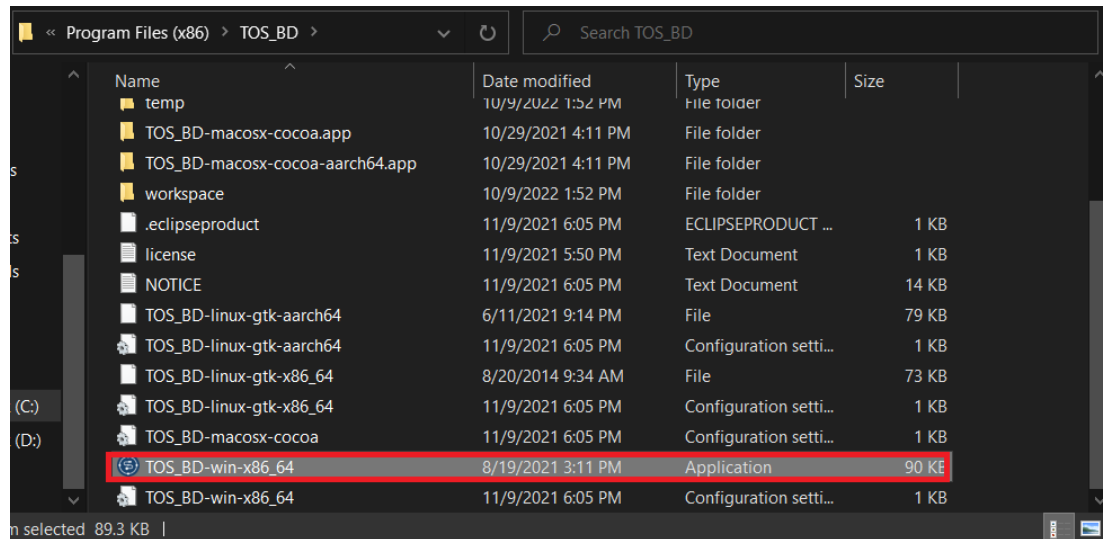


## Module 2 Project (Part 2)

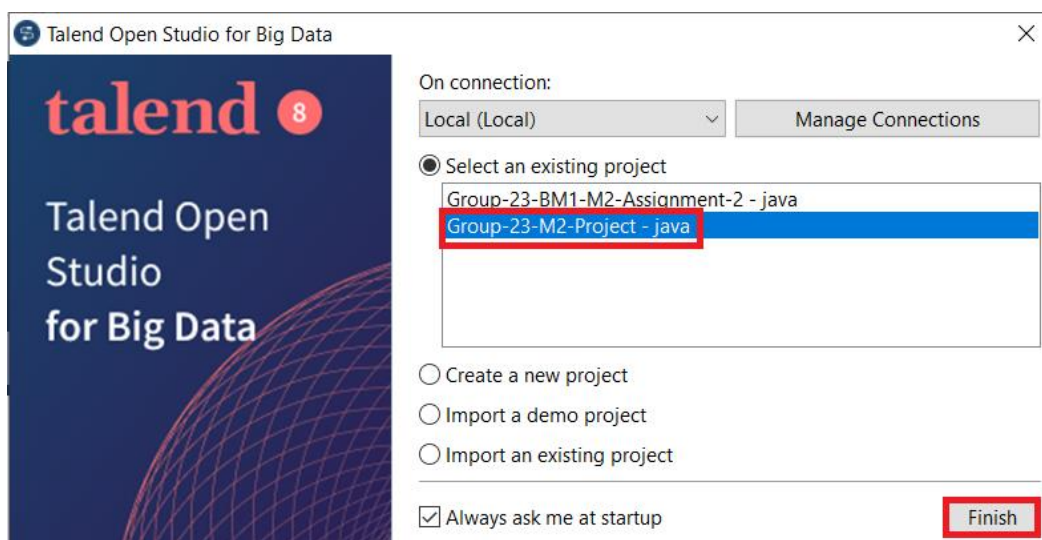
### I. Opening Talend for Big Data

Once again, we will be using Talend for this part of the project. To open it, you simply navigate to the directory folder containing your Talend for Big Data.



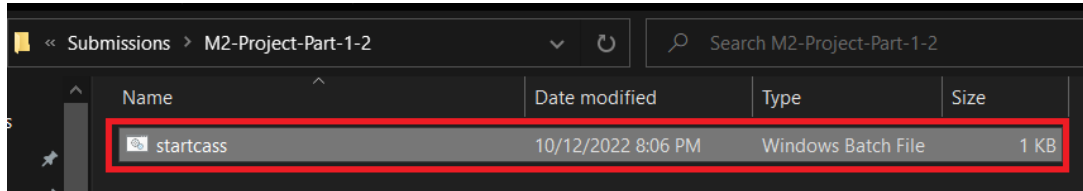
**Note:** If you continue to encounter issues such as "Incompatible JVM," you may need to update your system's environment variables or follow the previous steps from Module 2 Project Part 1.

Once your Talend for Big Data is open, you can proceed to selecting or creating a new project. Since we already have a project (Group-23-M2-Project) made for our "Module 2 Project Part 1" we can use that. We selected that project and clicked on Finish to open Talend.



## II. Initializing Cassandra

In our Module 2 Project Part 1, we created a file named **startcass.bat** that automated the process to opening Cassandra. Since we will be using Cassandra again, we need to execute the file.



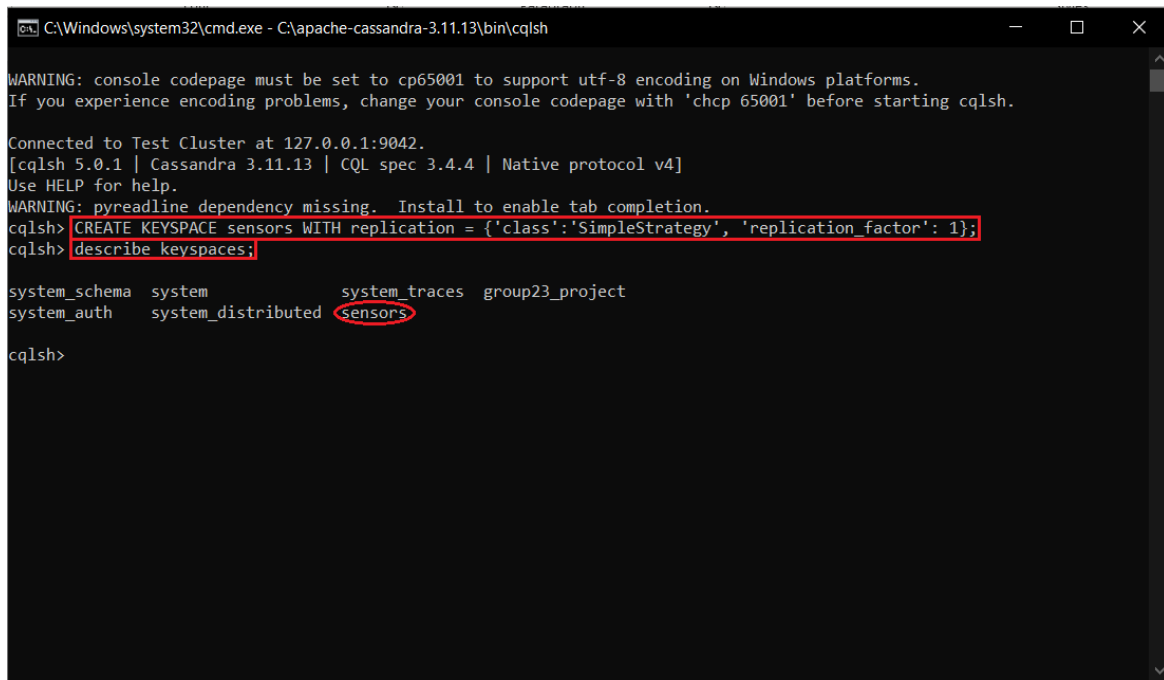
Following execution, two terminals were running in our background: one for the Cassandra database and one for the CQL shell.



In the terminal where the CQL shell was initialized, we ran the following command:

```
CREATE KEYSPACE sensors WITH replication =  
{'class':'SimpleStrategy', 'replication_factor': 1};
```

This command is used to create a keyspace named “sensors” where we will place the table that will contain data from the dim\_sensors.xlsx that was given. To verify if the keyspace was created, we use the command “describe keyspaces;”.



```
C:\Windows\system32\cmd.exe - C:\apache-cassandra-3.11.13\bin\cqlsh

WARNING: console codepage must be set to cp65001 to support utf-8 encoding on Windows platforms.
If you experience encoding problems, change your console codepage with 'chcp 65001' before starting cqlsh.

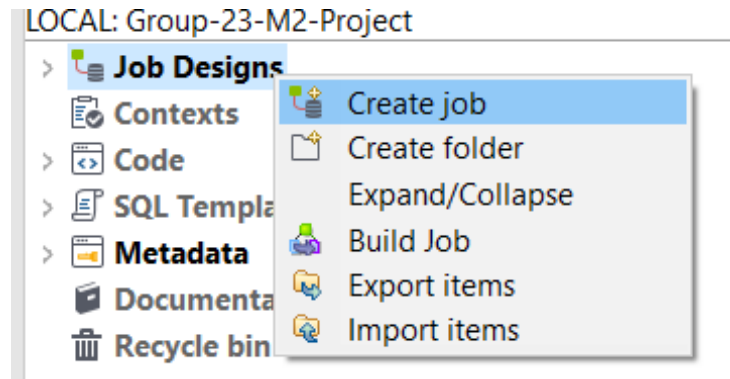
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.13 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
WARNING: pyreadline dependency missing. Install to enable tab completion.
cqlsh> CREATE KEYSPACE sensors WITH replication = {'class':'SimpleStrategy', 'replication_factor': 1};
cqlsh> describe keyspaces;

system_schema  system          system_traces  group23_project
system_auth    system_distributed sensors
```

Cassandra is now ready for the next steps.

### III. Creating a Job and Establishing a NoSql Connection

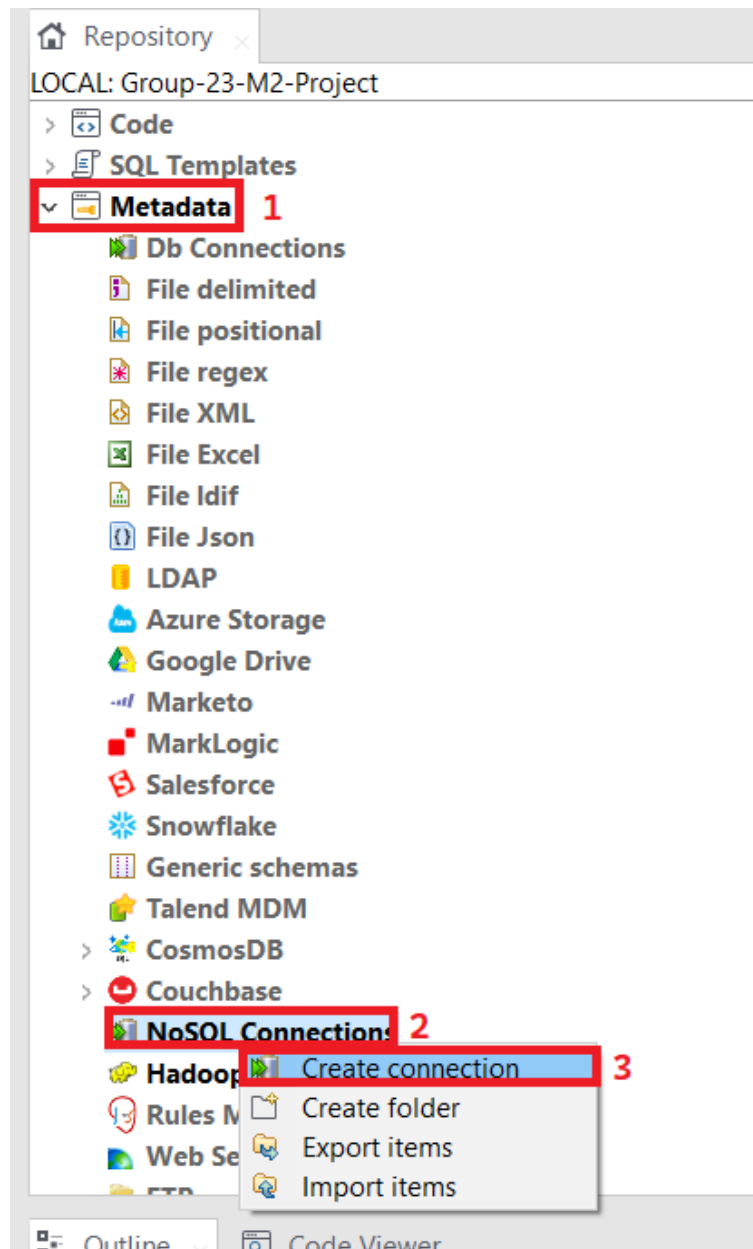
Go back to your open Talend Studio for Big Data. Under the Repository tab, right click *Job Designs* and click on *Create Job*.



In our example, we named our job *ImporttoCass* but you may name it however you wish.

A screenshot of the 'New job' dialog box in Talend Studio. The dialog has a title bar with a Talend logo and the text 'New job'. Below the title bar, there's a section titled 'New job' with a warning icon and the text 'It is inadvisable to leave the purpose blank.' To the right of this text is a small icon of a database cylinder. The main area of the dialog contains several fields: 'Name' (filled with 'ImporttoCass'), 'Purpose' (empty), 'Description' (empty), 'Author' (filled with 'user@talend.com'), 'Locker' (empty), 'Version' (filled with '0.1'), 'Status' (empty), and 'Path' (empty). There are also small buttons for 'M' and 'm' next to the 'Version' field, and a 'Select' button next to the 'Path' field. At the bottom right, there are two buttons: 'Finish' (highlighted with a blue border) and 'Cancel'.

Previously we made a NoSql Connection in this Talend project; however, for the sake of documentation, we will reinstate it again. To establish the connection between the Cassandra database and Talend, select *Metadata* > Right click *NoSQL Connections* > *Create connection*.



The new NoSQL Connection window should appear, prompting the user to enter a name, purpose, and description for the new NoSQL connection. After filling out all of the required fields, click "Next >." For this project, we named the connection *NoSQLCass* and left the remaining text field and area blank.

**NoSQL Connection**

**New NoSQL Connection on repository**

⚠ It is inadvisable to leave the purpose blank.

Name: **NoSQLCass**

Purpose:

Description:

Author: user@talend.com

Locker:

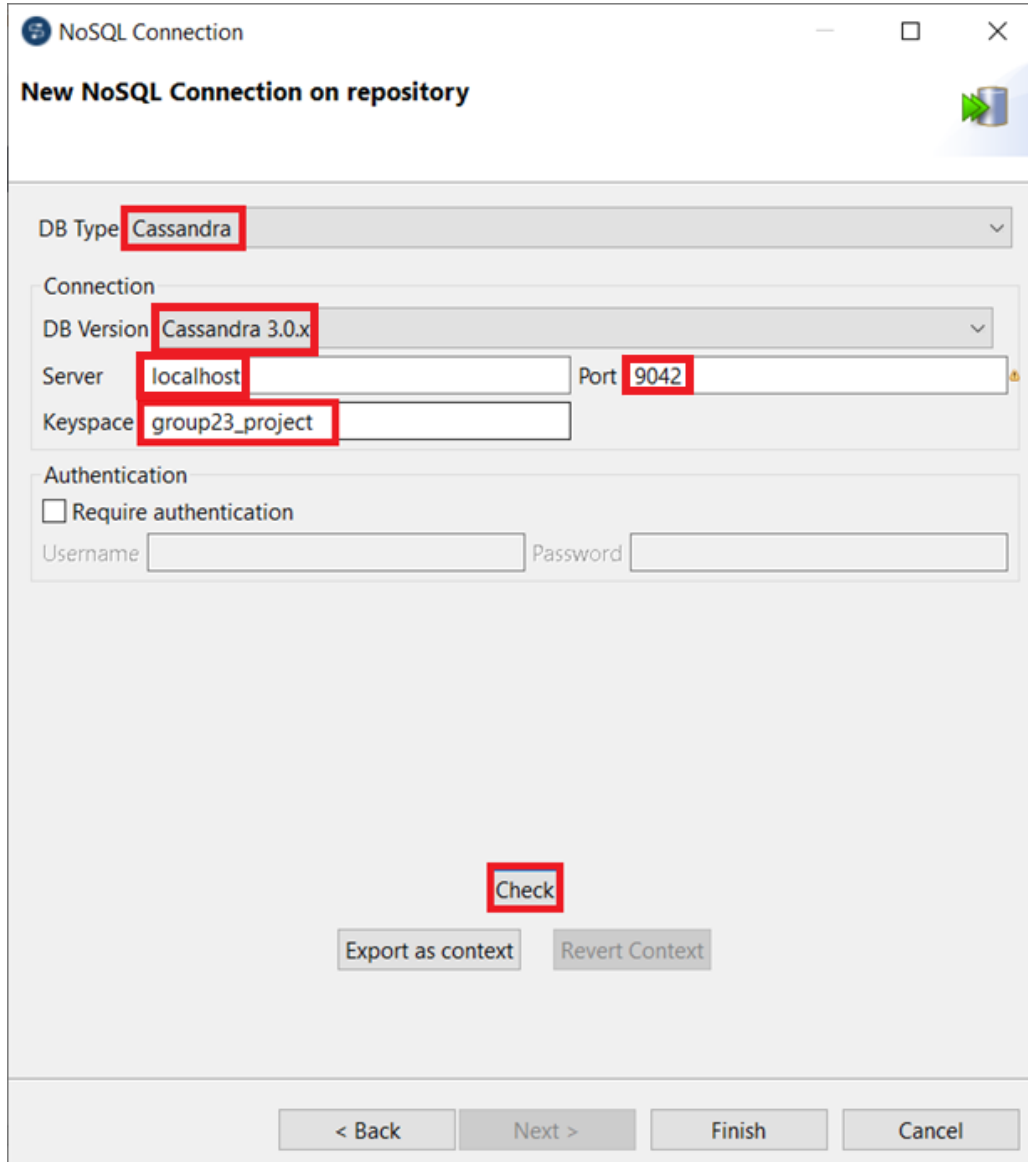
Version: 0.1

Status:

Path:

< Back   **Next >**   Finish   Cancel

The next step is to select the database (DB Type) that you wish to establish a Talend connection with. Select *Cassandra* as the **DB Type** and *Cassandra 3.0.x* as the **DB Version**. Specify *localhost* or *127.0.0.1* as the **Server** with *9042* as its **Port**. For the **Keyspace**, you may name it however you like. Once everything is finished, click on “Check” to test the connection.



**NoSQL Connection**

**New NoSQL Connection on repository**

DB Type: **Cassandra**

Connection

DB Version: **Cassandra 3.0.x**

Server: **localhost** Port: **9042**

Keyspace: **group23\_project**

Authentication

☐ Require authentication

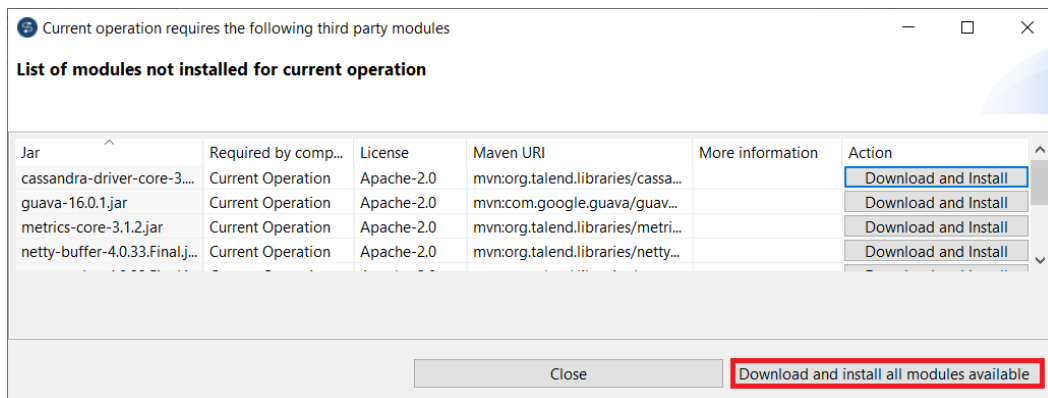
Username: Password:

**Check**

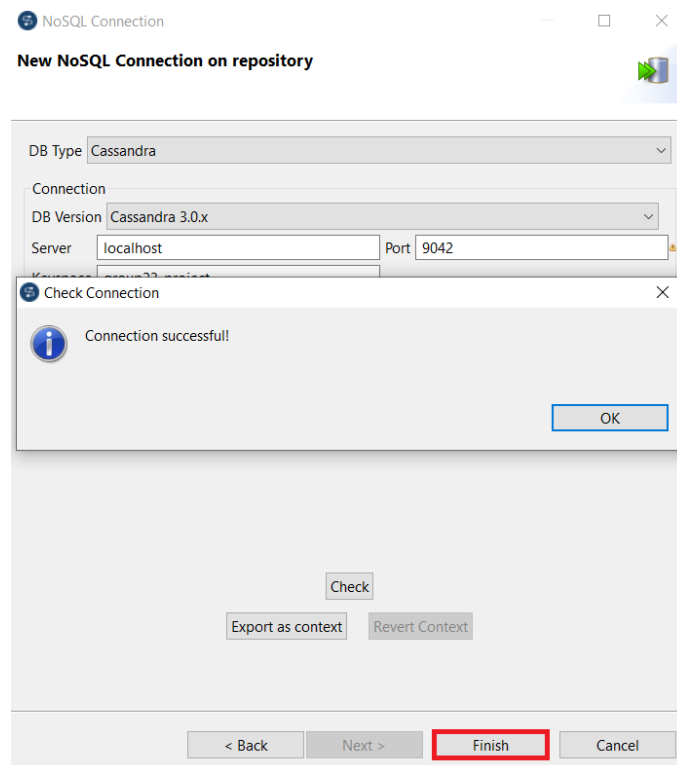
Export as context Revert Context

< Back Next > Finish Cancel

You will then encounter a message that states that some list of modules should be downloaded for you to continue. Just click on “Download and install all modules available”.



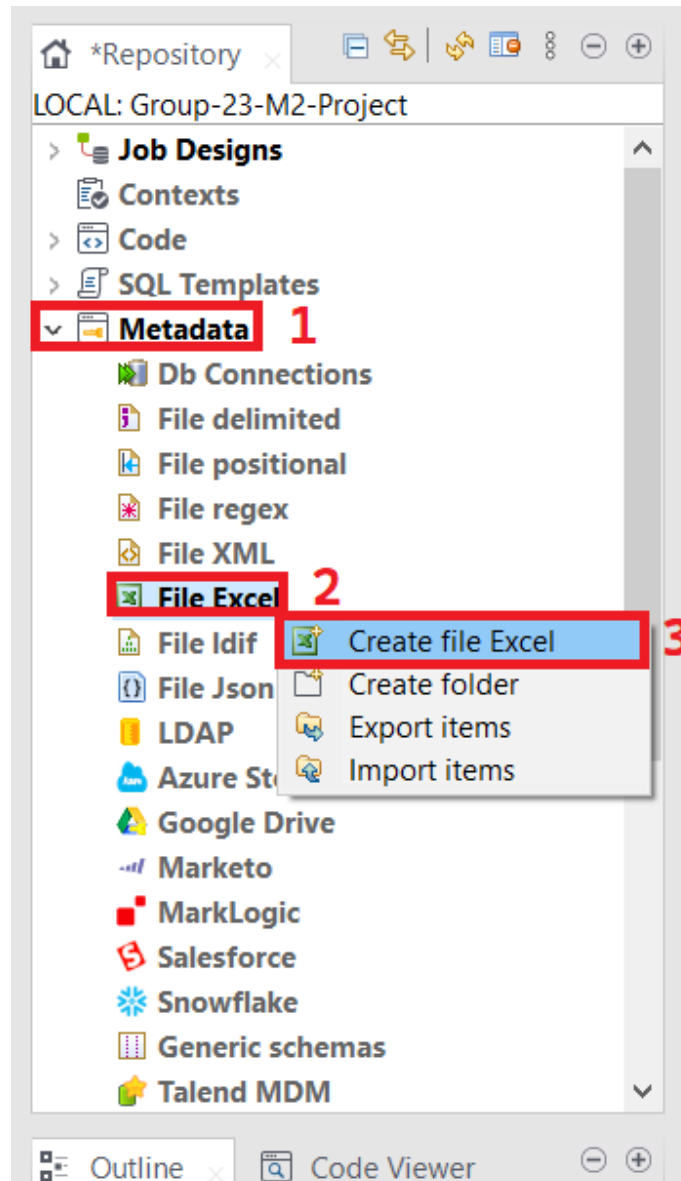
After downloading the modules needed, your connection should be successful. Click “Finish” once everything is all set.





#### IV. Importing Data to Cassandra using Talend

To import the provided dataset into Cassandra using Talend, select the “>” next to **Metadata**, then right click on **File Excel** > **Create file Excel**.



The *New Excel File* window should appear. Here we specified “dim\_sensors” as the **Name**, however, you may choose a different one. Then, click *Next >*.

New Excel File

**File - Step 1 of 4**

⚠ It is inadvisable to leave the purpose blank.

Name

Purpose

Description

Author

Locker

Version

Status

Path

< Back **Next >** Finish Cancel

Then, select the “Browse...” button on the right of the *File* text field, and find the *dim\_sensors* file. The file’s directory should appear like the one in the screenshot below. Select the *Read excel2007 file format(xlsx)* checkbox to ensure that all Excel versions of the dataset will be read. The File Viewer should automatically detect the columns and its contents, as shown in the bottom half of the window. No changes are required so click *Next >* to proceed.

**New Excel File**

**File - Step 2 of 4**

Add a Metadata File on repository  
Define the path of the file and the format settings

**File Settings**

Server: Localhost 127.0.0.1

File: C:/Users/ASUS/Documents/Mapua/Third Year - 1st Term/CS172 DATA SCIENCE 2/Submissions/M2-Project-Part-1-2 **Browse...**

☒ Read excel2007 file format(xlsx)

Generation mode: Memory-consuming(User mode)

**File Viewer and Sheets setting**

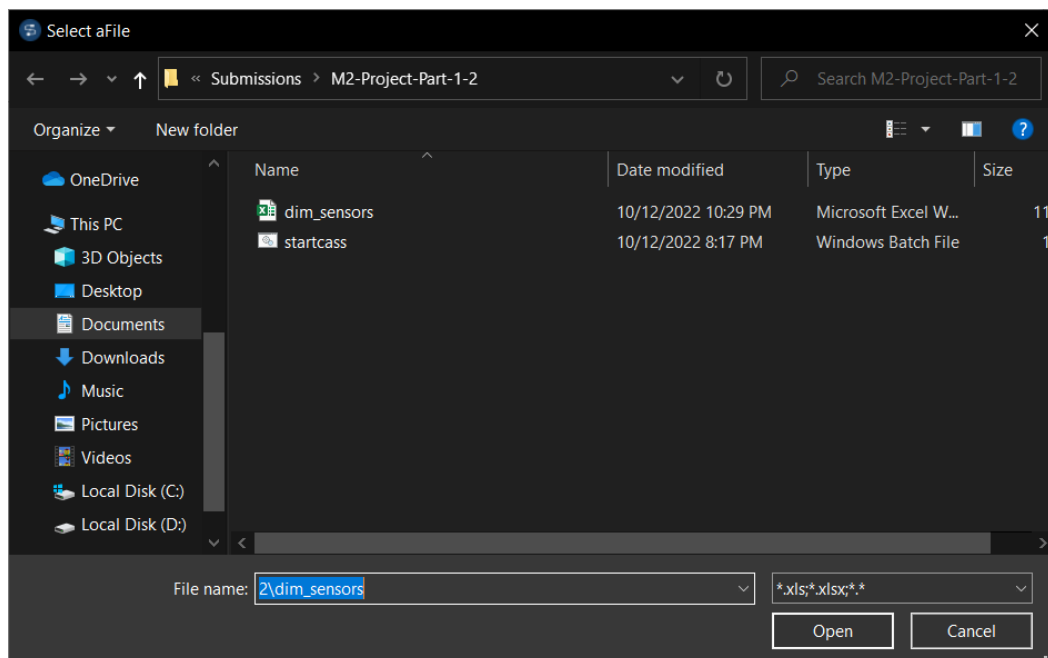
Set sheets parameters

☒ All sheets/DSelect sheet  
☒ dim\_sensors-1

Please select sheet (Sheet structure as schema guide) **nsors-1**

A	B	C	D	E	F	G
senso...	zone...	node...	senso...	senso...	surve...	road_...
CCTV_...	9.0	1015.0	CCTV_...	Digo...	Surve...	Digo...
CCTV_...	5.0	1299.0	CCTV_...	Digo...	Surve...	Digo...
CCTV_...	1.0	1300.0	CCTV_...	Pend...	Surve...	Pend...
CCTV_...	2.0	1331.0	CCTV_...	Pres J...	Surve...	Roxa...
CCTV_...	1.0	1330.0	CCTV_...	Digo...	Surve...	Digo...
CCTV_...	5.0	1340.0	CCTV_...	Digo...	Surve...	Digo...
CCTV_...	5.0	1340.0	CCTV_...	Digo...	Surve...	Digo...

< Back **Next >** Finish Cancel



For the next step, make sure to check mark the box next to *Set heading row as column names*. After, click on “Refresh Preview” to check if the excel output looks right. Click “Finish” once everything is done.

New Excel File

File - Step 3 of 4

Add a Metadata File on repository  
Define the setting of the parse job

File Settings

Encoding

UTF-8

☐ Advanced separator(for number)

Thousands separator:

,

Decimal separator:

.

Rows To Skip

If any rows must be ignored, specify the following parameters

Header☒

1

Footer☐

Metadata column setting

First column:

1

Last column:

Limit Of Rows

If the number of lines must be limited, specify this number

Limit☐

Preview

Output

☒ Set heading row as column names

Refresh Preview

sensor_id	zone_id	node_id	sensor_type	sensor_location	survey_point	road_mon
CCTV_01	9	1015	CCTV camera	Digos - Makar Rd - E. Bulaong Ave Intersection	Survey Point 1	Digos-Ma
CCTV_02	5	1299	CCTV camera	Digos - Makar Rd - Pendatun Ave Intersection	Survey Point 2	Digos-Ma
CCTV_03	1	1300	CCTV camera	Pendatun Ave - Pres JP Laurel Ave Intersection	Survey Point 3	Pendatum
CCTV_04	2	1331	CCTV camera	Pres JP Laurel Ave - Roxas Ave Intersection	Survey Point 4	Roxas Ave

Export as context

Revert Context

< Back

Next >

Finish

Cancel

There is nothing else to change in the last step so click on “Finish” to complete the process.

**New Excel File**

**File - Step 4 of 4**

Add a Schema on repository  
Define the Schema

Name: metadata

Comment:

Schema

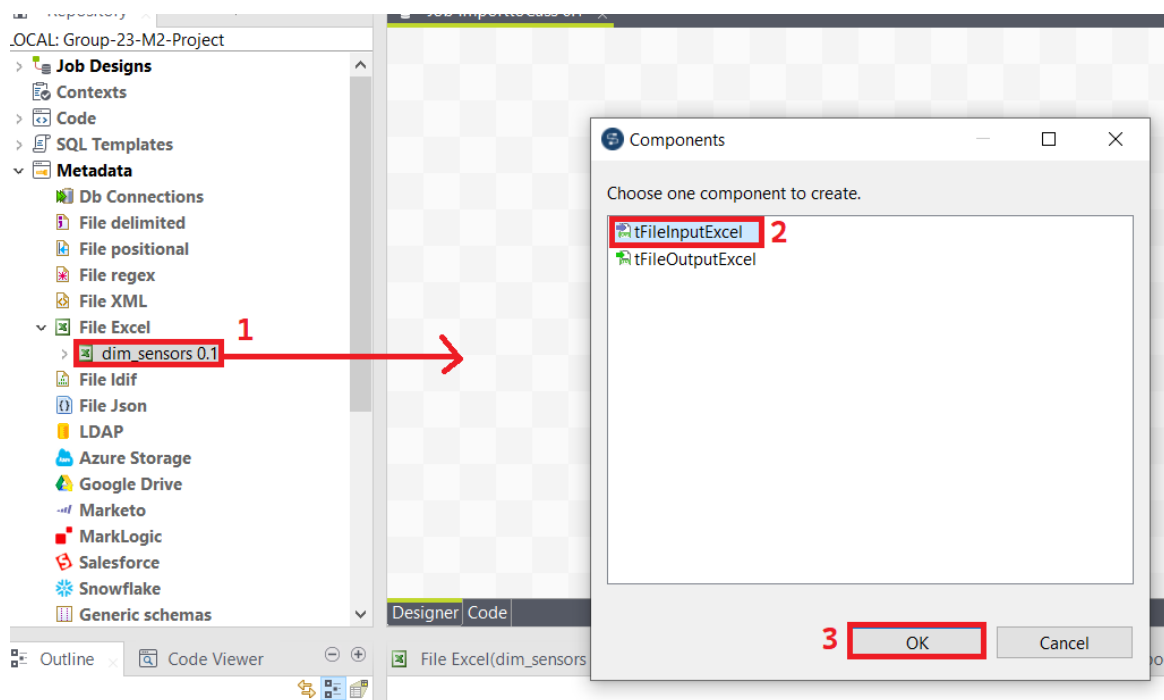
Click to update schema preview Guess

Description of the Schema

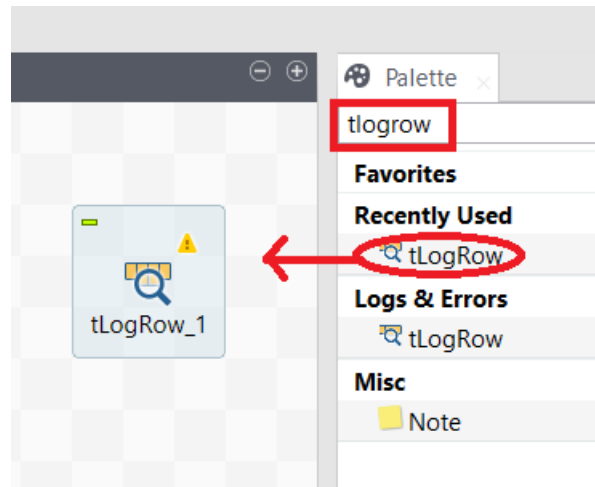
Column	K...	Type	<input checked="" type="checkbox"/> N...	Date Pattern (Ctrl...	Length	Precision	Default	Comment
sensor_id	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		7	0		
zone_id	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		1	0		
node_id	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		4	0		
sensor_type	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		11	0		
sensor_location	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		49	0		
survey_point	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		15	0		

< Back Next > **Finish** Cancel

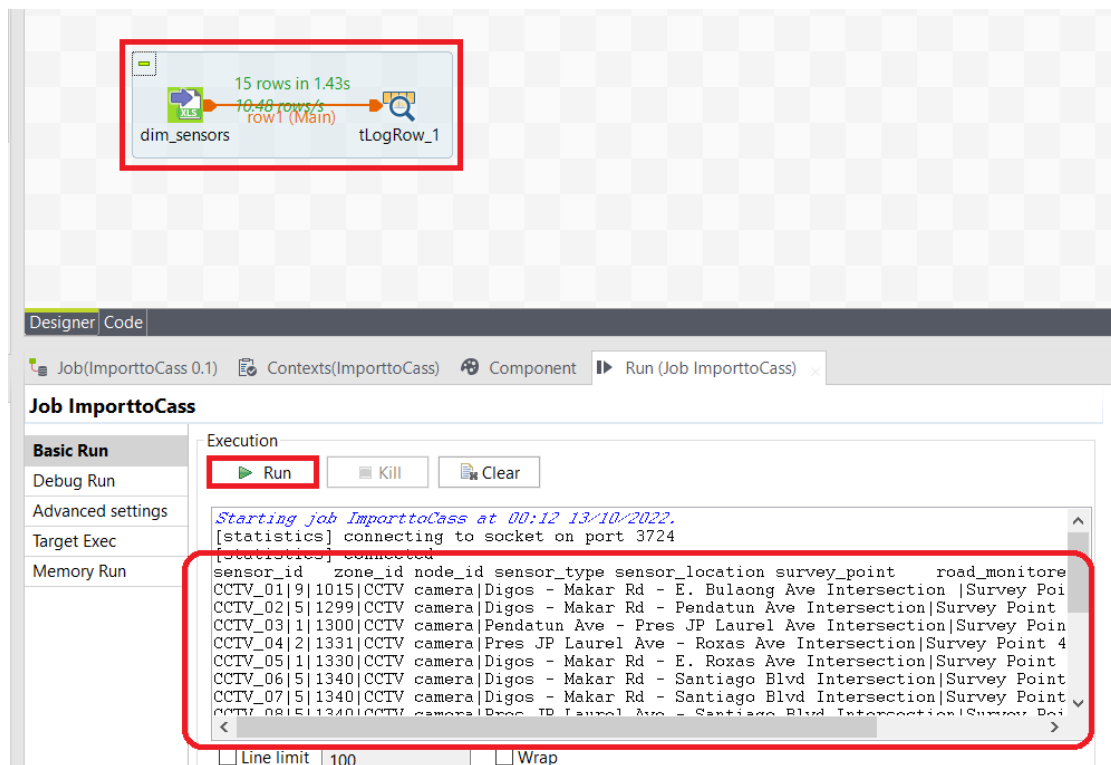
Next, drag the newly made File excel component *dim\_sensors* to the job. Choose tFileInputExcel and click “OK”.



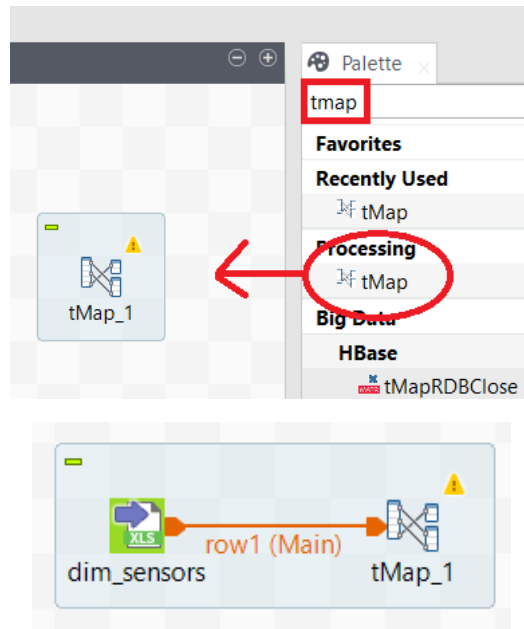
Then, in the Palette tab to your right, search for **tLog Row**, then select and drag it. This component will help verify the data. Drag the component to the workspace, as indicated.



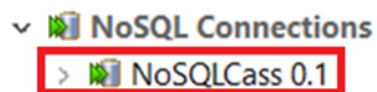
Now, drag an arrow from the *dim\_sensors* component that was created earlier to the *tLogRow\_1* component. To run the execution, go to the *Run (Job ImporttoCass)* tab and select the *Run* command. The emphasized portion of the screenshot should appear; here we see the data from *dim\_sensors*. Delete this *tLogRow\_1* component since we do not need it for importing.



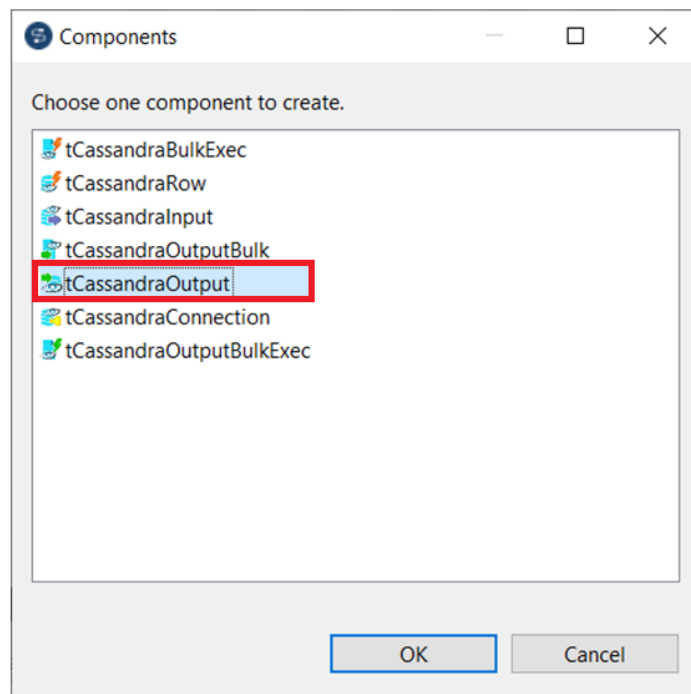
On the Palette tab located at the right-hand side area of the Talend window, search for **tMap**, then select and drag it to the workspace, as indicated. Then, drag an arrow from the *dim\_sensors* component that was created earlier to the *tMap\_1* job.



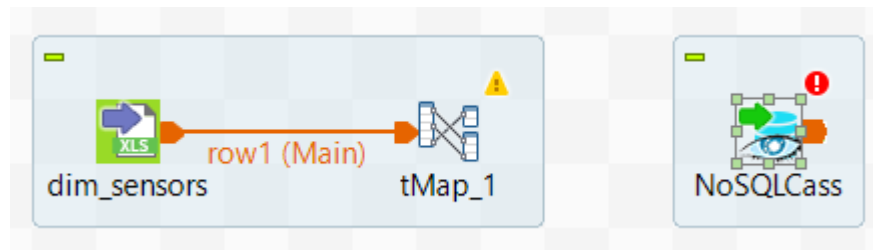
Then, under the Repository tab click on NoSQL Connections and drag the NoSQLCass component we made earlier to the job. Select *tCassandraOutput* and click “OK.”



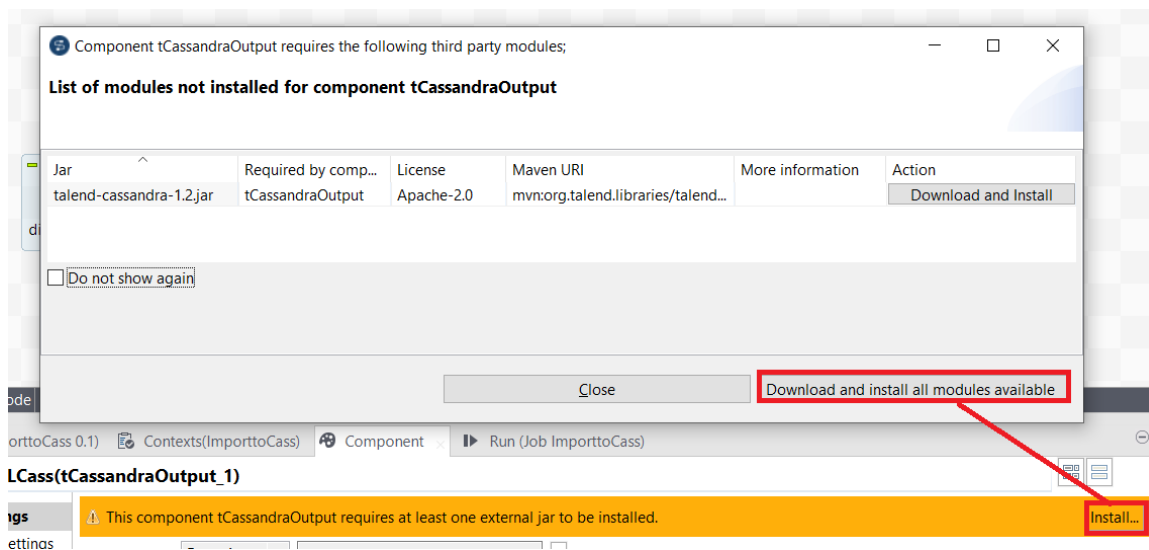
Drag this to  
the job



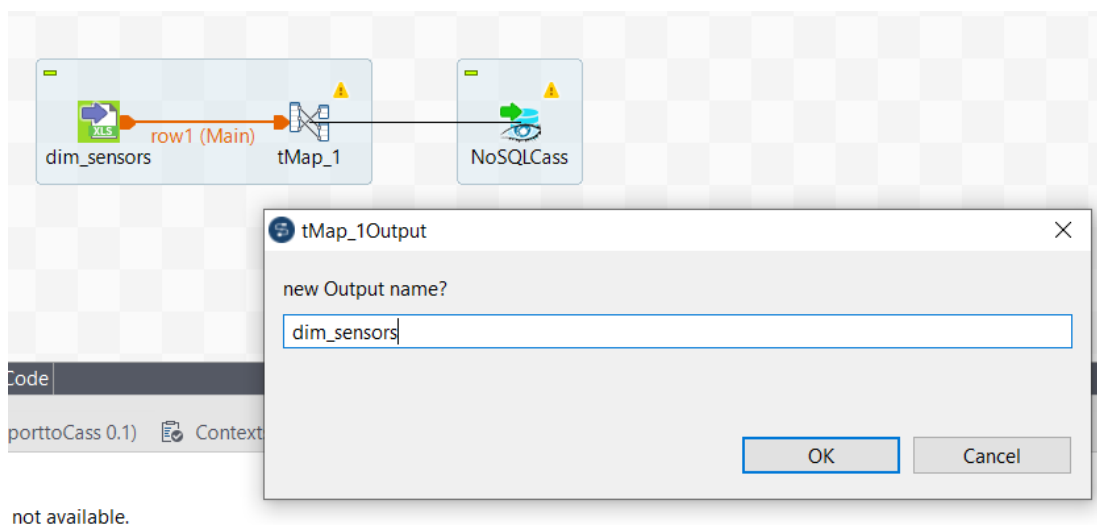
The workspace should appear like the figure below.



Since the *tCassandraOutput* component requires at least one external jar file to be installed, select *Install...* then click on the *Download and install all modules available* button to download the required modules.



Next, drag an arrow from the *tMap* component to the *tCassandraOutput*. There will be a prompt to name the tMap output. We named this output as *dim\_sensors* but you may name it whatever you want.





Next, double click on the *tCassandraOutput* component to check its basic settings. Make sure you selected “Built-In” for **PROPERTY**. After, under Keyspace configuration, properly indicate the **Keyspace** we made earlier which would be *sensors*. Then, under Column family configuration, indicate the **Column family** which would be *dim\_sensors* and select “Create column family if does not exist” for **Action on column family**. Lastly, click on the button with the three dots to edit the table schema.

The screenshot shows the configuration window for the **NoSQLCass(tCassandraOutput\_1)** component. The following elements are annotated with red boxes and numbers:

- 1:** A red circle around the **NoSQLCass** component icon in the workflow designer.
- 2:** A red box around the **PROPERTY** dropdown menu, which is set to **Built-In**.
- 3:** A red box around the **Keyspace** text field, which contains the value **"sensors"**.
- 4:** A red box around the **Column family** text field, which contains the value **"dim\_sensors"**.
- 5:** A red box around the **Action on column family** dropdown menu, which is set to **Create column family if does not exist**.
- 6:** A red box around the **Edit schema** button, which is represented by a small square icon with three dots.

In the schema, we need to make sure that *sensor\_id* is our primary key. Click on “OK” once finished.

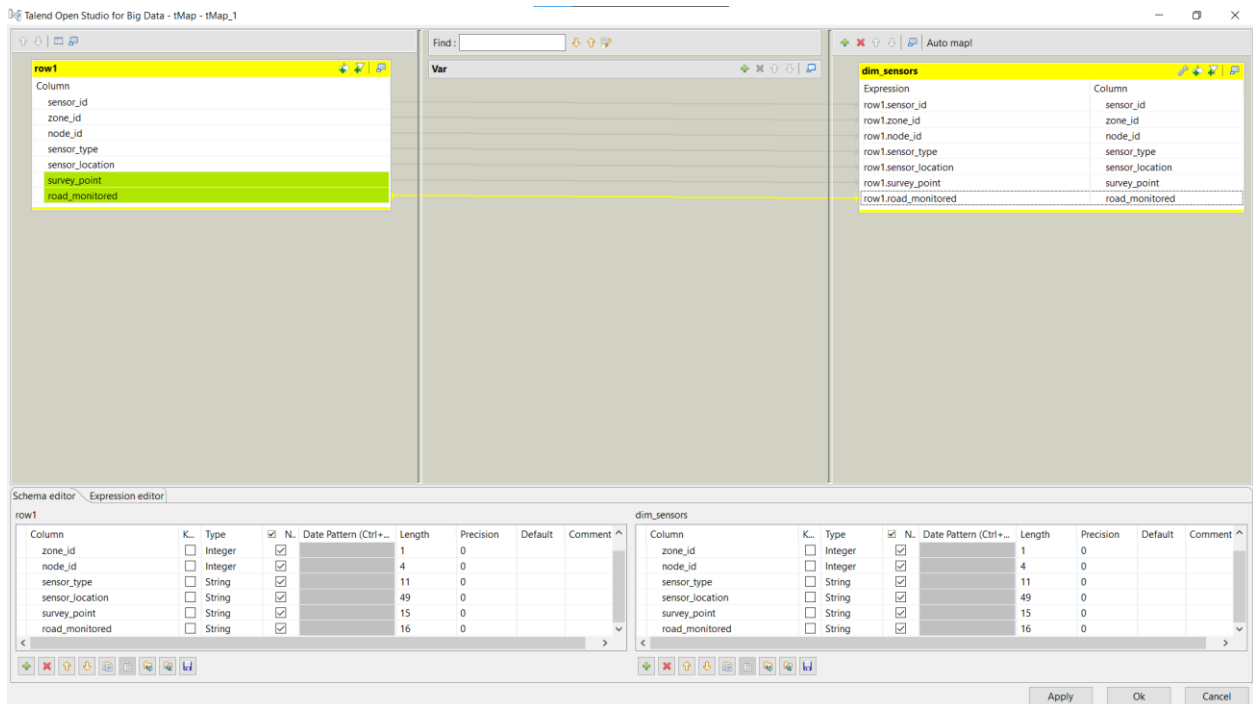
The screenshot shows the **Schema of NoSQLCass** dialog box. It displays two tables: **dim\_sensors (Input)** and **NoSQLCass (Output)**. The **sensor\_id** column in the output table is highlighted with a red box and a key icon, indicating it is the primary key.

Column	K...	Ty...	✓	N..	Date ...	Le...	Pr...	D...	Co...
sensor_id	<input type="checkbox"/>	St...	<input checked="" type="checkbox"/>			7	0		
zone_id	<input type="checkbox"/>	Int...	<input checked="" type="checkbox"/>			1	0		
node_id	<input type="checkbox"/>	Int...	<input checked="" type="checkbox"/>			4	0		
sensor_t...	<input type="checkbox"/>	St...	<input checked="" type="checkbox"/>			11	0		
sensor_l...	<input type="checkbox"/>	St...	<input checked="" type="checkbox"/>			49	0		
survey_p...	<input type="checkbox"/>	St...	<input checked="" type="checkbox"/>			15	0		
road_m...	<input type="checkbox"/>	St...	<input checked="" type="checkbox"/>			16	0		

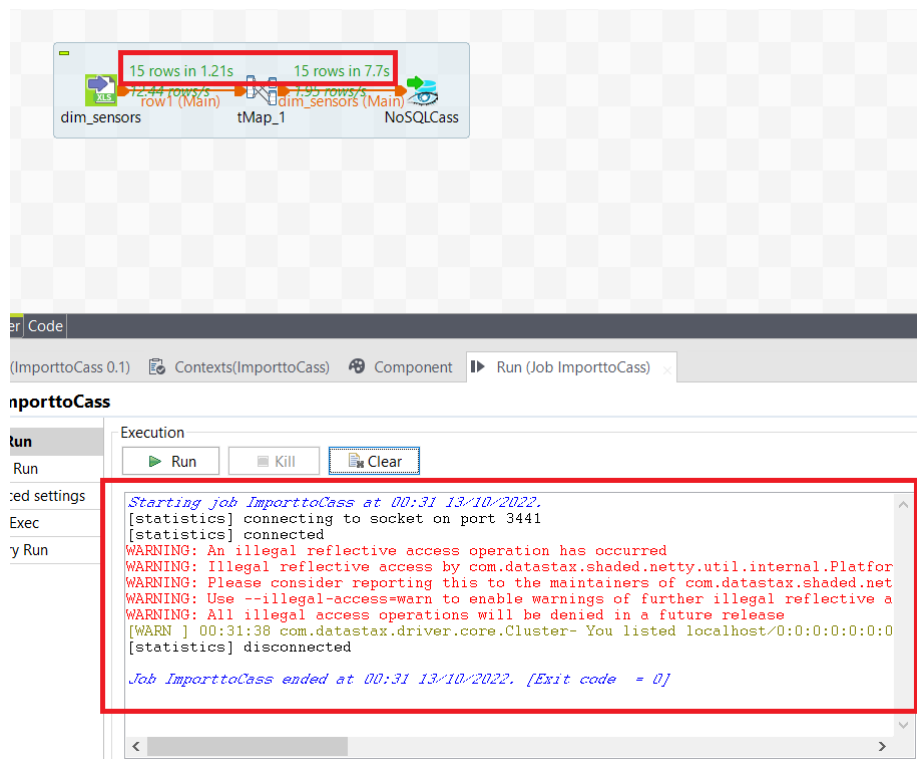
  

Column	Db Colu...	Key	Type	DB Type	✓	N..	Dat...
sensor_id	sensor_id	<input checked="" type="checkbox"/>	String	ascii	<input checked="" type="checkbox"/>		
zone_id	zone_id	<input type="checkbox"/>	Integer	int	<input checked="" type="checkbox"/>		
node_id	node_id	<input type="checkbox"/>	Integer	int	<input checked="" type="checkbox"/>		
sensor_t...	sensor_t...	<input type="checkbox"/>	String	ascii	<input checked="" type="checkbox"/>		
sensor_l...	sensor_l...	<input type="checkbox"/>	String	ascii	<input checked="" type="checkbox"/>		
survey_p...	survey_...	<input type="checkbox"/>	String	ascii	<input checked="" type="checkbox"/>		
road_mo...	road_m...	<input type="checkbox"/>	String	ascii	<input checked="" type="checkbox"/>		

Go back to your job and double-click on the tMap component to open the Schema editor. Drag every component from row1 to the table (dim\_sensors) we have for Cassandra.



Now, we will run the job. The emphasized portion of the screenshot below indicates that all 15 rows from the excel file were imported successfully to our Cassandra database.



## V. Verifying the Importing of data to Cassandra

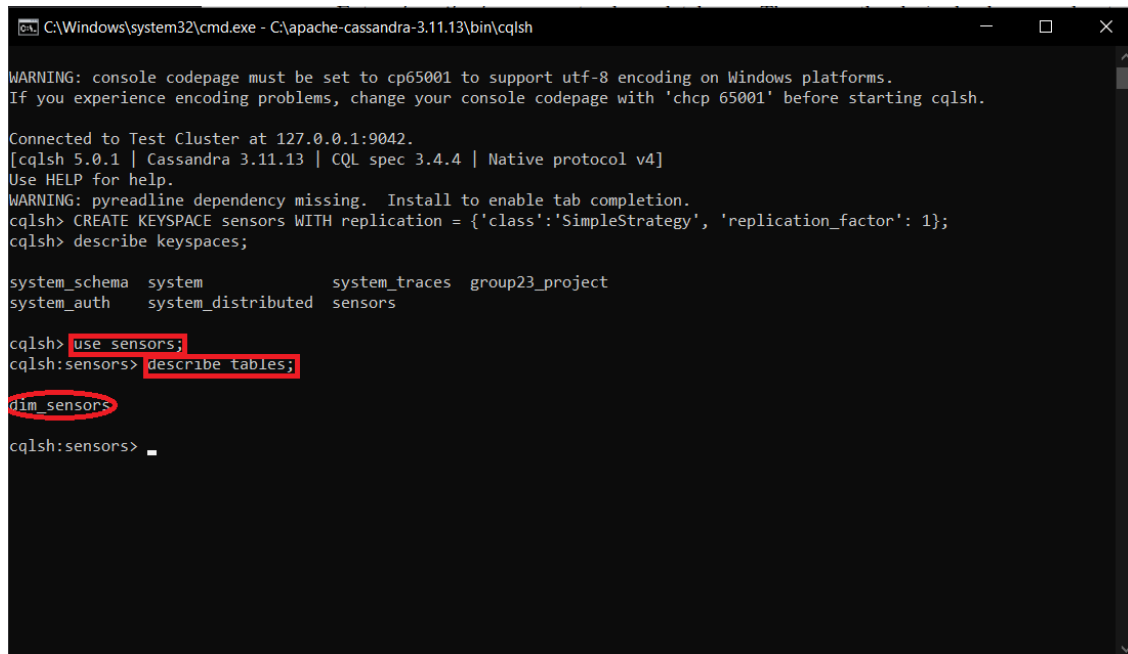
To verify that the import from the dataset to a Cassandra table via Talend was successful, run the following commands on Cassandra's CQL shell:

**use sensors;**

*This command opens or uses the database in which the table was saved.*

**describe tables;**

*This command shows or describes the tables saved in the database sensors.*



```
C:\Windows\system32\cmd.exe - C:\apache-cassandra-3.11.13\bin\cqlsh

WARNING: console codepage must be set to cp65001 to support utf-8 encoding on Windows platforms.
If you experience encoding problems, change your console codepage with 'chcp 65001' before starting cqlsh.

Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.13 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
WARNING: pyreadline dependency missing. Install to enable tab completion.
cqlsh> CREATE KEYSPACE sensors WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};
cqlsh> describe keyspaces;

system_schema  system              system_traces  group23_project
system_auth    system_distributed  sensors

cqlsh> use sensors;
cqlsh:sensors> describe tables;

dim_sensors

cqlsh:sensors> _
```

Note: If all the configurations made previously were correct, the table `dim_sensors` should appear after executing the `describe tables;` command.

Now, run the following command to see if table dim\_sensors has been populated:

```
select * from dim_sensors;
```

```
C:\Windows\system32\cmd.exe - C:\apache-cassandra-3.11.13\bin\cqlsh
WARNING: console codepage must be set to cp65001 to support utf-8 encoding on Windows platforms.
If you experience encoding problems, change your console codepage with 'chcp 65001' before starting cqlsh.

Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.13 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
WARNING: pyreadline dependency missing. Install to enable tab completion.
cqlsh> CREATE KEYSPACE sensors WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};
cqlsh> describe keyspaces;

system_schema  system              system_traces  group23_project
system_auth    system_distributed sensors

cqlsh> use sensors;
cqlsh:sensors> describe tables;

dim_sensors

cqlsh:sensors> select * from dim_sensors;
```

sensor_id	node_id	road_monitored	sensor_location	sensor_type	survey_point	zone_id
CCTV_07	1340	Digos Makar Rd	Digos - Makar Rd - Santiago Blvd Intersection	CCTV camera	Survey Point 6b	5
CCTV_08	1340	Santiago Blvd.	Pres JP Laurel Ave - Santiago Blvd Intersection	CCTV camera	Survey Point 7	5
CCTV_09	1370	Digos Makar Rd	Digos Makar Rd - Jose Catolico Ave Intersection	CCTV camera	Survey Point 8a	7
CCTV_10	1370	J. Catolico Ave.	Digos Makar Rd - Jose Catolico Ave Intersection	CCTV camera	Survey Point 8b	7
CCTV_02	1299	Digos-Makar Rd	Digos - Makar Rd - Pendatun Ave Intersection	CCTV camera	Survey Point 2	5
CCTV_11	1370	J. Catolico Ave.	Digos Makar Rd - Jose Catolico Ave Intersection	CCTV camera	Survey Point 8c	7
CCTV_14	1360	Leon Llido St.	Salvani - Leon Llido St Intersection	CCTV camera	Survey Point 11	7
CCTV_12	1364	Digos Makar Rd	Digos Makar Rd - Honorio Arriola St. Intersection	CCTV camera	Survey Point 9	7
CCTV_04	1331	Roxas Ave.	Pres JP Laurel Ave - Roxas Ave Intersection	CCTV camera	Survey Point 4	2
CCTV_15	1396	Digos Makar Rd	Digos Makar Rd. - NLSA Road (Lagao Public Market)	CCTV camera	Survey Point 12	7
CCTV_05	1330	Digos Makar Rd	Digos - Makar Rd - E. Roxas Ave Intersection	CCTV camera	Survey Point 5	1
CCTV_13	1310	Aparente St.	Salvani - Apparante St Intersection	CCTV camera	Survey Point 10	1
CCTV_06	1340	Digos Makar Rd	Digos - Makar Rd - Santiago Blvd Intersection	CCTV camera	Survey Point 6a	5
CCTV_03	1300	Pendatun Ave.	Pendatun Ave - Pres JP Laurel Ave Intersection	CCTV camera	Survey Point 3	1
CCTV_01	1015	Digos-Makar Rd	Digos - Makar Rd - E. Bulaong Ave Intersection	CCTV camera	Survey Point 1	9

```
(15 rows)
cqlsh:sensors>
```

This proves that our Talend project was a success, and we were able to meet the objectives of the activity.

## REFERENCES

*CREATE KEYSPACE*. (n.d.). DataStax. Retrieved from [https://docs.datastax.com/en/cql-oss/3.3/cql/cql\\_reference/cqlCreateKeyspace.html](https://docs.datastax.com/en/cql-oss/3.3/cql/cql_reference/cqlCreateKeyspace.html)

*Managing NoSQL metadata*. (n.d.). Talend. Retrieved from <https://help.talend.com/r/en-US/7.3/studio-user-guide-big-data/managing-nosql-metadata>

*Read Excel File in Talend*. (n.d.). Tutorial Gateway. Retrieved from <https://www.tutorialgateway.org/read-excel-file-in-talend/>