

Tugas Eksplorasi Knime Big Data



Aurelia Fairuz Rachmadi

05111740000141

Dosen :

Abdul Munif, S.Kom, M.Sc..

DEPARTEMEN INFORMATIKA

INSTITUT TEKNOLOGI SEPULUH NOPEMBER

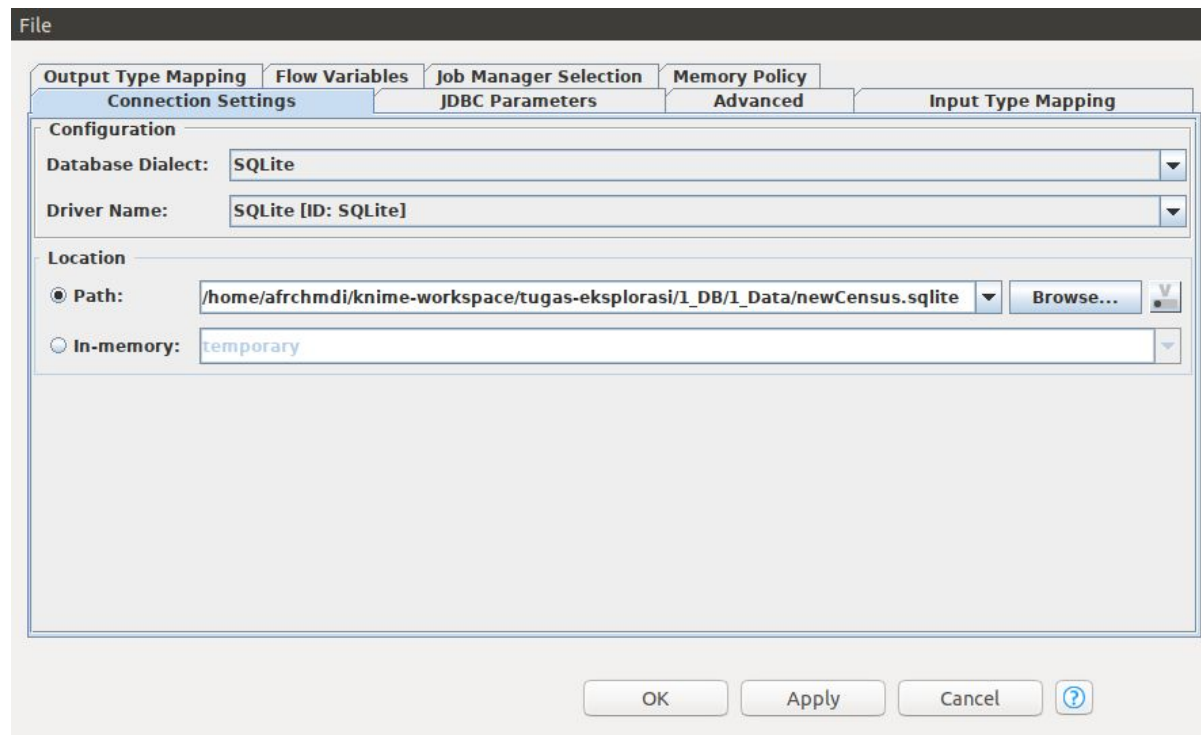
SURABAYA

1_DB

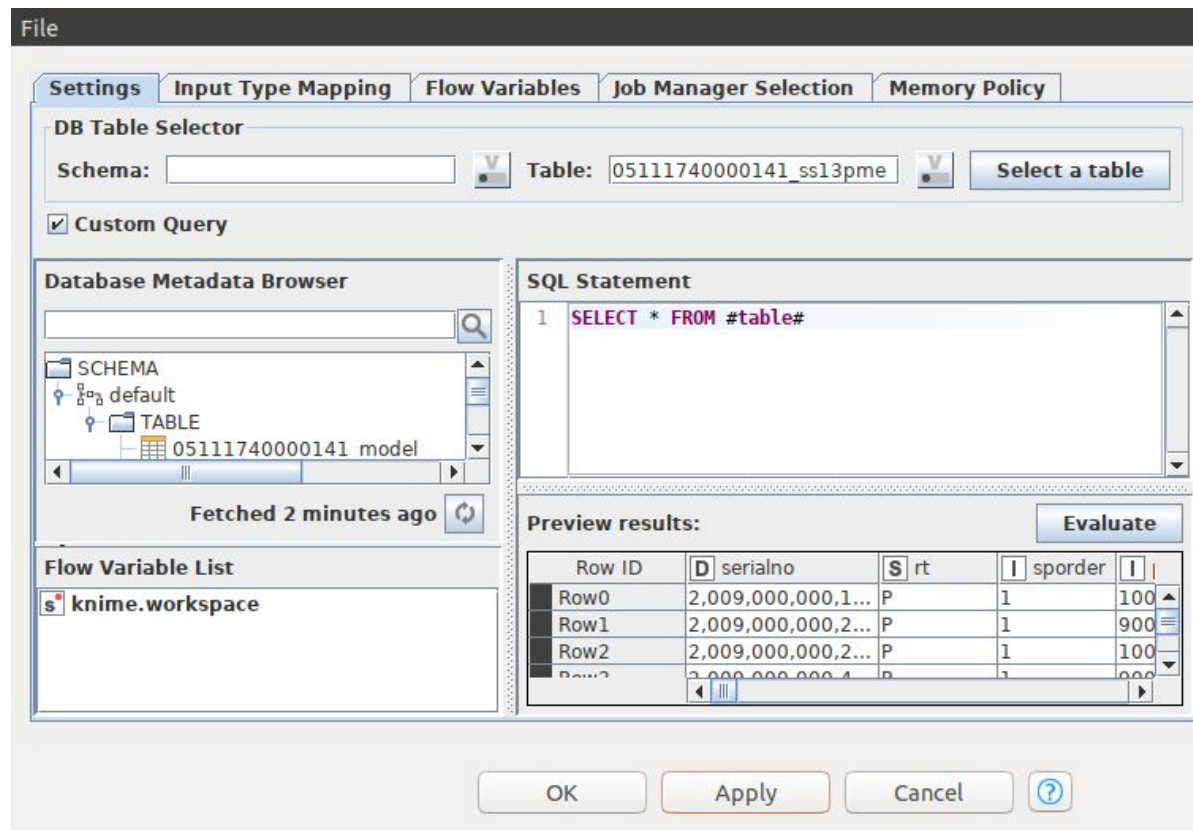
a. 01_DB_Connect

Langkah-langkah:

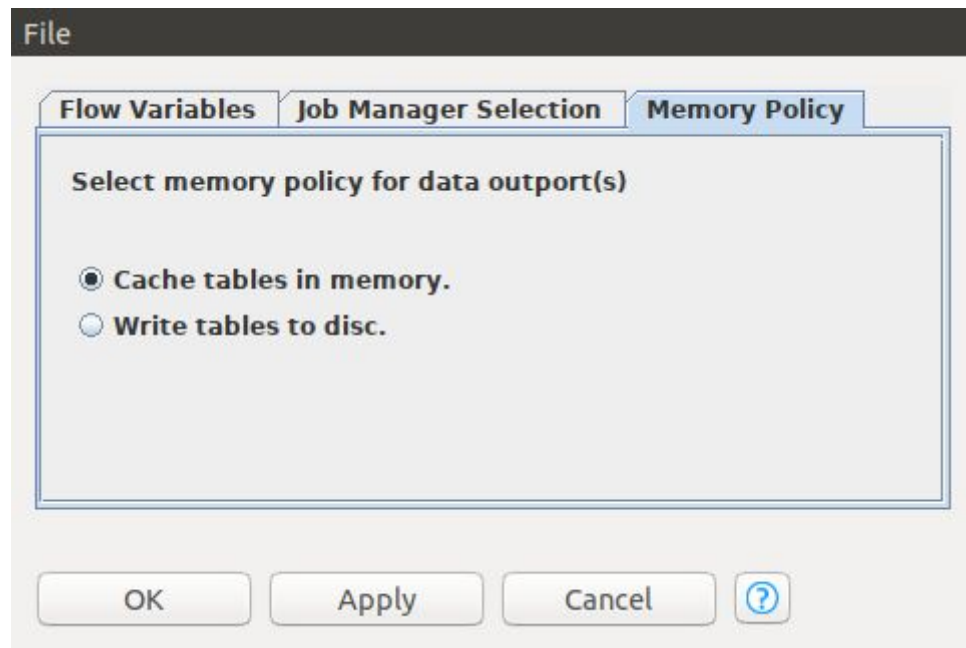
- Membuat koneksi SQLite pada file newCensus.sqlite. Pilih “SQLite Connector” dengan konfigurasi:



- Pilih tabel “05111740000141_ss13pme”. Pilih “DB Table Selector” lalu klik “select a table” dengan konfigurasi



- Import data yang tersebut pada KNIME data table
Pilih “DB Reader“dengan konfigurasi:



Setelah itu klik kanan lalu pilih “KNIME data table” untuk melihat data

File Hilite Navigation View

Table "database" - Rows: 30372 Spec - Columns: 295 Properties Flow Variables

Row ID	D serialno	S rt	I sporder	I puma00	I puma10	I st	I adjinc	I pwgtp	I agep
Row0	2,009,000,000,1...	P	1	1000	-9	23	1085467	8	68
Row1	2,009,000,000,2...	P	1	900	-9	23	1085467	7	51
Row2	2,009,000,000,2...	P	1	1000	-9	23	1085467	8	57
Row3	2,009,000,000,4...	P	1	900	-9	23	1085467	32	22
Row4	2,009,000,001,0...	P	1	500	-9	23	1085467	23	57
Row5	2,009,000,001,0...	P	1	900	-9	23	1085467	23	43
Row6	2,009,000,001,0...	P	1	1000	-9	23	1085467	5	27
Row7	2,009,000,001,2...	P	1	200	-9	23	1085467	23	22
Row8	2,009,000,001,3...	P	1	500	-9	23	1085467	7	69
Row9	2,009,000,001,4...	P	1	700	-9	23	1085467	17	47
Row10	2,009,000,001,4...	P	1	900	-9	23	1085467	29	43
Row11	2,009,000,002,0...	P	1	800	-9	23	1085467	9	48
Row12	2,009,000,002,2...	P	1	200	-9	23	1085467	15	45
Row13	2,009,000,002,7...	P	1	200	-9	23	1085467	18	68
Row14	2,009,000,003,0...	P	1	800	-9	23	1085467	13	46
Row15	2,009,000,003,1...	P	1	900	-9	23	1085467	20	66
Row16	2,009,000,003,2...	P	1	100	-9	23	1085467	20	66
Row17	2,009,000,003,6...	P	1	800	-9	23	1085467	11	84
Row18	2,009,000,004,3...	P	1	1000	-9	23	1085467	20	53
Row19	2,009,000,004,3...	P	1	800	-9	23	1085467	23	94
Row20	2,009,000,004,4...	P	1	500	-9	23	1085467	5	66
Row21	2,009,000,004,5...	P	1	400	-9	23	1085467	93	32
Row22	2,009,000,005,3...	P	1	800	-9	23	1085467	7	59
Row23	2,009,000,005,6...	P	1	800	-9	23	1085467	6	44
Row24	2,009,000,006,3...	P	1	100	-9	23	1085467	26	48
Row25	2,009,000,007,2...	P	1	500	-9	23	1085467	9	67
Row26	2,009,000,007,5...	P	1	500	-9	23	1085467	24	60
Row27	2,009,000,007,8...	P	1	900	-9	23	1085467	20	37
Row28	2,009,000,008,1...	P	1	100	-9	23	1085467	30	50
Row29	2,009,000,008,7...	P	1	1000	-9	23	1085467	19	59
Row30	2,009,000,009,2...	P	1	900	-9	23	1085467	23	42
Row31	2,009,000,009,5...	P	1	1000	-9	23	1085467	47	79
Row32	2,009,000,009,6...	P	1	1000	-9	23	1085467	5	86
Row33	2,009,000,009,6...	P	1	800	-9	23	1085467	7	47
Row34	2,009,000,010,0...	P	1	800	-9	23	1085467	12	68
Row35	2,009,000,010,4...	P	1	1000	-9	23	1085467	6	66

b.

c. 02_DB_InDB_Processing

1. Melakukan koneksi database (SQLite) newCensus.sqlite untuk membaca tabel 05111740000141_ss13hme (house data) dan 05111740000141_ss13pme (person data) menggunakan "SQLite Connector" dengan konfigurasi berikut ini:

File

Output Type Mapping | Flow Variables | Job Manager Selection | Memory Policy

Connection Settings | JDBC Parameters | Advanced | Input Type Mapping

Configuration

Database Dialect: SQLite

Driver Name: SQLite [ID: SQLite]

Location

☒ Path: /home/afrchmdi/knime-workspace/tugas-eksplorasi/1_DB/1_Data/newCensus.sqlite Browse...

☐ In-memory: temporary

OK Apply Cancel ?

Setelah itu pilih “DB Table Selector” untuk memilih tabel 05111740000141_ss13hme and 05111740000141_ss13pme

Settings Input Type Mapping Flow Variables Job Manager Selection Memory Policy

DB Table Selector

Schema: Table: 1740000141_ss13pme

☒ Custom Query

Database Metadata Browser

05111740000141_m
05111740000141_s
05111740000141_s
05111740000141_s

Fetches 1 minute ago

Flow Variable List

knime.workspace

SQL Statement

```
1 SELECT * FROM #table#
```

Preview results:

Row ID	D serialno	S rt	I sporder
Row0	2,009,000,000,1...	P	1
Row1	2,009,000,000,2...	P	1
Row2	2,009,000,000,2...	P	1
Row3	2,009,000,000,4...	P	1

Settings Input Type Mapping Flow Variables Job Manager Selection Memory Policy

DB Table Selector

Schema: Table: 05111740000141_ss13hme

☒ Custom Query

Database Metadata Browser

05111740000141_model
05111740000141_ss13hme
05111740000141_ss13pme
05111740000141_ss13pme_d

Fetches 1 minute ago

Flow Variable List

knime.workspace

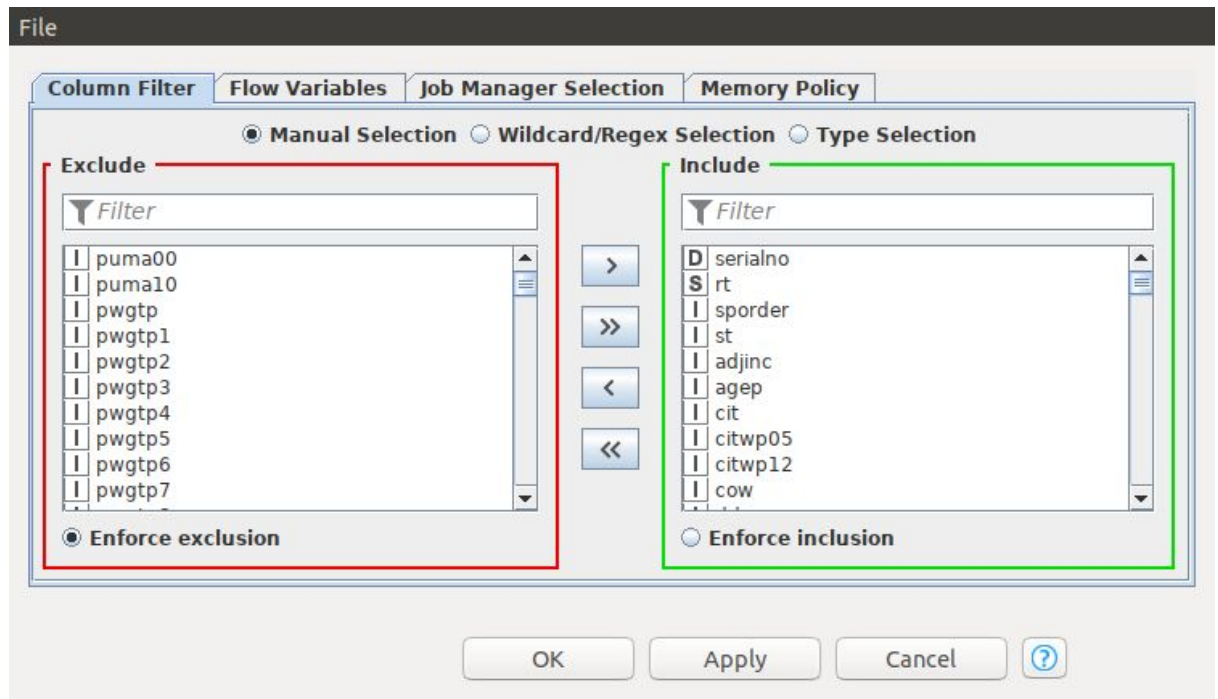
SQL Statement

```
1 SELECT * FROM #table#
```

Preview results:

Row ID	I insp	S rt	D serialno	I
Row0	500	H	2,009,000,000,1...	1
Row1	?	H	2,009,000,000,1...	1
Row2	720	H	2,009,000,000,2...	1
Row3	?	H	2,009,000,000,2...	1

2. Me-remove semua kolom yang bernama PUMA* dan PWGTP* menggunakan “DB Column Filter” dengan konfigurasi



Awalnya table 05111740000141_ss13pme memiliki 295 kolom. Tetapi setelah dilakukan filter menjadi 212 kolom.

Melakukan Join tabel 05111740000141_ss13hme and 05111740000141_ss13pme pada SERIALNO dengan menggunakan “DB Joiner” dan konfigurasi

File

Joiner Settings Column Selection Flow Variables Job Manager Selection Memory Policy

Join Mode

Join mode: Inner Join

Joining Columns

☒ Match all of the following ☐ Match any of the following

Top Input ('left' table)	Bottom Input ('right' table)		
D serialno	D serialno	+	-
		+	

OK Apply Cancel ?

Dengan menggunakan inner join serialno pada kedua tabel maka hasil penggabungannya menjadi 416 kolom

File

Table Preview DB Spec - Columns: 416 DB Query DB Session Flow Variables

Cache no. of rows: 100

Row ID	I insp	S rt	D serialno	I division	I puma00	I puma10	I region	I st	I adjhsg	I adjinc	I wgtp	I np	I type	I acr	I ags
Row0	500	H	2,009,000,000.1...	1	1000	-9	1	23	1086032	1085467	8	1	1	1	?
Row1	720	H	2,009,000,000.2...	1	900	-9	1	23	1086032	1085467	6	4	1	1	?
Row2	380	H	2,009,000,000.2...	1	1000	-9	1	23	1086032	1085467	8	2	1	2	2
Row3	?	H	2,009,000,000.4...	1	900	-9	1	23	1086032	1085467	33	6	1	1	?
Row4	1200	H	2,009,000,001.0...	1	500	-9	1	23	1086032	1085467	23	2	1	1	?
Row5	?	H	2,009,000,001.0...	1	900	-9	1	23	1086032	1085467	23	1	1	?	?
Row6	400	H	2,009,000,001.0...	1	1000	-9	1	23	1086032	1085467	5	5	1	1	?
Row7	?	H	2,009,000,001.2...	1	200	-9	1	23	1086032	1085467	0	1	3	?	?
Row8	1200	H	2,009,000,001.3...	1	500	-9	1	23	1086032	1085467	7	2	1	3	1
Row9	500	H	2,009,000,001.4...	1	700	-9	1	23	1086032	1085467	18	3	1	1	?
Row10	?	H	2,009,000,001.4...	1	900	-9	1	23	1086032	1085467	29	2	1	?	?
Row11	600	H	2,009,000,002.0...	1	800	-9	1	23	1086032	1085467	10	2	1	2	1
Row12	1900	H	2,009,000,002.2...	1	200	-9	1	23	1086032	1085467	15	5	1	1	?
Row13	800	H	2,009,000,002.7...	1	200	-9	1	23	1086032	1085467	18	2	1	2	1
Row14	0	H	2,009,000,003.0...	1	800	-9	1	23	1086032	1085467	13	2	1	2	1
Row15	580	H	2,009,000,003.1...	1	900	-9	1	23	1086032	1085467	20	1	1	1	?
Row16	330	H	2,009,000,003.2...	1	100	-9	1	23	1086032	1085467	19	3	1	2	1
Row17	600	H	2,009,000,003.6...	1	800	-9	1	23	1086032	1085467	11	3	1	2	1
Row18	430	H	2,009,000,004.3...	1	1000	-9	1	23	1086032	1085467	20	5	1	2	1
Row19	1200	H	2,009,000,004.3...	1	800	-9	1	23	1086032	1085467	24	2	1	1	?
Row20	800	H	2,009,000,004.4...	1	500	-9	1	23	1086032	1085467	5	2	1	3	1
Row21	400	H	2,009,000,004.5...	1	400	-9	1	23	1086032	1085467	94	1	1	2	1
Row22	1000	H	2,009,000,005.3...	1	800	-9	1	23	1086032	1085467	7	3	1	3	1
Row23	1000	H	2,009,000,005.6...	1	800	-9	1	23	1086032	1085467	6	3	1	2	1
Row24	0	H	2,009,000,006.3...	1	100	-9	1	23	1086032	1085467	27	2	1	2	1
Row25	800	H	2,009,000,007.2...	1	500	-9	1	23	1086032	1085467	8	2	1	2	1
Row26	1200	H	2,009,000,007.5...	1	500	-9	1	23	1086032	1085467	24	1	1	1	?
Row27	?	H	2,009,000,007.8...	1	900	-9	1	23	1086032	1085467	21	1	1	2	1
Row28	320	H	2,009,000,008.1...	1	100	-9	1	23	1086032	1085467	30	1	1	2	1
Row29	0	H	2,009,000,008.7...	1	1000	-9	1	23	1086032	1085467	20	3	1	1	?
Row30	600	H	2,009,000,009.2...	1	900	-9	1	23	1086032	1085467	22	3	1	2	1
Row31	?	H	2,009,000,009.5...	1	1000	-9	1	23	1086032	1085467	47	3	1	?	?
Row32	2500	H	2,009,000,009.6...	1	1000	-9	1	23	1086032	1085467	6	2	1	2	1
Row33	?	H	2,009,000,009.6...	1	800	-9	1	23	1086032	1085467	8	2	1	2	1

3. Melakukan filter all rows from ss13pme where COW is NULL menggunakan “DB Row Filter”. Konfigurasinya:

File

Conditions Flow Variables Job Manager Selection Memory Policy

Query View

cow IS NULL

Edit condition

Delete

I cow IS NULL

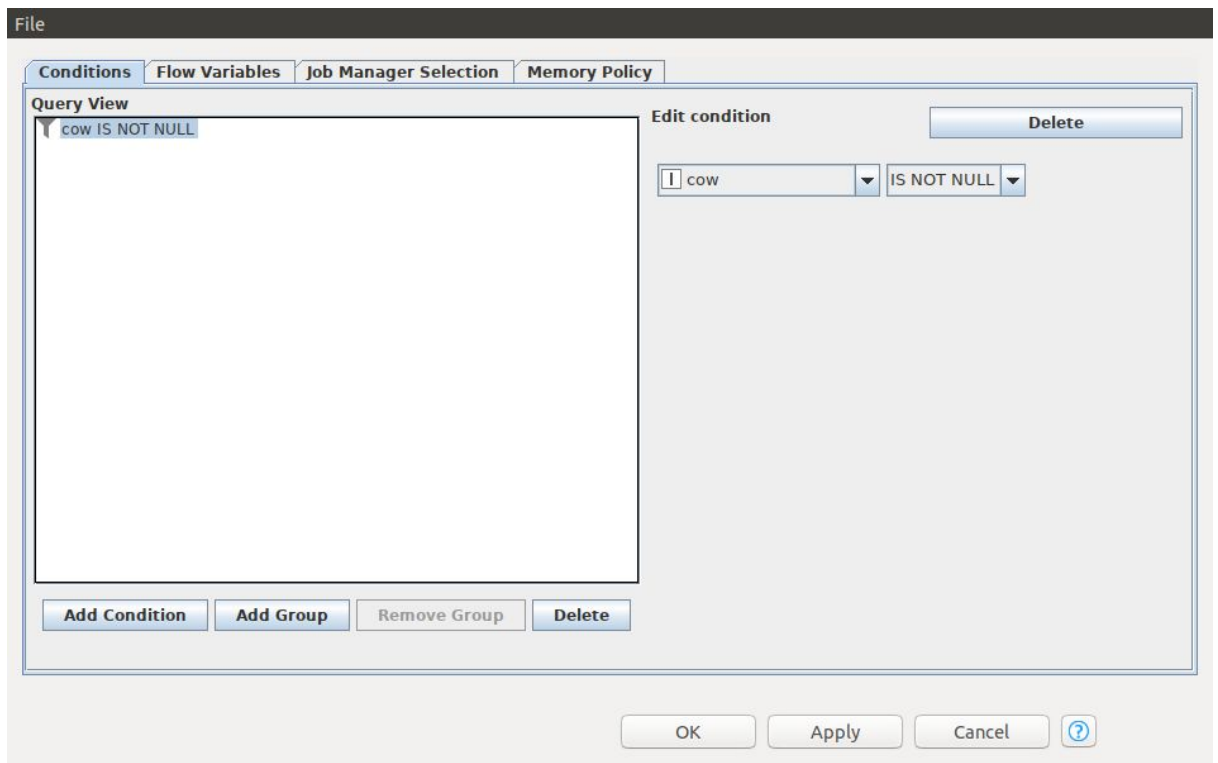
Add Condition Add Group Remove Group Delete

OK Apply Cancel ?

Hasil yang didapat adalah ada sebanyak 8662 data yang COW nya NULL

File								
Table Preview DB Spec - Columns: 212 DB Query DB Session Flow Variables								
Cache no. of rows: 100								
Row ID	D serialno	S rt	I sporder	I st	I adjinc	I agep	I cit	I citwp05
Row0	2,009,000,001,2...	P	1	23	1085467	22	5	?
Row1	2,009,000,001,3...	P	1	23	1085467	69	1	?
Row2	2,009,000,002,7...	P	1	23	1085467	68	1	?
Row3	2,009,000,003,0...	P	1	23	1085467	46	1	?
Row4	2,009,000,003,2...	P	1	23	1085467	66	1	?
Row5	2,009,000,003,6...	P	1	23	1085467	84	1	?
Row6	2,009,000,004,3...	P	1	23	1085467	94	1	?
Row7	2,009,000,004,4...	P	1	23	1085467	66	1	?
Row8	2,009,000,007,2...	P	1	23	1085467	67	1	?
Row9	2,009,000,009,5...	P	1	23	1085467	79	1	?
Row10	2,009,000,010,6...	P	1	23	1085467	56	1	?
Row11	2,009,000,011,1...	P	1	23	1085467	72	4	1987
Row12	2,009,000,013,4...	P	1	23	1085467	75	1	?
Row13	2,009,000,013,4...	P	1	23	1085467	75	1	?
Row14	2,009,000,013,6...	P	1	23	1085467	87	1	?
Row15	2,009,000,016,1...	P	1	23	1085467	86	4	1946
Row16	2,009,000,018,0...	P	1	23	1085467	89	1	?
Row17	2,009,000,018,4...	P	1	23	1085467	77	1	?
Row18	2,009,000,018,7...	P	1	23	1085467	73	1	?
Row19	2,009,000,019,2...	P	1	23	1085467	63	1	?
Row20	2,009,000,019,8...	P	1	23	1085467	94	1	?
Row21	2,009,000,020,6...	P	1	23	1085467	67	1	?
Row22	2,009,000,021,6...	P	1	23	1085467	88	3	?
Row23	2,009,000,021,6...	P	1	23	1085467	69	1	?
Row24	2,009,000,022,4...	P	1	23	1085467	85	1	?
Row25	2,009,000,024,5...	P	1	23	1085467	66	1	?
Row26	2,009,000,025,1...	P	1	23	1085467	64	1	?
Row27	2,009,000,025,5...	P	1	23	1085467	67	1	?
Row28	2,009,000,025,5...	P	1	23	1085467	70	1	?
Row29	2,009,000,025,8...	P	1	23	1085467	75	1	?
Row30	2,009,000,026,4...	P	1	23	1085467	83	4	2006
Row31	2,009,000,029,2...	P	1	23	1085467	75	1	?
Row32	2,009,000,030,2...	P	1	23	1085467	72	1	?
Row33	2,009,000,031,5...	P	1	23	1085467	82	1	?

4. Melakukan filter all rows from ss13pme where COW is NOT NULL menggunakan “DB Row Filter”. Konfigurasinya:



Hasil yang didapat adalah ada sebanyak 21710 data yang COW nya NULL

File

Table Preview DB Spec - Columns: 212 DB Query DB Session Flow Variables

Cache no. of rows: 100

Row ID	D serialno	S rt	I sporder	I st	I adjinc	I agep	I cit	I c
Row0	2,009,000,000,1...	P	1	23	1085467	68	1	?
Row1	2,009,000,000,2...	P	1	23	1085467	51	1	?
Row2	2,009,000,000,2...	P	1	23	1085467	57	1	?
Row3	2,009,000,000,4...	P	1	23	1085467	22	1	?
Row4	2,009,000,001,0...	P	1	23	1085467	57	1	?
Row5	2,009,000,001,0...	P	1	23	1085467	43	1	?
Row6	2,009,000,001,0...	P	1	23	1085467	27	1	?
Row7	2,009,000,001,4...	P	1	23	1085467	47	1	?
Row8	2,009,000,001,4...	P	1	23	1085467	43	1	?
Row9	2,009,000,002,0...	P	1	23	1085467	48	1	?
Row10	2,009,000,002,2...	P	1	23	1085467	45	3	?
Row11	2,009,000,003,1...	P	1	23	1085467	66	1	?
Row12	2,009,000,004,3...	P	1	23	1085467	53	1	?
Row13	2,009,000,004,5...	P	1	23	1085467	32	1	?
Row14	2,009,000,005,3...	P	1	23	1085467	59	1	?
Row15	2,009,000,005,6...	P	1	23	1085467	44	1	?
Row16	2,009,000,006,3...	P	1	23	1085467	48	1	?
Row17	2,009,000,007,5...	P	1	23	1085467	60	1	?
Row18	2,009,000,007,8...	P	1	23	1085467	37	1	?

5. calculate average AGEp for the different SEX groups menggunakan “DB GroupBY” dengan konfigurasi berikut ini:

File

Settings Description Flow Variables Job Manager Selection Memory Policy

Groups Manual Aggregation Pattern Based Aggregation Type Based Aggregation

Group settings

Available column(s)

Filter

serialno
rt
sporder
st
adjinc
agep
cit
citwp05
citwp12
cow
ddrs
dear
deye
dout
dphy
drat
dratx

Group column(s)

Filter

sex

Advanced settings

Column naming: Aggregation method (column name) Add COUNT(*) column name: COUNT(*)

OK Apply Cancel ?

Dan hasil yang diperoleh:

File

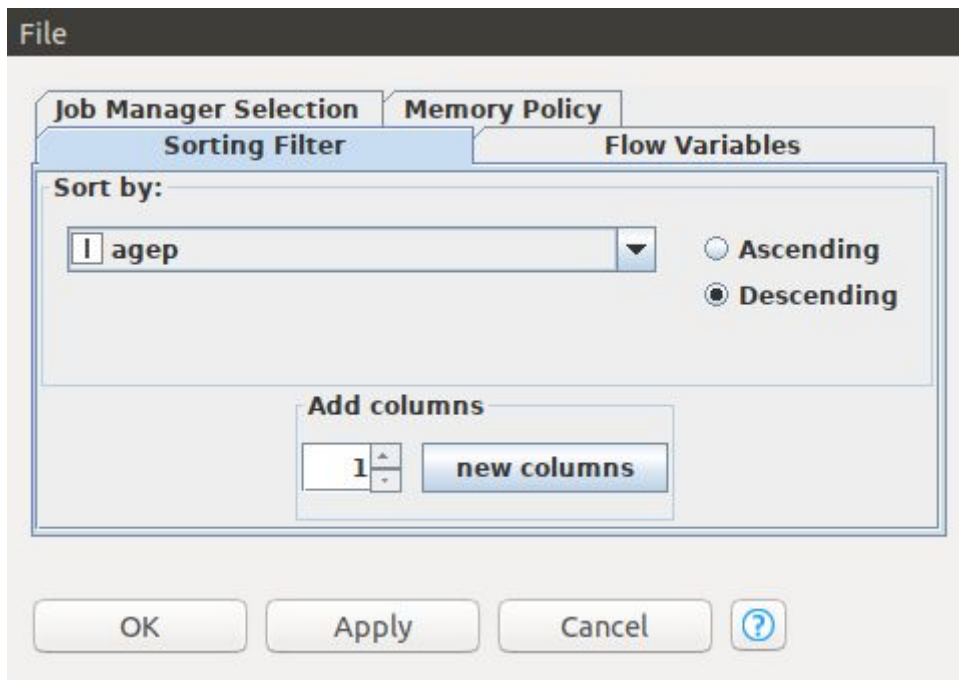
DB Query DB Session Flow Variables

Table Preview DB Spec - Columns: 3

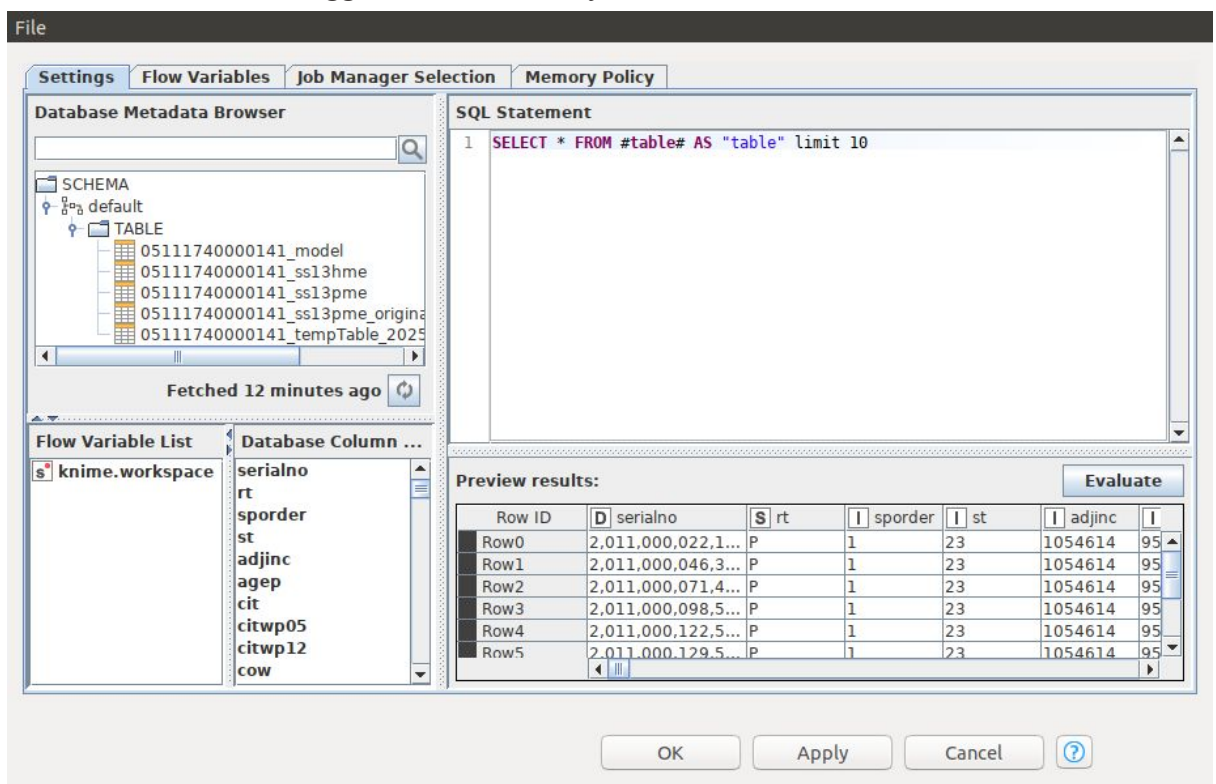
Cache no. of rows: 100

Row ID	I sex	D AVG(agep)	D CUSTOM(agep)
Row0	1	53.493	73.668
Row1	2	53.87	67.949

6. Optional. Urutkan baris data secara descending AGEp menggunakan “DB Sorter”



dan ambil 10 teratas menggunakan “DB Query”



Hasilnya:

File																
Table Preview DB Spec - Columns: 212 DB Query DB Session Flow Variables																
Cache no. of rows: 100																
Row ID	D serialno	S rt	I sporder	I st	I adjinc	I agep	I cit	I citwp05	I citwp12	I cow	I ddrs	I dear	I deye	I dout	I dphy	
Row0	2,011,000,022,1...	P	1	23	1054614	95	1	?	?	?	1	2	2	1	1	
Row1	2,011,000,046,3...	P	1	23	1054614	95	1	?	?	?	2	2	2	2	2	
Row2	2,011,000,071,4...	P	1	23	1054614	95	1	?	?	?	2	2	1	1	2	
Row3	2,011,000,098,5...	P	1	23	1054614	95	1	?	?	?	1	2	2	1	1	
Row4	2,011,000,122,5...	P	1	23	1054614	95	1	?	?	?	1	1	1	1	1	
Row5	2,011,000,129,5...	P	1	23	1054614	95	4	2007	-9	?	1	2	2	1	1	
Row6	2,011,000,149,6...	P	1	23	1054614	95	1	?	?	?	1	2	2	1	1	
Row7	2,011,000,185,1...	P	1	23	1054614	95	1	?	?	?	1	2	2	1	1	
Row8	2,011,000,210,3...	P	1	23	1054614	95	1	?	?	?	1	1	1	1	1	
Row9	2,011,000,216,9...	P	1	23	1054614	95	1	?	?	?	2	2	2	2	2	

d. 03_DB_Modelling

Pada kasus ini mirip seperti “02_DB_InDB_Processing”.

Yang membedakan adalah proses modelling menggunakan algoritma desicion tree untuk melakukan prediksi nilai COW pada kolom COW yang NULL.

Jika kolom COW nya NULL maka kolom COW tersebut di remove terlebih dahulu dengan menggunakan “DB Column Filter”. Konfigurasinya:

File			
Column Filter	Flow Variables	Job Manager Selection	Memory Policy
<input checked="" type="radio"/> Manual Selection <input type="radio"/> Wildcard/Regex Selection <input type="radio"/> Type Selection			
Exclude <div> <input type="text" value="Filter"/> <div> <input checked="" type="radio"/> Enforce exclusion </div> </div>		Include <div> <input type="text" value="Filter"/> <div> <input type="radio"/> Enforce inclusion </div> </div>	
<div> <input type="text" value="I cow"/> </div>		<div> <input type="text" value="D serialno"/> <input type="text" value="S rt"/> <input type="text" value="I sporder"/> <input type="text" value="I st"/> <input type="text" value="I adjinc"/> <input type="text" value="I agep"/> <input type="text" value="I cit"/> <input type="text" value="I citwp05"/> <input type="text" value="I citwp12"/> <input type="text" value="I ddrs"/> </div>	
<input type="button" value="OK"/>		<input type="button" value="Apply"/> <input type="button" value="Cancel"/> <input type="button" value="Help"/>	

Maka hasilnya tidak ada kolom row pada tabel tersebut:

File									
Table "database" - Rows: 8662			Spec - Columns: 211		Properties	Flow Variables			
Columns: ...	Column Type	Column I...	Color Han...	Size Hand...	Shape Ha...	Filter Han...	Lower Bound	Upper Bound	Value 0
serialno	Number (double)	0					2,009,000,001,2...	2,013,001,492,2...	?
rt	String	1					?		P
sporder	Number (integer)	2					1	1	?
st	Number (integer)	3					23	23	?
adjinc	Number (integer)	4					1,007,549	1,085,467	?
agep	Number (integer)	5					0	95	?
cit	Number (integer)	6					1	5	?
citwp05	Number (integer)	7					-9	2,009	?
citwp12	Number (integer)	8					-9	2,012	?
ddrs	Number (integer)	9					1	2	?
dear	Number (integer)	10					1	2	?
deye	Number (integer)	11					1	2	?
dout	Number (integer)	12					1	2	?
dphy	Number (integer)	13					1	2	?
drat	Number (integer)	14					1	6	?
dratx	Number (integer)	15					1	2	?
drem	Number (integer)	16					1	2	?
eng	Number (integer)	17					1	4	?
fer	Number (integer)	18					1	2	?
gcl	Number (integer)	19					1	2	?
gcm	Number (integer)	20					1	5	?
gcr	Number (integer)	21					1	2	?
hins1	Number (integer)	22					1	2	?
hins2	Number (integer)	23					1	2	?
hins3	Number (integer)	24					1	2	?
hins4	Number (integer)	25					1	2	?
hins5	Number (integer)	26					1	2	?
hins6	Number (integer)	27					1	2	?
hins7	Number (integer)	28					1	2	?
intp	Number (integer)	29					-5,300	185,000	?
jwmnp	Number (integer)	30					?	?	?
jrwp	Number (integer)	31					?	?	?
jrwr	Number (integer)	32					?	?	?
lanx	Number (integer)	33					1	2	?
mar	Number (integer)	34					1	5	?
marhd	Number (integer)	35					1	2	?

Sedangkan data yang memiliki nilai COW yang semula berformat Integer diubah menjadi String dengan menggunakan “Number to String”. Konfigurasinya:

File

Options Flow Variables Job Manager Selection Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

- ☒ serialno
- ☐ sporder
- ☐ st
- ☐ adjinc
- ☐ agep
- ☐ cit
- ☐ citwp05
- ☐ citwp12
- ☐ ddrs
- ☐ dear

☒ Enforce exclusion

Include

Filter

cow

☐ Enforce inclusion

OK Apply Cancel ?

Setelah itu melakukan modelling menggunakan algoritma desicion tree menggunakan “Desicion Tree Learner”. Konfigurasinya:

File

Job Manager Selection Memory Policy

Options PMMLSettings Flow Variables

General

Class column S cow

Quality measure Gini index

Pruning method No pruning

☒ Reduced Error Pruning

Min number records per node 2

Number records to store for view 10,000

☒ Average split point

Number threads 8

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column I fyoeep

Binary nominal splits

☐ Binary nominal splits

Max #nominal 10

☐ Filter invalid attribute values in child nodes

OK Apply Cancel ?

Dari hasil training tersebut dapat dilakukan prediksi menentukan nilai COW yang NULL menggunakan “Desicion Tree Predictor”. Berikut konfigurasinya:

File

Options

Flow Variables

Job Manager Selection

Memory Policy

Maximum number of stored patterns for HiLite-ing: 10,000

☒ Change prediction column name

COW

☐ Append columns with normalized class distribution

Suffix for probability columns

OK

Apply

Cancel

?

Hasilnya:

File HiLite Navigation View

Table "database" - Rows: 8662		Spec - Columns: 212		Properties	Flow Variables																								
Row ID	ID serialno	S	rt	I	sporder	I	st	I	adjnc	I	agep	I	cit	I	citwp05	I	citwp12	I	ddrs	I	dear	I	deye	I	dout	I	dphy	I	drat
Row0	2,009,000,001.2...	P		1	23		1085467	22	5		?		?		?	2	2	2	2	2	2	2	2	2	2	?			
Row1	2,009,000,001.3...	P		1	23		1085467	69	1		?		?		?	2	2	2	2	2	2	2	2	2	2	?			
Row2	2,009,000,002.7...	P		1	23		1085467	68	1		?		?		?	2	2	2	2	2	2	2	2	2	2	?			
Row3	2,009,000,003.0...	P		1	23		1085467	46	1		?		?		?	2	2	2	2	2	2	2	2	2	1	?			
Row4	2,009,000,003.2...	P		1	23		1085467	66	1		?		?		?	2	2	2	2	2	2	2	2	2	1	?			
Row5	2,009,000,003.6...	P		1	23		1085467	84	1		?		?		?	2	1	2	2	1	1	1	1	1	?				
Row6	2,009,000,004.3...	P		1	23		1085467	94	1		?		?		?	1	1	1	1	1	1	2	2	2	?				
Row7	2,009,000,004.4...	P		1	23		1085467	66	1		?		?		?	2	1	2	2	1	1	1	1	?					
Row8	2,009,000,007.2...	P		1	23		1085467	67	1		?		?		?	2	2	2	2	2	2	2	2	?					
Row9	2,009,000,009.5...	P		1	23		1085467	79	1		?		?		?	1	1	2	2	1	1	1	1	?					
Row10	2,009,000,010.6...	P		1	23		1085467	56	1		?		?		?	2	2	2	2	2	2	2	2	1	?				
Row11	2,009,000,011.1...	P		1	23		1085467	72	4		?		1987		-9	2	2	2	2	2	2	2	2	?					
Row12	2,009,000,013.4...	P		1	23		1085467	75	1		?		?		?	2	1	2	2	2	2	2	2	1	3				
Row13	2,009,000,013.4...	P		1	23		1085467	75	1		?		?		?	2	1	2	2	2	2	2	2	2	?				
Row14	2,009,000,013.6...	P		1	23		1085467	87	1		?		?		?	2	2	2	1	1	1	1	1	5					
Row15	2,009,000,016.1...	P		1	23		1085467	86	4		?		1946		-9	2	2	2	2	2	2	2	2	?					
Row16	2,009,000,018.0...	P		1	23		1085467	89	1		?		?		?	2	2	2	2	2	1	2	2	?					
Row17	2,009,000,018.4...	P		1	23		1085467	77	1		?		?		?	2	2	2	2	2	2	2	2	?					
Row18	2,009,000,018.7...	P		1	23		1085467	73	1		?		?		?	2	2	2	2	2	2	2	2	?					
Row19	2,009,000,019.2...	P		1	23		1085467	63	1		?		?		?	2	2	2	2	2	2	2	2	?					
Row20	2,009,000,019.8...	P		1	23		1085467	94	1		?		?		?	1	1	1	1	1	1	1	?						
Row21	2,009,000,020.6...	P		1	23		1085467	67	1		?		?		?	2	1	1	2	2	2	2	1	?					
Row22	2,009,000,021.6...	P		1	23		1085467	88	3		?		?		?	2	1	2	2	2	2	2	1	?					
Row23	2,009,000,021.6...	P		1	23		1085467	69	1		?		?		?	2	2	2	2	2	2	2	1	?					
Row24	2,009,000,022.4...	P		1	23		1085467	85	1		?		?		?	2	2	2	2	2	2	2	2	?					
Row25	2,009,000,024.5...	P		1	23		1085467	66	1		?		?		?	1	1	1	1	1	1	1	?						
Row26	2,009,000,025.1...	P		1	23		1085467	64	1		?		?		?	2	1	2	2	2	2	2	2	?					
Row27	2,009,000,025.5...	P		1	23		1085467	67	1		?		?		?	2	2	2	2	2	2	2	2	?					
Row28	2,009,000,025.5...	P		1	23		1085467	70	1		?		?		?	2	2	2	2	2	2	2	2	?					
Row29	2,009,000,025.8...	P		1	23		1085467	75	1		?		?		?	2	2	2	2	2	2	2	1	?					
Row30	2,009,000,026.4...	P		1	23		1085467	83	4		?		2006		-9	1	2	2	2	2	1	1	1	?					
Row31	2,009,000,029.2...	P		1	23		1085467	75	1		?		?		?	2	2	2	2	2	1	1	1	?					
Row32	2,009,000,030.2...	P		1	23		1085467	72	1		?		?		?	2	2	2	2	2	2	2	2	?					
Row33	2,009,000,031.5...	P		1	23		1085467	83	1		?		?		?	2	2	2	2	2	1	1	1	?					
Row34	2,009,000,032.4...	P		1	23		1085467	79	1		?		?		?	2	1	2	2	2	2	2	1	?					
Row35	2,009,000,032.4...	P		1	23		1085467	85	1		?		?		?	1	2	2	2	2	1	1	1	?					

e. 04_DB_WritingToDB

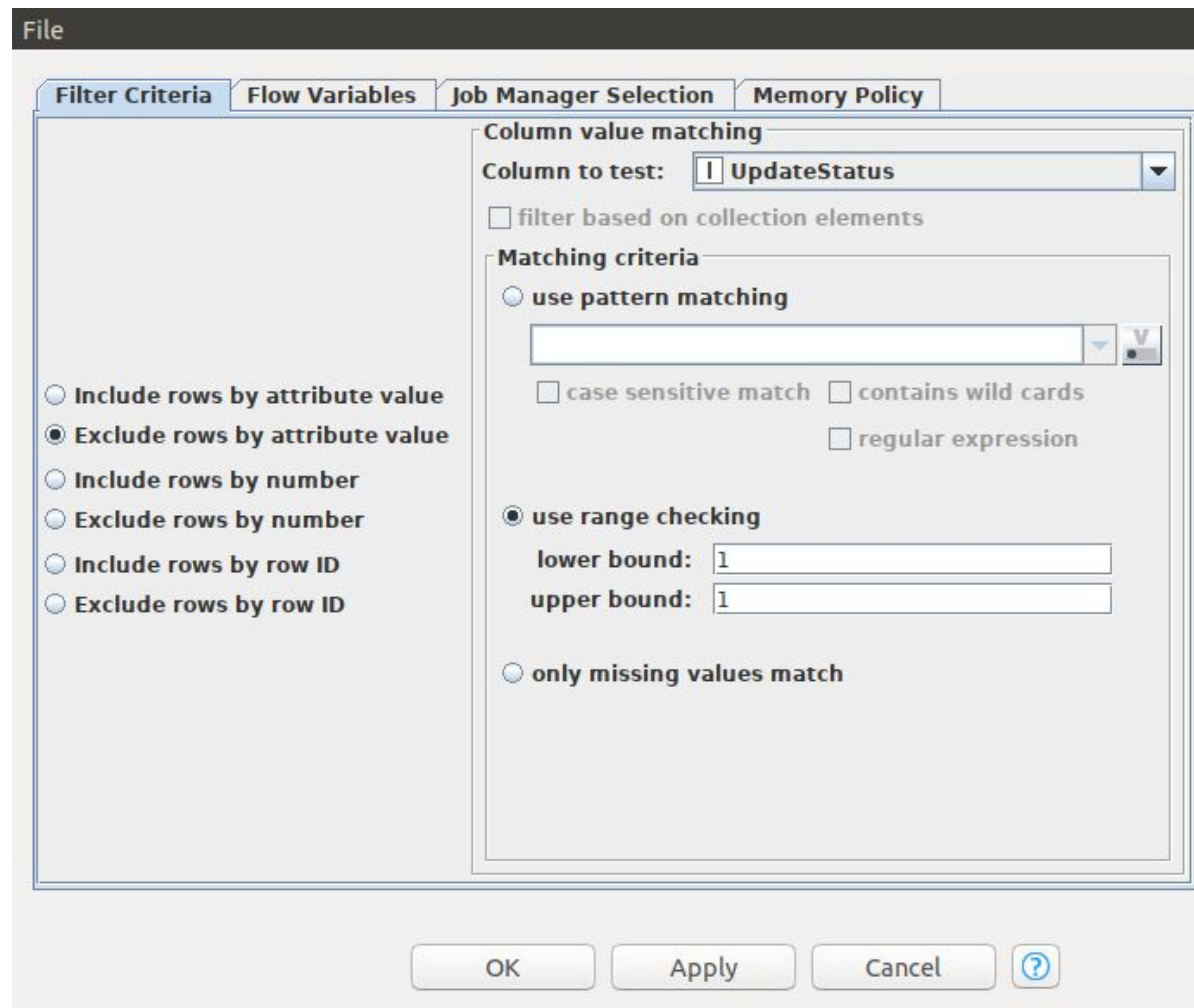
Pada bagian ini merupakan penambahan dari bagian “03_DB_Modelling”. Yaitu:

1. Tulis prediksi COW yang nilai COW nya NULL menggunakan “DB Update”
Konfigurasinya:

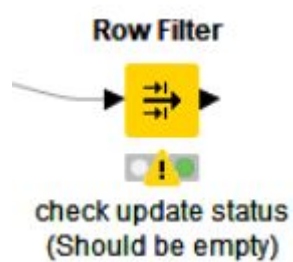
The screenshot shows the 'DB Update' configuration window. The 'Table to update' section is set to '05111740000141_ss13pme'. The 'Batch Size' is 1,000. The 'Fail on error' and 'Append update status columns' options are checked. The 'Select the columns to update (SET in SQL)' section is configured with 'Manual Selection'. The 'Exclude' list contains columns: serialno, rt, sporder, st, adjinc, agep, cit, citwp05, citwp12. The 'Include' list contains the column: cow. The 'Select identification columns (WHERE in SQL)' section is also configured with 'Manual Selection'. The 'Exclude' list contains columns: rt, sporder, st, adjinc, agep, cit, citwp05, citwp12, ddrs. The 'Include' list contains the column: serialno.

Hasilnya kolom ROW tidak ada yang NULL

2. Tulis original table dengan nama tabel 05111740000141_ss13pme_original menggunakan “DB Connection Table Writer”.
Untuk memastikan update berhasil maka lakukan konfigurasi berikut dengan menggunakan “Row Filter”:



Hasilnya empty:



3. writes model and timestamp with a Database Writer node. Pertama lakukan setting timestamps dengan menggunakan 'timestamp & model'

File

Options **Job Manager Selection**

Select the job manager for this node

<<default>>

Settings for selected job manager

OK Apply Cancel ?

Lalu dengan menggunakan “DB Writer” lakukan konfigurasi berikut ini:

File

Settings **Output Type Mapping** **Flow Variables** **Job Manager Selection** **Memory Policy**

Table to write

Schema: Table: 05111740000141_ss13pme_original Select a table

Batch Size: 1,000 ☒ Fail on error ☒ Append write status columns ☐ Disable DB Data output port ☐ Remove existing table

Select the columns to write (SET in SQL)

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude

Filter

No columns in this list

☒ Enforce exclusion

Include

Filter

☒ Date&Time ☒ PMML

☐ Enforce inclusion

OK Apply Cancel ?

2_Hadoop\2_Exercise

1. 00_Setup_Hive_Table

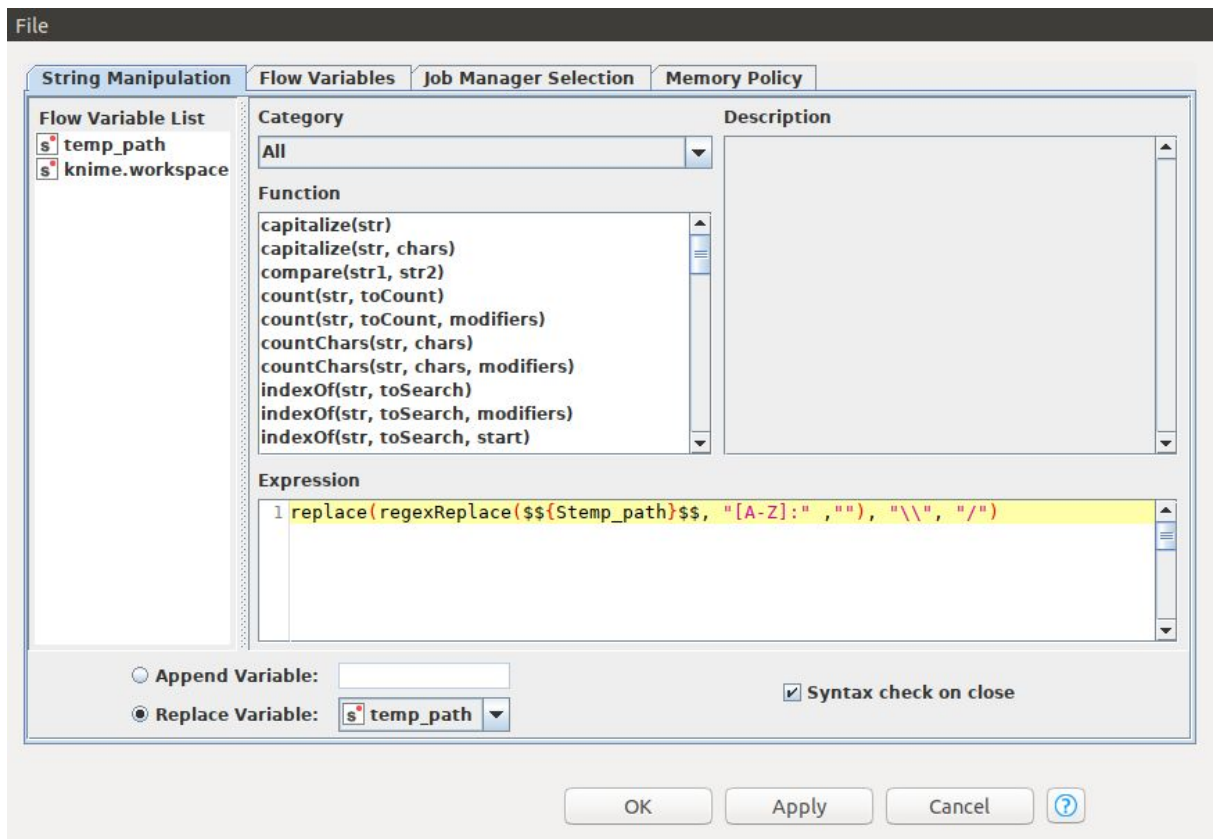
Melakukan sets up lingkungan lokal big data lalu load subset pada tabel yang berasal dari file ss13hme.csv dan ss13pme.csv

Pertama kali buat HDFS compatible path dengan membuat “Create Temp Dir”. Konfigurasinya:

The screenshot shows a 'File' dialog box with the following configuration:

- Job Manager Selection** and **Memory Policy** tabs are visible at the top.
- The **Configuration** tab is selected, showing the following fields:
 - Directory base name:** knime_tc_
 - Export path as (variable name):** temp_path
 - ☐ **Delete directory on reset**
 - ☐ **Create temp directory in workflow folder**
- Below the configuration fields is a section for **Additional path variables**, which contains a table with two columns: **Variable** and **File**. The table is currently empty.
- To the right of the table are **Add** and **Remove** buttons.
- At the bottom of the dialog are **OK**, **Apply**, **Cancel**, and a help icon button.

Lalu sambungkan dengan “String Manipulation”. Konfigurasinya:



Output dari hasil eksekusi tersebut nantinya akan disambungkan ke “DB Loader”

Kemudian buat “Create Local Big Data Environment” dengan konfigurasi:

File

Job Manager Selection Memory Policy

Local Big Data Environment Settings Flow Variables

Context name: knimeSparkContext

Number of threads: 2

Action to perform on disp...

- ☐ Destroy Spark context
- ☐ Delete Spark DataFrames
- ☒ Do nothing

☐ Use custom Spark settings


Custom Spark settings:

```
spark.jars: /path/to/some.jar  
spark.sql.shuffle.partitions: 100
```


SQL Support

- ☐ Spark SQL only
- ☐ HiveQL
- ☒ HiveQL and provide JDBC connection

☐ Use custom Hive data folder (Metastore DB & Warehouse)

Browse... 

☐ Hide warning about an existing local Spark context

OK Apply Cancel 

Output nya adalah menghasilkan Hive Connection dan HDFS Connection yang akan connect ke masing-masing “DB Loader”

Untuk membaca file CSV nya maka pertama kali load data menggunakan “Table Reader”. Konfigurasinya:

File

Options Flow Variables Job Manager Selection Memory Policy

Schema: default

Table: 05111740000141_ss13pme

Target folder: /tmp/knime_00_Setup_Hive_T36840/knime_t

Select a table

Browse...

The "targetFolder" parameter is controlled by a variable.

OK Apply Cancel ?

Kemudian membuat DB Table baru dari data sebelumnya dengan konfigurasi:

File

Table Creator Settings Flow Variables Job Manager Selection Memory Policy

Settings Columns Keys Additional Options Dynamic Type Settings Dynamic Keys Settings

Table Settings

Schema:

Table name: 05111740000141_ss13pme

☐ Create temporary table

☐ Fail if table exists

Dynamic Settings

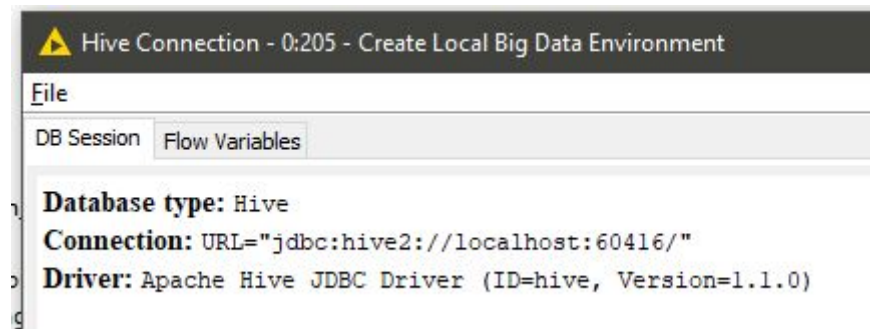
☒ Use dynamic settings

OK Apply Cancel ?

Setelah itu dengan menggunakan DB Loader maka data akan dapat ditampilkan pada database.

Untuk melihat hasilnya:

- a. Buka DBeaver
- b. Buat koneksi “New Database Connection”
- c. Pilih “Apache Hive”
- d. Setting port yang digunakan. Sesuaikan dengan port yang ada pada “Hive Connection” pada Create Local Big Data Environment yaitu port 60416



Generic JDBC Connection Settings

Database connection settings.



Main Driver properties SSH Proxy

JDBC URL: jdbc:hive2://localhost:45543

Host: localhost Port: 45543

Database/Schema:

User name:

Password: ☒ Save password locally

i You can use variables in connection parameters. Connection details (name, type, ...)

Driver name: Hadoop / Apache Hive Edit Driver Settings

Test Connection ... < Back Next > Cancel Finish

- e. Lakukan Test Connection untuk memastikan koneksi dapat dilakukan



Generic JDBC Connection Settings

Database connection settings.

Main

Driver properties

SSH

Proxy

JDBC URL:

jdbc:hive2://localhost:45543

Host:

localhost

Port:

45543

Database/Schema:

User name:

Password:

password locally

Connection Test

i

Connected (2260 ms)

Server:

Spark SQL 2.4.3

Driver:

org.apache.hive.jdbc.HiveDriver
2.7.3.2.6.5.0-292

OK

i

You can use variables in connection parameters.

Connection details (name, type, ...)

Driver name:

Hadoop / Apache Hive

Edit Driver Settings

Test Connection ...

< Back

Next >

Cancel

Finish

2. 01_Hive_Modelling

Setelah koneksi berhasil dibuat dan data telah tersimpan dalam database maka pada tahap ini akan dilakukan modelling data dengan menggunakan algoritma Decision Tree

Pertama lakukan “Create Local Big Data Environment”. Konfigurasinya:

File

Job Manager Selection **Memory Policy** **Local Big Data Environment Settings** **Flow Variables**

Context name:

Number of threads:

Action to perform on disp...

- ☐ Destroy Spark context
- ☐ Delete Spark DataFrames
- ☒ Do nothing

☐ Use custom Spark settings

Custom Spark settings:

```
spark.jars: /path/to/some.jar
spark.sql.shuffle.partitions: 100
```

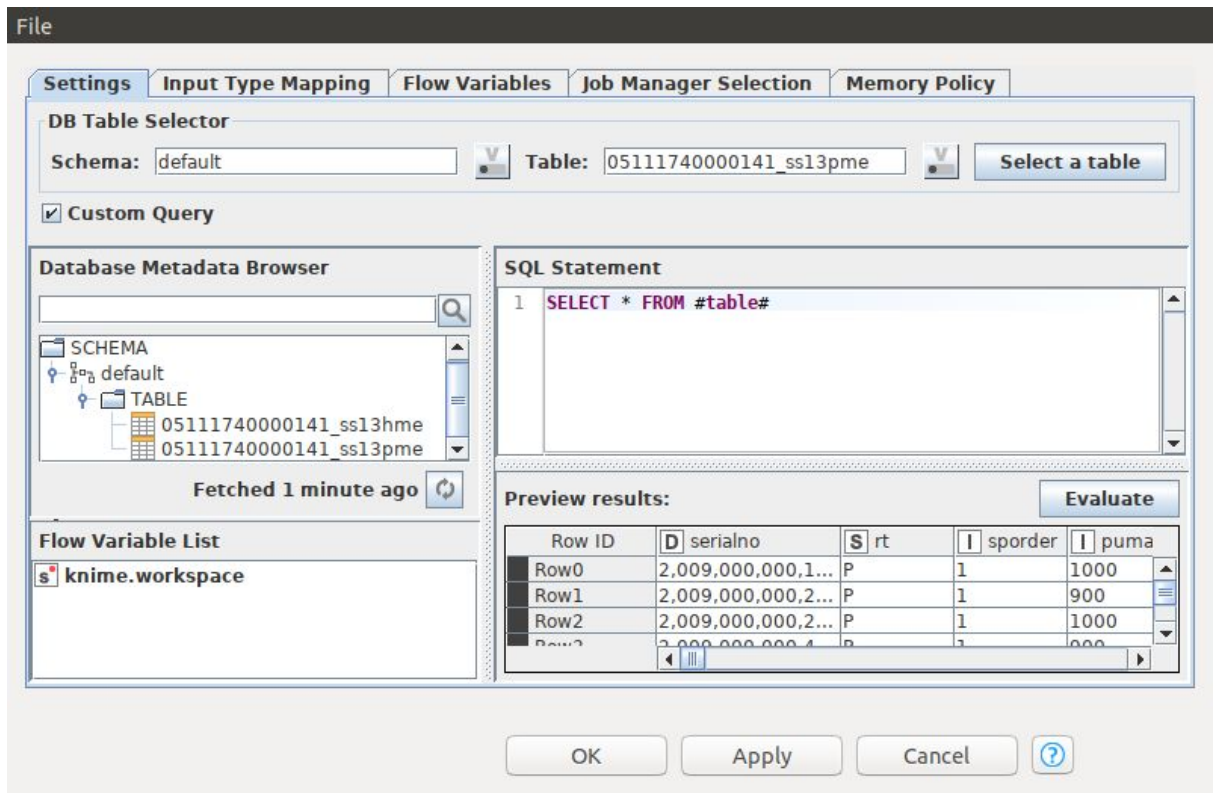
SQL Support

- ☐ Spark SQL only
- ☐ HiveQL
- ☒ HiveQL and provide JDBC connection

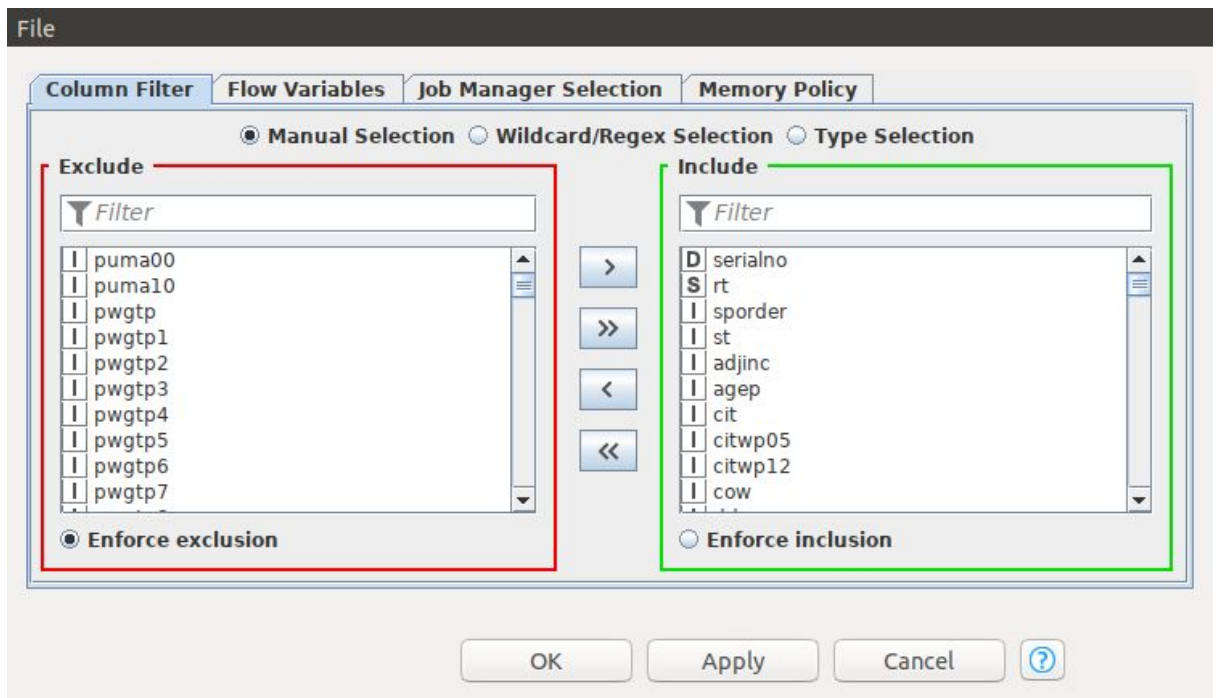
☐ Use custom Hive data folder (Metastore DB & Warehouse)

☐ Hide warning about an existing local Spark context

Setelah itu pilih tabel 05111740000141_ss13pme dengan menggunakan “DB Table Selector”.



Lakukan filter kolom untuk me remove kolom PUMA* & PWGTP*. Konfigurasinya:



Awalnya table 05111740000141_ss13pme memiliki 295 kolom. Tetapi setelah dilakukan filter menjadi 212 kolom.

Selanjutnya melakukan filter all rows from ss13pme where COW is NULL menggunakan “DB Row Filter”.

Hasil yang didapat adalah ada sebanyak 8662 data yang COW nya NULL

Lalu melakukan filter all rows from ss13pme where COW is NOT NULL menggunakan “DB Row Filter”.

Hasil yang didapat adalah ada sebanyak 21710 data yang COW nya NULL

Jika kolom COW nya NULL maka kolom COW tersebut di remove terlebih dahulu dengan menggunakan “DB Column Filter”.

Maka hasilnya tidak ada kolom row pada tabel tersebut.

Sedangkan data yang memiliki nilai COW yang semula berformat Integer diubah menjadi String dengan menggunakan “Number to String”.

Setelah itu melakukan modelling menggunakan algoritma desicion tree menggunakan “Desicion Tree Learner”.

Dari hasil training tersebut dapat dilakukan prediksi menentukan nilai COW yang NULL menggunakan “Desicion Tree Predictor”. Berikut konfigurasinya:

3. 02_Hive_WritingToDB

1. It reads data from local Hive;
2. it selects table ss13pme;
3. it isolates rows with missing cow value and rows with not missing cow values;
4. it imports the data into KNIME;
5. it creates a model to predict values for cow;
6. it uses cow predictions to override missing income values;
7. it rebuilds the datasets with the predicted cow values instead of the missing values
8. it writes the results back into Hive.

Pada subbab ini akan dilakukan override missing income values menggunakan cow predictions dan menyimpannya bersama data training cow learner ke dalam database Hive

Pertama kali buat HDFS compatible path dengan membuat “Create Temp Dir”.

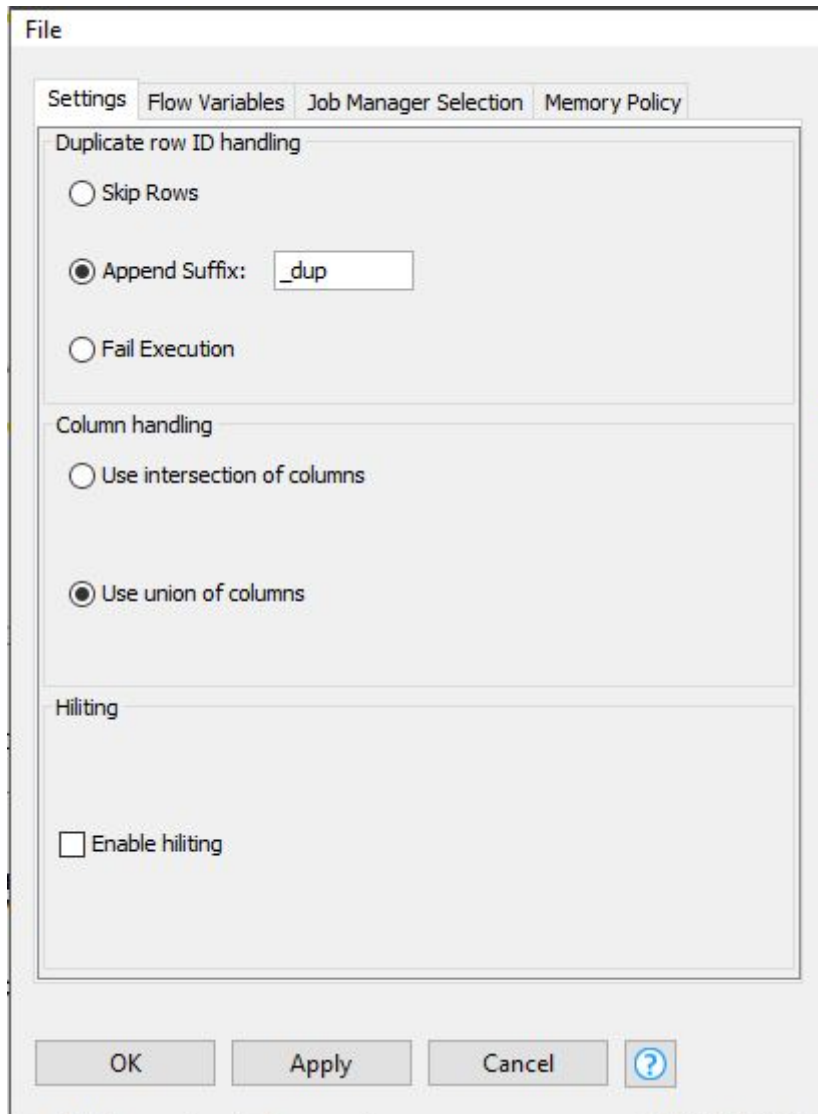
Lalu sambungkan dengan “String Manipulation”.

Output dari hasil eksekusi tersebut nantinya akan disambungkan ke “DB Loader”

Kemudian buat “Create Local Big Data Environment” dengan konfigurasi:

Output nya adalah menghasilkan Hive Connection dan HDFS Connection yang akan connect ke “DB Loader” dan “DB Table Creator”

Hasil data train dan data test akan digabungkan dengan menggunakan “Concatenate”.



Setelah itu akan dibuat tabel baru menggunakan “DB Table Creator” dengan input hasil concatenate (Data table) dan DB Connection pada Local Big Data Environment. Tabel baru tersebut saya beri nama “05111740000141_resultssh3pme”. Konfigurasinya:

File

Table Creator Settings | Flow Variables | Job Manager Selection | Memory Policy

Settings | Columns | Keys | Additional Options | Dynamic Type Settings | Dynamic Keys Settings

Table Settings

Schema:

Table name:

☐ Create temporary table

☐ Fail if table exists

Dynamic Settings

☒ Use dynamic settings

OK Apply Cancel ?

Setelah table database berhasil dibuat maka outputnya akan dijadikan input pada “DB Loader” sebagai DB Connection bersama dengan Create Local Big Data Environment (Connection information port), Concatenate (Data to load into database), dan String Manipulation (Variable Inport). Konfigurasinya:

File

Options Flow Variables Job Manager Selection Memory Policy







Schema: default 

Table: 05111740000141_resultssh3pme  **Select a table**

Target folder: /tmp/knime_02_Hive_Writing36848/knime_tc  **Browse...** 

 The "targetFolder" parameter is controlled by a variable.

OK Apply Cancel 

Link Github <https://github.com/afrchmdi/Tugas-Eksplorasi-KNIME-Big-Data>