

ENABLING DIGITAL TRANSFORMATION



Onepoint Ltd Talend Kudu Components





ENABLING DIGITAL TRANSFORMATION



Contents

Introduction	3
Abstract	3
About Apache Kudu	3
Pre-Requisites	3
Kudu Installation	3
Talend Installation	3
Talend Components Folder Setup	3
Kudu Components Installed	4
Support Materials	5
Example Schema	5
tKuduOutput	5
Example Job 1	5
Step by step instructions	5
Example Job 2	
Step by step instructions	
tKuduInput	
Example Job 1	
Step by step instructions	
Example Job 2	22
Step by step instructions	22
Common Errors	
Requested Replication Factor	
Solution	
Connection Failure	27





INTRODUCTION

ABSTRACT

In this tutorial you can learn how to use the Talend Kudu components created by One point Ltd. These components are:

	Name	Description
	tKuduInput	This is the component used to read data from Apache Kudu.
1808	tKuduOutput	This is the component used to save data from Apache Kudu.

These components are free and can be downloaded from <u>Talend Exchange</u>.

ABOUT APACHE KUDU

Apache Kudu is a revolutionary distributed columnar store for **Hadoop** that enables the powerful combination of fast analytics on fast data. Kudu complements the existing Hadoop storage options, **HDFS** and Apache **HBase**. Additional information on Apache Kudu, its architecture and use cases can be found at (<u>http://getkudu.io/</u>).

At the time of this creation of this document (June 2016) the Apache Kudu is still in beta stage. Onepoint Ltd is planning to release a new version of the components as soon as Apache Kudu 1.0 is released.

PRE-REQUISITES

Kudu Installation

You will need to have Apache Kudu installed in order to be able to use the components. Apache Kudu runs on multiple Linux distributions and can be installed following the instructions on this page:

http://getkudu.io/docs/installation.html

A developer friendly option to be able to develop on one single machine would be to use a Cloudera VM with Linux on which you run Kudu and then have Talend running on the hosting OS.

Talend Installation

You will also need to have at least Talend Open Source 6.0 installed on your machine, in order to be able to use the components. Any of the Talend Enterprise versions would of course also work for this tutorial.

Talend Components Folder Setup

Finally you will need to have the components folder properly setup, so that you can install the components from <u>Talend Exchange</u>. Here are the instructions to do so:

https://help.talend.com/display/KB/Installing+a+custom+component





Components 🔅 🕈 🗸 🛪							
User component folder: D:\dev\onepoint\etg-new\talend\components Browse Data Viewer Row limit 1000							
Mapper Default mapping links display as : Auto							
tRunJob							
Joblet Don't show corresponding job after double click on a Joblet Component.							
Component Assist							
Restore Defaults Apply							

Kudu Components Installed

Finally you should have the Kudu components installed in your Talend Components folder. The easiest way to find the components in Talend Exchange is simply by searching for "Kudu":

Search Results for Latest

Kudu		×
. Oregoint	Kudu Output Onepoint Ltd The Onepoint Kudu components allow integration with Apache Kudu directly from Talend.	압습습습습(0) + 1
• Orapolez	Kudu Input Onepoint Ltd The Onepoint Kudu components allow integration with Apache Kudu directly from Talend.	prec 습습습습습(0) ± 1
		Free





Support Materials

EXAMPLE SCHEMA

The schema used in the examples is always the same. It represents the data of a customer and might be tedious to create manually. For this reason we provide an xml export of the schema which you can use in this tutorial.



kudu_tutorial_schema.xml

In order to import the schema into any of the components mentioned in the examples, please use this button:

kuduinput_1									
Column	Db Column	K	Ту	☑ N	Date	Le	Pr	D	C
		•	_						
* X 0		GQ [R 19						

tKuduOutput

This component allows you to write data to Apache Kudu. It accepts one input flow connection. Furthermore it also supports optional output and reject flow connections.

Optionally the component allows you to create and delete Kudu tables too.

EXAMPLE JOB 1

In this job we will write some dummy data to a Kudu table which will be created in case the Kudu table does not exist yet.

Step by step instructions

1. We will start by creating a standard Talend job (if you are using the "Enterprise version"). If you are using the open source version of Talend you just typically create a normal job.





a. Enterprise version

b La kudu	
Þ 🗋 po 😭	Create Standard Job
🗟 Big B 🔩	Create Job from temprates
🕞 Big D 🗎	Create folder
🚱 Services 🗀	Rename folder
🥃 Joblet D	Expand/Collapse
🛛 🔂 Context 🗙	Delete
🛛 🗔 Code 🛛 🍐	Build Job
🕞 🗐 SQL Ten 😫	Generate Doc As HTML
🛛 🔄 Metadat 🔍	Export items
🖻 🖬 Docume 🗟	Import items
N 🛗 Recycle bin	
b. TOS version	IVUCIJ
b. TOS version	Create job Create folder Expand/Collapse Build Job Import items Export items
b. TOS version	Create job Create folder Expand/Collapse Build Job Import items Export items

2. We will fill the details of the New Job dialogue.

New Job Add a job i	in the repository	
Name	KudoTutorial1	
Purpose	Show how tKuduOutput works	
Description	Show how tKuduOutput works	*
Author	user@talend.com	×
Locker		
Version	0.1	Mm
Status		+
Path	kudu	Select





3.We select the tFixedFlowInput component from the Palette and drop it on the job view panel.

9	9.9	
8		
tFixedF	lowInput_1	

4. We click on the created tFixedFlowInput component and click on the "Edit schema" button.

÷						-
	Schema	Built-In	-	Edit schema		
					_	ŧ

- 5. The schema we are going to create describes a customer. It contains the following fields:
 - a. Email (the primary key)
 - b. Surname
 - c. Given name
 - d. Age
 - e. Country
 - f. Married
 - g. Weight
 - h. Photo
 - i. Profession

j. Insertion Date

Column K email surname age country	Type String String String Integer	N	Date Pattern (Ctr	Length	Precision	Default	Comment
sumame age Country	String String String Integer	 V V 					
surname name age age a	String String Integer	V					
name age age age	String Integer	1					
age Country	Integer						
country		1					
and the second s	String	1				TUKT	
married	Boolean	1					
weight 📃	Double	1					
photo 📰	byte[]	1					
profession 📃	String	1					
insertionDate	Date	1	"dd-MM-yyyy"			new ja	
• × + + = = = = =							





Please note that Kudu always needs a primary key which is in this case the email field.

Hint: alternatively you can import the schema file provided in this tutorial (see chapter Support Materials).

6. Now we create the data for this same component. For this purpose we are going to use an inline table.

* tFixedFlowIn	put_1										100
Rasic settings	Schema Rath Str + Editablema int										
Advanced settings	a Number of roses 1										
Dynamic settings	Mode										
Vev.	O Use Single	Table									
Dolumentation	Of Use Svine 1	Talohe									
Validation Rules	Inline Table	enal	sanana	Name.	401	country	married	angit	photo.	professore	insertionDates
		"giff@gmail.com"	"Femandes"	"Gitterto"	-86	14"	. faiter	72.3	new torte(30%) TL	"Coder"	
		"viatiminbarquiere.	"limitor"	"Valmo"	-45		-false	80,23	netw byte[](7), T1,	"Architect"	
		"juste@gmail.com"	"Vale"	"hequin"	40		false .	73.33	new tayle(3(70, TL		
		"wscheeltz@gmail	"Schweitz"	"West"	- 28	100	THE.	80.12	new byte(30%, 75		
		"Iemando menaton	"Mendonps"	"Fernando"	- 46	*	714	98.12	new tyte()(7), 7)+	"Arihitet"	
			6	the variable artic	hed to this parame	we - NUME					

7. At this point in time we have a fully configured tFixedFlowInput component which can be linked to a tKuduOutput component. Now we search in the palette for the tKuduOutput component which you can typically find in the category "**Databases/Kudu**".



8. We select the tKuduOutput component from the Palette and drop it on the job view panel.



9. Now we connect the tFixedFlowInput component with the tKuduOutput component.







10. The tKuduOutput connection needs to be configured. We click on the tKuduOutput component and change the data in the "Basic settings" view. You have to set all parameters on this panel:

a. Server – The name of the server on which Apache Kudu is running. Please note that on test environments you might have to change the hosts file to map the name to a specific IP address.

b. Port – The port on which Apache Kudu is running.

c. Table name – The name of the table which is going to store the data.

d. Create table – The table creation options. We have chosen "Delete if exists and create again", because we want to guarantee that this example runs without errors.

e. Operation – The data operation to be executed by this component. In this case we are going to insert data.

🖕 JobeKudoTutoria	(101) 🔂 Cor	itexts(KudoTutorial1)	49 Component	I Run Uob KudoTutorial1)	Test Cases	Integration Action	Θ
KudoTutorial1	0.1					# 8	
Basic settings	Schema	Built-In * Edi	t schema 🔲 Syne	columns			
Advanced settings	Datasource	-	_				
Dynamic settings	Server	"quickstart.cloudera		Port 7051			
View	Table param	eters					
Documentation	Table name	"customers"		Create toble	ete if exists and crea	ste again	
Validation Rules	Operation	Insert -					

11. (Optional) If you have started Kudu on a Cloudera distribution VM or on a simple VM, most probably you will need to set the number of replicas to 1.

🚛 Job(KudoTutorial	10.1) 🔂 Contexts(KudoTutorial1)	Component	I► Run (lob KudoTutorial1)	📋 Test Cases	Integration Action	•
🚝 tKuduOutput	12					
Basic settings Advanced settings	Replica number 1	1				-
Dynamic settings View Documentation	Rood session configuration Flush action Auto flush Auto flush background - Every write Auto flush - Every write sent to the	+ *	n the background			
Validation Rules	Ignore all duplicate rows					
	Fail on operation error tStatCatcher Statistics Enable parallel execution					

12. Now we can run the job and see, if everything is ok.

🐛 Job(KudoTutorial1.0.1) 🔓 Contexts(KudoTutorial1) 🕫 Component 🕩 Run (Job Ku	idoTutorial1) 📋 Test Cases 🛛	Integration Act	ion ©
Job KudoTutorial1		Default	
Basic Run Execution		Name	Value
Debug Run 🔛 Kill 🗽 Clear		1.12141.020	
Advanced settings	1		
Target Exec			
Memory Run		>	





13. In case of success you should see something like this on Talend Studio:

- tFixedFlo	5 rows in 1.12s
Designer Code Jobs	cript
Job(KudoTutorial	1 0.1) 🛃 Contexts(KudoTutorial1) 🔂 Component 🕩 Run (Job KudoTutorial1) - 🗌 Test Case
JOD KUGO I UTORIA	
Basic Run	Execution
Debug Run	Run Kill 🕅 Clear
Advanced settings	Starting job KudoTutorial1 at 14:59 22/06/2016.
Target Exec	
Memory Run	[statistics] connected 312 [New I/O worker #1] INFO org.kududb.client.AsyncKuduClient - Discovered tablet Kudu Master for table Kudu Master with partition ["", "") 945 [New I/O worker #1] INFO org.kududb.client.AsyncKuduClient - Discovered tablet caddf171bac44e8bb198e05293377485 for table customers with partitics] disconnected Jab KudoTutorial1 ended at 14:59 22/08/2016. [exit code=0]

In case of errors, please check the Common Errors chapter

EXAMPLE JOB 2

In this job we will write some dummy data to a Kudu table. Some of this data will be correct and some of this data will violate the primary key contract and will be rejected.

Step by step instructions

1. We will start by creating a standard Talend job (if you are using the "Enterprise version"). If you are using the open source version of Talend you just typically create a normal job.

a. Enterprise version

Image: Participation of the second	
Þ 🗋 po 📬	Create Standard Job
🗟 Big D 🖢	Create Job nom templates
🔓 Big D 🗋	Create folder
🚱 Services 🗋	Rename folder
🥃 Joblet D	Expand/Collapse
🖻 🔂 Context 🗙	Delete
🛛 🐼 Code 🛛 💩	Build Job
🕞 🗐 SQL Ten 😫	Generate Doc As HTML
🖻 🗮 Metadat 🔍	Export items
🕞 🖬 Docume 🧕	Import items
-	
b. TOS version	1
b. TOS version	mouch
b. TOS version	Create job
b. TOS version	° Create job
b. TOS version	Create job Create folder
b. TOS version	Create job Create folder Expand/Collapse
b. TOS version	 Create job Create folder Expand/Collapse Build Job
b. TOS version	 Create job Create folder Expand/Collapse Build Job
b. TOS version	 Create job Create folder Expand/Collapse Build Job Import items
b. TOS version	 Create job Create folder Expand/Collapse Build Job Import items Export items
b. TOS version	 Create job Create folder Expand/Collapse Build Job Import items Export items





2. We will fill the details of the New Job dialogue.

Name KudoTutorial2 Purpose Show how tKuduOutput rejections work Description Show how tKuduOutput rejections work Author user@talend.com Locker	New Job Add a job i	n the repository	
Purpose Show how tKuduOutput rejections work Description Show how tKuduOutput rejections work Author user@talend.com Locker	Name	KudoTutorial2	
Description Show how tKuduOutput rejections work Author user@talend.com Locker Version 0.1 M Status Path kudu Sele	Purpose	Show how tKuduOutput rejections work	
Author user@talend.com Locker 0.1 M Status Path kudu Sele	Description	Show how tKuduOutput rejections work	*
Author user@talend.com Locker 0.1 M Status 2014 Sele			Ŧ
Locker Version 0.1 M Status Path kudu Sele	Author	user@talend.com	
Version 0.1 M Status Path kudu Sele	Locker		
Status Sele	Version	0.1	Mm
Path kudu Sele	Status		+
	Path	kudu	Select

3. We select the tFixedFlowInput component from the Palette and drop it on the job view panel.



4. We click on the created tFixedFlowInput component and click on the "Edit schema" button.

iput_t



5. The schema we are going to create describes a customer. It contains the following fields:

- a. Email (the primary key)
- b. Surname
- c. Given name
- d. Age
- e. Country
- f. Married
- g. Weight
- h. Photo





i. Profession

		D .
	Incortion	1)0+0
	INSELION	LIAIP
- I +	1113011011	Duic

Reariowinput								
Column	K	Туре	🗹 N.	Date Pattern (Ctr	Length	Precision	Default	Comment
🥄 email	1	String	E3					
sumame	10	String	1					
name		String	1					
age	13	Integer	1					
country	10	String	1				TUKT	
married	83	Boolean	1					
weight	10	Double	1					
photo	10	byte[]	~					
profession		String	1					
insertionDate	13	Date	1	"dd-MM-yyyy"			new ja	
	R							

Please note that Kudu always needs a primary key which is in this case the email field.

If you have completed the first job in this tutorial you can simply copy / paste the schema fields using the copy / paste buttons (
). Or you can simply import the schema file provided in this tutorial (see chapter Support Materials).

6. Now we create the data for this same component. For this purpose we are going to use an inline table.

asis settings	Schema	Bull-In	athema 💷								
duarword settings	Number of cos	#1.									
ynumic wettings	Mode										
ew.	Use Single	fable									
and a feature of the local data	Of Use Svine T	able									
Dolumentation											
Indebox Rules	Inine Table	enal	sanane	7474	401	country	married	aright	photo	professor	inentorDat
Indedices Rules	Inine Table	enal "gittilgnat.com"	sanane "Versaides"	"Giberto"	404 44	country '\A'	married fater		photo new byte()(70, TL	profession "Coder"	inantonDat
initation Rufes	brine Table	enall "giff@gnat.com" "Vallmir.bars.kow.	sature "ferundes" "Basukov"	"Giberto" "Vadimo"	404 45	country "\A"	manied fater fater	949ght 72.3 90.23	photo new tyte()(70, TL new byte()(70, TL	profession "Coder" "Architect"	nartorDa
distation Rules	biline Table	enall "gift@gnat.com" "statimicbanukou" "yvere@gnat.com"	sinane "ferandes" "Sanakov" "Vae"	Name "Giberts" "Viatims" "Naquim"	404 46 45 40	country Var	married False False	800ght 72.3 8023 73.33	photo new tryte(3(75; TL new tryte(3(75; TL new tryte(3(75; TL	profession "Coder" "Architect"	nertorDa
entación Rutes	bline Table	enal "gitt@gnat.com" Vistimir.banukou" "yiste@gnat.com" "wicheetu@gnat.	sanane "fenantes" "Sanatov" "Vale" "Schweitt"	Name "Giberts" "Valtimu" "Nagum" "Wat"	404 41 40 21	country 'W'	manied fator fator fator fator fator	940ght 72.3 9023 73.23 80.12	photo new tyte(2)70; TL new tyte(2)70; TL new tyte(2)70; TL new tyte(2)70; TL	profession "Coder" "Architect"	nertorDe

If you have completed the first job in this tutorial you can simply copy / paste the data fields using the copy / paste buttons (

7. Now we are going to duplicate the first row of the tFixedFlowInput inline table component. We

-				and the second se	and the second se	The second se	Contraction of the local division of the loc	A Real Property lies and the lies of the l	and the second se
gill@gnat.co	s" "Feinandes"	"Gitberto"	46	Tak'	false	72.3	new byte()(70, 71,	"Coder"	
VILLENIN COME	55. 1015000V	10500	45		1955	1027	NEW BYREDING PL -	wesser	
"ywte@gmail.c	m" "Vale"	"niupeol"	40	1000	Table	73.23	new byte()(20, 71,		
*wstweitz@gr *lemando.men	uel. "Schweitz" Ion. "Mendor	2. Click on the	copy button	'dk' Tin'	true true	1. Click on this	row styte[](70, 71, .	"Avchitect"	
Ne Table									
re Table email	antane	name "Witherto"	404 45	country	married faber		photo any bytef000, 71	profession	insertionDa
e Table e Table e Table f Gittigenalize f gittigenalize	Luname T [*] 'Senados' S [*] 'Tenados'	name (Vidherto) "Gilberto"	age Ali Ali	country "cat" "sat"	married Solar Salar	12.3 72.3	photo one futer(10), 71 new byter](70, 71, _	profession "Coder"	mertonDa
re Table e Table e mai 'oit9omatao Vatimitiani	surtaine n° <u>formandes'</u> n° Temandes' ios. fiturisaov'	name "Oritecto" "Gilberto" Wadmur	म्प्रस <u>क</u> म्ह	country "ad" "ad"	married fater fater	mmight 72.3 70.21	photo any bytef(120, 71 new bytef(120, 71, - new bytef(120, 71, -	profession "Coder" "Coder" "Architect"	insertionDa
e Table email 'n Montal of 'n M	sunane s' Tensedes' s' Tensedes' co. Tensedos' cor 'Vala' set Sinuciti	name "Vitterto" "Giberto" Viddmu "Koaqu "Naqu 3. C	ege 46 45 lick on the past	e button	married failer failer failer	122 723 6071 4. A new row sho	photo one batel(20, 7) new bytel(20, 7), new bytel(20, 7), totol (20, 7), totol (profession "Coder" "Coder" "Architect"	insertionDe





8. At this point in time we have a fully configured tFixedFlowInput component which can be linked to a tKuduOutput component. Now we search in the palette for the tKuduOutput component which you can typically find in the category **"Databases/Kudu**".

🛿 Palette 🛛	₹ ⊖ €
tKuduOutput	Q
Favorites	
Recently Used	
💐 tKuduOutput	
Databases	
KUDU	
💐 tKuduOutput	
Misc	
note	

9. We select the tKuduOutput component from the Palette and drop it on the job view panel.



10. Now we connect the tFixedFlowInput component with the tKuduOutput component.



11. The tKuduOutput connection needs to be configured. We double-click the tKuduOutput component and change the data in the "Basic settings" view. You have to set all parameters on this panel:

a. Server – The name of the server on which Apache Kudu is running. Please note that on test environments you might have to change the hosts file to map the name to a specific IP address.





b. Port – The port on which Apache Kudu is running.

c. Table name – The name of the table which is going to store the data.

d. Create table – The table creation options. We have chosen "Delete if exists and create again", because we want to guarantee that this example runs without errors.

e. Operation – The data operation to be executed by this component. In this case we are going to insert data.

KudoTutorial1	01			
Basic settings	Schema	Built-In * Edit schema	Sync columns	
Advanced settings	Datasource			
Dynamic settings	Server	"quickstart.cloudera"	Por 7051	
View	Table param	eters		
Documentation	Table name	"customers"	Create t. ple Delete if exists and create again 🐨	
Validation Rules	Operation	Insert -		

12. (Optional) If you have started Kudu on a Cloudera distribution VM or on a simple VM, most probably you will need to set the number of replicas to 1.

- Joo(kudo rutonal	TOT) EQ CONCENTS(KD00 (DEOUBIL)	o Component	Ruh (Job Kubo Lutonal1)	L] Test Cases	sa integration Action	0.0
👫 tKuduOutput	2					
Basic settings Advanced settings	Replica number 1	1				
Dynamic settings View Documentation	Flush action Auto flush Auto flush background - Every write	+ *	n the background			
Validation Rules	Ignore all duplicate rows	CIVCI				
	Fail on operation error tStatCatcher Statistics Enable parallel execution					

13. Now double-click on the tKuduOutput component and select the "Advanced Settings" tab.

tFis	row1 (Main) kedFlowInput_1
Designer Code Job	script
Job(KudoTutoria	al2 0.1) 🔀 Contexts(KudoTutorial2) 🔗 Component 🔹 🕩 Run (Job KudoTu
🚝 tKuduOutpu	t_1
Basic settings	Use defaults
Advanced settings	Replica number 3
Dynamic settings	Kudu session configuration
View	Flush action Auto flush
Documentation	Auto flush background - Every write sent to the server in the background
Validation Rules	Ignore all duplicate rows
	Fail on operation error
	TStatCatcher Statistics
	Enable parallel execution





14.Now untick the "Fail on operation error" checkbox.

tKuduOutput	L <u>1</u>
Basic settings	Use defaults
Advanced settings	Replica number 1
Dynamic settings	Kudu session configuration
View	Flush action Auto flush
Documentation	Auto flush background - Every write sent to the server in the background Auto flush - Every write sent to the server
Validation Rules	Ignore all duplicate rows
	Fail on operation error
	tStatCatcher Statistics
	Enable parallel execution

15. Now search in the palette for a tLogRow, select it and drop it on the job view panel.



16. Now click with the right mouse on the tKuduOutput component and select "Reject". After that, drag the reject connector onto the tLogRow component.







-								
	a	row1.0	(aic)		·			
	tFixedFlowInput_1	TOWLO	VIGILIY	tKuduOutpu	ut_1			
						row2 (Rej	ect)	
						1	0	0.
							tL	ogRow_1
Deciment Code	labergint							
Designer Code	ouscript		254			_	_	_
te Job(KudoTute	orial2 0.1) 🐻 Co	ontexts(KudoTut	orial2) 🕫	Component	. I≻	Run (Job Ki	udoTutorial2)	Test
tLogRow_	1							
Basic settings	Schema	Built-In	* Edit sci	nema 📄 Syno	c colum	ns		
	s Mode							
Advanced setting	and the second se							
Advanced setting	Basic							
Advanced setting Dynamic setting View	Basic Table (pr	int values in cell	s of a table)					
Advanced setting Dynamic setting View Documentation	Basic Table (pri Vertical (e)	int values in cell each row is a ke	s of a table y/value list)					

17. Now you can double click on the tLogRow component and select the "Table" mode.

18. Now we are going to add another tLogRow to this job and connect the tKuduOutput component to it using a regular row connector.

	tKuduQi	Row	Main	
		Trigger	•	
		Undo Redo		
	3	Copy Paste Delete Select All	v_1	
-			- Ting	
tFixe	edFlowInput_1	row1 (Main)	tKuduOutput_1	tain order:1) tLogRow_2
			row2 (Reject on	der:2)





	0	6 rows in 1.18s 5.09 pwg/s	0	5 rows in 1.18s	0
tFix	edFlowInput_1	row1 (Main)	tKuduOutput_1	1 rows in 1.18s 2.85 roms/s 2 (Reject order:2)	tLogRow_2
				tLogRov	v_1
Code Jobs	script				
Code Job KudoTutoria	script (2.0.1) 🔂 Conte	xts(KudoTutorial2)	Component	tun (Job KudoTutorial2) 📋	Test Cases 🛛 In
Code Job (udoTutoria gRow_2 ttings	script (2 0.1) 🔁 Conte Schema	xts(KudoTutorial2)	Component IF F	tun (Job KudoTutorial2) 📑	Test Cases 🛛 In
Code Jobs (udoTutoria gRow_2 ttings d settings	Schema Mode Basic	xts(KudoTutorial2) 6	Component IF F	lun (Job KudoTutorial2) 📑	Test Cases 🛛 In
Code Jobs KudoTutoria gRow_2 ttings td settings : settings	Script 12 0.1) Conte Schema Mode Basic Table (print v	xts(KudoTutorial2)	Component IP F	lun (Job KudoTutorial2) 📋	Test Cases 🛛 DI In
r Code Jobs KudoTutoria gRow_2 sttings ed settings c settings intation	Schema Schema Mode Basic Verticar (eacl	xts(KudoTutorial2)	Component IF F	tun (Job KudoTutorial2) 📑	Test Cases 🛛 In

19. Also double-click on the tLogRow_2 component and please choose the "Table" mode.

20. Now we run the job and if everything goes well, you should see that most of the rows except one are printed out by the tLogRow_2 component. One row will be rejected though, due to a duplicate primary key.

	C.R.	6 rows in 1.14s	-	CINK-	Sito	vis in 1.15:		100 kg			
tFixe	edFlowInput_1	Tow I (Mdin)	tKudu	Output_1	1 rows in 1: 0.82 roserva www.Report of	Main order 15s (der:2)	di di di di di di di di di di	tLogRow			
se Jobs Tutorial	cript	ts(KudoTutorial2)	6 Component	ent 🕨 Ru	ın (Job KudoT	(utorial2)	🗇 Test (Cases D3	Integrat	ion Action	
utoria	12										
	Execution		201002								
	Execution Run	iii Kiil	Clear								
tings	Execution Run 957 [Now 1/2	worker #1] IN	Clear F0 org.ku	dudb.clie	nt.AsyncKu	duClien	t - Discov	ered tab	let 4be	a6156c189	455bbfcb3
tings	Execution Run 957 [Now 1/ 990 [main] 1	worker #1] IN ERROR component	Clear F0 org.ku s_develops	dudb.clie ment.kudo	nt.AsyncKu tutorial2_	duClien 0_1.Kud	t - Discow oTutorial?	red tab - Row e	let 4be rror fo	a6156c189 or primary	455bbfcb3 key-"gil
tings	Execution Run 957 [Now 1/0 990 [noin] 1 1	Worker #1] IN RROR component	Clear FO org.ku s_develop	dudb.clie ment.kudo	nt.AsyncKu tutorial2_ tLo	duClien 0_1.Kud gRow_2	t - Discov oTutorial2	rred tab - Row e	let 4be rror fo	a6156c189 primary	455bbfcb3 key-"gil
ttings	Execution Run 957 [New Ivi 990 [moin] 1 1 1 1 1 1 1 1 1 1 1 1 1 1	Worker #1] IN RROR component	Clear FD org.ku s_develops	dudb.clie sent.kudo surname	st.AsyncKu tutorial2 tLo	duClien 0_1.Kud gRow_2 age cou	t - Discow oTutorial2 atry[marrie	rred tab - Row e d weigh	let 4be rror fo t photo	a6156c189 pr primary	455bbfcb3 key-~gil
ttings	Execution Run 957 [Now 1× 960 [main] 1 email gilf@gmail vladimir.pr jvale@gmail Hechneitz@gmail	com blev@gmail.com ndonca@onepoin	Clear S_develops	dudb.clie sent.kudo surname Fernandes Polev Vale Schweitz Mendonça	nt.AsyncKu tutorial2 tLo bame Gilberto Vladimir Joaquim Wim Fernando	duClien 0_1.Kud gRow_2 age cou 46 luk 45 lUK 40 LUK 35 ldk 46 lin	t - Discov oTutorial2 htty morriv false false false false true	red tab - Row e d weigh 172.3 190.23 173.23 180.12 198.12	let 4be rror fo IFGH IFGH IFGH IFGH IFGH	of professi professi Coder Architec Inull Inull Architec	455bbfcb3 key-*gil 123-06- 123-06- 123-06- 123-06- 123-06- 123-06-
ttings	Execution Run 957 [New I× 990 [main] 1 email silf@gmail Vladimir.pr jvale@gmail kenne tz@gmail	com >lev@gmail.com should component: .com should component: .com mail.com	Clear Boorg.kur Solevelops I I I I I I I I I I I I I I I I I I I	dudb.clie sent.kudo surname Pernandes Polev Vale Schweitz Mendonça	nt.AsyncKu tutorial2 ttc bame Gilberto Vladimir Joaquim Wim Fernando	duClien 0_1.Rud gRow_2 age cou 46 uk 45 UK 45 UK 45 UK 35 dk 46 in	t - Discov oTutorial2 htry[morri- [felse [felse [felse [true]true	rred tab - Row e d weigh 72.3 90.23 173.23 80.12 98.12	let 4be rror fo FGH FGH FGH FGH FGH	of profession (profession) (Coder (Architec Inul) Inul) (Architec	455bbfcb3 key-"g11 11 123-06- 123-06- 123-06- 123-06- 123-06- 123-06-
ttings	Execution Run 957 [New 1/2 990 [main] 1 email gilf@gmail yvale@gmail yvale@gmail	COM COM COM COM COM COM COM COM	Clear S_develops	dudb.clie ment.kudo surname Pernandes Polev Vale Schweitz Mendonça	nt.AsyncKu tutorial2 tLo bame Gilberto Vladimir Joaquim Fernando	duClien 0_1.Kud gRow_2 age cou 46 luk 45 lUK 40 lUK 35 ldk 46 lin	t - Discov oTutorial2 htry marrin [false false false true true	rred tab - Row e 	let 4be rror fo IFGH IFGH IFGH IFGH	offscille professi Coder Architec Inull Inull Architec	455bbfcb3 key-"gil anlinsert (23-06- (23-06- (23-06- (23-06-
ttings	Execution Run 957 [New 1/2 990 [main] 1 email ruladimir.pc formando.me email	com pail.com mail.com jsurname j	Clear S_develops	dudh.clie ment.kudo surname Pernandes Polev Zehweitz Mendonça	nt.AsyncKu tutorial2 tLo laame Gilberto Vladimir Joaquim Fernando	duClien 0_1.Kud gRow_2 age cou 46 luk 45 UK 45 UK 35 ldk 46 lin +	t - Discov oTutorial2 ntry marrin [felse [felse [felse [true [true] [photo]pro	red tab - Row e rd weigh 72.3 90.23 73.23 80.12 98.12 98.12 	let 4be rror fo FGH FGH FGH FGH FGH FGH	osDate er	455bbfcb3 key-"gil on insert [23-06- [23-06- [23-06- [23-06- [23-06-





tKuduInput

This component allows you to read tabular data from Apache Kudu tables. You can either scan through the whole table or you can use query filters. This tutorial contains two example jobs, one demonstrating a scan and another one demonstrating how the end user can use the query fields.

Warning: you should have executed before proceeding either Example Job 1 or Example Job 2.

EXAMPLE JOB 1

In this example job you will learn how to setup the tKuduInput component and how to perform a full table scan on a Kudu component.

Step by step instructions

1. We will start by creating a standard Talend job (if you are using the "Enterprise version"). If you are using the open source version of Talend you just typically create a normal job.

a. Enterprise version







2. Here are the details of the created job.

New Job Add a job i	n the repository
Name	KudoTutorial3
Purpose	Demonstrate how to create a job which performs a full scan on a Kudu table
Description	Demonstrates how to create a job which performs a full scan on a Kudu table
Author	user@talend.com
Locker	
Version	0.1 M m
Status	
Path	kudu Select

3. We select the tKuduInput component from the Palette and drop it on the job view panel.

Palette × Y ² ○ ●	
tKuduInput Q ^	
Favorites	1 m
Recently Used	
🎼 tKuduInput	tKuduInput_1
Databases	
KUDU KuduInput Misc	
note	

4. We double-click on the created tKuduInput component and click on the "Edit schema" button.

-

Schema	Built-In	-	Edit schema	 l
Schema	Dunt In		Luit schema	ł.

5. The schema we are going to create describes a customer. It contains the following fields:

- a. Email (the primary key)
- b. Surname
- c. Given name
- d. Age
- e. Country
- f. Married





- g. Weight
- h. Photo
- i. Profession
- j. Insertion Date

Column	Db Column	K	Type	28 N.	Date Pattern (Ctrl+Sp_	Length	Precision	Default	Comment
🤽 email	email	12	String	13	1				
sumame	sumarrie	10	String	192					
name	name	0	String	2					
age	age	0	Integer	2					
country	country	10	String	925				"UK"	
married	married	0	Boolean	125					
weight	weight	10	Double	.(2)					
photo	photo	173	byte[]	R.					
profession	profession	123	String	125					
insertionDate	insertionDate	10	Date .	4	"dd-MM-yyyy"			new java.	-

Hint: if you have already gone through the previous example jobs (Example Job 1, Example Job 2) you can simply copy the schema from the tFixedInputFlow component. Or you can simply the schema file provided in this tutorial (see chapter Support Materials).

6. The tKuduInput component needs to be configured. We double-click the tKuduInput component and change the data in the "Basic settings" view. You have to set all parameters on this panel:

a. Server – The name of the server on which Apache Kudu is running. Please note that on test environments you might have to change the hosts file to map the name to a specific IP address.

- b. Port The port on which Apache Kudu is running.
- c. Table name The name of the table which is going to store the data.

d. Query type – The selected value should be "Scan the whole table".

👫 tKuduInput_	L			
Basic settings	Schema	Built-In * Edit schema		
Advanced settings	Data source			
Dynamic settings	Server	"quickstart.cloudera"	* Port 7051	
View	Table param	eters		
Documentation	Table name	"customers"		
Validation Rules	Query type Scan the w	hole table D User defined query		
	Limit			

Please note: the limit is currently in the server side.

7. We select the tLogRow component from the Palette and drop it on the job view panel.

🛛 Palette 🛛	Y 😑 🖯	6	
tLogRow	Q		
Favorites		The second	
Recently Used ⁻ tLogRow		tKuduInput_1	tLogRow_1
Logs & Error ^전 tLogRov	Displays the flow ontent (rows) on the		
Misc R	lun Job console		
note			





8. Now double-click on the tLogRow component and select the "Table" mode.

- tKudu	Input_1	row1 (Main)		tLogRow	1	
Decigner Code John	crint					10
Job(KudoTutoria	13 0.1) 🐻 C	ontexts(Kudo]	utorial	3) 49 Com	ponent	I► Run ()
tLogRow_1						
Basic settings	Schema	Built-In	•	Edit schema	Sync	columns
Advanced settings	Mode					
Dynamic settings	Basic			_		
View	Table (p	rint values in c	ells of a	a table)		
Documentation	Vertical	(each row is a	key/val	ue list)		
Validation Rules	Print con	tent with log4j				

9. Now we create a row connection from the tKuduInput component to the tLogRow component.



10. The job can now be executed. In case of success you will see the following:

tKudul	5 rows in 0.79s row1 0Main) hput_1 tLogRow_1
Designer Code Jobs	copt
ta Job(KudoTutorial	3 0.1) 🚯 Contexts(KudoTutorial3) 🗝 Component 🕨 Run (Job KudoTutorial3) 👘 Test Cases 🕫 Integration Action
Job KudoTutoria	13
Basic Run	Execution
Advanced settings	Characteristic Characterist 2 and 12
Target Exec	Starting job Kuudulutariais at 12225 2000-2010.
Memory Run	[statistics] connected 345 [New I/O worker #1] INFO org.kududb.client.AsyncKuduClient - Discovered tablet Kudu Master for table 555 [New I/O worker #1] INFO org.kududb.client.AsyncKuduClient - Discovered tablet 4bea5156c189455bbfcb3
	tLogRow_1
	email surname name age country married weight photo profession insert
	fernando.mendonca@onepointltd.com Mendonça Fernando 46 in true 98.12 FGH Architect 23-06- gilf@gmail.com Fernandes[Gilberto 46 uk false 72.3 FGH [Coder 23-06- jvale@gmail.com Vale Joquim 40 UK false 73.23 FGH [Null 23-06- vladimir.polev@gmail.com Polev Vladimir 45 UK false 90.23 FGH Architect 23-06- wschweitz@gmail.com Schweitz Wim 35 dk true 80.12 FGH [null 23-06-





EXAMPLE JOB 2

In this example job you will learn how to setup the tKuduInput component and how to perform a user defined queries scan with the tKuduInput component.

Step by step instructions

1. We will start by creating a standard Talend job (if you are using the "Enterprise version"). If you are using the open source version of Talend you just typically create a normal job.

a. Enterprise version



2. Here are the details of the created job.

New Job Add a job i	n the repository	
Name	KudoTutorial4	
Purpose	Demonstrate how to use user defined queries in the tKuduInput component	
Description	Demonstrates how to use user defined queries in the tKuduInput component	*
		-
Author	user@talend.com	
Locker		
Version	0.1	Mm
Status		
Path	kudu	Select





3. We select the tKuduInput component from the Palette and drop it on the job view panel.

\varTheta Palette 🛛	¶≘ ⊖ ⊕	
tKuduInput	Q ^	
Favorites		
Recently Used		
💐 tKuduInput		tKuduInput_1
Databases		
KUDU KuduInput Misc		
note		

4. We double-click on the created tKuduInput component and click on the "Edit schema" button.

Schema	Built-In	-	Edit schema	

5. The schema we are going to create describes a customer. It contains the following fields:

- a. Email (the primary key)
- b. Surname
- c. Given name
- d. Age
- e. Country
- f. Married
- g. Weight
- h. Photo
- i. Profession
- j. Insertion Date

Column	Db Column	K.,	Type	2 N.	Date Pattern (Ctrl+Sp_	Length	Precision	Default	Comment	
🔍 email	email	12	String	13	20.181					
sumame	sumarrie	10	String	192						
name	name	10	String	2						
age	age	- 83	Integer	2						
country	country	83	String	92				"UK"		
married	married	10	Boolean	8						
weight	weight	10	Double	12						
photo	photo	103	byte[]	98						
profession	profession	123	String	121	in the second					
insertionDate	insertionDate	0	Date .	1	"dd-MM-yyyy"			new java.	3	

Hint: if you have already gone through the previous example jobs (Example Job 1, Example Job 2) you can simply copy the schema from the tFixedInputFlow component. Or you can import simply the schema file provided in this tutorial (see chapter Support Materials).

6. We are going first to create a query which filters out all customers which are older than 40. In order to create such a query, double click on the tKuduInput component and select "User defined query"





/ fields	Column	Operator	Value	

7. Now add one line to the query by pressing the " is button. Select the "age" column and the "GREATER" operator. Write into the "Value" field "40" (with no quotes).

Column	Operator	Value	
age	GREATER	40	
	The variable attache	ed to this parameter is:QUERY_FIELDS	

8. We select the tLogRow component from the Palette and drop it on the job view panel.



9. Now double-click on the tLogRow component and select the "Table" mode.

	nu o	row1 (Main)		
tKudu	Input_1			tLogRow_1	
Designer Code Jobs	script				
Job(KudoTutoria	13 0.1) 🔂 C	ontexts(Kudo	Tutorial	3) 🗝 Compo	onent 🕨 🕨 Run (
tLogRow_1					
Basic settings	Schema	Built-In	•	Edit schema	Sync columns
Advanced settings	Mode				
Dynamic settings	O Basic		10100	_	
View	Table (p)	rint values in o	cells of a	a table)	
Documentation	Vertical ((each row is a	key/val	ue list)	
	Print cont	tent with log4	1		





10. Now we create a row connection from the tKuduInput component to the tLogRow component.



11. Now you can run the job for the first time and you will see that all customers listed on the console are 40+ of age:

3 rows in 0.55s 5 row1 (Main) tKudulnput_1	tLogRow_1						
obscript crial4 0.31 🐻 Contexts(KudoTutorial4)	A Component D R	un Uob Kudo	Tutorial4)	FT Test G	ases Ø3	Integrat	ion Action
orial4	the second second second						
Execution	Clear						
Starting job KudoTutoria. [statistics] connecting the [statistics] connected 347 [New I/O worker #1] I 394 [New I/O worker #1] I	4 at 13:06 23:06/2 o socket on port 34 NFO org.kududb.clis NFO org.kududb.clis	1016. 462 ent.AsyncKv ent.AsyncKv	uduClien uduClien	- Discove - Discove	red tabl	et Kud et 4be	lu Master fo a6156c18945
		tL	ogRow_1				
email	surname	name	age cou	try marrie	d weight	photo	profession
fernando.mendonca@onepoi gilf@gmail.com vladimir.polev@gmail.com	ntltd.com Mendonça Fernandes Polev	Fernando s Gilberto Vladimir	46 in 46 uk 45 UK	true false false	98.12 72.3 90.23	FGH FGH FGH	Architect Coder

12. Now let us change the existing filter and try to find a user by email address. Double click on the tKuduInput component and remove the existing filter and add the following filter:

Query fields	Column	Operator	Value	
	email	EQUALS	"gilf@gmail.com"	
			The variable attached to this parameter is:QUERY	FIELDS_
			The variable anacles to this parameter is _Quert.	JACKNO

13. Now run the job again and you will see that there is only one single entry in the output.

xecution							
► Run	III Kill 🕅 Kill						
Starting jo	b KudoTutorial4 at	lear performance	tracking and	trace trac	king		
[statistics [statistics 393 [New I/0 437 [New I/0] connecting to soc] connected 0 worker #1] INFO o 0 worker #1] INFO o	ket on port 4 rg.kududb.clip rg.kududb.clip	028 ent.Asynck ent.Asynck	uduClie uduClie	nt - D nt - D	iscovered iscovered	tablet Kudu Mast tablet 4bea6156c
			+				
email	surname name	age coun	try marrie	d weigh	t phot	o professi	ion insertionDate
gilf@gmail	.com Fernandes Gilb	erto 46 uk	false	172.3	FGH	Coder	23-06-2016
[statistics Job KudoTut] disconnected corial4 ended at 13:	12 23/06/2016	. [exit c	de=0j			





14. Let us now create a combined filter which filters by age and by country. Add the following lines to the query fields:

Query fields	Column	Operator	Value	
	age	GREATER	40	
	country	EQUALS	"in"	

15. Now run the job again and you will see all customers which are associated to "in" and over 40.

tKuduinput_1	1 rows in 0.55s	tLogRow_1					
Jobscript			_				
orial4 0.1) 🖪 Con	ntexts(KudoTutorial4) 🛛 🏟	Component I> Run (Job	KudoTutorial4)	Test Cases 🔊 Integ	gration Action		
orial4							
Execution Run	Kill 🕄 Kill	ar					
95 Starting [statisti [statisti 341 [New 386 [New	Starting job KudoTutorial4 at 13:18 23/06/2016. [statistics] connecting to socket on port 3990 [statistics] connected 341 [New I/O worker #1] INFO org.kududb.client.AsyncKuduClient - Discovered tablet Kudu Master for table 386 [New I/O worker #1] INFO org.kududb.client.AsyncKuduClient - Discovered tablet 4bea6156c189455bbfcb3 tLogRow_1						
email		surname name	age country	married weight pho	to profession inserti		
fernando	.mendonca@onepoint1	td.com Mendonça Ferna	ando 46 in	true 98.12 FGH	Architect 23-06-2		





Common Errors

REQUESTED REPLICATION FACTOR

One of the most common errors that you will probably get when you run the job for the first time on a test environment using a single virtual machine is:

"org.kududb.client.MasterErrorException: Server[Kudu Master - quickstart.cloudera:7051] INVALID_ARGUMENT[code 4]: Not enough live tablet servers to create a table with the requested replication factor 3. 1 tablet servers are alive."



SOLUTION

Simply set the requested replication number in the **Advanced Settings** tab to 1 in this case:

Connection Failure

This problem occurs when the Kudu services have not been started properly. Typically this is what you see on your screen:

TFixedFlo	Starting Errer of row1 (Main) winput_1 tKeduOutput_2			
igner Code Jobs	mpt			
Job(KudoTutorial	1 0.1) 🔀 Contexts(KudoTutorial1) 🐠 Component 🕩 Run (Job KudoTutorial1) 👘 Test Cases, 📾 Integration Action			
> KudoTutoria	11			
sic Run bug Run vanced settings get Exec mory Run	Execution			





Typically there is nothing you can do in Talend about this. You should in this case check, if the two Apache Kudu services are running on the Apache Kudu server:



We suggest in this case to try to start the services with:

service kudu-master start service kudu-tserver start

More information about Kudu administration can be found on this page: <u>http://getkudu.io/docs/troubleshooting.html</u>

One Point Consulting Ltd: Alpha House, Unit 14, 100 Villiers Road, London, NW2 5PJ, United Kingdom Phone: +44 (0) 203-198-6699 | www.onepointltd.com | Email: contact@onepointltd.com