

ASSIGNMENT 1

1st transformation extracts an Excel file and destructively load the contents to a MySQL table. You can use any Excel file of your choice.

2nd transformation extracts a JSON file and destructively load some of the contents in the file to another MySQL table. You can use any JSON file other than the one that was used in class.

3rd transformation reads one of the MySQL targets tables above as a source table and incrementally load the data into a new table.

Build a Kettle job that includes all those three transformations. The name of the job must be assign1_main_[FirstName].kjb.

ASSIGNMENT 1

The screenshot displays the Spoon - Excel extraction (changed) application window. The main workspace shows a workflow diagram with a 'Microsoft Excel input' connector linked to a 'Table output' connector. A configuration dialog for the 'Table output' connector is open, showing the following settings:

- Step name: Table output
- Connection: mysql
- Target schema: (empty)
- Target table: Assignment1_week1
- Commit size: 1000
- Truncate table: ☐
- Ignore insert errors: ☐
- Specify database fields: ☐
- Main options: Database fields
- Partition data over tables: ☐
- Partitioning field: (empty)
- Partition data per month: ☐
- Partition data per day: ☐
- Use batch update for inserts: ☒

Below the configuration dialog, there are buttons for 'Help', 'OK', 'Cancel', and 'SQL'. To the right of the main workspace, a 'Simple SQL editor' window is open, displaying the following SQL statement:

```
CREATE TABLE Assignment1_week1
(
  Churnr TINYTEXT
)
```

The 'Results of the SQL statement...' window is also open, showing the results of the SQL execution:

```
SQL executed: CREATE TABLE Assignment1_week1
(
  Churnr TINYTEXT
, AccountWeeks TINYTEXT
, ContractRenewal TINYTEXT
, DataPlan TINYTEXT
, DataUsage TINYTEXT
, CustServCalls TINYTEXT
, DayMins TINYTEXT
, DayCalls TINYTEXT
, MonthlyCharge TINYTEXT
, OverageFee TINYTEXT
, RoadMins TINYTEXT
)
1 SQL statements executed
```

Excel extraction

ASSIGNMENT 1

The screenshot displays the Apache NiFi web console. At the top, there are tabs for 'Welcome!', 'Excel extraction', 'Json Extraction', and 'Table to table'. Below the tabs is a toolbar with various icons and a '100%' zoom level. The main workspace shows a data flow diagram with two components: 'Microsoft Excel input' and 'Table output', connected by a directed arrow. Both components have a green checkmark icon above them. Below the workspace is the 'Execution Results' panel, which includes tabs for 'Logging', 'Execution History', 'Step Metrics', 'Performance Graph', 'Metrics', and 'Preview data'. The 'Logging' tab is selected, showing a list of log entries with timestamps and messages.

Execution Results

- 2021/09/20 14:59:26 - Spoon - Using legacy execution engine
- 2021/09/20 14:59:26 - Spoon - Transformation opened.
- 2021/09/20 14:59:26 - Spoon - Launching transformation [Excel extraction]...
- 2021/09/20 14:59:26 - Spoon - Started the transformation execution.
- 2021/09/20 14:59:26 - Excel extraction - Dispatching started for transformation [Excel extraction]
- 2021/09/20 14:59:26 - Table output.0 - Connected to database [mysql] (commit=1000)
- 2021/09/20 14:59:28 - Microsoft Excel input.0 - Finished processing (I=3333, O=0, R=0, W=3333, U=0, E=0)
- 2021/09/20 14:59:28 - Table output.0 - Finished processing (I=0, O=3333, R=3333, W=3333, U=0, E=0)
- 2021/09/20 14:59:28 - Spoon - The transformation has finished!!

Excel Transformation

ASSIGNMENT 1

The screenshot displays the Apache NiFi web console interface. At the top, there are tabs for 'Welcome!', 'Excel extraction', 'Json Extraction', and 'Table to table'. Below the tabs is a toolbar with various icons and a '100%' zoom level. The main workspace shows a data flow diagram with two components: 'JSON input' (represented by a document icon with a green checkmark) and 'json_extraction' (represented by a database icon with a green checkmark). An arrow points from 'JSON input' to 'json_extraction'. Below the workspace, the 'Execution Results' section is visible, showing a log of events. The log includes timestamps and messages such as 'Spoon - Using legacy execution engine', 'Spoon - Transformation opened.', 'Spoon - Launching transformation [Json Extraction]...', 'Spoon - Started the transformation execution.', 'Json Extraction - Dispatching started for transformation [Json Extraction]', 'json_extraction.0 - Connected to database [mysql] (commit=1000)', 'JSON input.0 - Finished processing (I=3, O=0, R=0, W=3, U=0, E=0)', 'json_extraction.0 - Finished processing (I=0, O=3, R=3, W=3, U=0, E=0)', and 'Spoon - The transformation has finished!!'.

JSON input → json_extraction

Execution Results

Logging Execution History Step Metrics Performance Graph Metrics Preview data

2021/09/20 15:52:43 - Spoon - Using legacy execution engine
2021/09/20 15:52:43 - Spoon - Transformation opened.
2021/09/20 15:52:43 - Spoon - Launching transformation [Json Extraction]...
2021/09/20 15:52:43 - Spoon - Started the transformation execution.
2021/09/20 15:52:43 - Json Extraction - Dispatching started for transformation [Json Extraction]
2021/09/20 15:52:43 - json_extraction.0 - Connected to database [mysql] (commit=1000)
2021/09/20 15:52:43 - JSON input.0 - Finished processing (I=3, O=0, R=0, W=3, U=0, E=0)
2021/09/20 15:52:43 - json_extraction.0 - Finished processing (I=0, O=3, R=3, W=3, U=0, E=0)
2021/09/20 15:52:43 - Spoon - The transformation has finished!!

JSON Transformation

ASSIGNMENT 1

The screenshot displays the Kettle (Pentaho Data Integration) interface. The top tab bar shows 'Welcome!', 'Excel extraction', 'Json Extraction', and 'Table to table'. The main workspace shows a transformation diagram with a 'Table input' step connected to a 'Table output' step. The 'Table output' step has a red error icon. The 'Execution Results' pane at the bottom shows a log of events, including an error message: 'org.pentaho.di.core.exception.KettleDatabaseException: Couldn't execute SQL: TRUNCATE TABLE cellphone_data'.

Execution Results

- 2021/09/20 16:11:10 - Spoon - Using legacy execution engine
- 2021/09/20 16:11:10 - Spoon - Transformation opened.
- 2021/09/20 16:11:10 - Spoon - Launching transformation [Table to table]..
- 2021/09/20 16:11:10 - Spoon - Started the transformation execution.
- 2021/09/20 16:11:10 - Table to table - Dispatching started for transformation [Table to table]
- 2021/09/20 16:11:10 - Table output.0 - Connected to database [localhost_backroom] (commit=1000)
- 2021/09/20 16:11:10 - Table output.0 - ERROR (version 8.3.0.0-371, build 8.3.0.0-371 from 2019-06-11 11.09.08 by buildguy) : Unexpected error
- 2021/09/20 16:11:10 - Table output.0 - ERROR (version 8.3.0.0-371, build 8.3.0.0-371 from 2019-06-11 11.09.08 by buildguy) : org.pentaho.di.core.exception.KettleDatabaseException:
- 2021/09/20 16:11:10 - Table output.0 - Couldn't execute SQL: TRUNCATE TABLE cellphone_data

Error while loading table incrementally

ASSIGNMENT 1

The screenshot displays the Apache NiFi web interface. At the top, there are tabs for 'Welcome!', 'Excel extraction', 'Json Extraction', and 'Table to table'. Below the tabs is a toolbar with various icons and a '100%' zoom level. The main workspace shows a flow diagram with two components: 'Table input' and 'Table output', connected by a directed arrow. Both components have a green checkmark icon above them. Below the workspace is the 'Execution Results' panel, which includes tabs for 'Logging', 'Execution History', 'Step Metrics', 'Performance Graph', 'Metrics', and 'Preview data'. The 'Logging' tab is selected, showing a list of log entries with timestamps and messages.

Execution Results

- 2021/09/20 16:20:12 - Spoon - Using legacy execution engine
- 2021/09/20 16:20:12 - Spoon - Transformation opened.
- 2021/09/20 16:20:12 - Spoon - Launching transformation [Table to table]...
- 2021/09/20 16:20:12 - Spoon - Started the transformation execution.
- 2021/09/20 16:20:13 - Table to table - Dispatching started for transformation [Table to table]
- 2021/09/20 16:20:13 - Table output.0 - Connected to database [mysql] (commit=1000)
- 2021/09/20 16:20:13 - Table input.0 - Finished reading query, closing connection.
- 2021/09/20 16:20:13 - Table input.0 - Finished processing (I=3333, O=0, R=0, W=3333, U=0, E=0)
- 2021/09/20 16:20:13 - Table output.0 - Finished processing (I=0, O=3333, R=3333, W=3333, U=0, E=0)
- 2021/09/20 16:20:13 - Spoon - The transformation has finished!!

Table to Table Transformation

ASSIGNMENT 1

The screenshot displays the Apache Spark Job1 execution results in Databricks. The top section shows a workflow diagram with four steps: Start, Excel extraction, Json Extraction, and Table to Table. The bottom section shows the 'Execution Results' tab with a log of events.

Execution Results

- 2021/09/20 16:28:28 - Table to Table - Using legacy execution engine
- 2021/09/20 16:28:28 - Table to table - Dispatching started for transformation [Table to table]
- 2021/09/20 16:28:28 - Table output.0 - Connected to database [mysql] (commit=1000)
- 2021/09/20 16:28:28 - Table input.0 - Finished reading query, closing connection.
- 2021/09/20 16:28:28 - Table input.0 - Finished processing (I=6666, O=0, R=0, W=6666, U=0, E=0)
- 2021/09/20 16:28:29 - Table output.0 - Finished processing (I=0, O=6666, R=6666, W=6666, U=0, E=0)
- 2021/09/20 16:28:29 - assign1_main_Narendra - Finished job entry [Table to Table] (result=[true])
- 2021/09/20 16:28:29 - assign1_main_Narendra - Finished job entry [Json Extraction] (result=[true])
- 2021/09/20 16:28:29 - assign1_main_Narendra - Finished job entry [Excel extraction] (result=[true])
- 2021/09/20 16:28:29 - assign1_main_Narendra - Job execution finished
- 2021/09/20 16:28:29 - Spoon - Job has ended.

Job1 screenshot