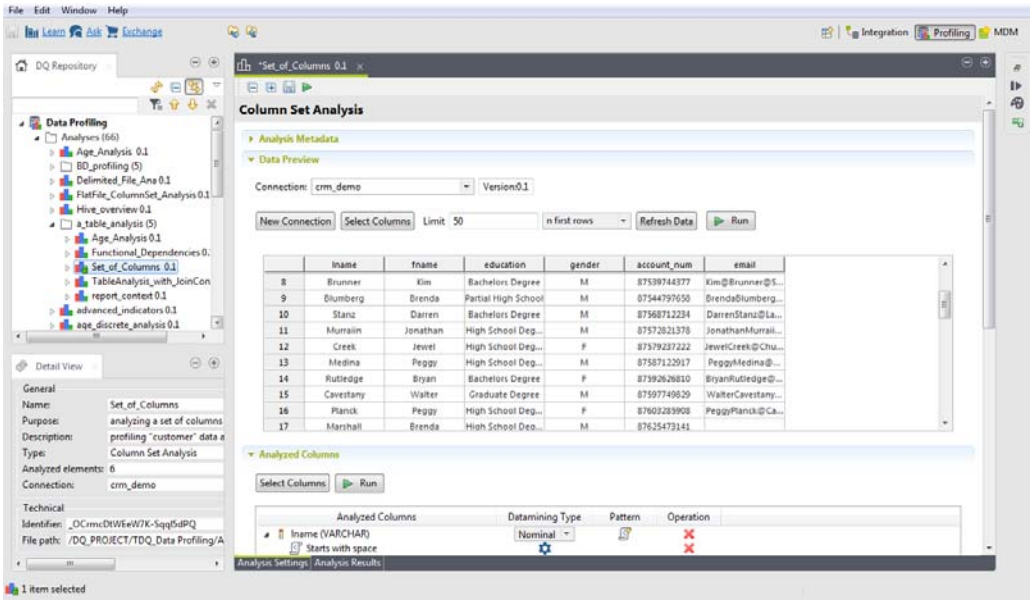
Talend 6.2 streamlines complex data processing on Spark and Hadoop so you can increase productivity and performance for Big Data, Real-Time Big Data, and Data Fabric integrations, thanks to the introduction of new components that leverage Talend Data Mapper. You can now parse, validate, and transform complex message formats on Spark without hand-coding in XML, CSV, IDoc, Avro, JSON, EDI, COBOL and more.

This lets you apply and test syntactic and semantic validation rules for big data integrations to ensure data accuracy and compliance, and get results fast by running everything at speed and scale on Spark.

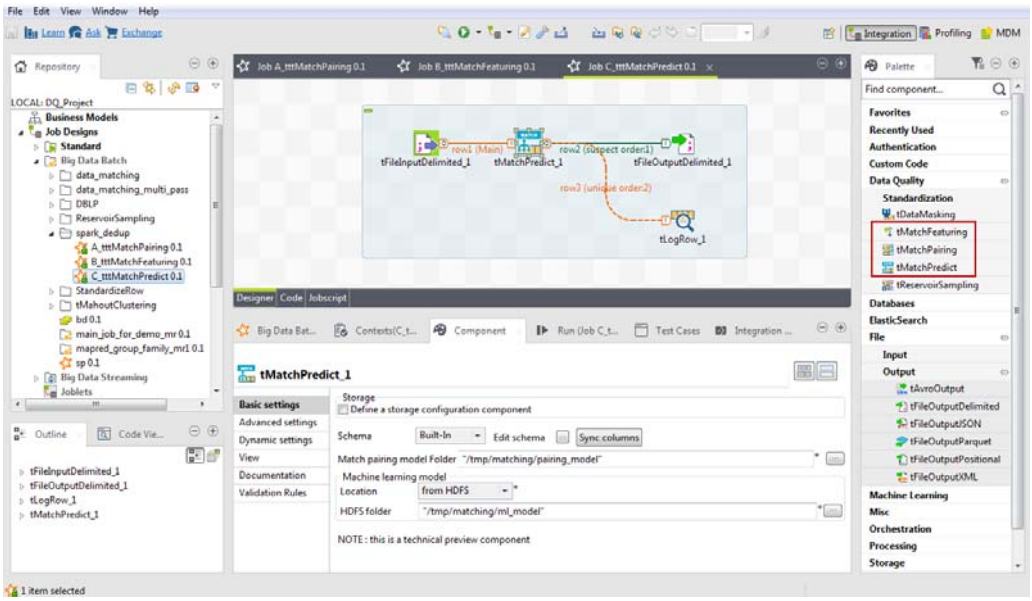
5. Data Quality

The Profiling perspective in the Studio now supports analyzing data in Amazon Redshift.

Analysis Editors have been enhanced with the Data Preview section and with new icons and buttons to optimize user experience when working with analyses. Additionally, running the analyses now automatically switches the editor to the Analysis Results view.

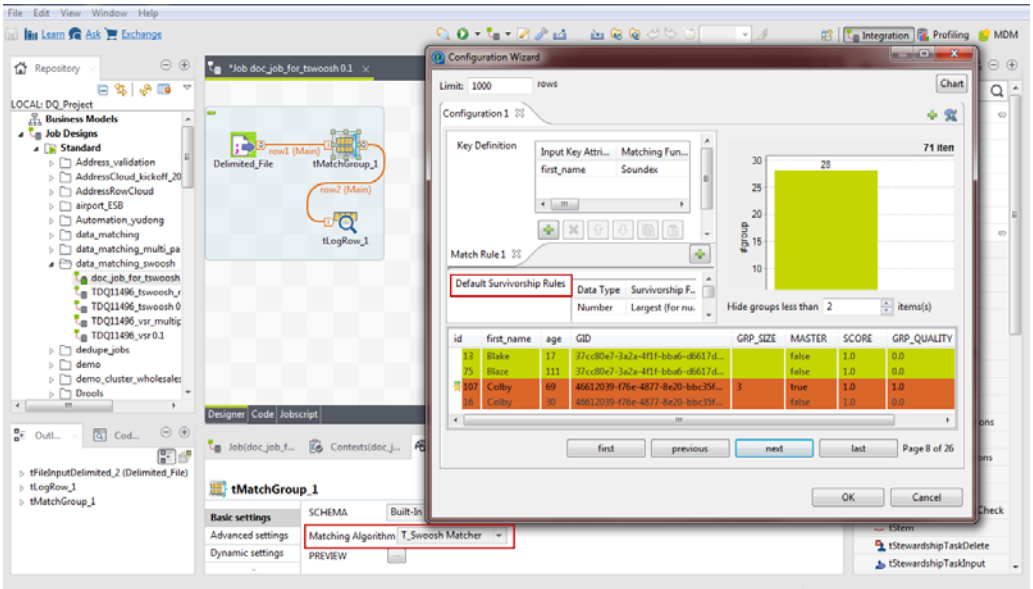


New matching components which work in Spark framework have been introduced (as a Technical Preview) in the studio: tMatchPairing, tMatchFeaturing and tMatchPredict.



Users can use now the components tStandardizeRow, tReservoirSampling and tTransliterate in a Spark framework.

The T-Swoosh algorithm is now supported in the standard tMatchGroup component.



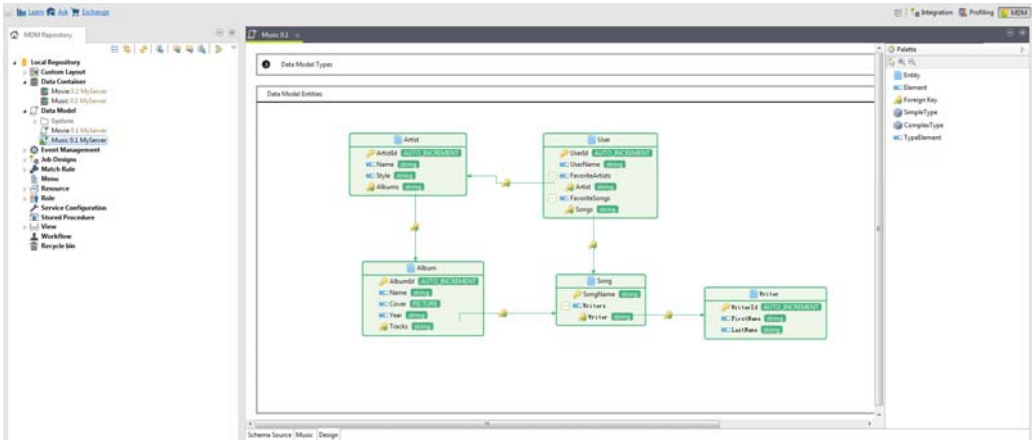
6.MDM

Talend 6.2 introduces machine learning-based data matching and deduplication on Spark, as a Technical preview. It provides a smarter and extremely scalable approach to connect your

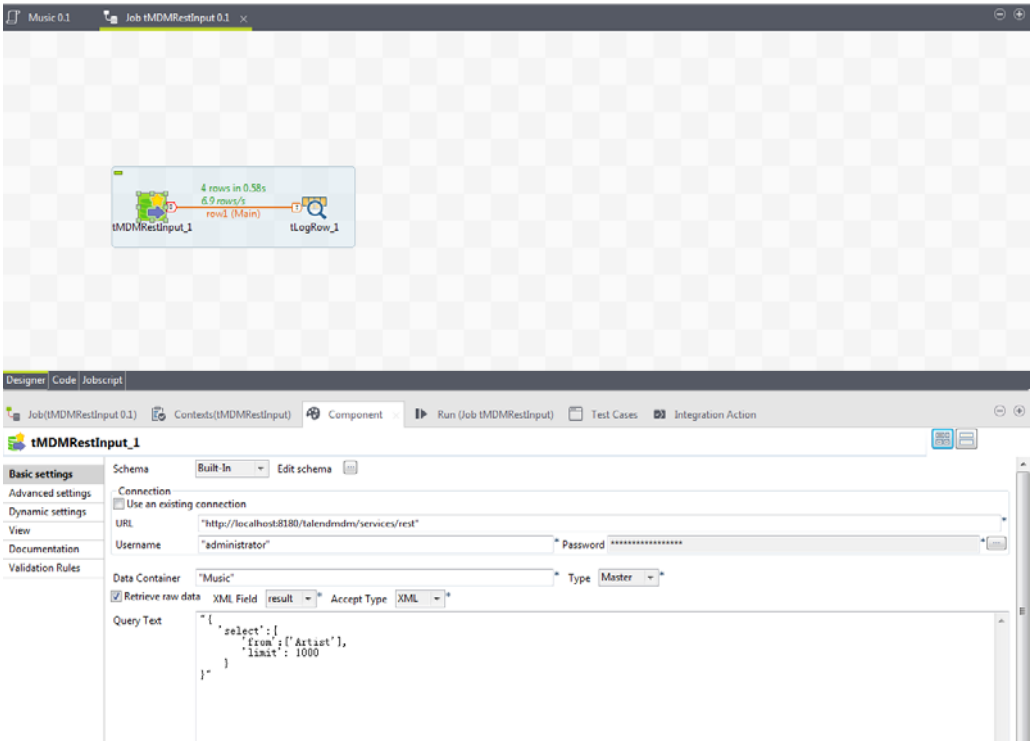
data, big data and master data through a new Spark based matching function leveraging Spark's elasticity, clustering and machine learning.

In the MDM Web User Interface, it is now possible to navigate through the relationships of a record, exploring both incoming and outgoing links, through the new relationship navigator.

It adds graphical modelling features for MDM and improved hierarchy exploration with filtering and graphs.



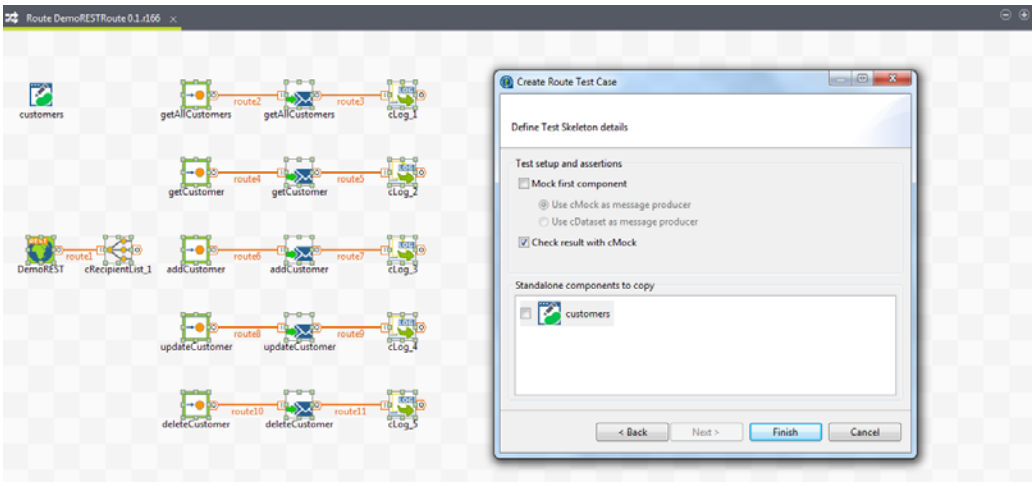
A new integration component, tMDMRestInput, based on the REST API, lets users extract master records with improved performance and a powerful query language.



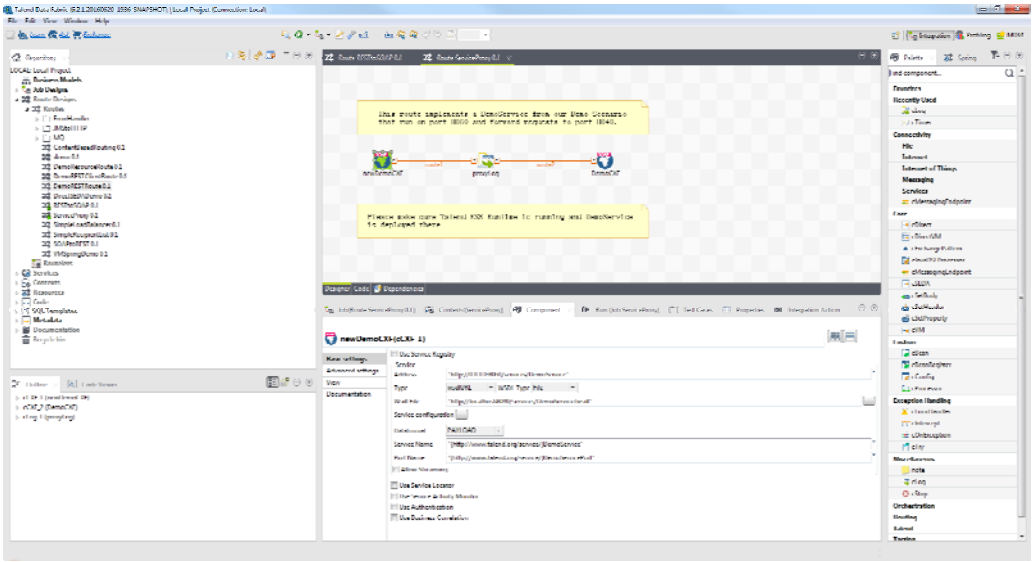
7.ESB

Talend 6.2 allows ESB routes to be run as a Spring Boot microservice, which is beneficial for large teams doing modular development where services are easier to deploy since they are autonomous.

The Talend Studio now supports graphical testing for ESB Routes using the 'Test Case' creation and execution feature, extended to also support specific Route-related features with cMock and the support of ‘Producer Templates’ for testing.

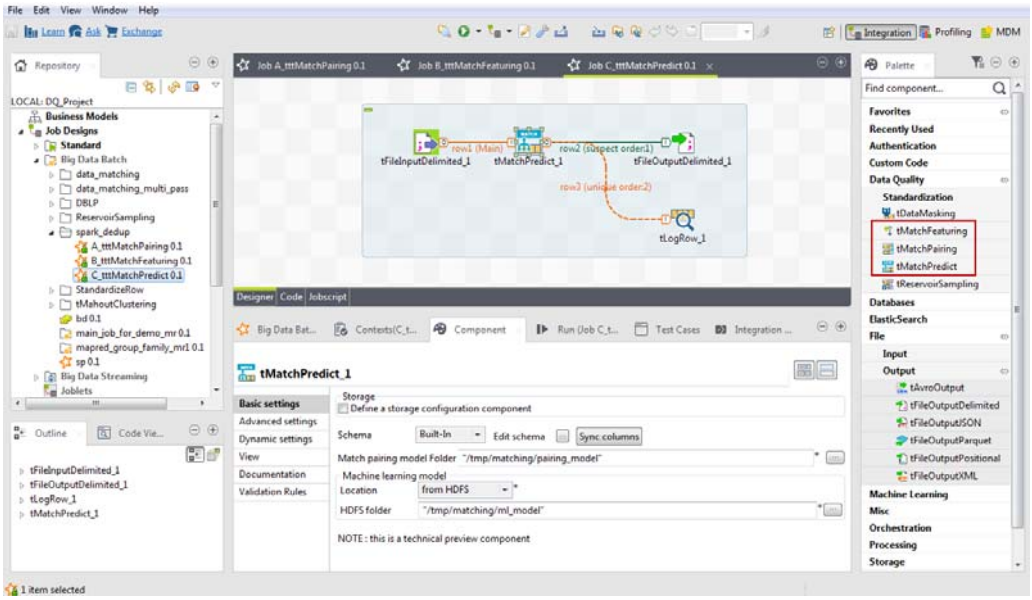


It unifies the Data Integration and ESB interface perspectives in order to improve productivity.



It ships with updated components for Apache Kafka and MQTT for enhanced ESB, cloud and big data interoperability.

It provides certification for AWS IoT Gateway with MQTT, ensuring compatibility for IoT scenarios in the cloud.



8.About Talend

Talend's integration solutions allow data-driven organizations to gain instant value from all their data. Through native support of modern big data platforms, Talend takes the complexity out of integration efforts and equips IT departments to be more responsive to the demands of the business, at a predictable cost. Based on open source technologies, Talend's scalable, future-proof solutions address all existing and emerging integration requirements. Talend is privately-held and headquartered in Redwood City, CA. For more information, please visit www.talend.com and follow us on Twitter: @Talend.