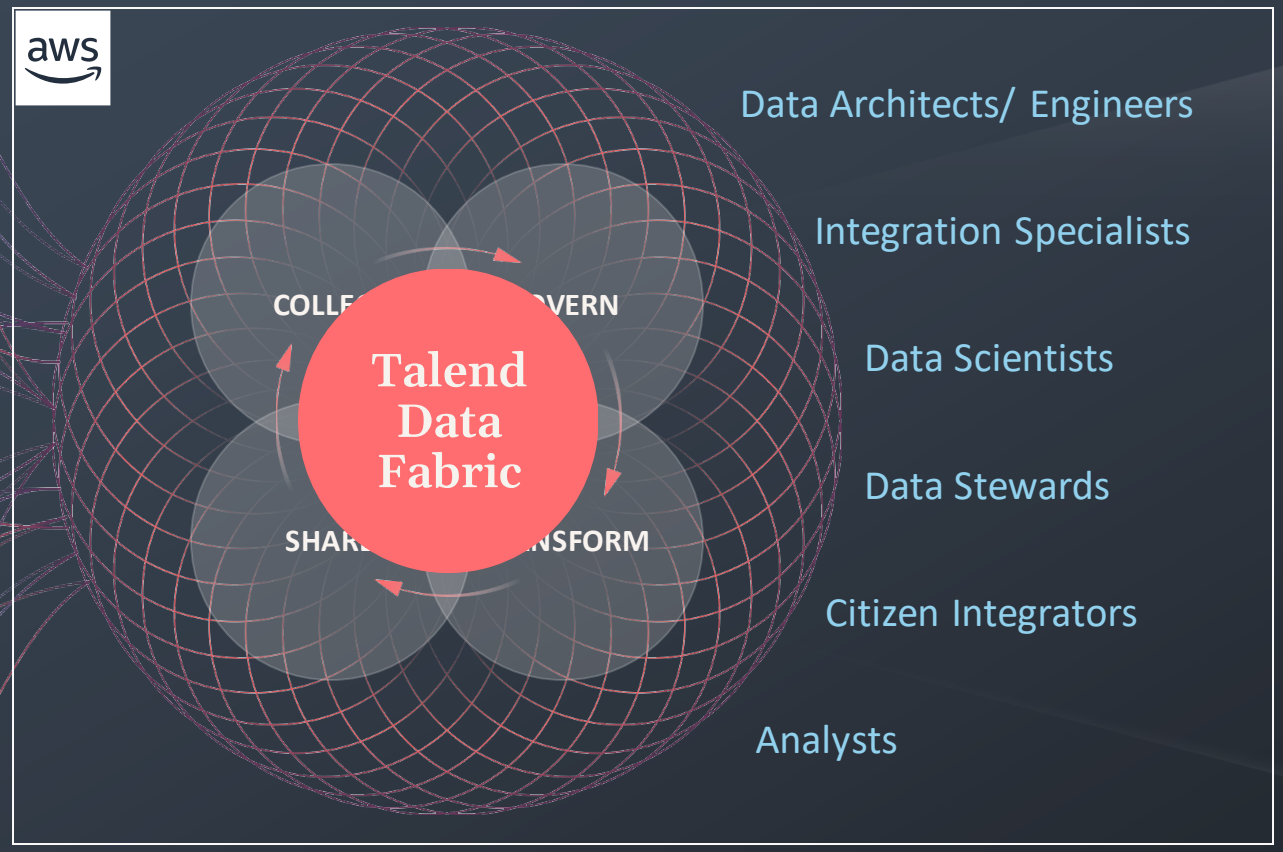
Talend and Amazon Redshift

May 2020

Talend Introduction

Cloud Data Management Platform: Integration, Quality & Governance

Data Chaos



Turn raw data into revenue

Innovate faster

Transform the customer experience

Dramatically reduce cost and risk

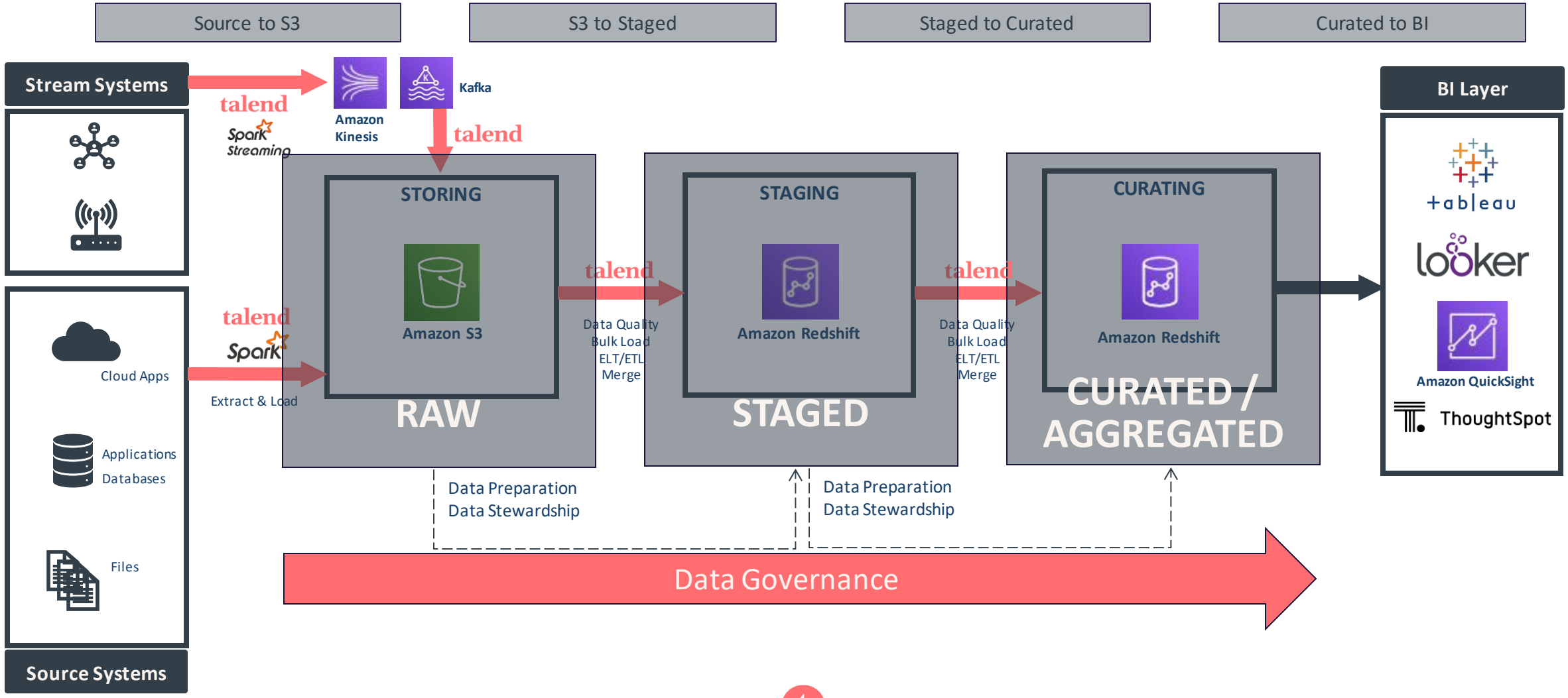
Data Lake | Data Warehouse | Data Processing | Migrations

Complete, Compliant and Timely Data

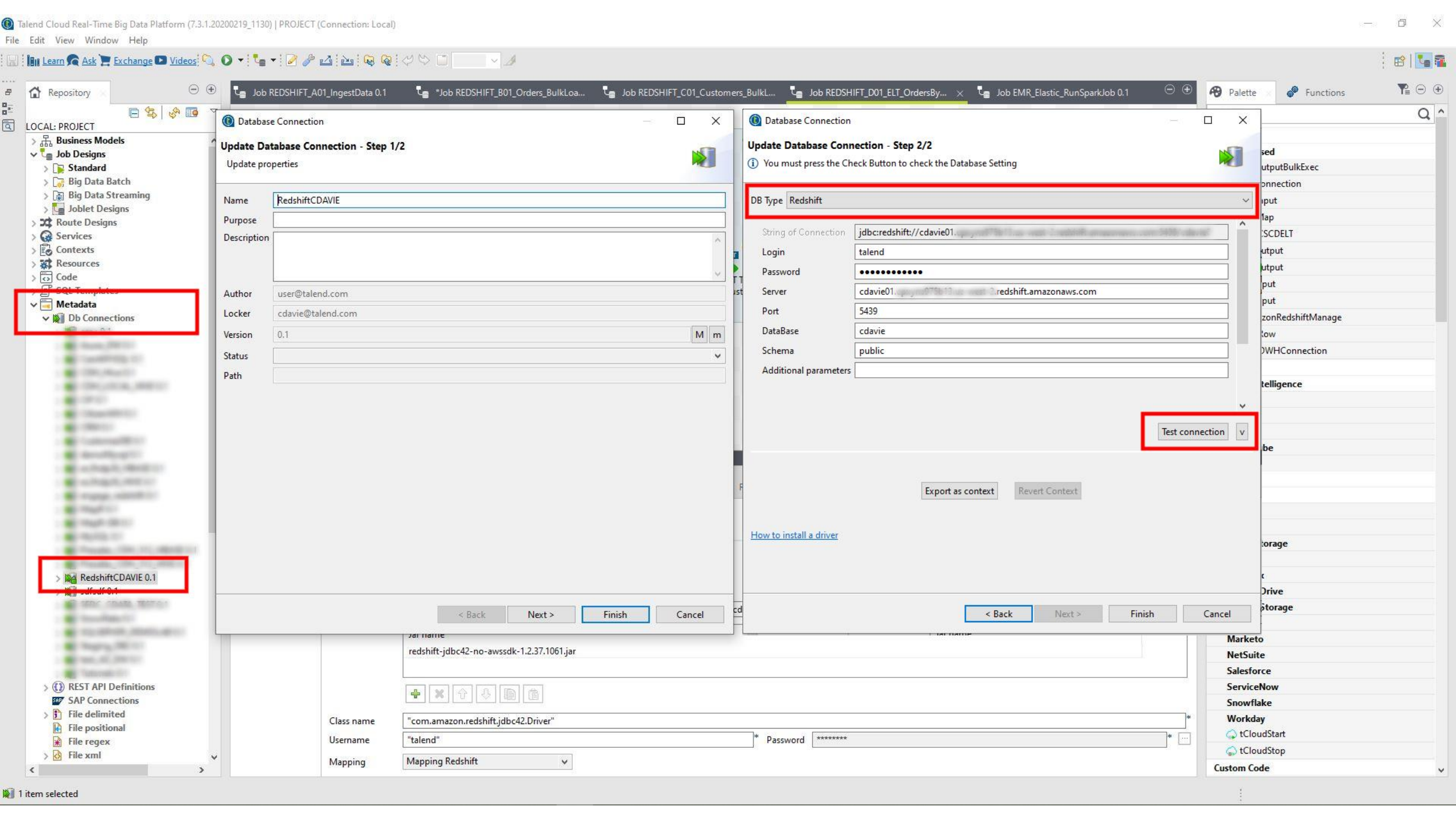
Gartner A recognized Magic Quadrant Leader in Data Integration Tools and Data Quality

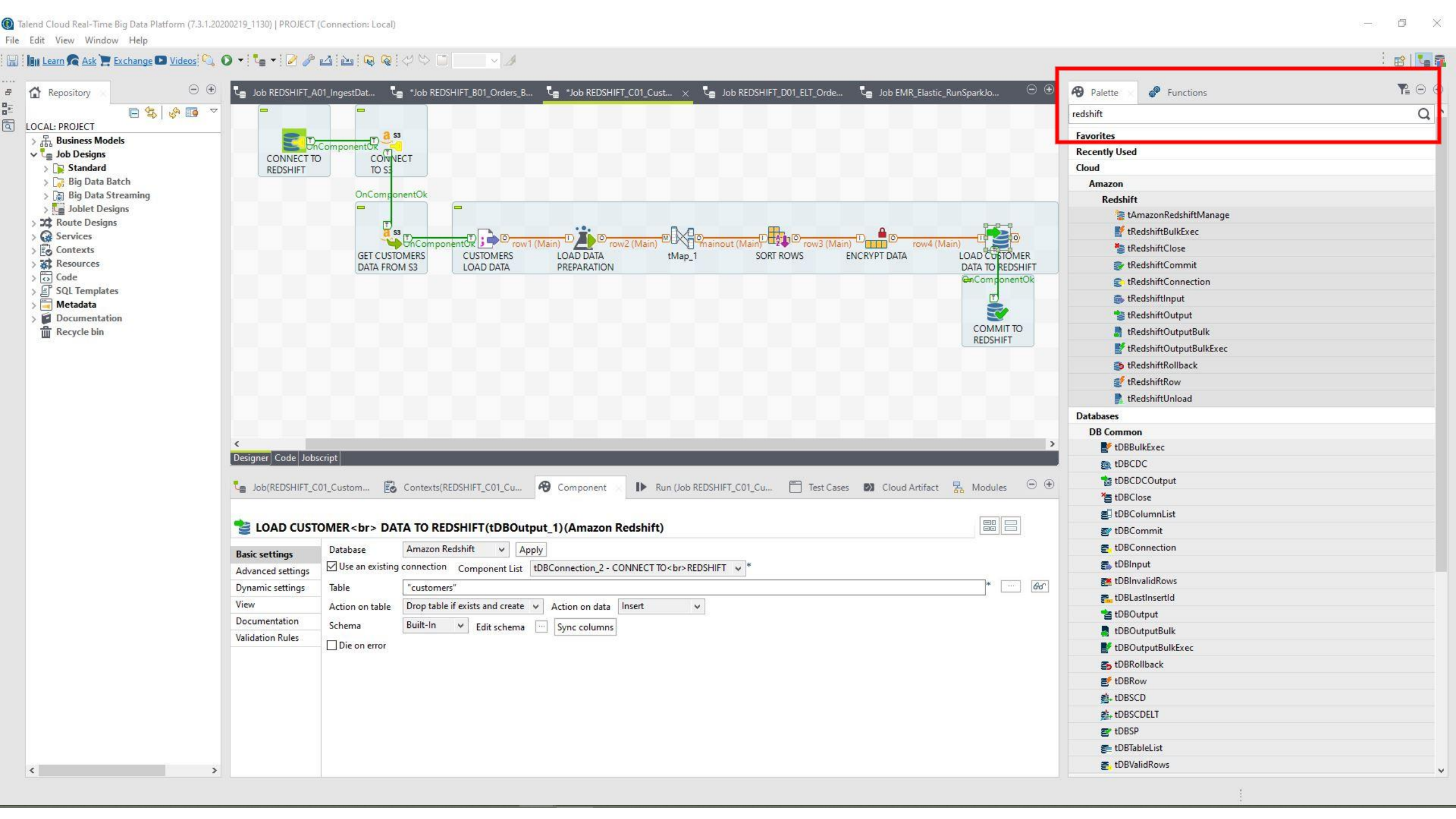
Demo Technical Reference Architecture

Best Practice Architecture for Cloud Data Warehousing

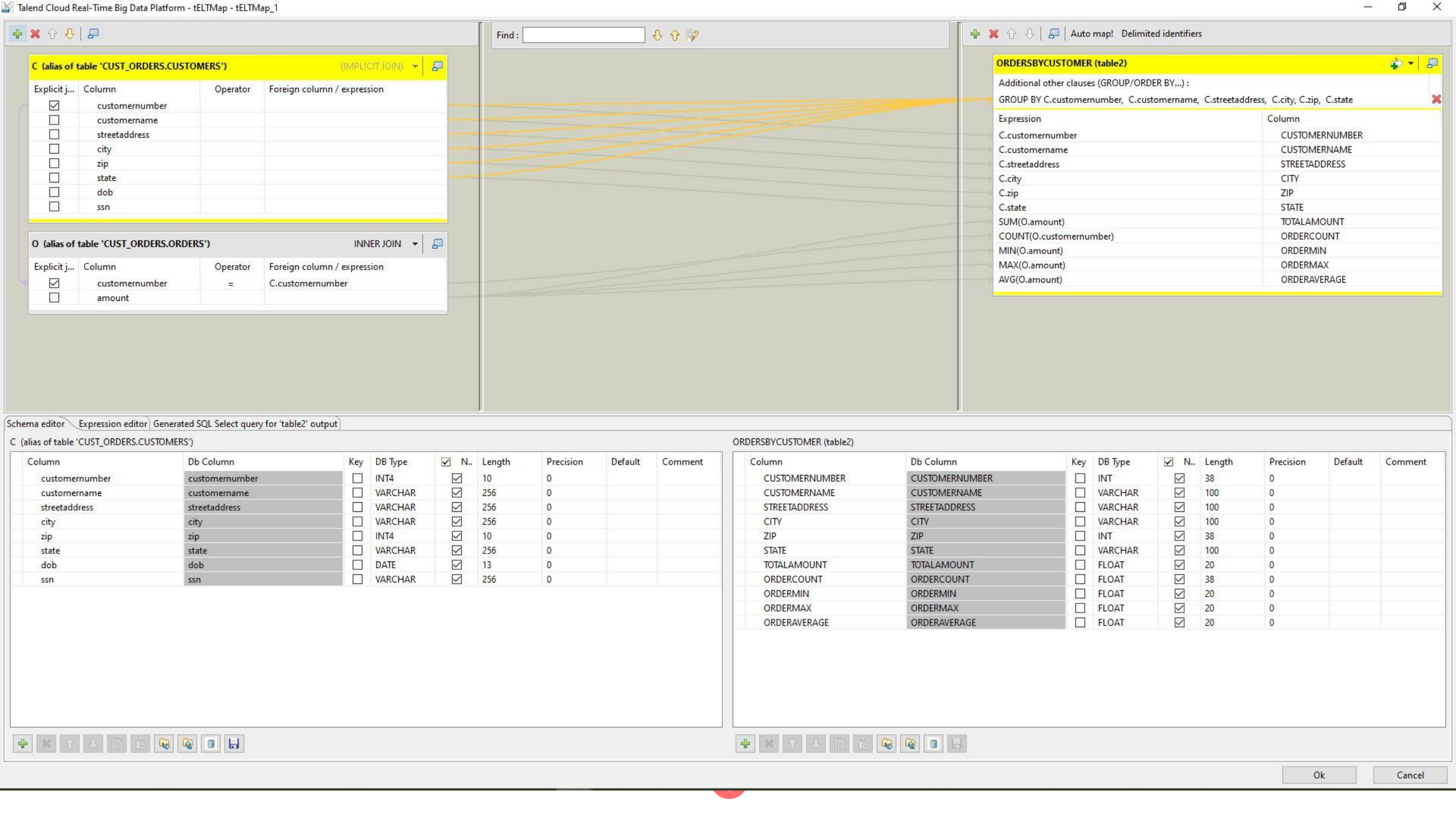


Amazon Redshift components in Talend Studio


Talend Cloud Real-Time Big Data Platform (7.3.1.20200219_1130) | PROJECT (Connection: Local)
File Edit View Window Help
Learn Ask Exchange Videos
Repository
LOCAL: PROJECT
Business Models
Job Designs
Standard
Big Data Batch
Big Data Streaming
Joblet Designs
Route Designs
Services
Contexts
Resources
Code
Metadata
Db Connections
RedshiftCDAVIE 0.1
REST API Definitions
SAP Connections
File delimited
File positional
File regex
File xml
Job REDSHIFT_A01_IngestData 0.1
*Job REDSHIFT_B01_Orders_BulkLoa...
Job REDSHIFT_C01_Customers_BulkL...
Job REDSHIFT_D01_ELT_OrdersBy...
Job EMR_Elastic_RunSparkJob 0.1
Palette
Functions
Database Connection
Update Database Connection - Step 1/2
Update properties
Name RedshiftCDAVIE
Purpose
Description
Author user@talend.com
Locker cdavie@talend.com
Version 0.1 M m
Status
Path
< Back Next > Finish Cancel
Database Connection
Update Database Connection - Step 2/2
You must press the Check Button to check the Database Setting
DB Type Redshift
String of Connection jdbc:redshift://cdavie01
Login talend
Password ••••••••••••
Server cdavie01 redshift.amazonaws.com
Port 5439
DataBase cdavie
Schema public
Additional parameters
Test connection
Export as context
Revert Context
How to install a driver
< Back Next > Finish Cancel
redshift-jdbc42-no-awssdk-1.2.37.1061.jar
Class name "com.amazon.redshift.jdbc42.Driver"
Username "talend" Password ********
Mapping Mapping Redshift
Marketo
NetSuite
Salesforce
ServiceNow
Snowflake
Workday
tCloudStart
tCloudStop
Custom Code
1 item selected


Talend Cloud Real-Time Big Data Platform (7.3.1.20200219_1130) | PROJECT (Connection: Local)
File Edit View Window Help
Learn Ask Exchange Videos
Repository
LOCAL: PROJECT
Business Models
Job Designs
Standard
Big Data Batch
Big Data Streaming
Joblet Designs
Route Designs
Services
Contexts
Resources
Code
SQL Templates
Metadata
Documentation
Recycle bin
Job REDSHIFT_A01_IngestDat...
*Job REDSHIFT_B01_Orders_B...
*Job REDSHIFT_C01_Cust...
Job REDSHIFT_D01_ELT_Orde...
Job EMR_Elastic_RunSparkJo...
CONNECT TO REDSHIFT
OnComponentOk
CONNECT TO S3
GET CUSTOMERS DATA FROM S3
CUSTOMERS LOAD DATA
row1 (Main)
LOAD DATA PREPARATION
row2 (Main)
tMap_1
mainout (Main)
SORT ROWS
row3 (Main)
ENCRYPT DATA
row4 (Main)
LOAD CUSTOMER DATA TO REDSHIFT
COMMIT TO REDSHIFT
Designer Code Jobscript
Job(REDSHIFT_C01_Custom...
Contexts(REDSHIFT_C01_Cu...
Component
Run (Job REDSHIFT_C01_Cu...
Test Cases
Cloud Artifact
Modules
LOAD CUSTOMER
 DATA TO REDSHIFT(tDBOutput_1)(Amazon Redshift)
Basic settings
Advanced settings
Dynamic settings
View
Documentation
Validation Rules
Database
Amazon Redshift
Apply
Use an existing connection
Component List
tDBConnection_2 - CONNECT TO
REDSHIFT
Table
"customers"
Action on table
Drop table if exists and create
Action on data
Insert
Schema
Built-In
Edit schema
Sync columns
Die on error
Palette
Functions
redshift
Favorites
Recently Used
Cloud
Amazon
Redshift
tAmazonRedshiftManage
tRedshiftBulkExec
tRedshiftClose
tRedshiftCommit
tRedshiftConnection
tRedshiftInput
tRedshiftOutput
tRedshiftOutputBulk
tRedshiftOutputBulkExec
tRedshiftRollback
tRedshiftRow
tRedshiftUnload
Databases
DB Common
tDBBulkExec
tDBCDC
tDBCDCOutput
tDBClose
tDBColumnList
tDBCommit
tDBConnection
tDBInput
tDBInvalidRows
tDBLastInsertId
tDBOutput
tDBOutputBulk
tDBOutputBulkExec
tDBRollback
tDBRow
tDBSCD
tDBSCDELT
tDBSP
tDBTableList
tDBValidRows

ELT in Amazon Redshift with Talend Studio


Talend Cloud Real-Time Big Data Platform - tELTMap - tELTMap_1
Find :
Auto map!
Delimited identifiers
C (alias of table 'CUST_ORDERS.CUSTOMERS')
(IMPLICIT JOIN)
Explicit j...
Column
Operator
Foreign column / expression
customernumber
customername
streetaddress
city
zip
state
dob
ssn
O (alias of table 'CUST_ORDERS.ORDERS')
INNER JOIN
customernumber
=
C.customernumber
amount
ORDERSBYCUSTOMER (table2)
Additional other clauses (GROUP/ORDER BY...) :
GROUP BY C.customernumber, C.customername, C.streetaddress, C.city, C.zip, C.state
Expression
Column
C.customernumber
CUSTOMERNUMBER
C.customername
CUSTOMERNAME
C.streetaddress
STREETADDRESS
C.city
CITY
C.zip
ZIP
C.state
STATE
SUM(O.amount)
TOTALAMOUNT
COUNT(O.customernumber)
ORDERCOUNT
MIN(O.amount)
ORDERMIN
MAX(O.amount)
ORDERMAX
AVG(O.amount)
ORDERAVERAGE
Schema editor
Expression editor
Generated SQL Select query for 'table2' output
C (alias of table 'CUST_ORDERS.CUSTOMERS')
Column
Db Column
Key
DB Type
N..
Length
Precision
Default
Comment
customernumber
INT4
10
customername
VARCHAR
256
streetaddress
city
zip
INT4
10
state
VARCHAR
256
dob
DATE
13
ssn
ORDERSBYCUSTOMER (table2)
CUSTOMERNUMBER
INT
38
CUSTOMERNAME
VARCHAR
100
STREETADDRESS
CITY
ZIP
INT
38
STATE
VARCHAR
100
TOTALAMOUNT
FLOAT
20
ORDERCOUNT
ORDERMIN
ORDERMAX
ORDERAVERAGE
Ok
Cancel

ELT By The Numbers

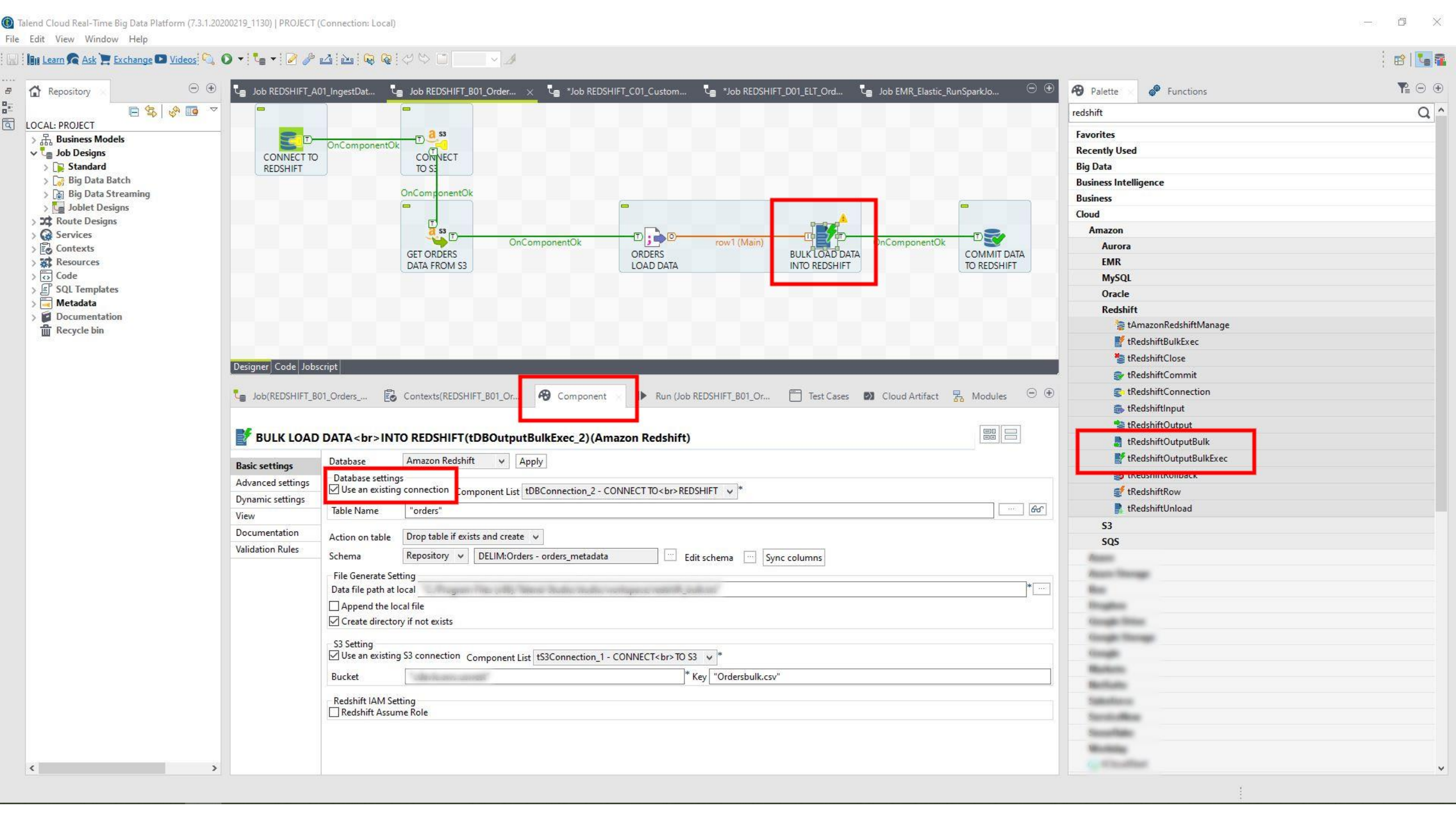
Cloud Data Warehouse Test with ELT vs standard loading

- 134m rows of data
- 26 Gb
- Jobs built in Talend Studio, run in Talend Cloud

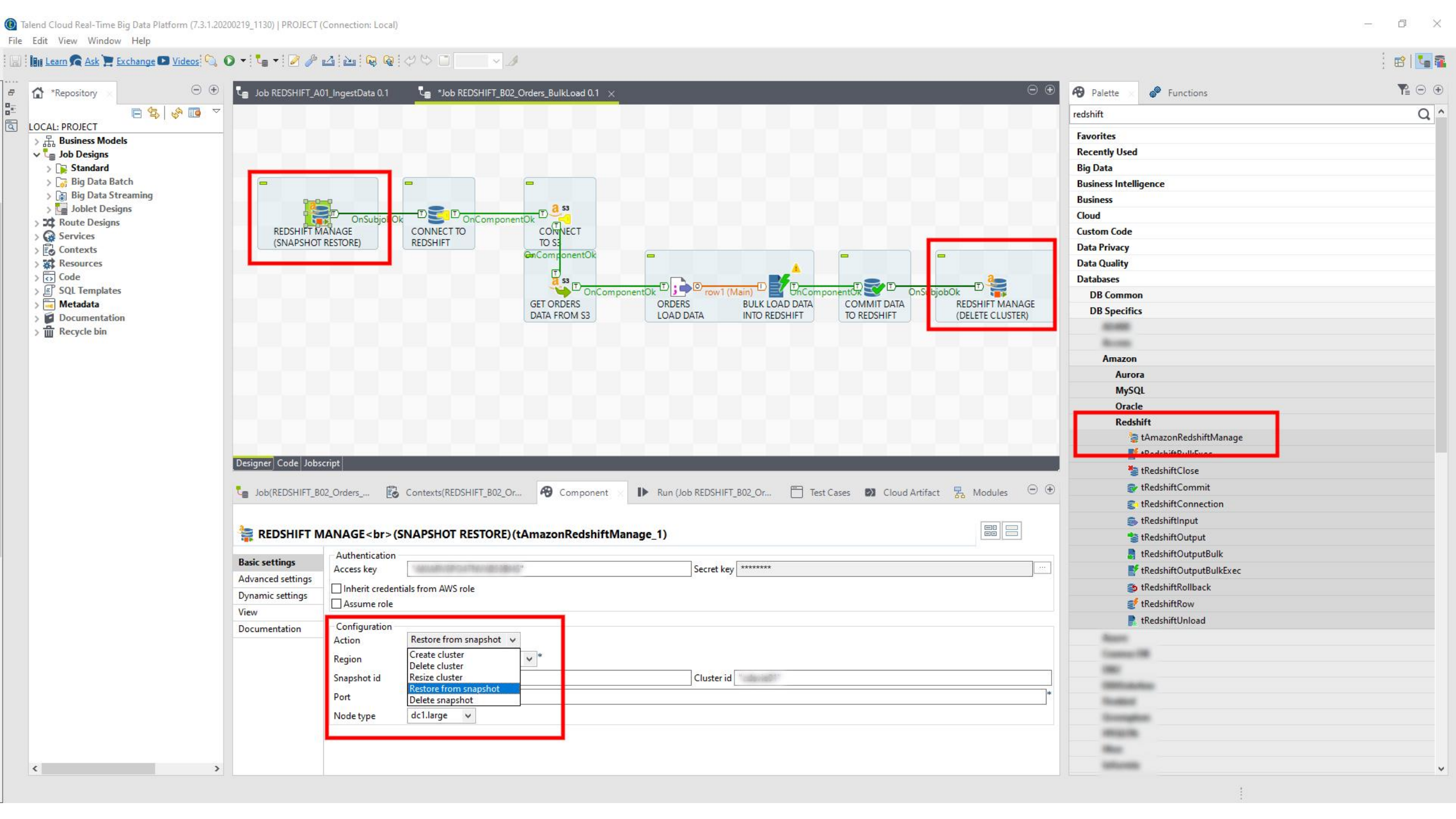
- ELT: 23 seconds
- Standard loading: >20 minutes

Your actual results may differ, however, it is clear that ELT loads data within a cloud data warehouse MUCH FASTER!

Bulk Loading into Amazon Redshift with Talend Studio


Talend Cloud Real-Time Big Data Platform (7.3.1.20200219_1130) | PROJECT (Connection: Local)
File Edit View Window Help
Learn Ask Exchange Videos
Repository
LOCAL: PROJECT
Business Models
Job Designs
Standard
Big Data Batch
Big Data Streaming
Joblet Designs
Route Designs
Services
Contexts
Resources
Code
SQL Templates
Metadata
Documentation
Recycle bin
Job REDSHIFT_A01_IngestDat...
Job REDSHIFT_B01_Order...
*Job REDSHIFT_C01_Custom...
*Job REDSHIFT_D01_ELT_Ord...
Job EMR_Elastic_RunSparkJo...
CONNECT TO REDSHIFT
OnComponentOk
CONNECT TO S3
OnComponentOk
GET ORDERS DATA FROM S3
OnComponentOk
ORDERS LOAD DATA
row1 (Main)
BULK LOAD DATA INTO REDSHIFT
OnComponentOk
COMMIT DATA TO REDSHIFT
Designer Code Jobscript
Job(REDSHIFT_B01_Orders_...
Contexts(REDSHIFT_B01_Or...
Component
Run (Job REDSHIFT_B01_Or...
Test Cases
Cloud Artifact
Modules
BULK LOAD DATA
INTO REDSHIFT(tDBOutputBulkExec_2)(Amazon Redshift)
Basic settings
Advanced settings
Dynamic settings
View
Documentation
Validation Rules
Database
Amazon Redshift
Apply
Database settings
Use an existing connection
Component List
tDBConnection_2 - CONNECT TO
REDSHIFT
Table Name
"orders"
Action on table
Drop table if exists and create
Schema
Repository
DELIM:Orders - orders_metadata
Edit schema
Sync columns
File Generate Setting
Data file path at local
Append the local file
Create directory if not exists
S3 Setting
Use an existing S3 connection
Component List
tS3Connection_1 - CONNECT
TO S3
Bucket
Key
"Ordersbulk.csv"
Redshift IAM Setting
Redshift Assume Role
Palette
Functions
redshift
Favorites
Recently Used
Big Data
Business Intelligence
Business
Cloud
Amazon
Aurora
EMR
MySQL
Oracle
Redshift
tAmazonRedshiftManage
tRedshiftBulkExec
tRedshiftClose
tRedshiftCommit
tRedshiftConnection
tRedshiftInput
tRedshiftOutput
tRedshiftOutputBulk
tRedshiftOutputBulkExec
tRedshiftRollback
tRedshiftRow
tRedshiftUnload
S3
SQS

Amazon Redshift Cluster Management with Talend Studio


Talend Cloud Real-Time Big Data Platform (7.3.1.20200219_1130) | PROJECT (Connection: Local)
File Edit View Window Help
Learn Ask Exchange Videos
*Repository
LOCAL: PROJECT
Business Models
Job Designs
Standard
Big Data Batch
Big Data Streaming
Joblet Designs
Route Designs
Services
Contexts
Resources
Code
SQL Templates
Metadata
Documentation
Recycle bin
Job REDSHIFT_A01_IngestData 0.1
*Job REDSHIFT_B02_Orders_BulkLoad 0.1
REDSHIFT MANAGE (SNAPSHOT RESTORE)
OnSubjobOk
CONNECT TO REDSHIFT
OnComponentOk
CONNECT TO S3
GET ORDERS DATA FROM S3
ORDERS LOAD DATA
row1 (Main)
BULK LOAD DATA INTO REDSHIFT
COMMIT DATA TO REDSHIFT
REDSHIFT MANAGE (DELETE CLUSTER)
Designer Code Jobscript
Job(REDSHIFT_B02_Orders_...
Contexts(REDSHIFT_B02_Or...
Component
Run (Job REDSHIFT_B02_Or...
Test Cases
Cloud Artifact
Modules
REDSHIFT MANAGE
(SNAPSHOT RESTORE) (tAmazonRedshiftManage_1)
Basic settings
Advanced settings
Dynamic settings
View
Documentation
Authentication
Access key
Secret key
****
Inherit credentials from AWS role
Assume role
Configuration
Action
Restore from snapshot
Create cluster
Delete cluster
Resize cluster
Restore from snapshot
Delete snapshot
Region
Snapshot id
Cluster id
Port
Node type
dc1.large
Palette
Functions
redshift
Favorites
Recently Used
Big Data
Business Intelligence
Business
Cloud
Custom Code
Data Privacy
Data Quality
Databases
DB Common
DB Specifics
Amazon
Aurora
MySQL
Oracle
Redshift
tAmazonRedshiftManage
tRedshiftClose
tRedshiftCommit
tRedshiftConnection
tRedshiftInput
tRedshiftOutput
tRedshiftOutputBulk
tRedshiftOutputBulkExec
tRedshiftRollback
tRedshiftRow
tRedshiftUnload

Data Quality with Amazon Redshift with Data Preparation and Talend Studio

Demonstration

Demonstration

Summary

Do you have Talend Studio?

- Go use those components on current or past projects
- ELT, Bulk Loading, Cluster Management

Are you a Talend Cloud customer?

- You have licenses for Data Preparation.
  Go connect to Amazon Redshift today.

Are you new to Talend?

- Go sign up for the Talend Cloud trial today.

talend