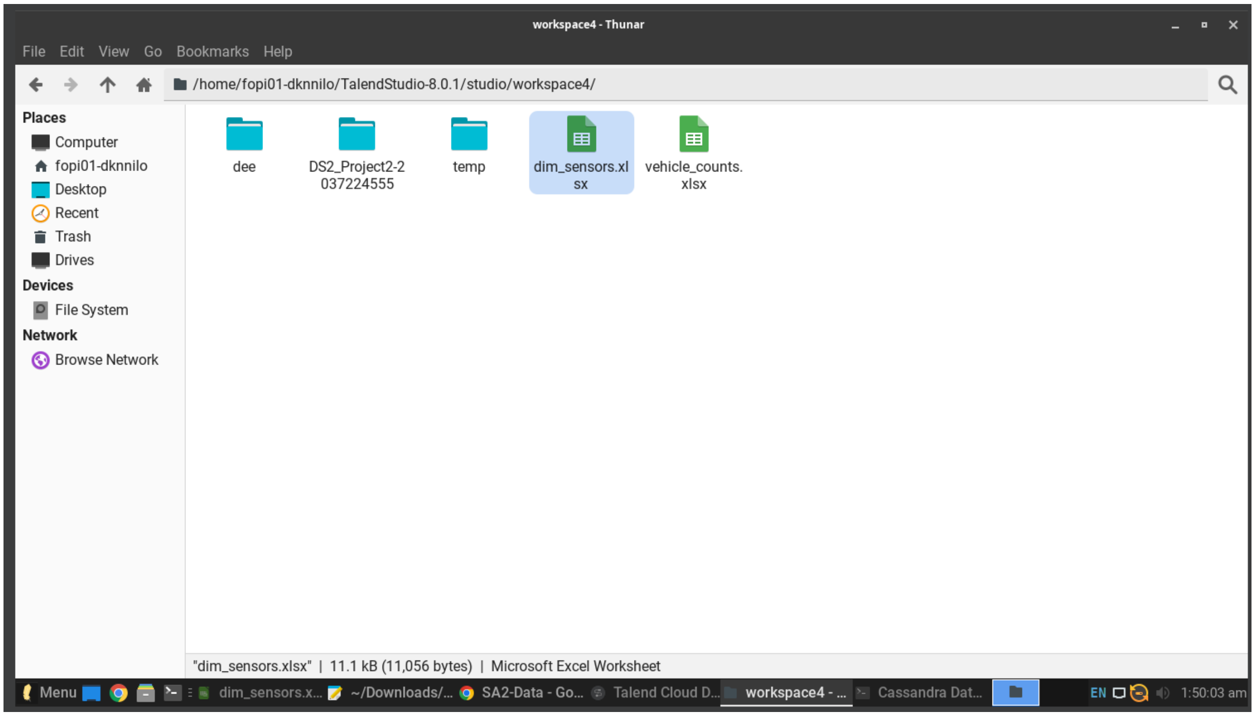
Group 5: Nilo, Daniella Kim N.
Peñaflor, Rowencell A.

Course/Section: CSS182-2/CS-O

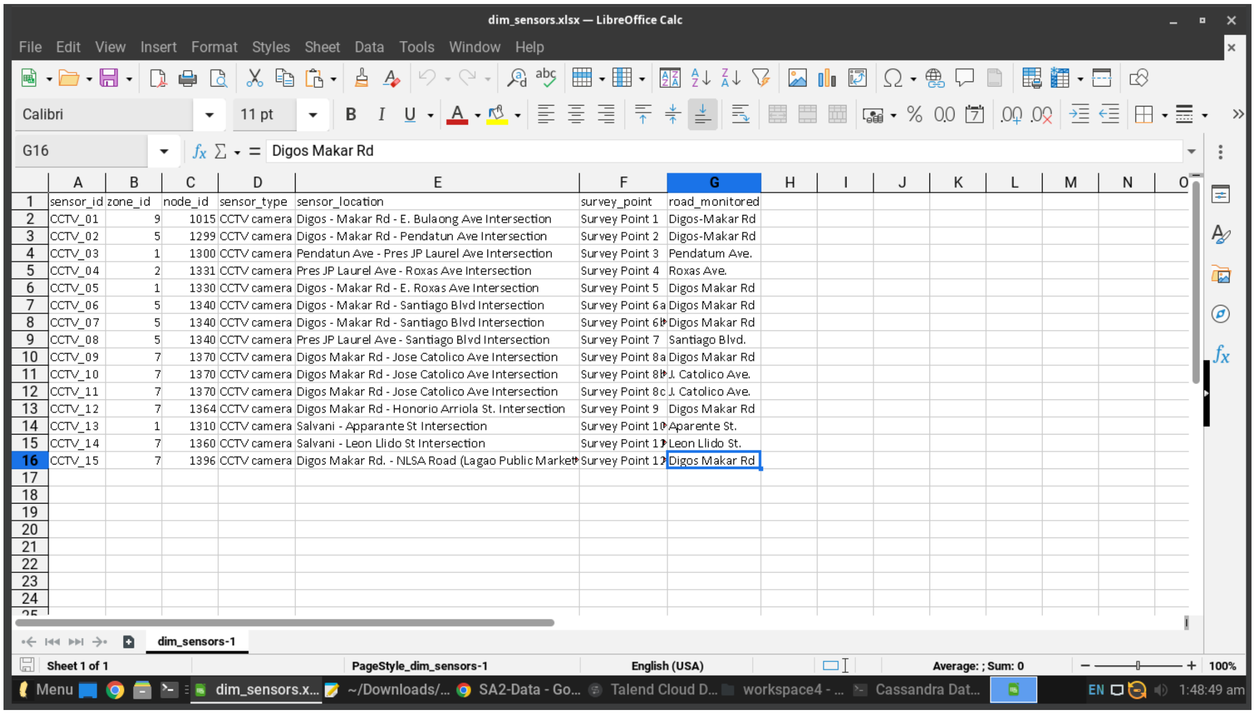
M2-SA2: Excel File to Cassandra Migration Process using Talend

DOWNLOADING AND EXAMINING THE EXCEL FILE

Before doing the activity, we downloaded the excel file, dim_sensor.xls, first and then checked its contents. The Excel file contains 7 columns (`sensor_id`, `zone_id`, `node_id`, `sensor_type`, `sensor_location`, `survery_point`, `and` `road_monitored`) and 15 entries of sensor_id.



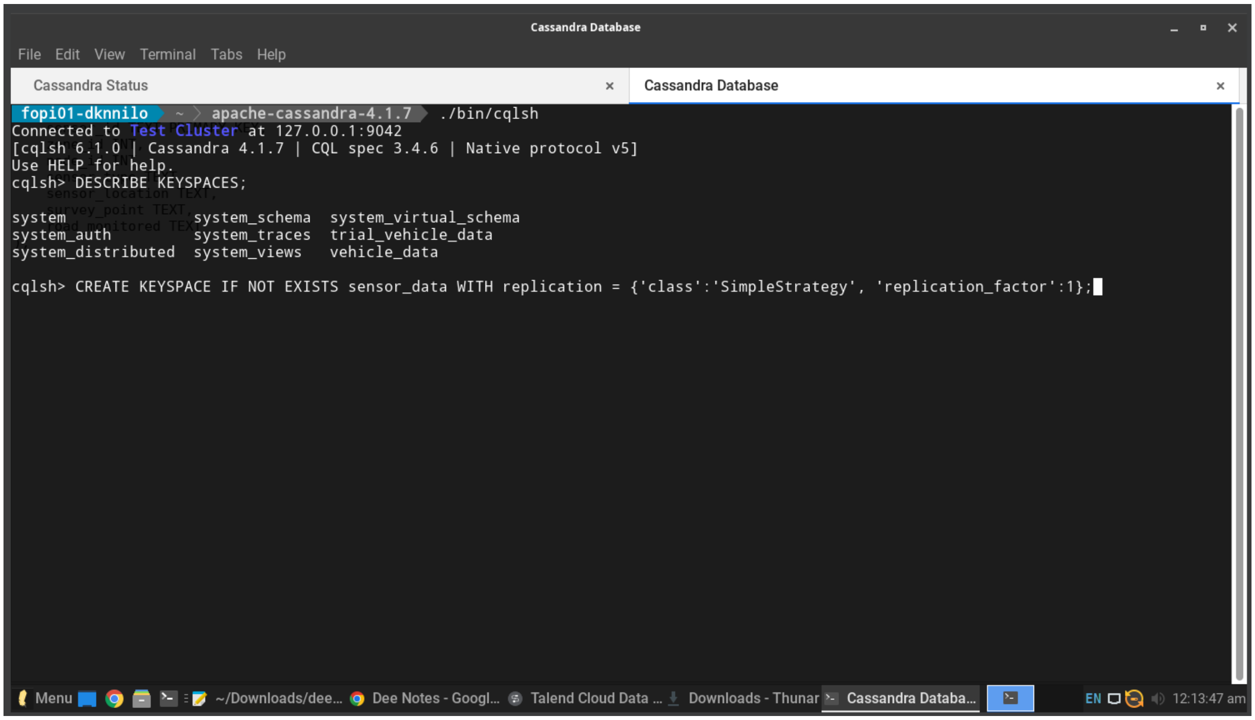
"dim_sensors.xlsx" | 11.1 kB (11,056 bytes) | Microsoft Excel Worksheet



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

CREATING CASSANDRA KEYSPACE AND TABLE

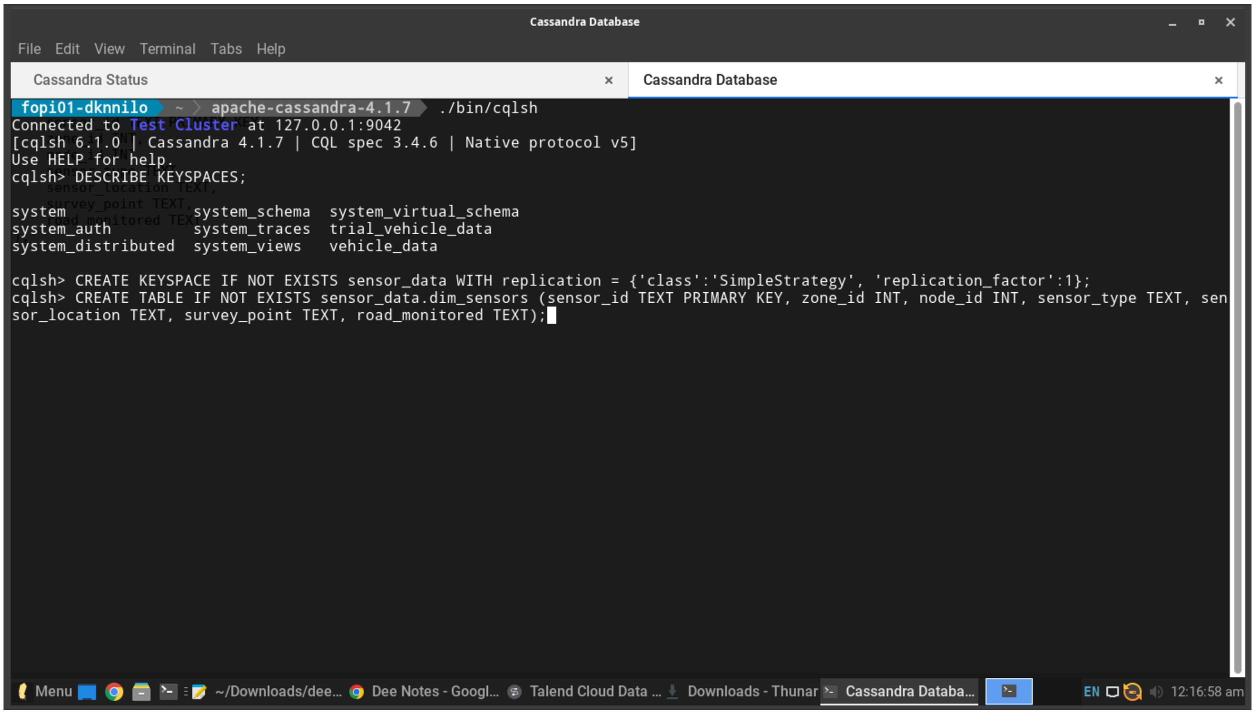
After checking the file, we utilized the `DESCRIBE KEYSPACES` command to key the available keyspaces. This will help determine if the `sensor_data` keyspace already exists. After making sure the keyspace did not exist, we created the keyspace using the command, `CREATE KEYSPACE IF NOT EXISTS sensor_data WITH replication = {'class': 'SimpleStrategy','replication_factor':1};`. We also created a table using the command, `CREATE TABLE IF NOT EXISTS sensor_data.dim_sensors (sensor_id TEXT PRIMARY KEY, zone_id INT, node_id INT, sensor_type TEXT, sensor_location TEXT, survey_point TEXT, road_monitored TEXT);`. The table columns used in the commands are taken from the columns on the Excel file.



```
fopi01-dknnilo  ~  apache-cassandra-4.1.7  ./bin/cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.7 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> DESCRIBE KEYSPACES;

system              system_schema  system_virtual_schema
system_auth         system_traces  trial_vehicle_data
system_distributed  system_views   vehicle_data

cqlsh> CREATE KEYSPACE IF NOT EXISTS sensor_data WITH replication = {'class':'SimpleStrategy', 'replication_factor':1};
```



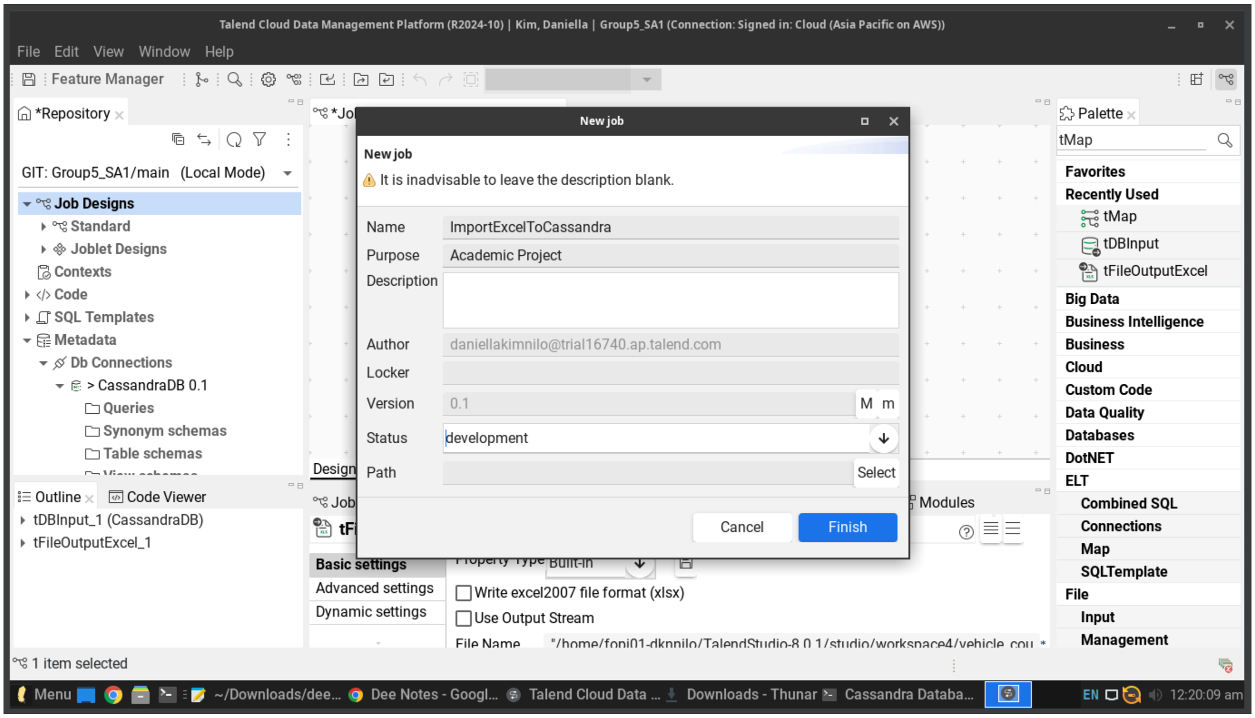
```
fopi01-dknnilo  ~  apache-cassandra-4.1.7  ./bin/cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.7 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> DESCRIBE KEYSPACES;

system              system_schema  system_virtual_schema
system_auth         system_traces  trial_vehicle_data
system_distributed  system_views   vehicle_data

cqlsh> CREATE KEYSPACE IF NOT EXISTS sensor_data WITH replication = {'class':'SimpleStrategy', 'replication_factor':1};
cqlsh> CREATE TABLE IF NOT EXISTS sensor_data.dim_sensors (sensor_id TEXT PRIMARY KEY, zone_id INT, node_id INT, sensor_type TEXT, sen
sor_location TEXT, survey_point TEXT, road_monitored TEXT);
```

CREATING A NEW JOB FOR EXCEL TO CASSANDRA

After creating a keyspace and table, we proceeded with creating a new job for importing Excel file to Cassandra database. In this step, we first click the new job button. We input the name of the project (ImportExcelToCassandra), its purpose (Academic Project), its status (development), and then clicked the Finish button.



Talend Cloud Data Management Platform (R2024-10) | Kim, Daniella | Group5_SA1 (Connection: Signed in: Cloud (Asia Pacific on AWS))

New job

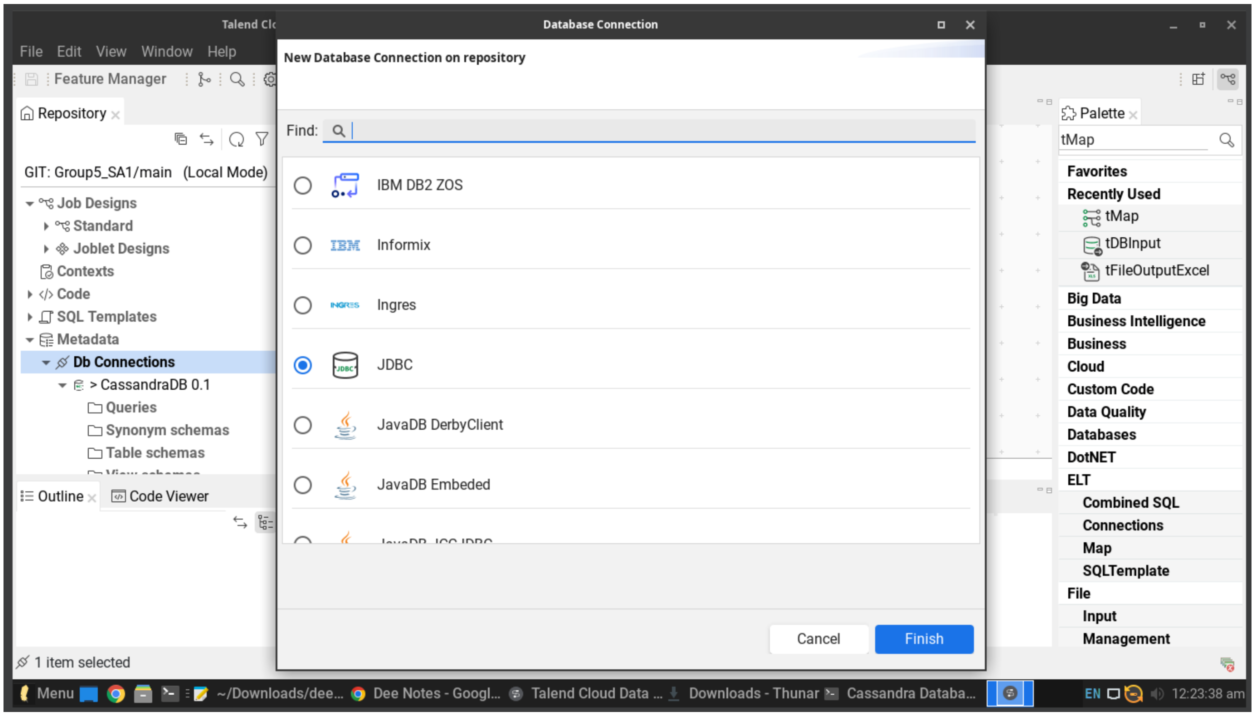
⚠ It is inadvisable to leave the description blank.

Name: ImportExcelToCassandra
Purpose: Academic Project
Description:
Author: daniellakimnilo@trial16740.ap.talend.com
Locker:
Version: 0.1
Status: development
Path:

☐ Write excel2007 file format (xlsx)
☐ Use Output Stream
File Name "/home/fopi01-dknnilo/TalendStudio-8.0.1/studio/workspace4/vehicle_cou

CREATING A JDBC DB CONNECTION

In creating this connection, we first clicked the JDBC button for the Database Connection window and entered `CassandrSensorDB` for name and Academic Project for Purpose in the Create New database window. We clicked the next button, and it redirected us to another page. On this page, we entered the JDBC URL (`jdbc:cassandra://127.0.0.1:9042/sensor_data`) and clicked the `cdata.jbdc.cassandra.jar` for the driver. We clicked the finish button to end the process. Another prompt showed our connection process is successful.



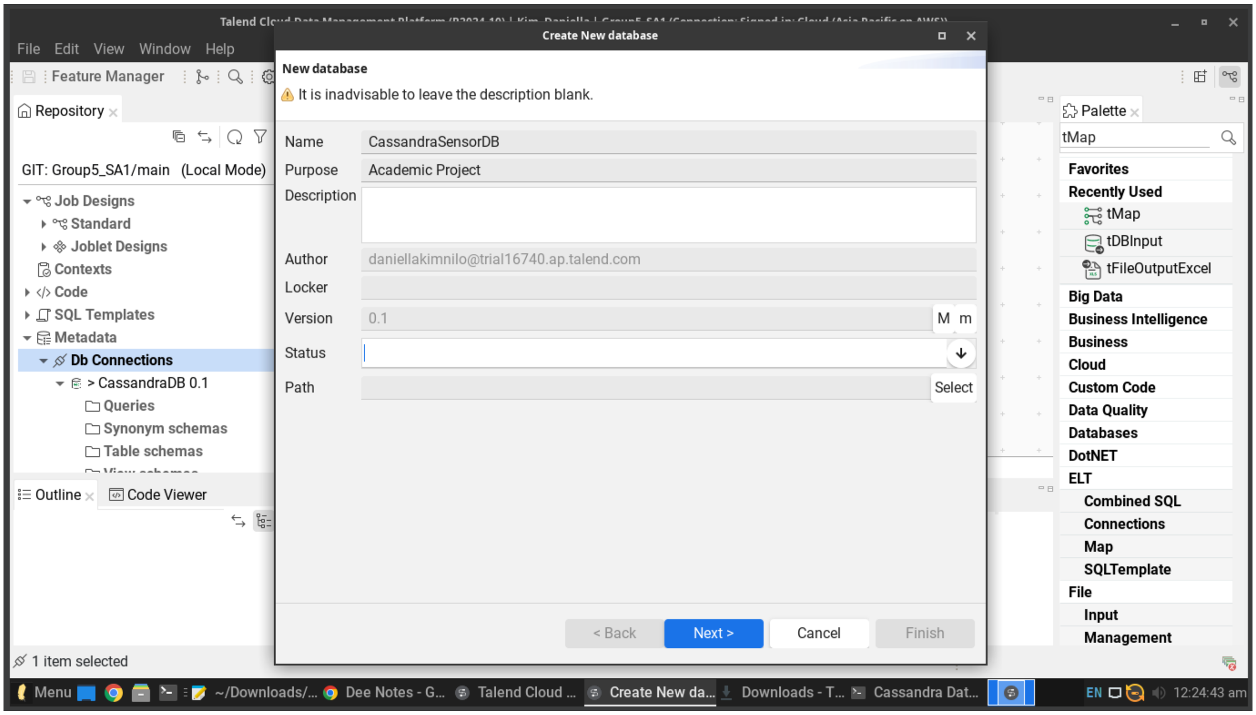
Database Connection

New Database Connection on repository

Find:

○ IBM DB2 ZOS
○ Informix
○ Ingres
● JDBC
○ JavaDB DerbyClient
○ JavaDB Embeded

GIT: Group5_SA1/main (Local Mode)



Create New database

New database

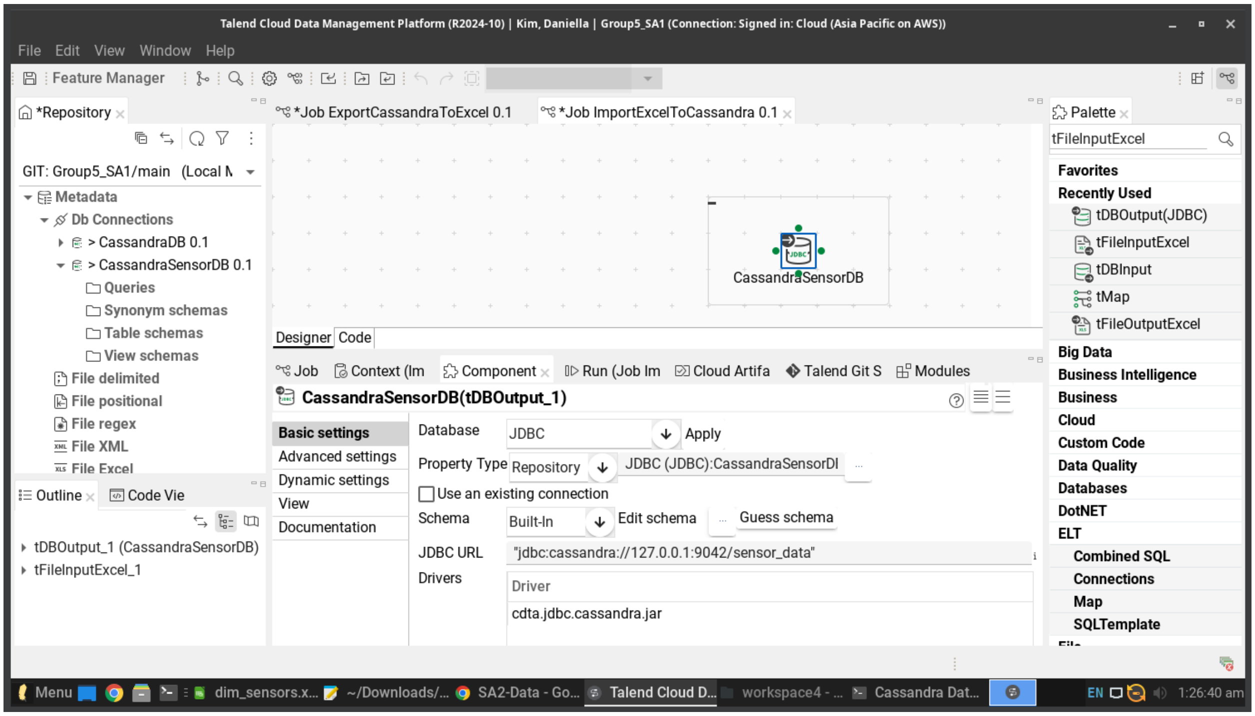
⚠ It is inadvisable to leave the description blank.

Name	CassandraSensorDB
Purpose	Academic Project
Description	
Author	daniellakimnilo@trial16740.ap.talend.com
Locker	
Version	0.1
Status	
Path	

GIT: Group5_SA1/main (Local Mode)

IMPORTING PROCESS

Clicking `OK` on the last prompt brought us back to the Job `ImportCassandraToExcel 0.1` workspace. We dragged the `CassandraSensorDB 0.1` and `tFileInputExcel` from the left palette to the workspace. The `tFileInputExcel` is utilized because this component reads an Excel file row by row, which splits them up into fields using regular expressions. It also sends the fields as defined in the Cassandra database schema to the next component in the Talend job.



Talend Cloud Data Management Platform (R2024-10) | Kim, Daniella | Group5_SA1 (Connection: Signed in: Cloud (Asia Pacific on AWS))

CassandraSensorDB(tDBOutput_1)

Basic settings

Database: JDBC

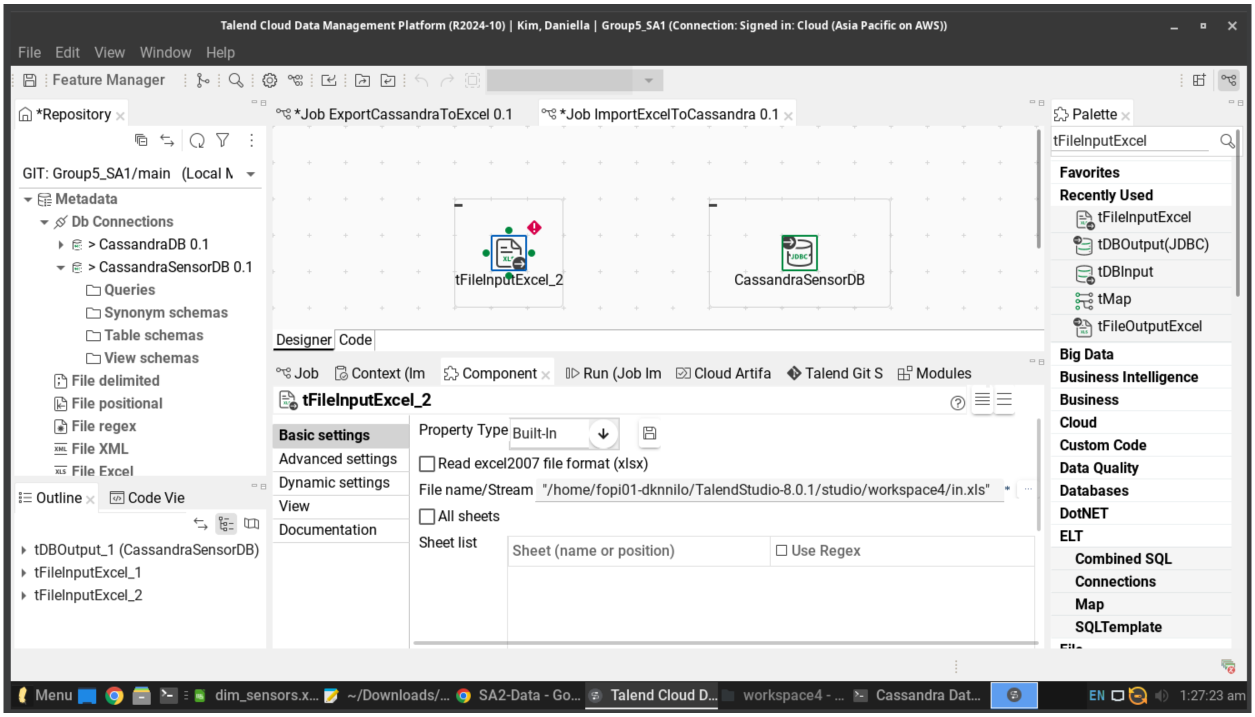
Property Type: Repository — JDBC (JDBC):CassandraSensorDI

□ Use an existing connection

Schema: Built-In

JDBC URL: "jdbc:cassandra://127.0.0.1:9042/sensor_data"

Drivers: Driver — cdta.jdbc.cassandra.jar



Talend Cloud Data Management Platform (R2024-10) | Kim, Daniella | Group5_SA1 (Connection: Signed in: Cloud (Asia Pacific on AWS))

tFileInputExcel_2

Basic settings

Property Type Built-In

□ Read excel2007 file format (xlsx)

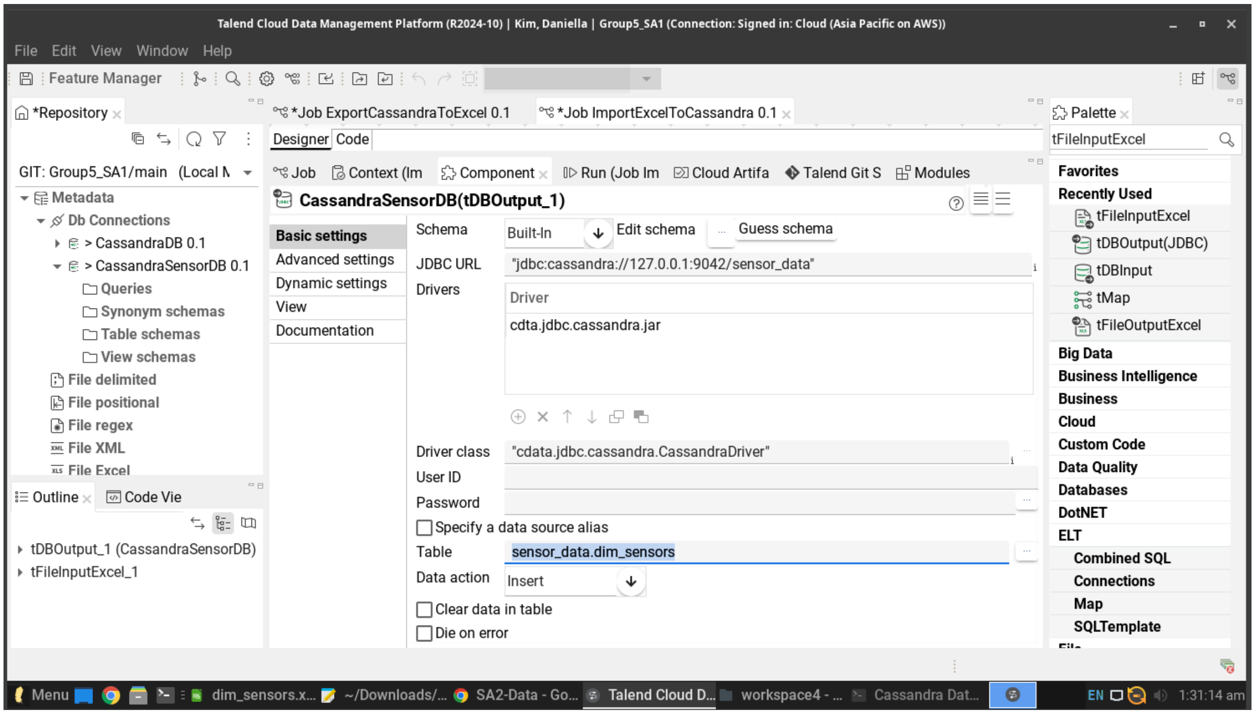
File name/Stream "/home/fopi01-dknnilo/TalendStudio-8.0.1/studio/workspace4/in.xls"

□ All sheets

Sheet list

Sheet (name or position)	□ Use Regex

The next process is to configure the `tDBOutput_1(CassandraSensorDB)` by clicking it and adjusting the `Designer` windows upwards and write `sensor_data.dim_sensors` in the `Table` textbox. After configuring the `tDBOutput_1(CassandraSensorDB)`, the component, `tFileInputExcel_1` , is configured by typing "`dim_sensors-1`" into the `Sheet list` textbox. Then, create a schema manually in `tFileInputExcel_1` to arrange column names and click the `OK` button.



Talend Cloud Data Management Platform (R2024-10) | Kim, Daniella | Group5_SA1 (Connection: Signed in: Cloud (Asia Pacific on AWS))

*Job ExportCassandraToExcel 0.1 | *Job ImportExcelToCassandra 0.1

CassandraSensorDB(tDBOutput_1)

Basic settings

Schema: Built-In

JDBC URL: "jdbc:cassandra://127.0.0.1:9042/sensor_data"

Drivers:

Driver
cdta.jdbc.cassandra.jar

Driver class: "cdata.jdbc.cassandra.CassandraDriver"

User ID:

Password:

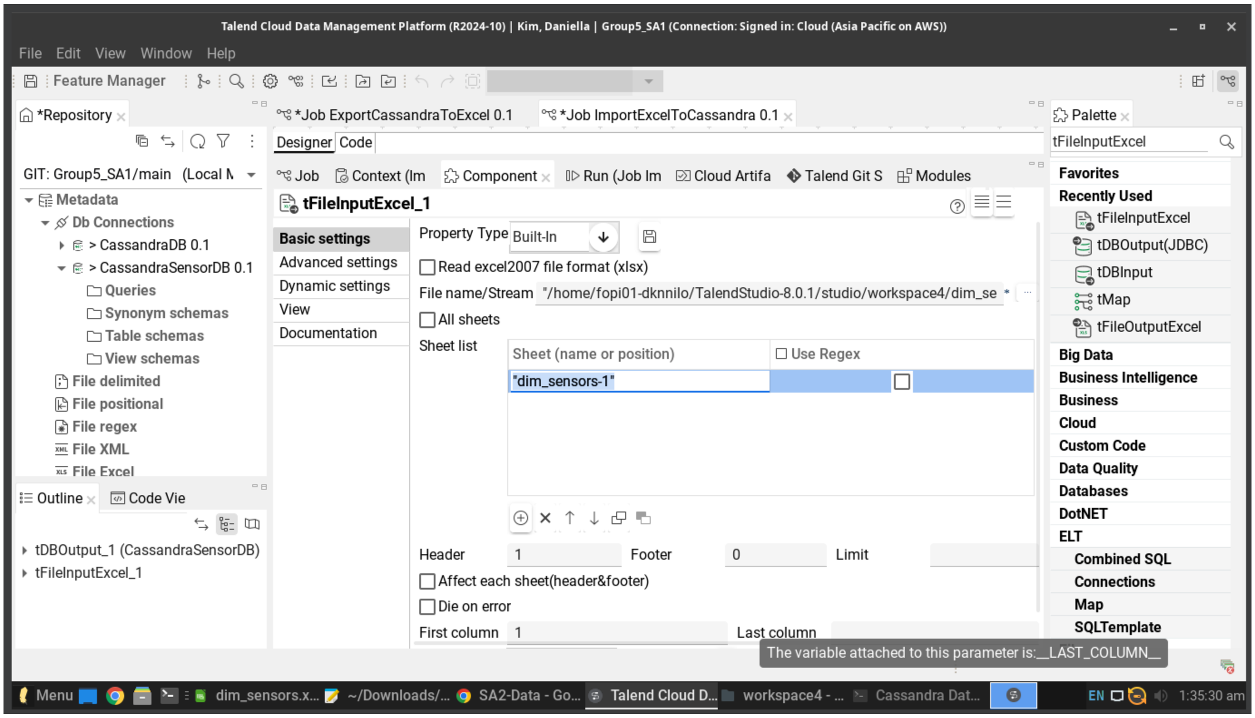
☐ Specify a data source alias

Table: sensor_data.dim_sensors

Data action: Insert

☐ Clear data in table

☐ Die on error



Talend Cloud Data Management Platform (R2024-10) | Kim, Daniella | Group5_SA1 (Connection: Signed in: Cloud (Asia Pacific on AWS))

*Job ExportCassandraToExcel 0.1 | *Job ImportExcelToCassandra 0.1

tFileInputExcel_1

Basic settings

Property Type: Built-In

☐ Read excel2007 file format (xlsx)

File name/Stream: "/home/fopi01-dknnilo/TalendStudio-8.0.1/studio/workspace4/dim_se

☐ All sheets

Sheet list:

Sheet (name or position)	☐ Use Regex
"dim_sensors-1"	☐

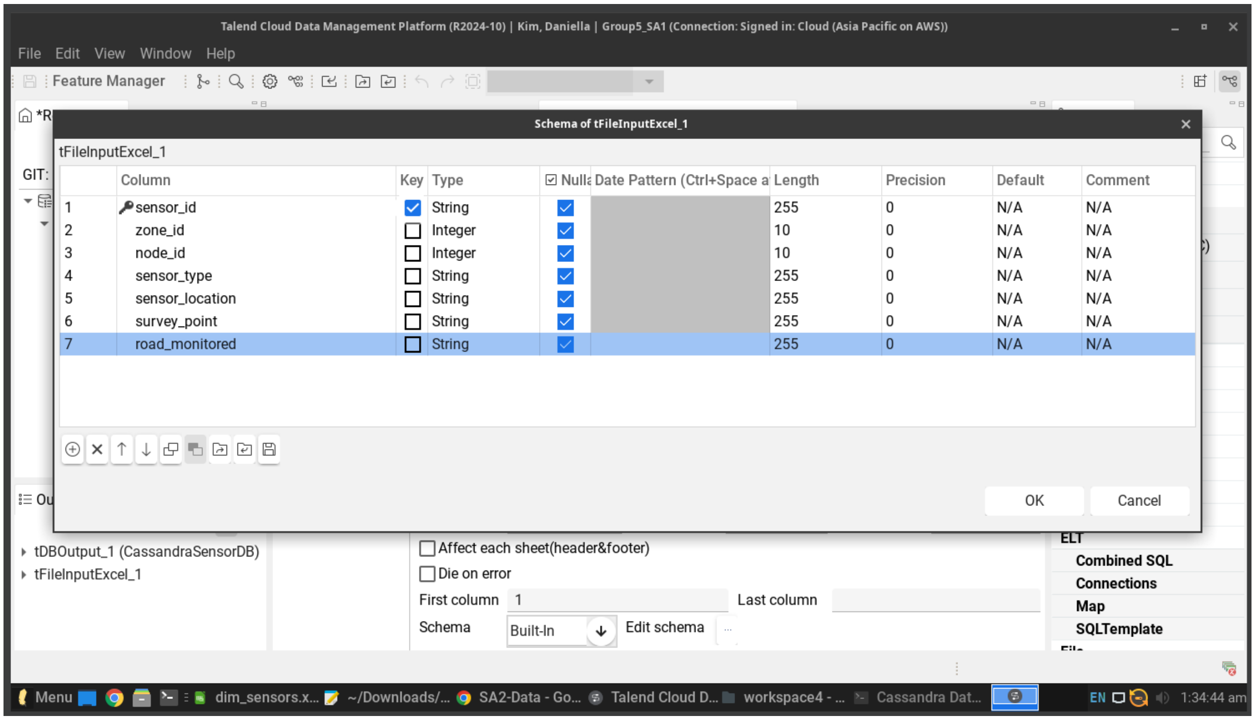
Header: 1 Footer: 0 Limit:

☐ Affect each sheet(header&footer)

☐ Die on error

First column: 1 Last column:

The variable attached to this parameter is:__LAST_COLUMN__



Talend Cloud Data Management Platform (R2024-10) | Kim, Daniella | Group5_SA1 (Connection: Signed in: Cloud (Asia Pacific on AWS))

Schema of tFileInputExcel_1

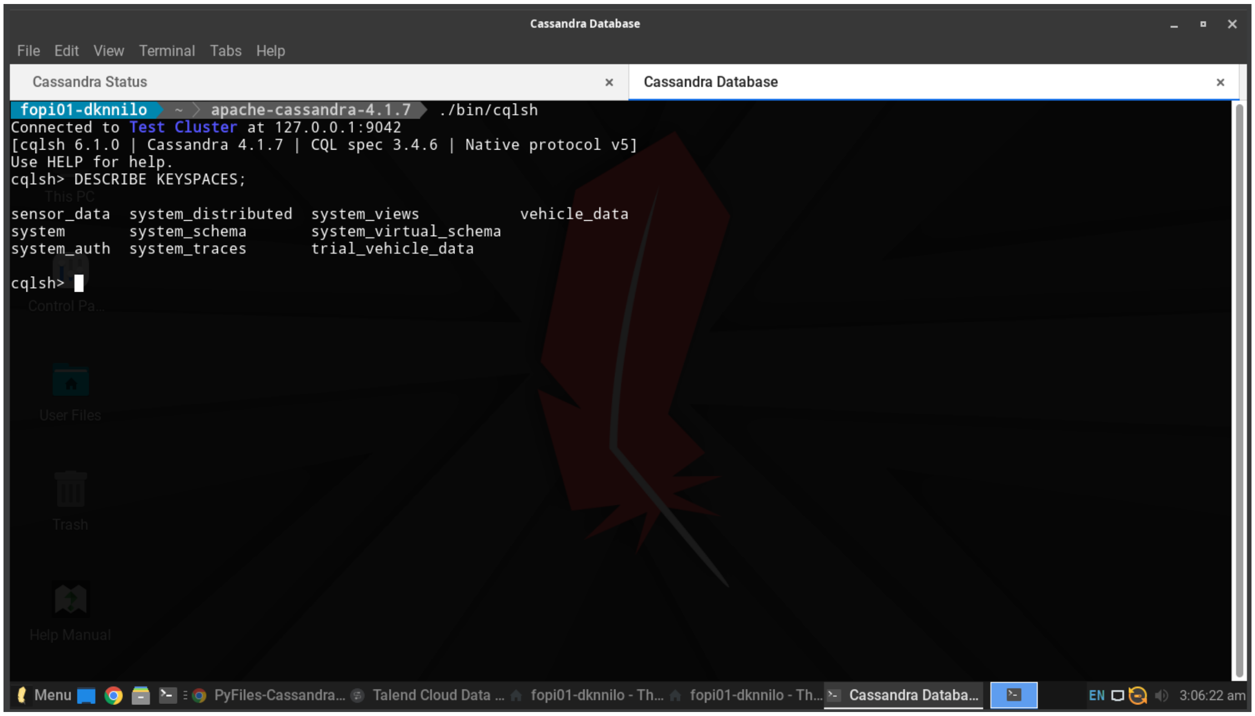
tFileInputExcel_1

	Column	Key	Type	☑ Nulla	Date Pattern (Ctrl+Space a	Length	Precision	Default	Comment
1	sensor_id	☑	String	☑		255	0	N/A	N/A
2	zone_id	□	Integer	☑		10	0	N/A	N/A
3	node_id	□	Integer	☑		10	0	N/A	N/A
4	sensor_type	□	String	☑		255	0	N/A	N/A
5	sensor_location	□	String	☑		255	0	N/A	N/A
6	survey_point	□	String	☑		255	0	N/A	N/A
7	road_monitored	□	String	☑		255	0	N/A	N/A

tDBOutput_1 (CassandraSensorDB)
tFileInputExcel_1

□ Affect each sheet(header&footer)
□ Die on error
First column: 1
Last column:
Schema: Built-In
Edit schema

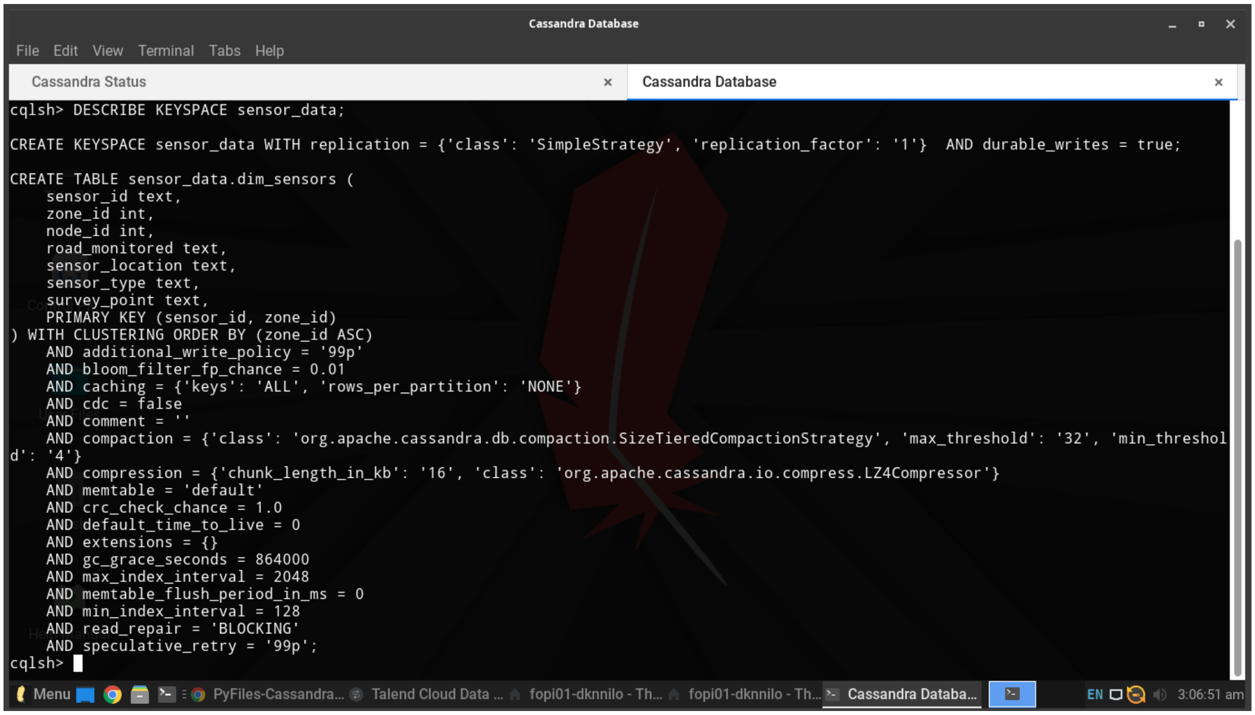
The next step is to connect the `tFileInputExcel_1` to `CassandraSensorDB(tDBOuput_1)` by manually establishing a connection line between the two components. Run the job and verify if the data is imported by querying in Cassandra. Use the command `DESCRIBE_KEYSPACE` to determine if the `sensor_data` keyspace is present. After confirming it is present, use the command `DESCRIBE KEYSPACE sensor_data;` to view the necessary information about the keyspace and the table present in the keyspace.



```
 fopi01-dknnilo  ~  apache-cassandra-4.1.7  ./bin/cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.7 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> DESCRIBE KEYSPACES;

sensor_data  system_distributed  system_views          vehicle_data
system       system_schema       system_virtual_schema
system_auth  system_traces       trial_vehicle_data

cqlsh> 
```

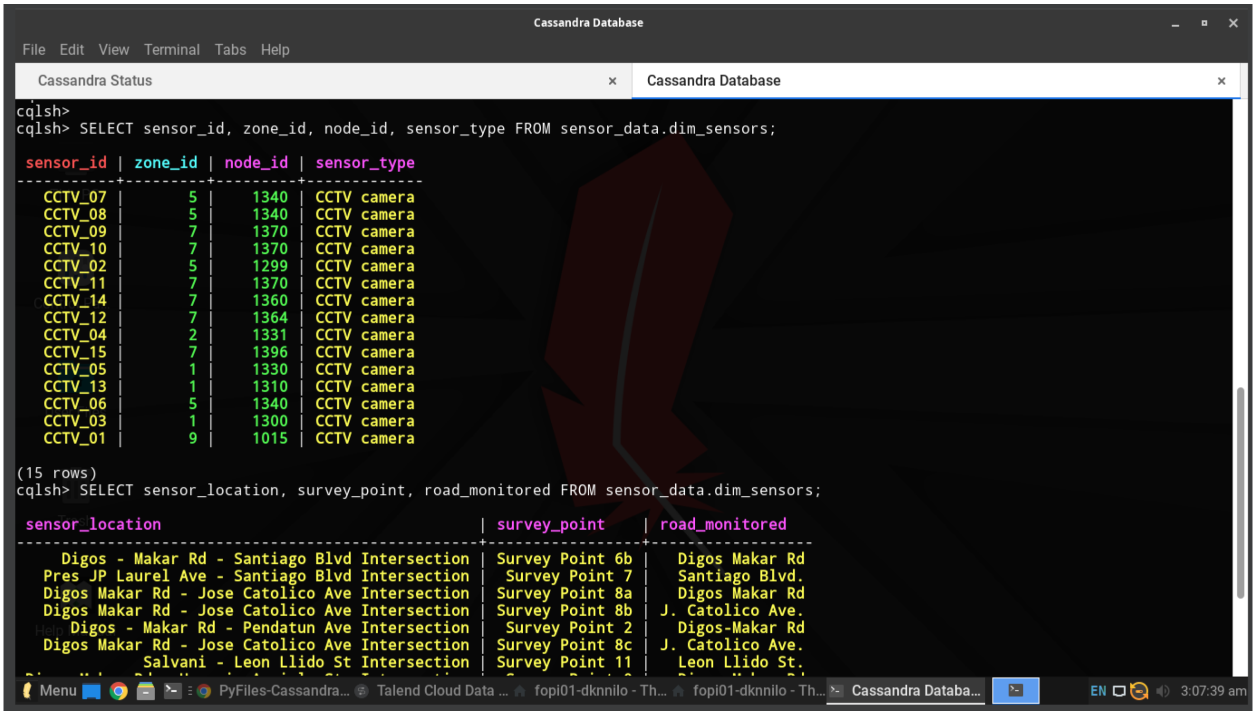


```
cqlsh> DESCRIBE KEYSPACE sensor_data;

CREATE KEYSPACE sensor_data WITH replication = {'class': 'SimpleStrategy', 'replication_factor': '1'}  AND durable_writes = true;

CREATE TABLE sensor_data.dim_sensors (
    sensor_id text,
    zone_id int,
    node_id int,
    road_monitored text,
    sensor_location text,
    sensor_type text,
    survey_point text,
    PRIMARY KEY (sensor_id, zone_id)
) WITH CLUSTERING ORDER BY (zone_id ASC)
    AND additional_write_policy = '99p'
    AND bloom_filter_fp_chance = 0.01
    AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
    AND cdc = false
    AND comment = ''
    AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshol
d': '4'}
    AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
    AND memtable = 'default'
    AND crc_check_chance = 1.0
    AND default_time_to_live = 0
    AND extensions = {}
    AND gc_grace_seconds = 864000
    AND max_index_interval = 2048
    AND memtable_flush_period_in_ms = 0
    AND min_index_interval = 128
    AND read_repair = 'BLOCKING'
    AND speculative_retry = '99p';
cqlsh> 
```

Afterwards, display the data using the commands `SELECT sensor_id, zone_id, node_id, sensor_type FROM sensor_data.dim_sensors;` to view the first 4 columns of the table and `SELECT sensor_location, survey_point, road_monitored FROM sensor_data.dim_sensors;` to view the location of the sensor, their survey point, and the road assigned to them for monitoring.

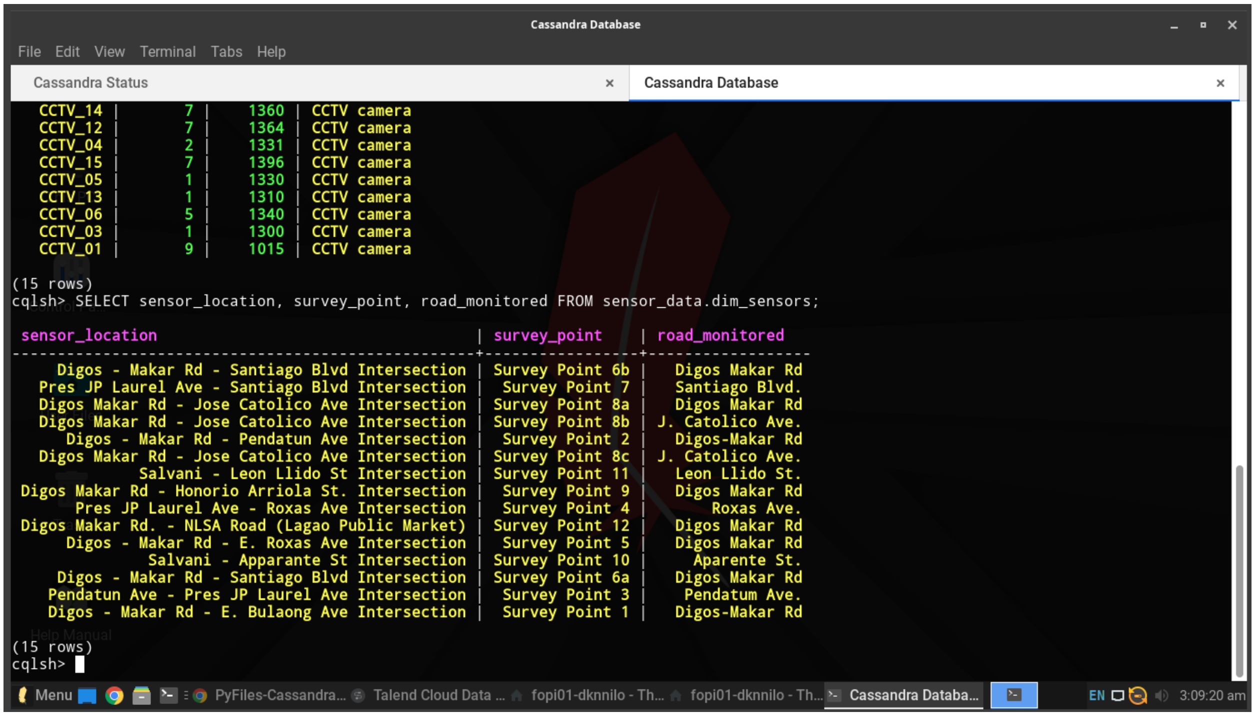


```
cqlsh>
cqlsh> SELECT sensor_id, zone_id, node_id, sensor_type FROM sensor_data.dim_sensors;

 sensor_id | zone_id | node_id | sensor_type
-----------+---------+---------+-------------
   CCTV_07 |       5 |    1340 | CCTV camera
   CCTV_08 |       5 |    1340 | CCTV camera
   CCTV_09 |       7 |    1370 | CCTV camera
   CCTV_10 |       7 |    1370 | CCTV camera
   CCTV_02 |       5 |    1299 | CCTV camera
   CCTV_11 |       7 |    1370 | CCTV camera
   CCTV_14 |       7 |    1360 | CCTV camera
   CCTV_12 |       7 |    1364 | CCTV camera
   CCTV_04 |       2 |    1331 | CCTV camera
   CCTV_15 |       7 |    1396 | CCTV camera
   CCTV_05 |       1 |    1330 | CCTV camera
   CCTV_13 |       1 |    1310 | CCTV camera
   CCTV_06 |       5 |    1340 | CCTV camera
   CCTV_03 |       1 |    1300 | CCTV camera
   CCTV_01 |       9 |    1015 | CCTV camera

(15 rows)
cqlsh> SELECT sensor_location, survey_point, road_monitored FROM sensor_data.dim_sensors;

 sensor_location                                   | survey_point     | road_monitored
---------------------------------------------------+------------------+------------------
     Digos - Makar Rd - Santiago Blvd Intersection | Survey Point 6b |    Digos Makar Rd
   Pres JP Laurel Ave - Santiago Blvd Intersection |  Survey Point 7 |    Santiago Blvd.
   Digos Makar Rd - Jose Catolico Ave Intersection | Survey Point 8a |    Digos Makar Rd
   Digos Makar Rd - Jose Catolico Ave Intersection | Survey Point 8b | J. Catolico Ave.
      Digos - Makar Rd - Pendatun Ave Intersection |  Survey Point 2 |    Digos-Makar Rd
   Digos Makar Rd - Jose Catolico Ave Intersection | Survey Point 8c | J. Catolico Ave.
              Salvani - Leon Llido St Intersection | Survey Point 11 |    Leon Llido St.
```



```
    CCTV_14 |        7 |     1360 | CCTV camera
    CCTV_12 |        7 |     1364 | CCTV camera
    CCTV_04 |        2 |     1331 | CCTV camera
    CCTV_15 |        7 |     1396 | CCTV camera
    CCTV_05 |        1 |     1330 | CCTV camera
    CCTV_13 |        1 |     1310 | CCTV camera
    CCTV_06 |        5 |     1340 | CCTV camera
    CCTV_03 |        1 |     1300 | CCTV camera
    CCTV_01 |        9 |     1015 | CCTV camera

(15 rows)
cqlsh> SELECT sensor_location, survey_point, road_monitored FROM sensor_data.dim_sensors;

 sensor_location                                   | survey_point    | road_monitored
---------------------------------------------------+-----------------+------------------
     Digos - Makar Rd - Santiago Blvd Intersection | Survey Point 6b |   Digos Makar Rd
   Pres JP Laurel Ave - Santiago Blvd Intersection |  Survey Point 7 |   Santiago Blvd.
   Digos Makar Rd - Jose Catolico Ave Intersection | Survey Point 8a |   Digos Makar Rd
   Digos Makar Rd - Jose Catolico Ave Intersection | Survey Point 8b | J. Catolico Ave.
      Digos - Makar Rd - Pendatun Ave Intersection |  Survey Point 2 |   Digos-Makar Rd
   Digos Makar Rd - Jose Catolico Ave Intersection | Survey Point 8c | J. Catolico Ave.
              Salvani - Leon Llido St Intersection | Survey Point 11 |   Leon Llido St.
 Digos Makar Rd - Honorio Arriola St. Intersection |  Survey Point 9 |   Digos Makar Rd
       Pres JP Laurel Ave - Roxas Ave Intersection |  Survey Point 4 |       Roxas Ave.
 Digos Makar Rd. - NLSA Road (Lagao Public Market) | Survey Point 12 |   Digos Makar Rd
      Digos - Makar Rd - E. Roxas Ave Intersection |  Survey Point 5 |   Digos Makar Rd
               Salvani - Apparante St Intersection | Survey Point 10 |     Aparente St.
     Digos - Makar Rd - Santiago Blvd Intersection | Survey Point 6a |   Digos Makar Rd
    Pendatun Ave - Pres JP Laurel Ave Intersection |  Survey Point 3 |    Pendatum Ave.
    Digos - Makar Rd - E. Bulaong Ave Intersection |  Survey Point 1 |   Digos-Makar Rd

(15 rows)
cqlsh> 
```

The order of the queries in the image is unsorted since Cassandra is designed for rapid distributed data access rather than preserving the order of the responses. The clustering columns sort the data in Cassandra. If the clustering columns are not defined, Cassandra is unable to sort rows within a partition. More specifically, if the table simply has a partition key, such as a `PRIMARY KEY`, the rows within the partition are not sorted in a certain order by default. Since the sorting between partitions is not supported, the queried data from different partitions appears to be unsorted.

REFERENCES:

Qlik Talend. (n.d.). *tFileInputExcel*. tFileInputExcel | Talend Components for Jobs Help. https://help.qlik.com/talend/en-US/components/8.0/excel/tfileinputexcel

Rowe, W. (2019, January 21). *Partition key vs composite key vs clustering columns in cassandra*. BMC Blogs. https://www.bmc.com/blogs/cassandra-clustering-columns-partition-composite-key/