



Data Warehousing using Talend

ivanmuhsiegfried@gmail.com

IVAN MUHAMMAD SIEGFRIED

©2021

 +62 838 201 73305

 @ivanmsiegfried

 [linkedin.com/in/ivanmsiegfried](https://www.linkedin.com/in/ivanmsiegfried)

DISCLAIMER

ALL ROW DATA DISPLAYED ARE SELF-CREATED AND NOT
COMPANY DATA





Problem 1

Using the travel_db database with the fact_flight_sales table, divide the following table according to the unique values in the payment_method column

```
F27135|2020-03-15 14:29:51|2020-03-15 14:42:43|SIN|DPS|RoundTrip|Transfer|8302020|542
F38033|2018-01-29 07:34:04|2018-01-29 07:35:32|KNO|CGK|OneWay|CreditCard|9044641|261
F46478|2020-03-21 09:14:37|2020-03-21 09:57:22|KNO|SIN|OneWay|CreditCard|2179595|856
F41237|2016-02-25 17:13:07|2016-02-25 17:25:48|BKK|KNO|RoundTrip|VirtualAccount|7867875|611
F40292|2019-06-27 17:26:42|2019-06-27 18:00:58|KNO|DPS|OneWay|Transfer|6200009|648
F88149|2020-04-02 23:39:11|2020-04-02 23:58:00|KNO|KNO|OneWay|CreditCard|7891221|200
F08929|2018-11-07 21:57:03|2018-11-07 22:30:15|CGK|BKK|OneWay|CreditCard|4657994|134
F35581|2017-09-29 05:46:28|2017-09-29 06:45:15|KNO|SUB|OneWay|MiniMarket|1247933|823
[statistics] disconnected
```



> Documents > local_hdfs > out

Name	Date modified	Type	Size
 payment_creditcard.xls	3/16/2021 11:32 AM	Microsoft Excel 97...	661 KB
 payment_minimarket.xls	3/16/2021 11:33 AM	Microsoft Excel 97...	675 KB
 payment_transfer.xls	3/16/2021 11:32 AM	Microsoft Excel 97...	666 KB
 payment_VA.xls	3/16/2021 11:32 AM	Microsoft Excel 97...	664 KB

Problem 1 Solution

Database Scheme

Schema

Filter for the Table.

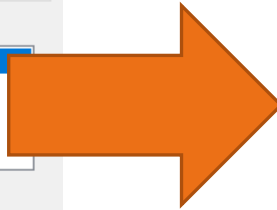
Select Filter Conditions
 Use the Name Filter Use the Sql Filter

Select Types
 TABLE VIEW SYNONYM

Set the Name Filter:
%

Set the Sql Filter:
SELECT TNAME FROM TAB WHERE TNAME LIKE 'BAL%'

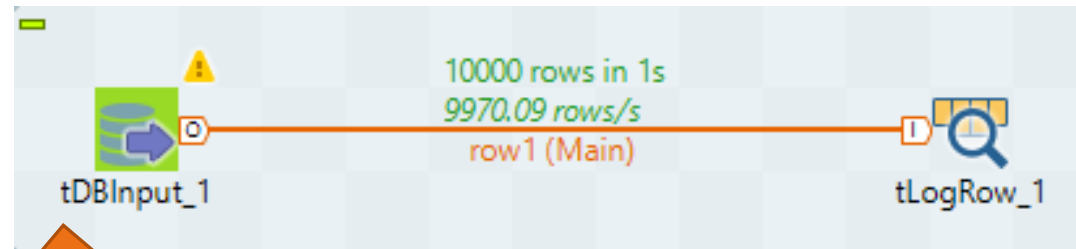
< Back Next > Finish Cancel



- DbInput 0.1
 - Queries
 - Synonym schemas
 - Table schemas
 - dim_flight_airport
 - Columns(4)
 - airport_id
 - airport_name
 - airport_city_name
 - airport_country_name
 - fact_flight_sales
 - Columns(9)
 - booking_id
 - booking_created_time
 - booking_paid_time
 - source_airport_id
 - destination_airport_id
 - trip_type
 - payment_method
 - booking_price_amount
 - user_id

Problem 1 Solution

First Insight of The Data



Job(filterpegepi 0.1) Contexts(filterpegepi) Component x Run (Job filterpegepi)

tDBInput_1(MySQL)

Basic settings

Database: MySQL [Apply]

Property Type: Built-In [Save]

Dynamic settings

DB Version: Mysql 8

View

Documentation

Use an existing connection

Host: "localhost" * Port: "3306" * Database: "travel_db"

Username: "root" * Password: "*****"

Schema: Built-In [Edit schema ...]

Table Name: ""

Query Type: Built-In [Guess Query] [Guess schema]

Query: "select * from fact_flight_sales"

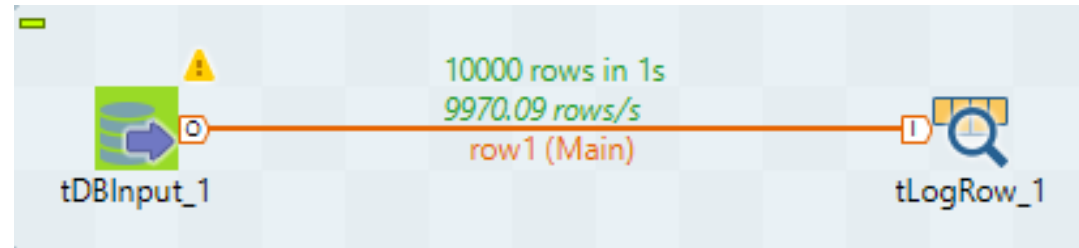
Schema of tDBInput_1

Column	Db Column	Key	Type	DB Type	<input checked="" type="checkbox"/>	N..	Date Pattern...	Length	Precis...	D
booking_id	booking_id	<input type="checkbox"/>	String	TEXT	<input checked="" type="checkbox"/>					
booking_created_...	booking_created_time	<input type="checkbox"/>	String	DATETIME	<input checked="" type="checkbox"/>					
booking_paid_time	booking_paid_time	<input type="checkbox"/>	String	DATETIME	<input checked="" type="checkbox"/>					
source_airport_id	source_airport_id	<input type="checkbox"/>	String	TEXT	<input checked="" type="checkbox"/>					
destination_airpor...	destination_airport_id	<input type="checkbox"/>	String	TEXT	<input checked="" type="checkbox"/>					
trip_type	trip_type	<input type="checkbox"/>	String	TEXT	<input checked="" type="checkbox"/>					
payment_method	payment_method	<input type="checkbox"/>	String	TEXT	<input checked="" type="checkbox"/>					
booking_price_a...	booking_price_amount	<input type="checkbox"/>	String	BIGINT	<input checked="" type="checkbox"/>			20		
user_id	user_id	<input type="checkbox"/>	String	BIGINT	<input checked="" type="checkbox"/>			20		

OK Cancel

Problem 1 Solution

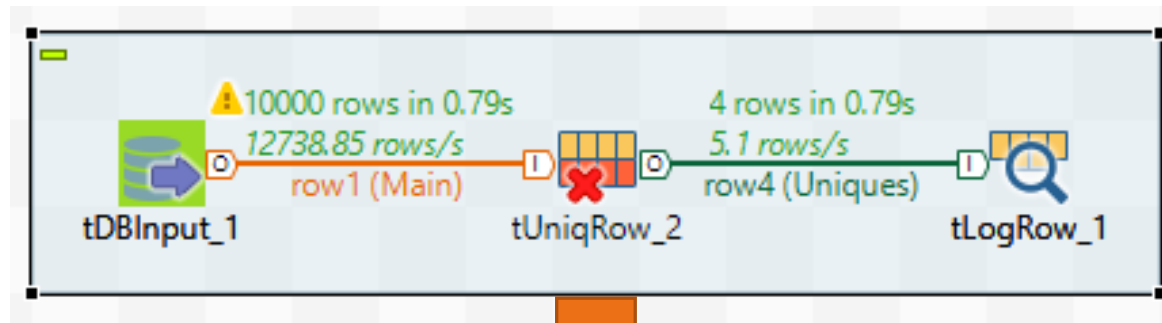
First Insight of The Data



```
F27135|2020-03-15 14:29:51|2020-03-15 14:42:43|SIN|DPS|RoundTrip|Transfer|8302020|542
F38033|2018-01-29 07:34:04|2018-01-29 07:35:32|KNO|CGK|OneWay|CreditCard|9044641|261
F46478|2020-03-21 09:14:37|2020-03-21 09:57:22|KNO|SIN|OneWay|CreditCard|2179595|856
F41237|2016-02-25 17:13:07|2016-02-25 17:25:48|BKK|KNO|RoundTrip|VirtualAccount|7867875|611
F40292|2019-06-27 17:26:42|2019-06-27 18:00:58|KNO|DPS|OneWay|Transfer|6200009|648
F88149|2020-04-02 23:39:11|2020-04-02 23:58:00|KNO|KNO|OneWay|CreditCard|7891221|200
F08929|2018-11-07 21:57:03|2018-11-07 22:30:15|CGK|BKK|OneWay|CreditCard|4657994|134
F35581|2017-09-29 05:46:28|2017-09-29 06:45:15|KNO|SUB|OneWay|MiniMarket|1247933|823
[statistics] disconnected
```

Problem 1 Solution

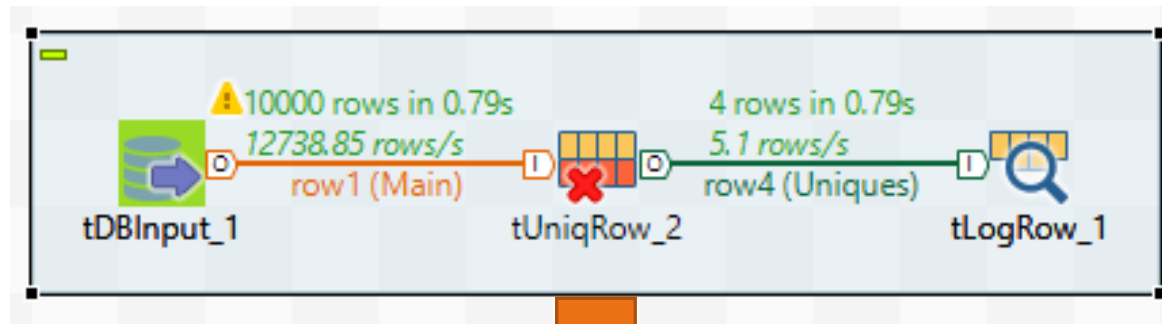
Find Unique Data from Payment Column



Column	<input type="checkbox"/> Key attribute	<input type="checkbox"/> Case Sensitive
booking_id	<input type="checkbox"/>	<input type="checkbox"/>
booking_created_time	<input type="checkbox"/>	<input type="checkbox"/>
booking_paid_time	<input type="checkbox"/>	<input type="checkbox"/>
source_airport_id	<input type="checkbox"/>	<input type="checkbox"/>
destination_airport_id	<input type="checkbox"/>	<input type="checkbox"/>
trip_type	<input type="checkbox"/>	<input type="checkbox"/>
payment_method	<input checked="" type="checkbox"/>	<input type="checkbox"/>
booking_price_amount	<input type="checkbox"/>	<input type="checkbox"/>
user_id	<input type="checkbox"/>	<input type="checkbox"/>

Problem 1 Solution

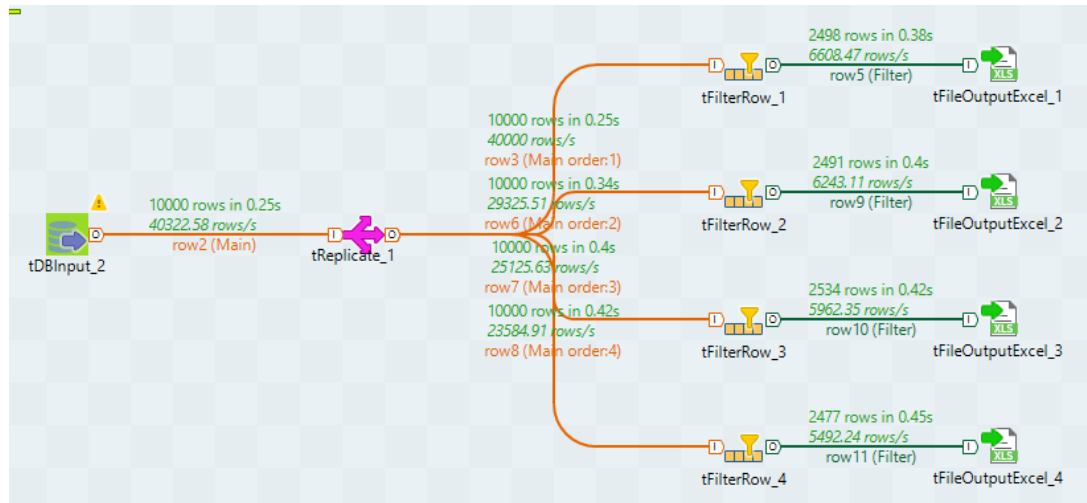
Find Unique Data from Payment Column



booking_id	booking_created_time	booking_paid_time	source_airport_id	destination_airport_id	trip_type	payment_method	booking_price_amount	user_id
F17442	2016-01-10 23:11:16	2016-01-10 23:59:06	BKK	BKK	RoundTrip	Transfer	390623	419
F60031	2019-06-24 22:48:13	2019-06-24 22:59:04	SUB	SIN	OneWay	VirtualAccount	9149768	371
F68215	2016-03-04 20:20:33	2016-03-04 20:57:14	BKK	CGK	RoundTrip	MiniMarket	387071	192
F38898	2019-08-25 06:32:17	2019-08-25 06:53:40	SIN	CGK	OneWay	CreditCard	369133	772

Problem 1 Solution

Data Splitting



tFilterRow_1

Schema: Built-In

Logical operator used to combine conditions: And

InputColumn	Function	Operator	Value
payment_method	Match	Equals	"Transfer"

Use advanced mode

tFileOutputExcel_1

Property Type: Built-In

Advanced settings: Write excel2007 file format (xlsx)

Dynamic settings: Use Output Stream

View: File Name: "C:/Users/HP/Documents/local_hdfs/out/payment_transfer.xls"

Documentation: Sheet name: "Sheet1"

Include header

Append existing file

Is absolute Y pos.

Font: Default

Define all columns auto size



Documents > local_hdfs > out

Name	Date modified	Type	Size
payment_creditcard.xls	3/16/2021 11:32 AM	Microsoft Excel 97...	661 KB
payment_minimarket.xls	3/16/2021 11:33 AM	Microsoft Excel 97...	675 KB
payment_transfer.xls	3/16/2021 11:32 AM	Microsoft Excel 97...	666 KB
payment_VA.xls	3/16/2021 11:32 AM	Microsoft Excel 97...	664 KB

booking_id	booking_created_time	booking_paid_time	source_airport_id	destination_airport_id	trip_type	payment_method	booking_price_amount	user_id
F17442	2016-01-10 23:11:16	2016-01-10 23:59:06	BKK	BKK	RoundTrip	Transfer	3790623	419
F55283	2015-03-17 04:17:38	2015-03-17 04:39:48	SUB	KNO	RoundTrip	Transfer	9244661	668
F75818	2019-11-11 08:15:18	2019-11-11 09:13:16	SUB	BKK	RoundTrip	Transfer	2485374	613
F77359	2018-12-26 18:17:58	2018-12-26 18:28:04	BKK	DPS	OneWay	Transfer	6733326	94
F68059	2020-03-26 06:53:17	2020-03-26 07:13:03	SIN	SIN	OneWay	Transfer	8102592	135
F02653	2018-04-15 20:19:58	2018-04-15 20:25:35	SIN	SIN	RoundTrip	Transfer	9585087	864
F30036	2018-11-15 02:17:55	2018-11-15 02:26:26	CGK	DPS	RoundTrip	Transfer	4470410	485
F26699	2019-11-08 18:26:33	2019-11-08 19:26:09	DPS	BKK	OneWay	Transfer	5336357	90
F49173	2019-09-12 14:50:53	2019-09-12 15:12:03	SUB	BKK	RoundTrip	Transfer	4282076	396
F34038	2015-11-03 13:37:44	2015-11-03 14:17:53	SIN	DPS	RoundTrip	Transfer	5173795	403
F22923	2015-12-02 04:23:19	2015-12-02 04:57:07	DPS	CGK	RoundTrip	Transfer	7413533	62
F05206	2018-01-15 12:38:06	2018-01-15 13:34:54	KNO	DPS	OneWay	Transfer	8139027	182
F71935	2016-01-21 07:06:05	2016-01-21 07:39:24	SIN	CGK	RoundTrip	Transfer	8108992	309
F78368	2017-06-25 06:19:41	2017-06-25 06:29:30	BKK	SUB	RoundTrip	Transfer	2765488	420
F34475	2020-02-12 08:33:06	2020-02-12 08:40:47	SIN	BKK	OneWay	Transfer	5946659	886
F19713	2015-07-18 10:04:45	2015-07-18 11:00:10	DPS	SIN	RoundTrip	Transfer	5752736	302

Problem 2

Using `travel_db` *database*, write a query and find a solution using Talend to find out popular international routes during Jan 2017 - Dec 2018. Sort from the most popular.

Problem 2 Solution

Using SQL Query

```
SELECT source_airport_id, destination_airport_id, COUNT(*)
FROM fact_flight_sales
WHERE (booking_paid_time between '2017-01-01 00:00:00' and '2018-12-
31 23:59:59' AND (source_airport_id<>destination_airport_id))
GROUP BY source_airport_id, destination_airport_id
ORDER BY count(*) DESC;
```

Problem 2 Solution

Database Scheme

Schema

Filter for the Table.

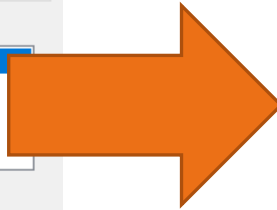
Select Filter Conditions
 Use the Name Filter Use the Sql Filter

Select Types
 TABLE VIEW SYNONYM

Set the Name Filter:
%

Set the Sql Filter:
SELECT TNAME FROM TAB WHERE TNAME LIKE 'BAL%'

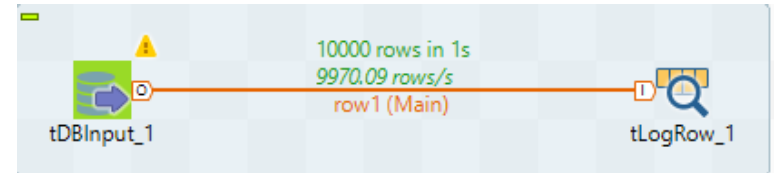
< Back Next > Finish Cancel



- DbInput 0.1
 - Queries
 - Synonym schemas
 - Table schemas
 - dim_flight_airport
 - Columns(4)
 - airport_id
 - airport_name
 - airport_city_name
 - airport_country_name
 - fact_flight_sales
 - Columns(9)
 - booking_id
 - booking_created_time
 - booking_paid_time
 - source_airport_id
 - destination_airport_id
 - trip_type
 - payment_method
 - booking_price_amount
 - user_id

Problem 2 Solution

First Insight of The Data



- DbInput 0.1
 - Queries
 - Synonym schemas
 - Table schemas
 - dim_flight_airport
 - Columns(4)
 - airport_id
 - airport_name
 - airport_city_name
 - airport_country_name
 - fact_flight_sales
 - Columns(9)
 - booking_id
 - booking_created_time
 - booking_paid_time
 - source_airport_id
 - destination_airport_id
 - trip_type
 - payment_method
 - booking_price_amount
 - user_id

Table: fact_flight_sales

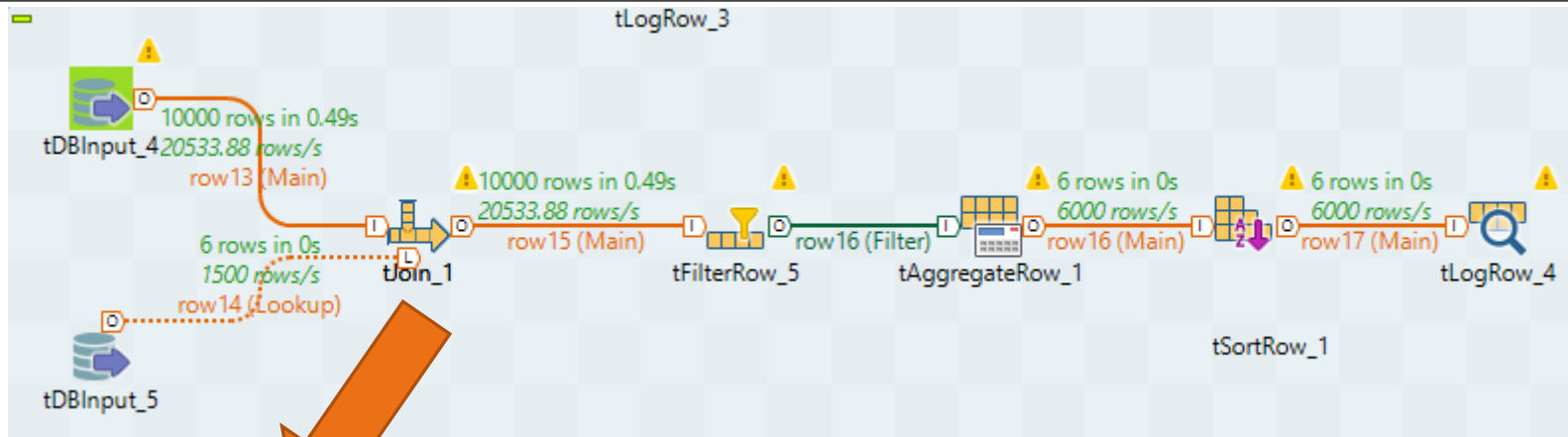
F27135	2020-03-15 14:29:51	2020-03-15 14:42:43	SIN	DPS	RoundTrip	Transfer	8302020	542
F38033	2018-01-29 07:34:04	2018-01-29 07:35:32	KNO	CGK	OneWay	CreditCard	9044641	261
F46478	2020-03-21 09:14:37	2020-03-21 09:57:22	KNO	SIN	OneWay	CreditCard	2179595	856
F41237	2016-02-25 17:13:07	2016-02-25 17:25:48	BKK	KNO	RoundTrip	VirtualAccount	7867875	611
F40292	2019-06-27 17:26:42	2019-06-27 18:00:58	KNO	DPS	OneWay	Transfer	6200009	648
F88149	2020-04-02 23:39:11	2020-04-02 23:58:00	KNO	KNO	OneWay	CreditCard	7891221	200
F08929	2018-11-07 21:57:03	2018-11-07 22:30:00	CGK	BKK	OneWay	CreditCard	4657994	134
F35581	2017-09-29 05:46:28	2017-09-29 05:50:55	KNO	SUB	OneWay	MiniMarket	1247933	823
[statistics] disconnected								

BKK	Suvarnabhumi Airport	Bangkok	Thailand
CGK	Soekarno-Hatta Airport	Jakarta	Indonesia
DPS	Ngurah Rai Airport	Denpasar-Bali	Indonesia
KNO	Kuala Namu Airport	Medan	Indonesia
SIN	Changi Airport	Singapore	Singapore
SUB	Juanda Airport	Surabaya	Indonesia

Table: dim_flight_airport

Problem 2 Solution

Join Process



Schema: Built-In Edit schema

Include lookup columns in output

Column mapping

Output column	Lookup column
source_airport_id	row14.airport_name
destination_airport_id	row14.airport_name

Key definition

Input key attribute	Lookup key attribute
source_airport_id	row14.airport_id
destination_airport_id	row14.airport_id

Schema: Built-In Edit schema Sync columns

Logical operator used to combine conditions: And

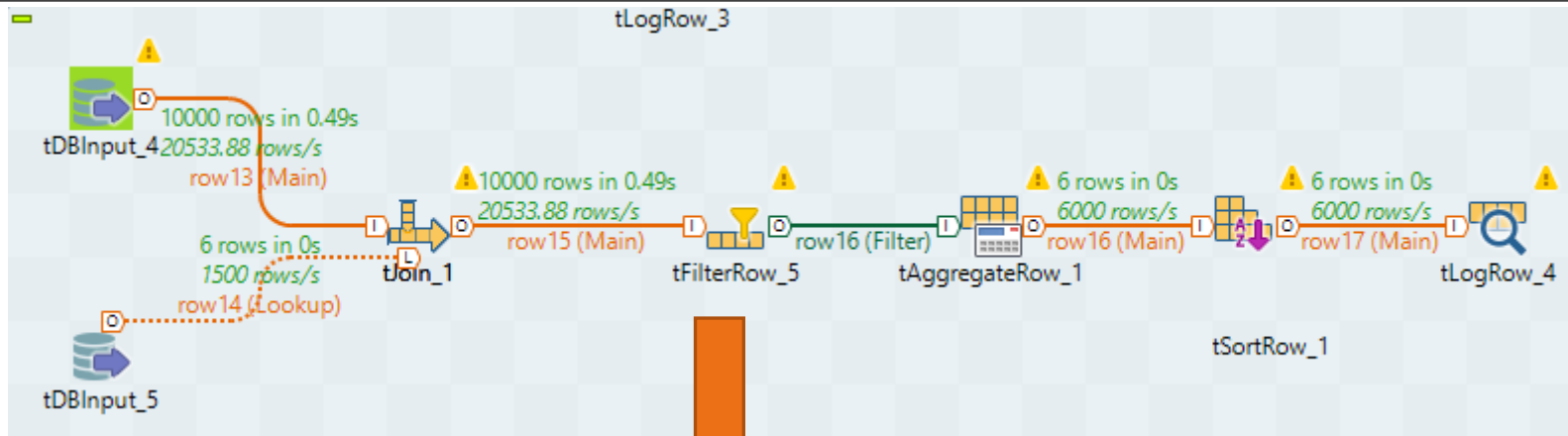
Conditions

InputColumn	Function	Operator	Value
booking_paid_time	Empty	Lower or equal to	TalendDate.parseDate("MM/dd/yyyy", "12/31/2018")
booking_paid_time	Empty	Greater or equal to	TalendDate.parseDate("MM/dd/yyyy", "01/01/2017")
booking_created_time	Empty	Lower or equal to	TalendDate.parseDate("MM/dd/yyyy", "12/31/2018")
booking_created_time	Empty	Greater or equal to	TalendDate.parseDate("MM/dd/yyyy", "01/01/2017")

Use advanced mode

Problem 2 Solution

Filtering Process



tFilterRow_5

Schema: Built-In | Edit schema | Sync columns

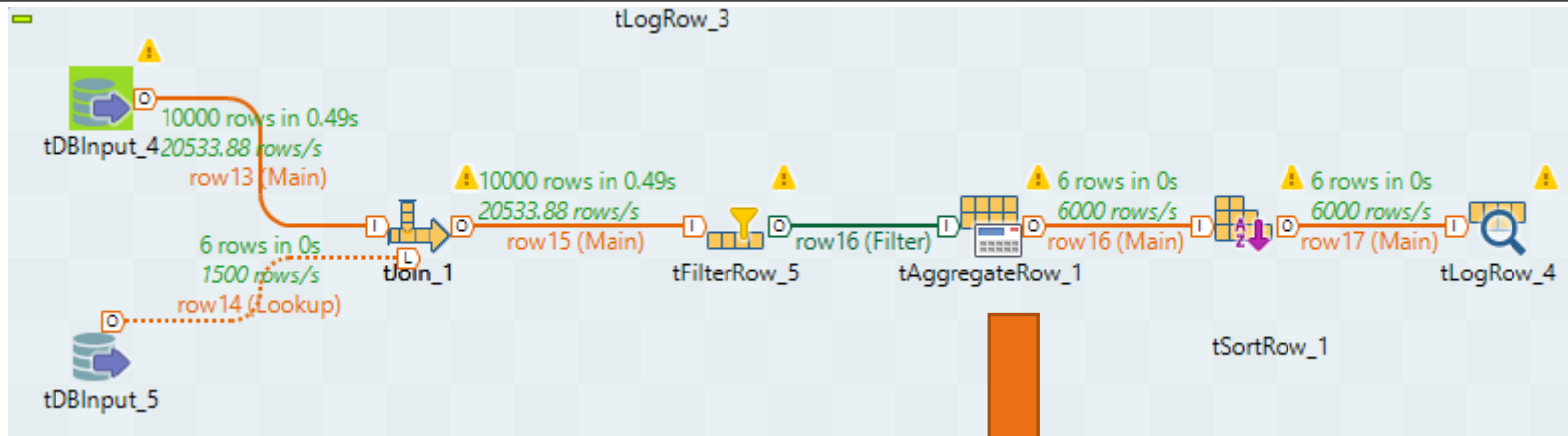
Logical operator used to combine conditions: And

InputColumn	Function	Operator	Value
booking_paid_time	Empty	Lower or equal to	TalendDate.parseDate("MM/dd/yyyy", "12/31/2018")
booking_paid_time	Empty	Greater or equal to	TalendDate.parseDate("MM/dd/yyyy", "01/01/2017")
booking_created_time	Empty	Lower or equal to	TalendDate.parseDate("MM/dd/yyyy", "12/31/2018")
booking_created_time	Empty	Greater or equal to	TalendDate.parseDate("MM/dd/yyyy", "01/01/2017")

Use advanced mode

Problem 2 Solution

Aggregating Process



tAggregateRow_1

Schema: Built-In | Edit schema | Sync columns

Group by

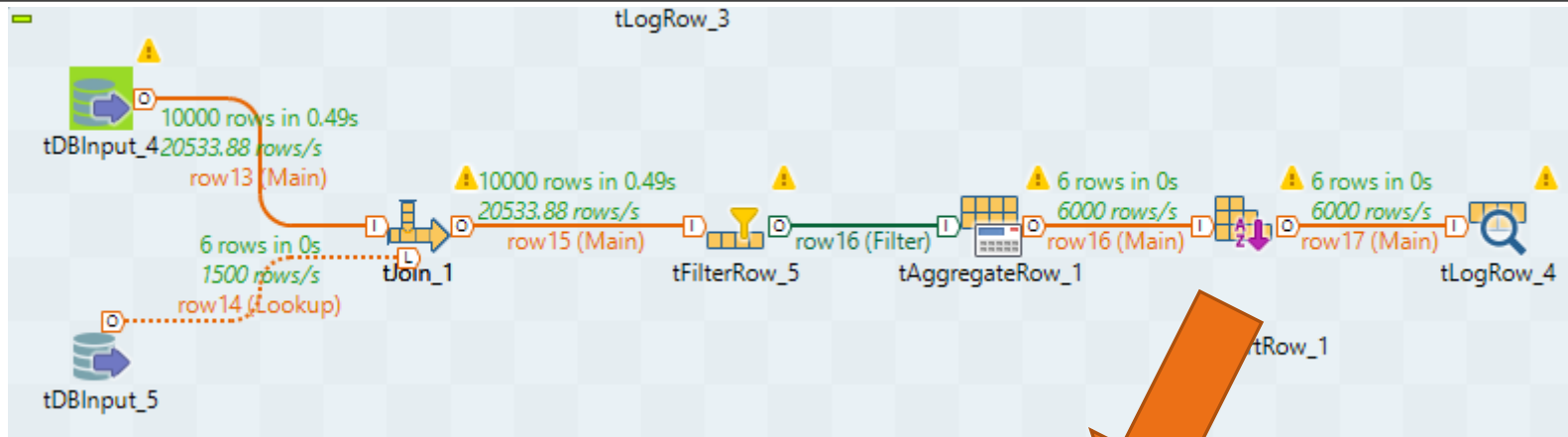
Output column	Input column position
source_airport_id	source_airport_id

Operations

Output column	Function	Input column position	Ignore null values
count_src_dest	count	destination_airport_id	<input type="checkbox"/>

Problem 2 Solution

Joining Process



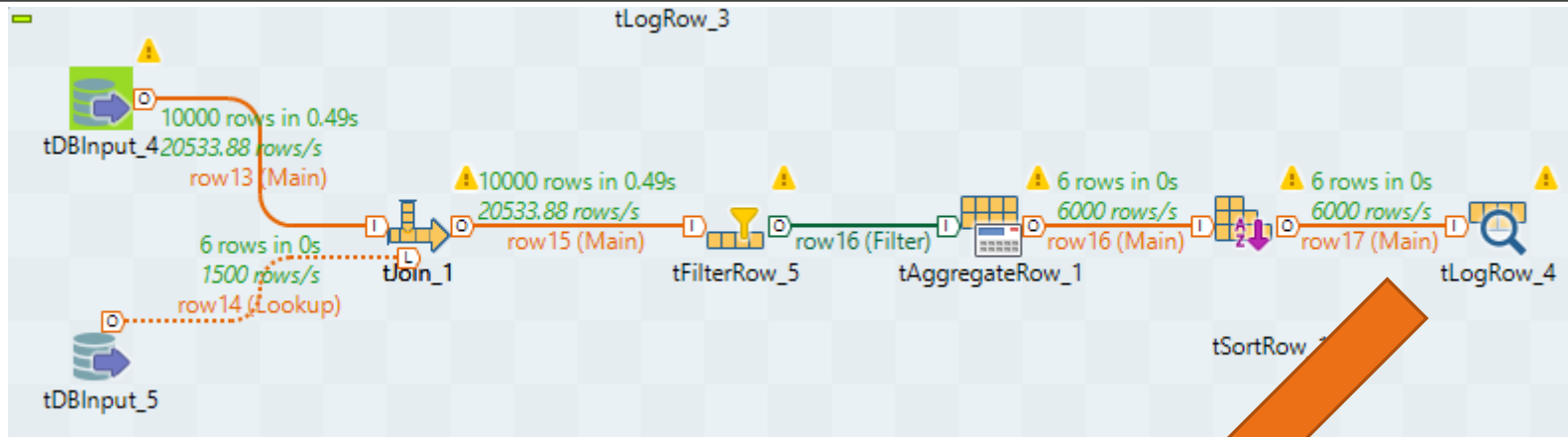
Schema Built-In Edit schema Sync columns

Schema column	sort num or alpha?	Order asc or desc?
count_src_dest	num	desc

+ × ↑ ↓ 📄 📋 +

Problem 2 Solution

Final Results



tLogRow_4		
source_airport_id	destination_airport_id	count_src_dest
Changi Airport	Juanda Airport	585
Ngurah Rai Airport	Juanda Airport	582
Soekarno-Hatta Airport	Juanda Airport	578
Kuala Namu Airport	Juanda Airport	554
Suvarnabhumi Airport	Juanda Airport	541
Juanda Airport	Juanda Airport	500



Data Warehousing using Talend

ivanmuhsiegfried@gmail.com

IVAN MUHAMMAD SIEGFRIED

©2021

 +62 838 201 73305

 @ivanmsiegfried

 [linkedin.com/in/ivanmsiegfried](https://www.linkedin.com/in/ivanmsiegfried)