traveloka

Data Modelling and Processing on a Travel Super App

Rendy B. Junior - Joshua Hendinata, Traveloka

Data Council Singapore, 17-18 July 2019

traveloka

#EmpoweringDiscovery

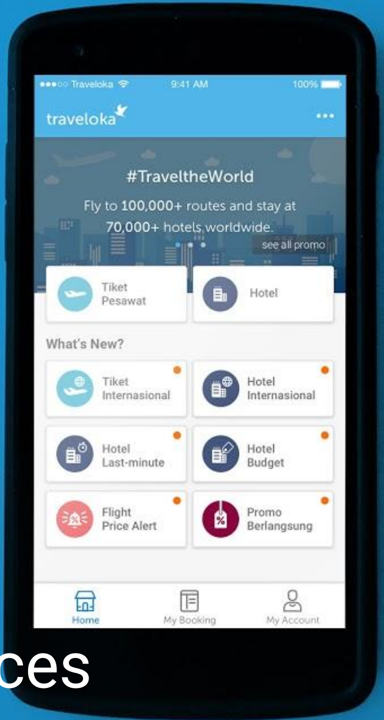
traveloka *first,*
then **THE WORLD**

A Travel Super App Company

Traveloka is an app that provides wide-range of travel-related product and services, #EmpoweringDiscovery, such as:

- Flight
- Hotel
- Theme parks
- International roaming package
- Activities
- Dine-in

traveloka



Our technology core has enabled us to scale Traveloka into

6 countries

across ASEAN rapidly in

less than 2 years.

8 offices

Incl. Singapore

1,000+

Global employees

400+

Engineers

Traveloka Data Challenges

Data Model Silos and Dirty Data Everywhere!

Data Model Silos was our Biggest Problem

- We democratized data wrangling
- Each business unit can create their own data model
- So **different** from one to another
- **Hard to analyze across business**

Who Suffers the Most

The one who suffers most are **cross business unit function**, such as:

- Marketing
- User Engagement
- Finance

Business case example: how can I make a **CLV (customer lifetime value) company-wide**, if **sales data** from each business unit is coming in **different schema**?

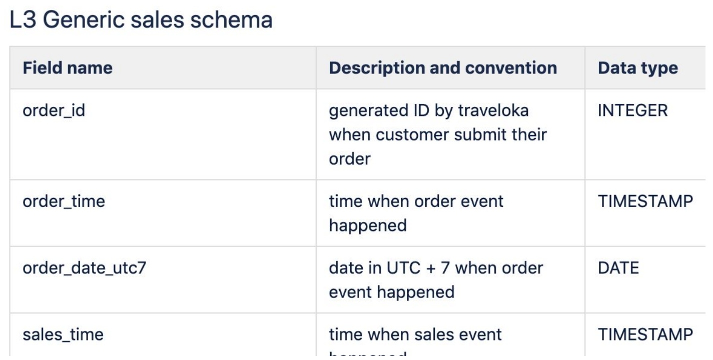
How do we solve Data Silos?

First rule, address the design, not the technology

So we address the design problem by designing **generic schema** across business units

Example: sales schema company-wide

So then..



L3 Generic sales schema

Field name	Description and convention	Data type
order_id	generated ID by traveloka when customer submit their order	INTEGER
order_time	time when order event happened	TIMESTAMP
order_date_utc7	date in UTC + 7 when order event happened	DATE
sales_time	time when sales event	TIMESTAMP

But how can we ensure everyone follows company-wide design?

But how can we ensure everyone follows company-wide design?

Framework come to the rescue

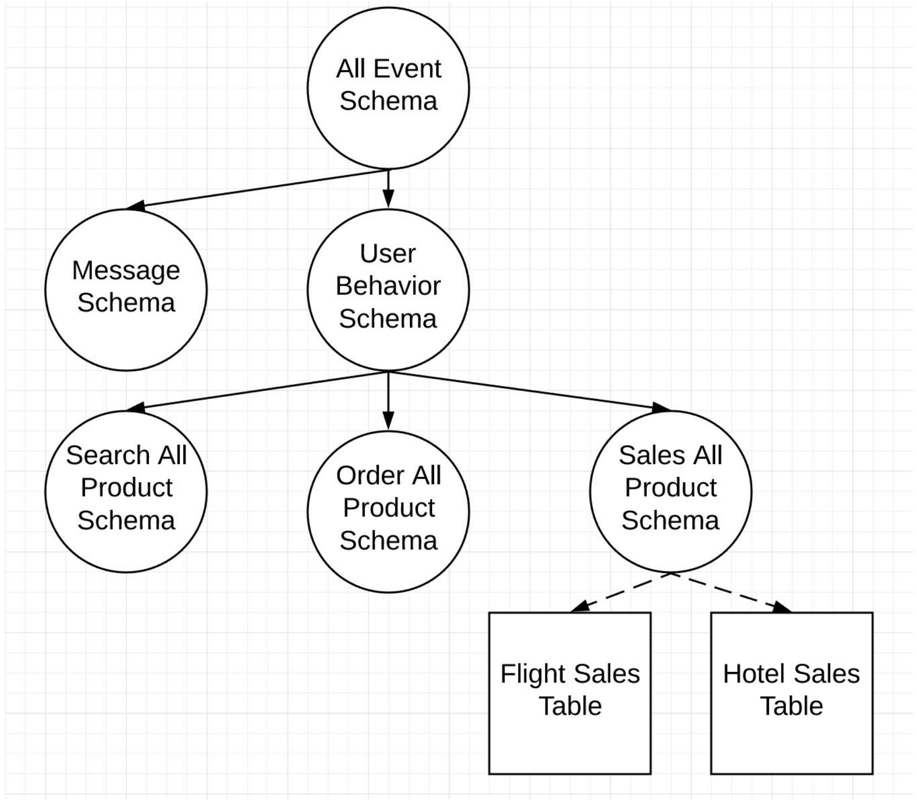
Super App Schema Framework with **Inheritance**

Schema Inheritance concept: child schema inherit properties of its parent.

Central team define the **parent schema**, all business units **must** follow

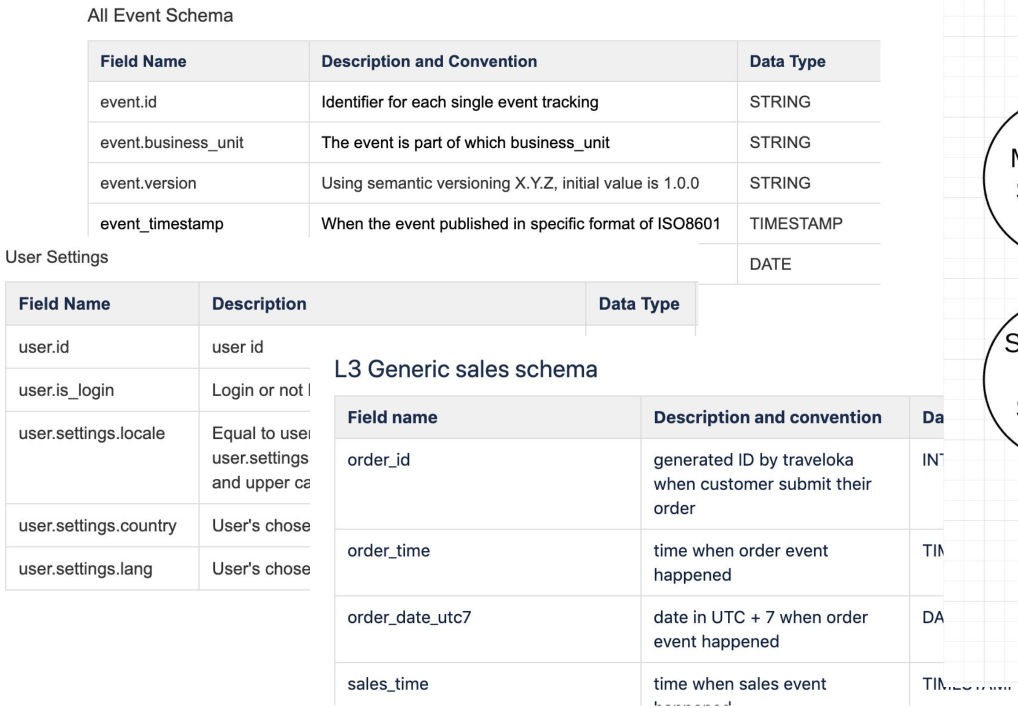
Schema Inheritance Concept

Example of inheritance tree.



Schema Inheritance Concept

Example of inheritance tree.



All Event Schema

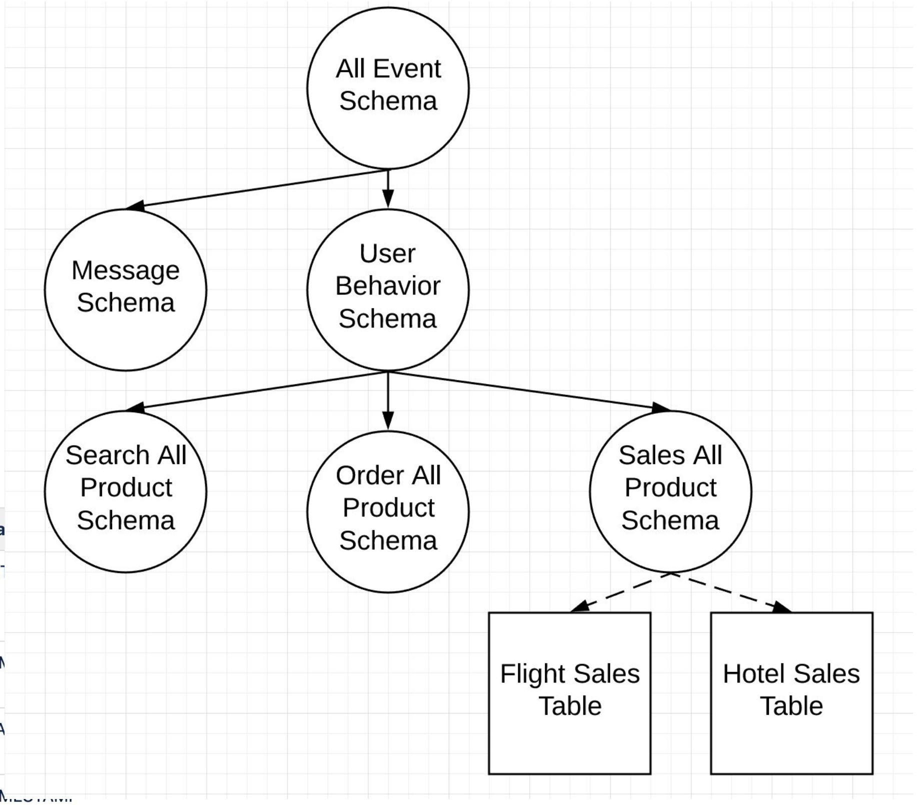
Field Name	Description and Convention	Data Type
event.id	Identifier for each single event tracking	STRING
event.business_unit	The event is part of which business_unit	STRING
event.version	Using semantic versioning X.Y.Z, initial value is 1.0.0	STRING
event_timestamp	When the event published in specific format of ISO8601	TIMESTAMP
		DATE

User Settings

Field Name	Description	Data Type
user.id	user id	
user.is_login	Login or not l	
user.settings.locale	Equal to use user.settings and upper ca	
user.settings.country	User's chose	
user.settings.lang	User's chose	

L3 Generic sales schema

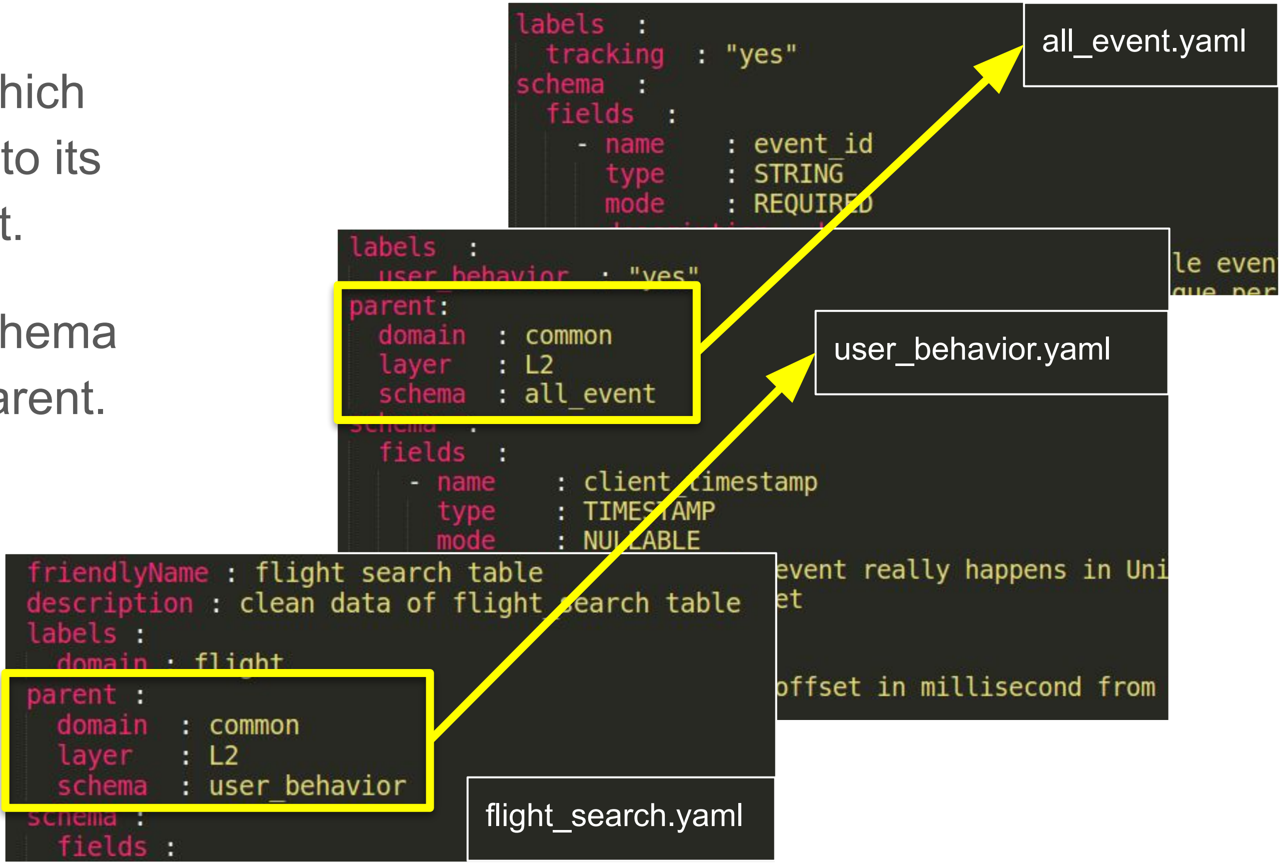
Field name	Description and convention	Da
order_id	generated ID by traveloka when customer submit their order	IN
order_time	time when order event happened	TIN
order_date_utc7	date in UTC + 7 when order event happened	DA
sales_time	time when sales event happened	TIMESTAMP



We put inheritance into schema

See `parent` field which ties child schema into its more generic parent.

It will resolve the schema recursively to the parent.



all_event.yaml

```
labels  :
  tracking  : "yes"
schema  :
  fields  :
    - name     : event_id
      type     : STRING
      mode     : REQUIRED
                                                    le even
                                                    que per
```

user_behavior.yaml

```
labels  :
  user_behavior  : "yes"
parent:
  domain  : common
  layer   : L2
  schema  : all_event
schema  :
  fields  :
    - name     : client_timestamp
      type     : TIMESTAMP
      mode     : NULLABLE
                                        event really happens in Uni
                                        et

                                        offset in millisecond from
```

flight_search.yaml

```
friendlyName : flight search table
description : clean data of flight search table
labels :
  domain : flight
parent :
  domain   : common
  layer    : L2
  schema   : user_behavior
schema :
  fields :
```

Sample Usage

It is very easy to analyse data across all business unit!

```
SELECT user.id, SUM(profit)

FROM fact_sales_*

GROUP BY 1
```

fact_order_* is equivalent to fact_order_flight UNION ALL fact_order_hotel and so on

traveloka

~~Data Model Silos~~ (solved!)
and
Dirty Data
Everywhere!

Pattern of Dirty Data

Business Rule Violation e.g.

- Min/max string length
- Min/max value
- String pattern
- Possible values (enumeration)

Repeated Process Everywhere

Those teams end up creating a process to make the data from each business unit **uniform** so that they can use it.

Repeated data processing → **waste of time, waste of money**

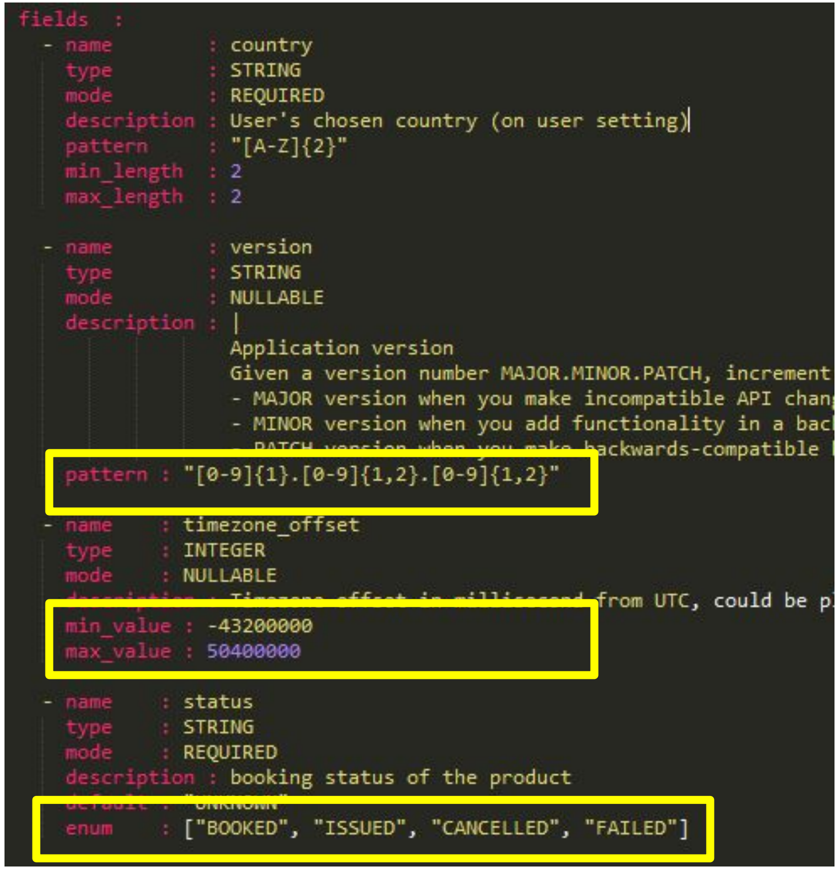
Now.. how to fix this situation?

So we add simple rules to the schema

Imagine you don’t have to implement code to do those

Write once use everywhere!

Executable spec concept enable collaboration



```
fields  :
  - name        : country
    type        : STRING
    mode        : REQUIRED
    description : User's chosen country (on user setting)
    pattern     : "[A-Z]{2}"
    min_length  : 2
    max_length  : 2

  - name        : version
    type        : STRING
    mode        : NULLABLE
    description : |
                  Application version
                  Given a version number MAJOR.MINOR.PATCH, increment
                  - MAJOR version when you make incompatible API chan
                  - MINOR version when you add functionality in a bac
                  - PATCH version when you make backwards-compatible
    pattern : "[0-9]{1}.[0-9]{1,2}.[0-9]{1,2}"

  - name    : timezone_offset
    type    : INTEGER
    mode    : NULLABLE
    description : Timezone offset in millisecond from UTC, could be p
    min_value : -43200000
    max_value : 50400000

  - name    : status
    type    : STRING
    mode    : REQUIRED
    description : booking status of the product
    default : "UNKNOWN"
    enum    : ["BOOKED", "ISSUED", "CANCELLED", "FAILED"]
```

~~Data Model Silos~~ (solved!)
and
~~Dirty Data~~ (solved!)
Everywhere!

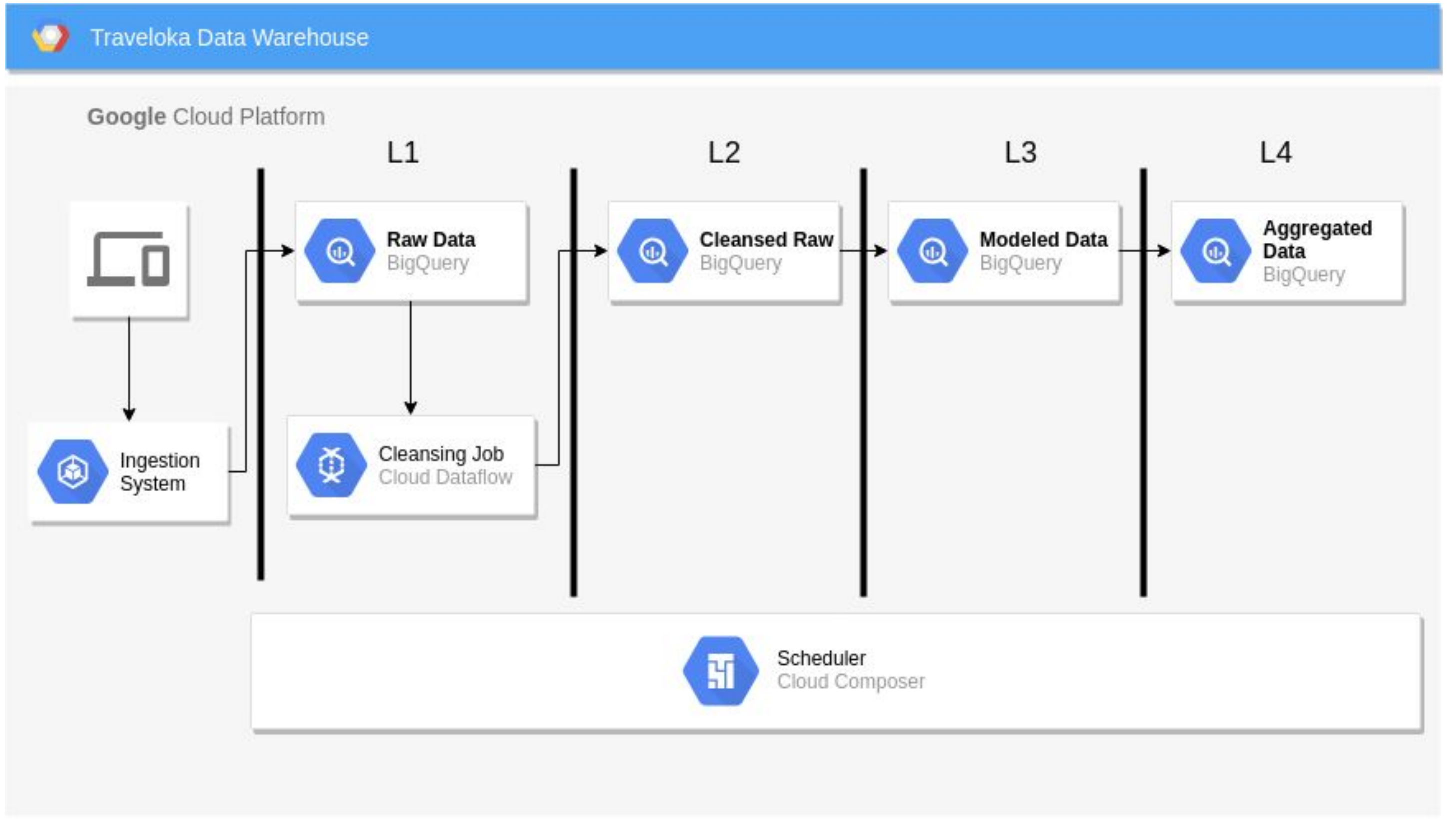
We call the framework NeoDDL

Just like normal DDL / schema (think CREATE TABLE command), but...

- **in YAML**, so it's easier to read both by human and machine
- **Support inheritance**, which is key to simpler ddl where we have so many fields duplication in many places (think session_id, cookie_id, etc.)
- **DDL & cleansing rule in one place**, you could specify simple cleansing rule in the DDL itself, think of adding regex to validate your STRING, or to check whether STRING value belong to certain enum or not. Eg.
- **Integrated to data catalog**

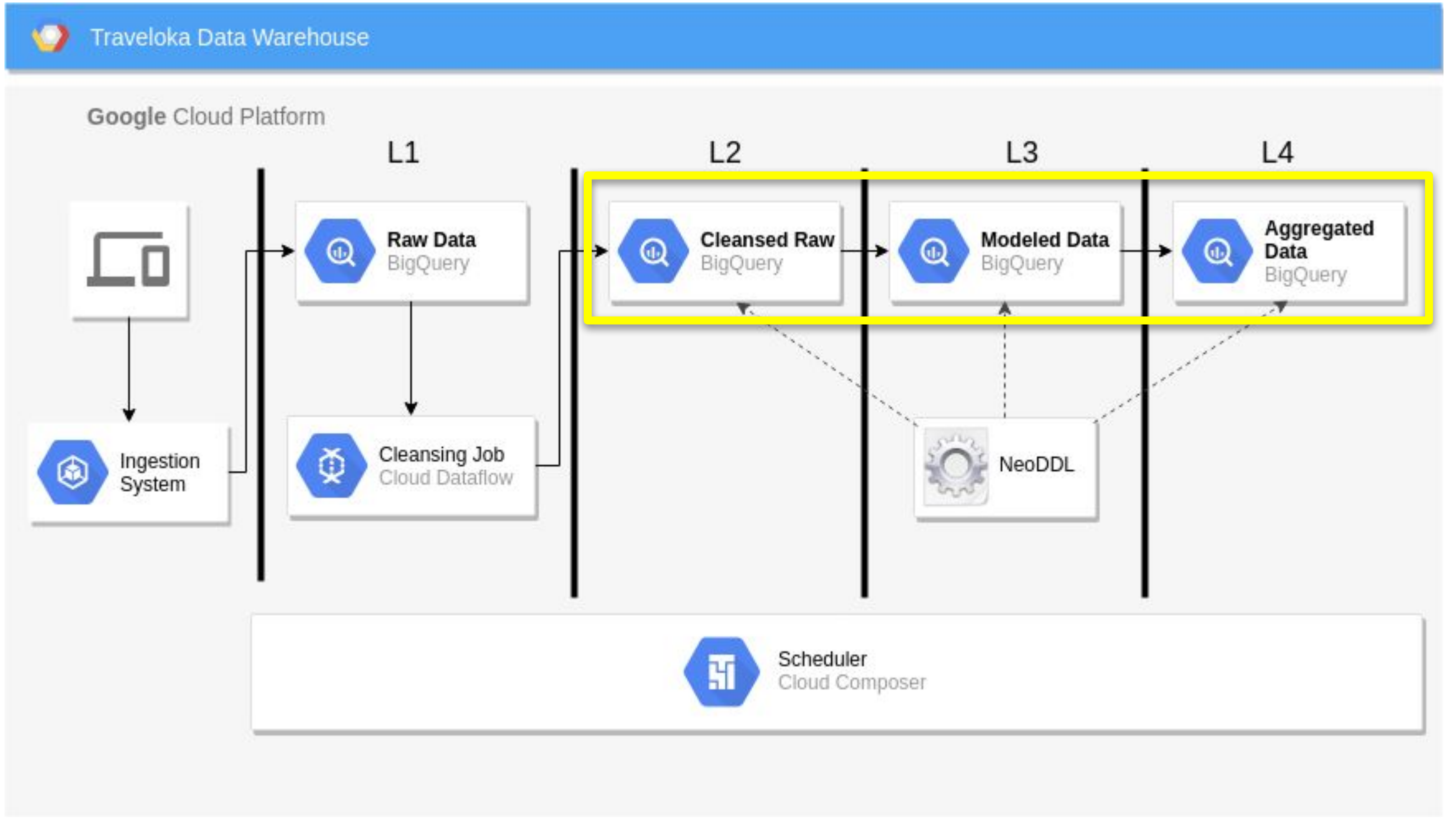
So how is NeoDDL being utilized in our data processing flow?

Our Current Data Warehouse



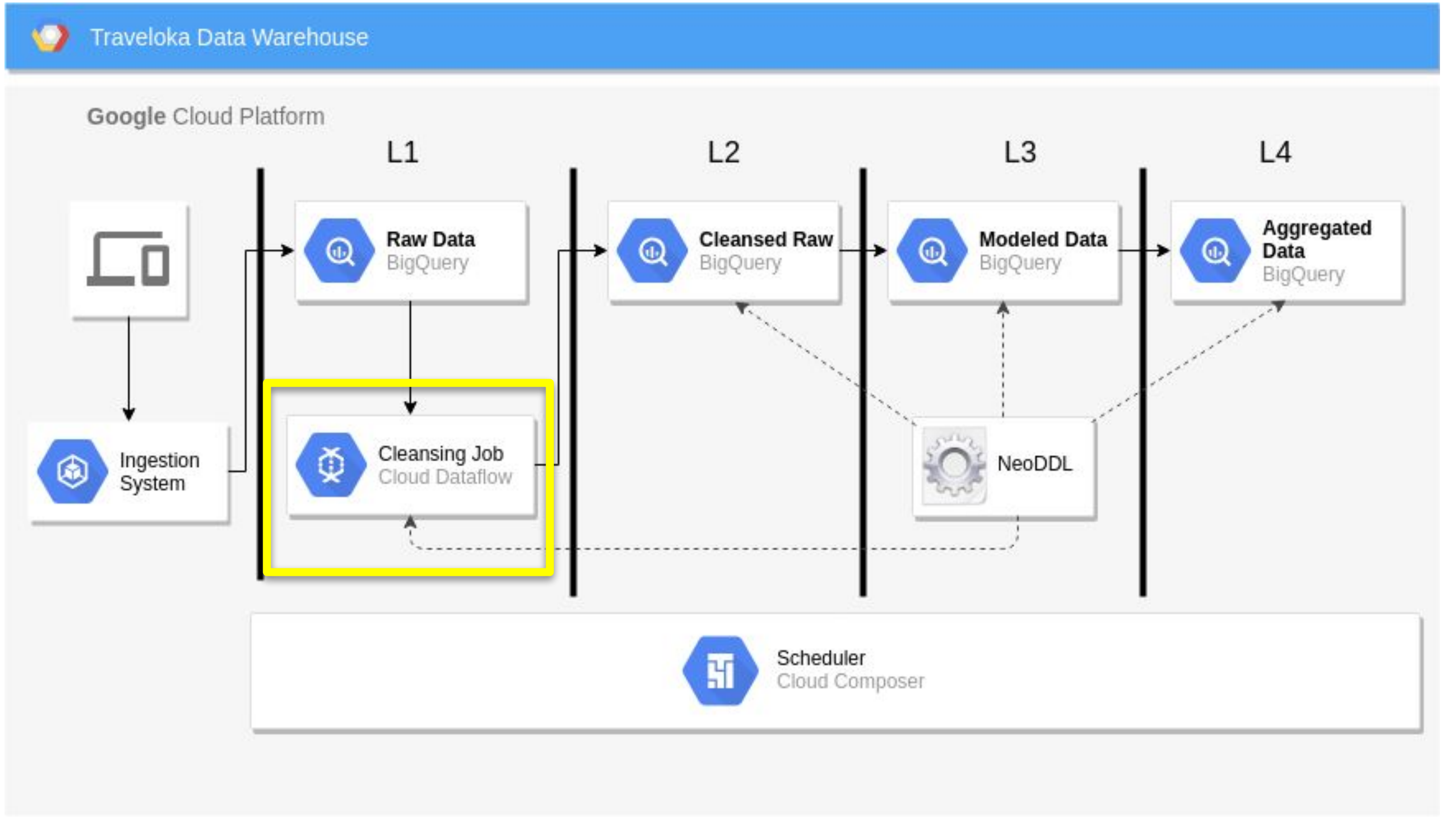
- Increasing data quality as the layer progresses
- Data staging area on L1 and L2
- Modeled data on L3 and L4

Our Current Data Warehouse

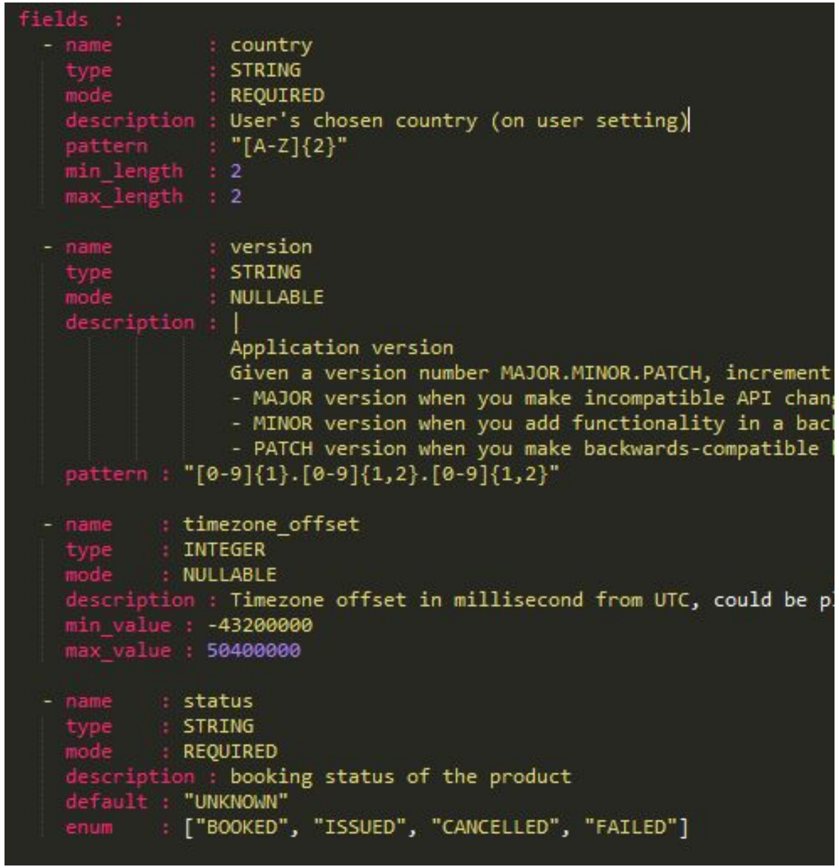


- NeoDDL is used in table creation
- Schema inheritance allows consistent **embedded** dimension schema across business units

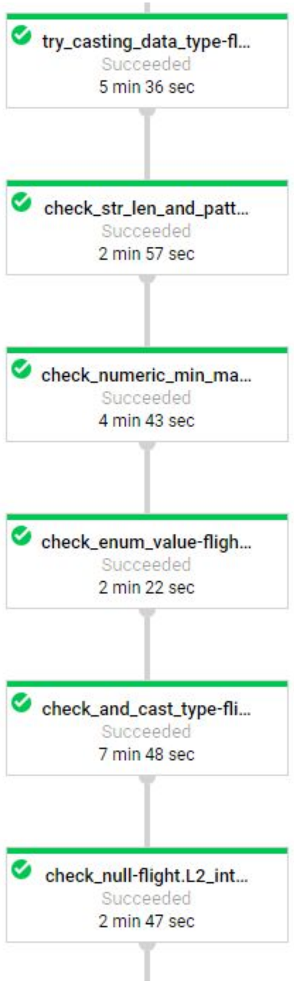
Our Current Data Warehouse



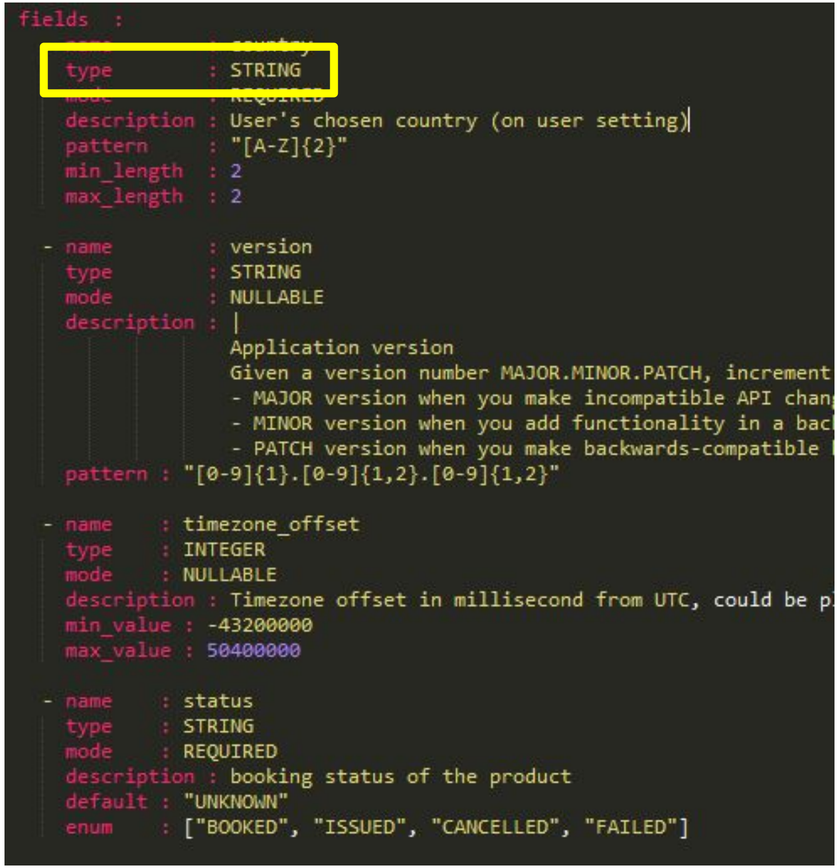
- NeoDDL is used during cleansing job in Cloud Dataflow
- Each rule is converted into dataflow step
- Consistent cleansing rule across business unit



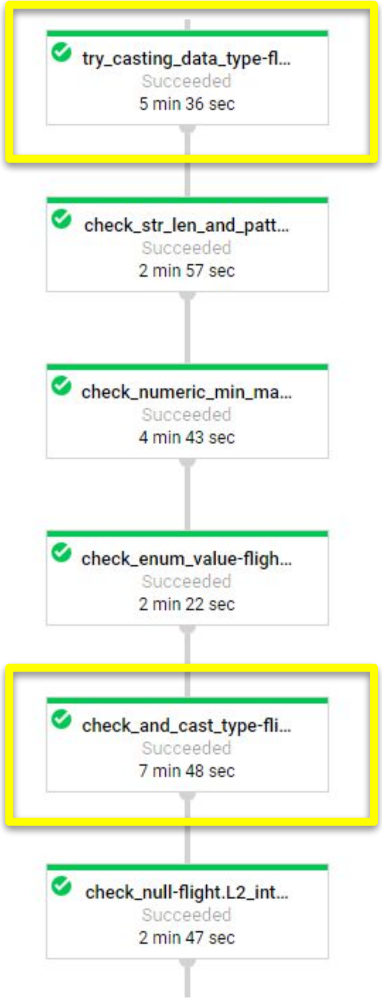
```
fields  :
  - name        : country
    type        : STRING
    mode        : REQUIRED
    description : User's chosen country (on user setting)
    pattern     : "[A-Z]{2}"
    min_length  : 2
    max_length  : 2

  - name        : version
    type        : STRING
    mode        : NULLABLE
    description : |
                  Application version
                  Given a version number MAJOR.MINOR.PATCH, increment
                  - MAJOR version when you make incompatible API chan
                  - MINOR version when you add functionality in a bac
                  - PATCH version when you make backwards-compatible
    pattern : "[0-9]{1}.[0-9]{1,2}.[0-9]{1,2}"

  - name    : timezone_offset
    type    : INTEGER
    mode    : NULLABLE
    description : Timezone offset in millisecond from UTC, could be p
    min_value : -43200000
    max_value : 50400000

  - name    : status
    type    : STRING
    mode    : REQUIRED
    description : booking status of the product
    default : "UNKNOWN"
    enum    : ["BOOKED", "ISSUED", "CANCELLED", "FAILED"]
```



```
try_casting_data_type-fl...
Succeeded
5 min 36 sec

check_str_len_and_patt...
Succeeded
2 min 57 sec

check_numeric_min_ma...
Succeeded
4 min 43 sec

check_enum_value-fligh...
Succeeded
2 min 22 sec

check_and_cast_type-fli...
Succeeded
7 min 48 sec

check_null-flight.L2_int...
Succeeded
2 min 47 sec
```



```
fields  :
    name         : country
    type         : STRING
    mode         : REQUIRED
    description : User's chosen country (on user setting)|
    pattern      : "[A-Z]{2}"
    min_length   : 2
    max_length   : 2

  - name         : version
    type         : STRING
    mode         : NULLABLE
    description : |
                   Application version
                   Given a version number MAJOR.MINOR.PATCH, increment
                   - MAJOR version when you make incompatible API chan
                   - MINOR version when you add functionality in a bac
                   - PATCH version when you make backwards-compatible
    pattern : "[0-9]{1}.[0-9]{1,2}.[0-9]{1,2}"

  - name    : timezone_offset
    type    : INTEGER
    mode    : NULLABLE
    description : Timezone offset in millisecond from UTC, could be p
    min_value : -43200000
    max_value : 50400000

  - name    : status
    type    : STRING
    mode    : REQUIRED
    description : booking status of the product
    default : "UNKNOWN"
    enum    : ["BOOKED", "ISSUED", "CANCELLED", "FAILED"]
```



```
✓ try_casting_data_type-fl...
  Succeeded
  5 min 36 sec

✓ check_str_len_and_patt...
  Succeeded
  2 min 57 sec

✓ check_numeric_min_ma...
  Succeeded
  4 min 43 sec

✓ check_enum_value-fligh...
  Succeeded
  2 min 22 sec

✓ check_and_cast_type-fli...
  Succeeded
  7 min 48 sec

✓ check_null-flight.L2_int...
  Succeeded
  2 min 47 sec
```

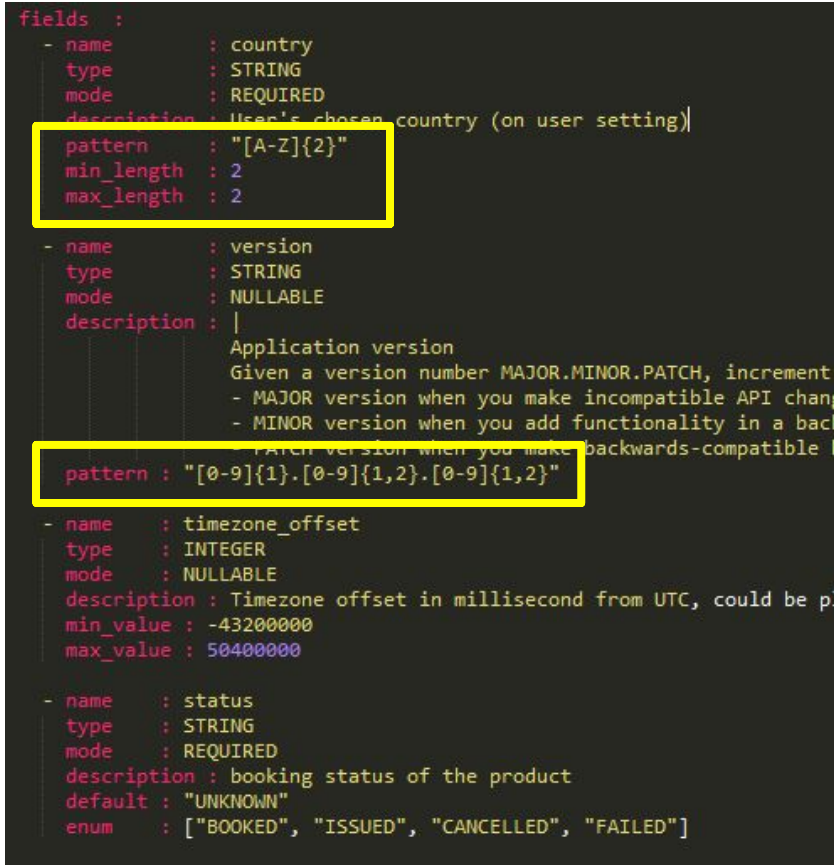
traveloka

First, try to cast the content.

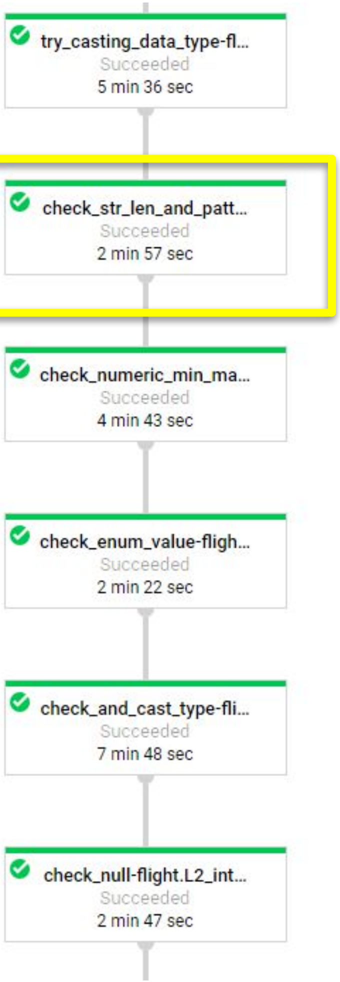
If cast-able, then validate..

Otherwise, **tag** the record and provide the default value

traveloka



```
fields  :
  - name        : country
    type        : STRING
    mode        : REQUIRED
    description : User's chosen country (on user setting)
    pattern     : "[A-Z]{2}"
    min_length  : 2
    max_length  : 2

  - name        : version
    type        : STRING
    mode        : NULLABLE
    description : |
                  Application version
                  Given a version number MAJOR.MINOR.PATCH, increment
                  - MAJOR version when you make incompatible API chan
                  - MINOR version when you add functionality in a bac
                  - PATCH version when you make backwards-compatible
    pattern : "[0-9]{1}.[0-9]{1,2}.[0-9]{1,2}"

  - name     : timezone_offset
    type     : INTEGER
    mode     : NULLABLE
    description : Timezone offset in millisecond from UTC, could be p
    min_value : -43200000
    max_value : 50400000

  - name     : status
    type     : STRING
    mode     : REQUIRED
    description : booking status of the product
    default : "UNKNOWN"
    enum     : ["BOOKED", "ISSUED", "CANCELLED", "FAILED"]
```



try_casting_data_type-fl...
Succeeded
5 min 36 sec

check_str_len_and_patt...
Succeeded
2 min 57 sec

check_numeric_min_ma...
Succeeded
4 min 43 sec

check_enum_value-fligh...
Succeeded
2 min 22 sec

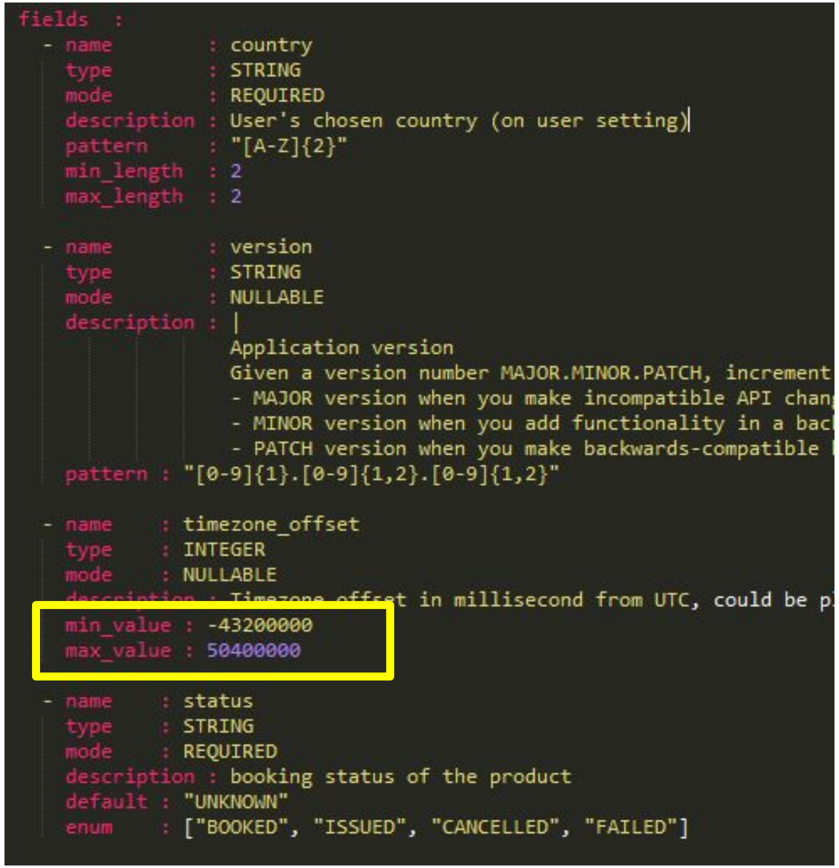
check_and_cast_type-fli...
Succeeded
7 min 48 sec

check_null-flight.L2_int...
Succeeded
2 min 47 sec

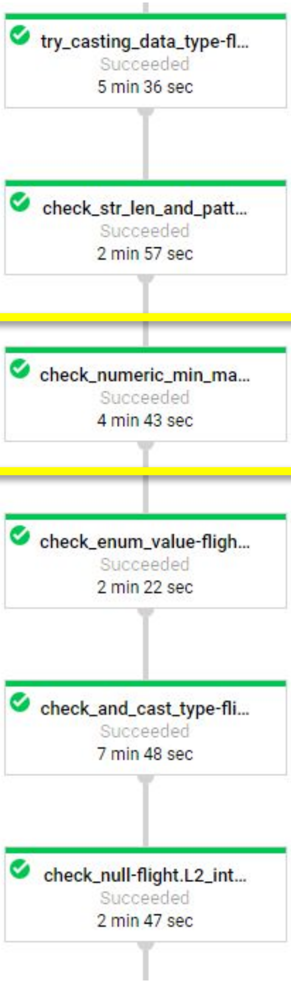
Violation will result in:

1. Error tagging

2. String padding or truncation for string length violation



```
fields  :
  - name        : country
    type        : STRING
    mode        : REQUIRED
    description : User's chosen country (on user setting)
    pattern     : "[A-Z]{2}"
    min_length  : 2
    max_length  : 2

  - name        : version
    type        : STRING
    mode        : NULLABLE
    description : |
                  Application version
                  Given a version number MAJOR.MINOR.PATCH, increment
                  - MAJOR version when you make incompatible API chan
                  - MINOR version when you add functionality in a bac
                  - PATCH version when you make backwards-compatible
    pattern : "[0-9]{1}.[0-9]{1,2}.[0-9]{1,2}"

  - name     : timezone_offset
    type     : INTEGER
    mode     : NULLABLE
    description : Timezone offset in millisecond from UTC, could be p
    min_value : -43200000
    max_value : 50400000

  - name     : status
    type     : STRING
    mode     : REQUIRED
    description : booking status of the product
    default : "UNKNOWN"
    enum     : ["BOOKED", "ISSUED", "CANCELLED", "FAILED"]
```



✓ try_casting_data_type-fl...
Succeeded
5 min 36 sec

✓ check_str_len_and_patt...
Succeeded
2 min 57 sec

✓ check_numeric_min_ma...
Succeeded
4 min 43 sec

✓ check_enum_value-fligh...
Succeeded
2 min 22 sec

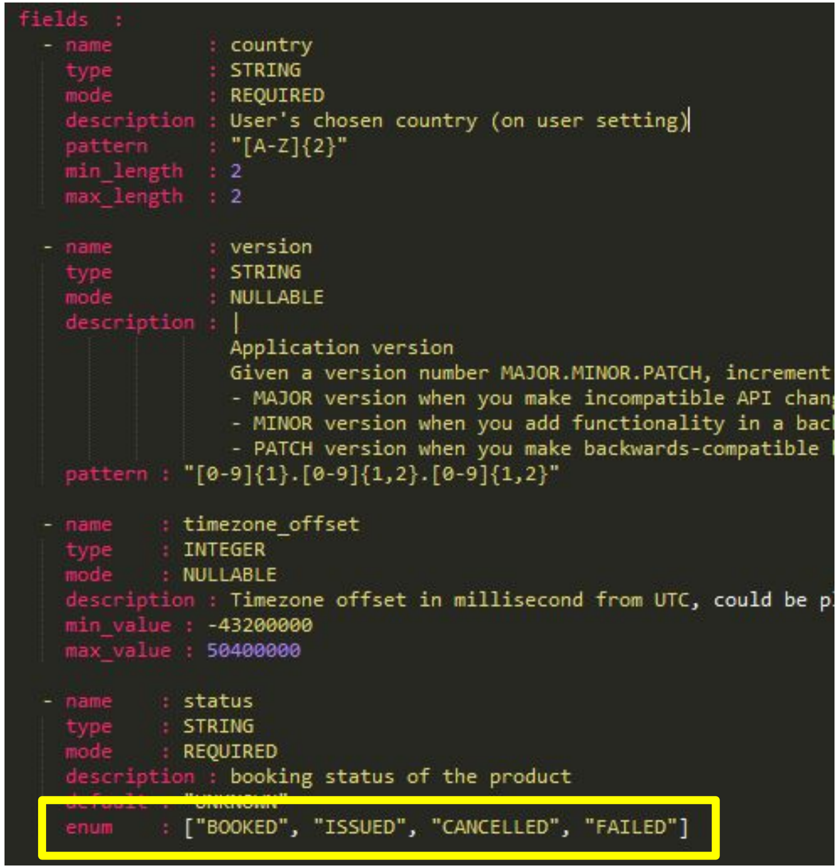
✓ check_and_cast_type-fli...
Succeeded
7 min 48 sec

✓ check_null-flight.L2_int...
Succeeded
2 min 47 sec

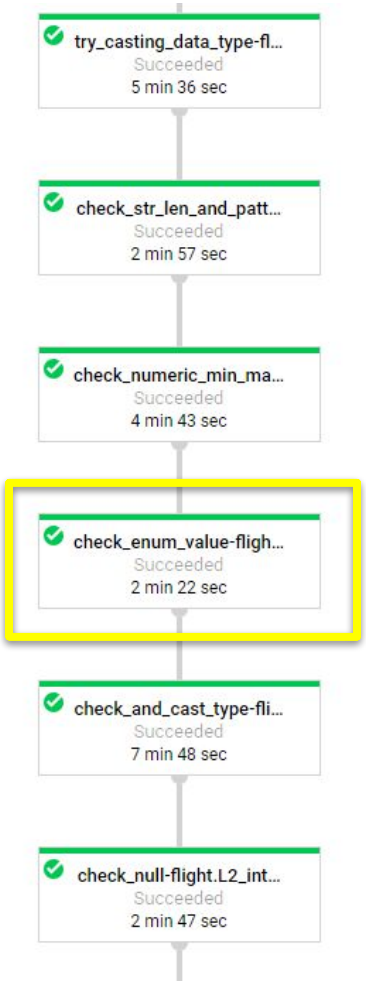
Violation will only result in error tagging

The data content will not be changed.

traveloka



```
fields  :
  - name        : country
    type        : STRING
    mode        : REQUIRED
    description : User's chosen country (on user setting)
    pattern     : "[A-Z]{2}"
    min_length  : 2
    max_length  : 2

  - name        : version
    type        : STRING
    mode        : NULLABLE
    description : |
                  Application version
                  Given a version number MAJOR.MINOR.PATCH, increment
                  - MAJOR version when you make incompatible API chan
                  - MINOR version when you add functionality in a bac
                  - PATCH version when you make backwards-compatible 
    pattern : "[0-9]{1}.[0-9]{1,2}.[0-9]{1,2}"

  - name    : timezone_offset
    type    : INTEGER
    mode    : NULLABLE
    description : Timezone offset in millisecond from UTC, could be p
    min_value : -43200000
    max_value : 50400000

  - name    : status
    type    : STRING
    mode    : REQUIRED
    description : booking status of the product
    default : "UNKNOWN"
    enum    : ["BOOKED", "ISSUED", "CANCELLED", "FAILED"]
```



try_casting_data_type-fl...
Succeeded
5 min 36 sec

check_str_len_and_patt...
Succeeded
2 min 57 sec

check_numeric_min_ma...
Succeeded
4 min 43 sec

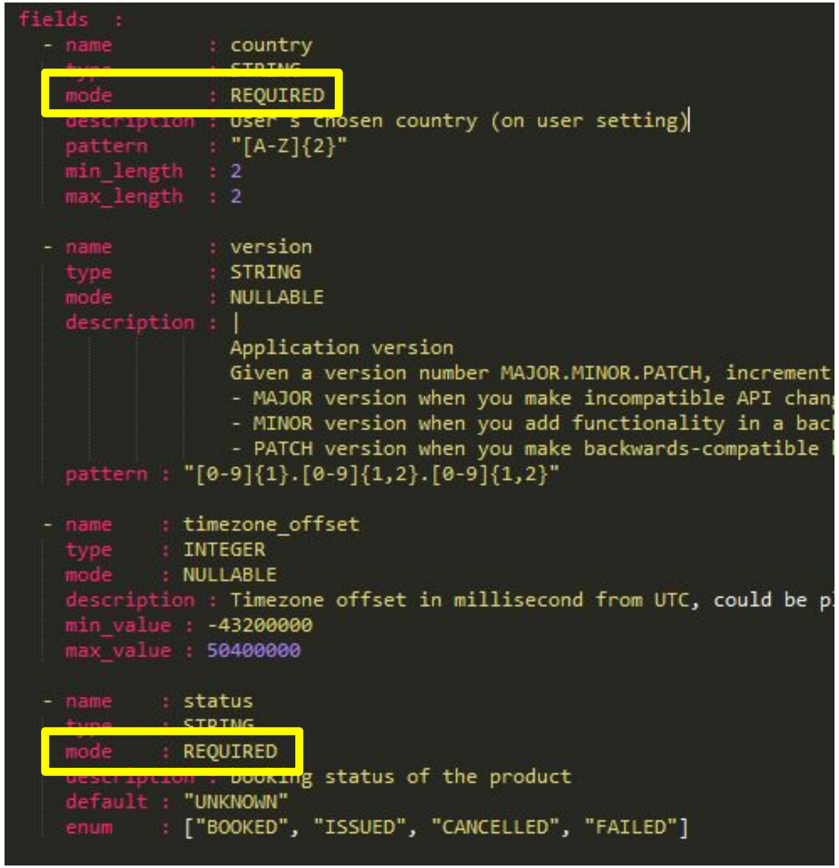
check_enum_value-fligh...
Succeeded
2 min 22 sec

check_and_cast_type-fli...
Succeeded
7 min 48 sec

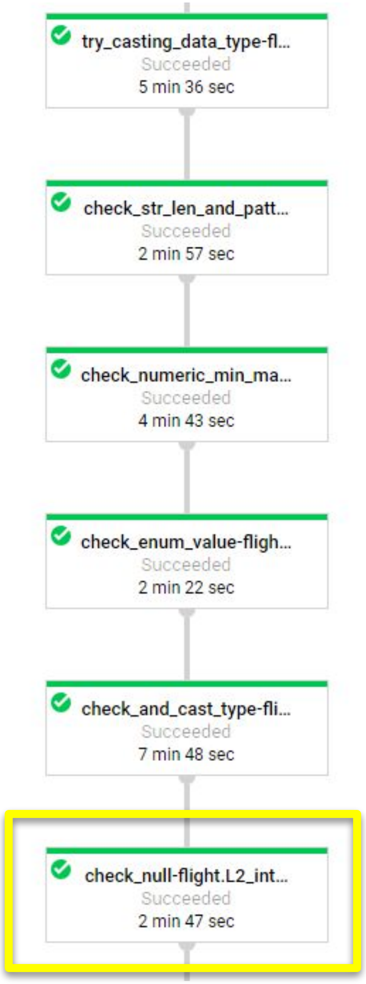
check_null-flight.L2_int...
Succeeded
2 min 47 sec

Violation will only result in error tagging

The data content will not be changed.



```
fields  :
  - name          : country
    type          : STRING
    mode          : REQUIRED
    description : User's chosen country (on user setting)
    pattern       : "[A-Z]{2}"
    min_length    : 2
    max_length    : 2

  - name          : version
    type          : STRING
    mode          : NULLABLE
    description : |
                    Application version
                    Given a version number MAJOR.MINOR.PATCH, increment
                    - MAJOR version when you make incompatible API chan
                    - MINOR version when you add functionality in a bac
                    - PATCH version when you make backwards-compatible
    pattern : "[0-9]{1}.[0-9]{1,2}.[0-9]{1,2}"

  - name      : timezone_offset
    type      : INTEGER
    mode      : NULLABLE
    description : Timezone offset in millisecond from UTC, could be p
    min_value : -43200000
    max_value : 50400000

  - name      : status
    type      : STRING
    mode      : REQUIRED
    description : Booking status of the product
    default : "UNKNOWN"
    enum      : ["BOOKED", "ISSUED", "CANCELLED", "FAILED"]
```



try_casting_data_type-fl...
Succeeded
5 min 36 sec

check_str_len_and_patt...
Succeeded
2 min 57 sec

check_numeric_min_ma...
Succeeded
4 min 43 sec

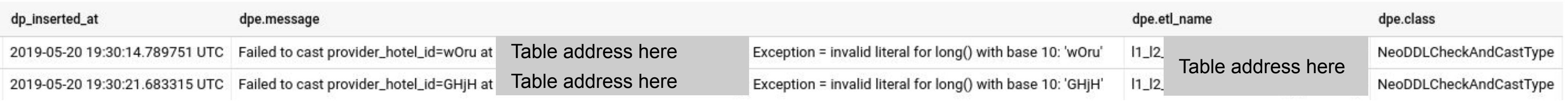
check_enum_value-fligh...
Succeeded
2 min 22 sec

check_and_cast_type-fli...
Succeeded
7 min 48 sec

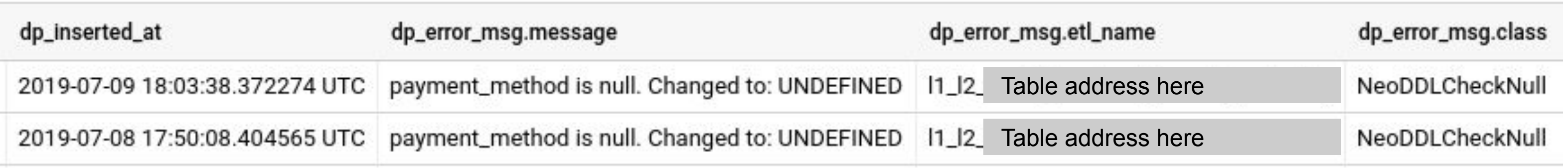
check_null-flight.L2_int...
Succeeded
2 min 47 sec

NULL value in REQUIRED field will be given its default value and tagged with error message

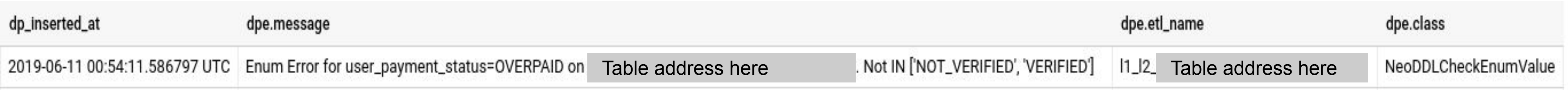
Sample Records with Error Tagging



dp_inserted_at	dpe.message	dpe.etl_name	dpe.class
2019-05-20 19:30:14.789751 UTC	Failed to cast provider_hotel_id=wOru at Table address here Exception = invalid literal for long() with base 10: 'wOru'	l1_l2_ Table address here	NeoDDLCheckAndCastType
2019-05-20 19:30:21.683315 UTC	Failed to cast provider_hotel_id=GHjH at Table address here Exception = invalid literal for long() with base 10: 'GHjH'	l1_l2_	NeoDDLCheckAndCastType

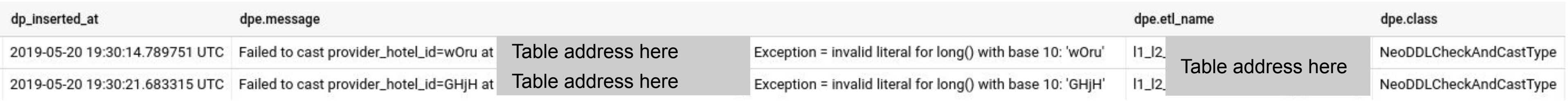


dp_inserted_at	dp_error_msg.message	dp_error_msg.etl_name	dp_error_msg.class
2019-07-09 18:03:38.372274 UTC	payment_method is null. Changed to: UNDEFINED	l1_l2_ Table address here	NeoDDLCheckNull
2019-07-08 17:50:08.404565 UTC	payment_method is null. Changed to: UNDEFINED	l1_l2_ Table address here	NeoDDLCheckNull

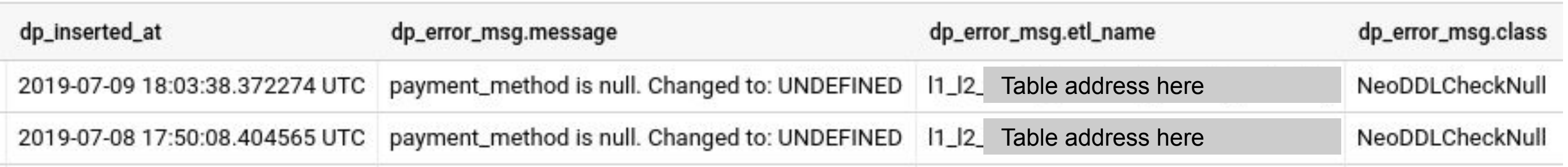


dp_inserted_at	dpe.message	dpe.etl_name	dpe.class
2019-06-11 00:54:11.586797 UTC	Enum Error for user_payment_status=OVERPAID on Table address here . Not IN ['NOT_VERIFIED', 'VERIFIED']	l1_l2_ Table address here	NeoDDLCheckEnumValue

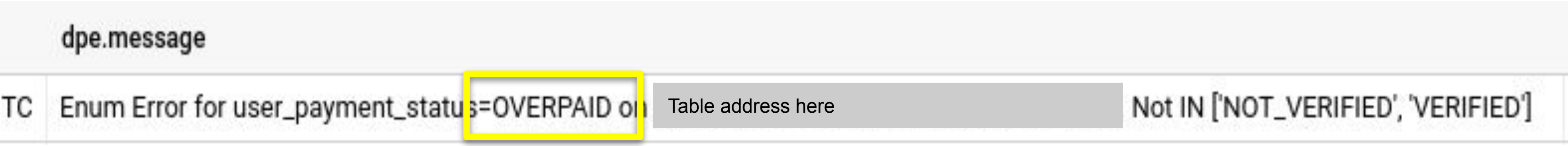
Sample Records with Error Tagging



dp_inserted_at	dpe.message	dpe.etl_name	dpe.class
2019-05-20 19:30:14.789751 UTC	Failed to cast provider_hotel_id=wOru at Table address here Exception = invalid literal for long() with base 10: 'wOru'	l1_l2_ Table address here	NeoDDLCheckAndCastType
2019-05-20 19:30:21.683315 UTC	Failed to cast provider_hotel_id=GHjH at Table address here Exception = invalid literal for long() with base 10: 'GHjH'	l1_l2_	NeoDDLCheckAndCastType



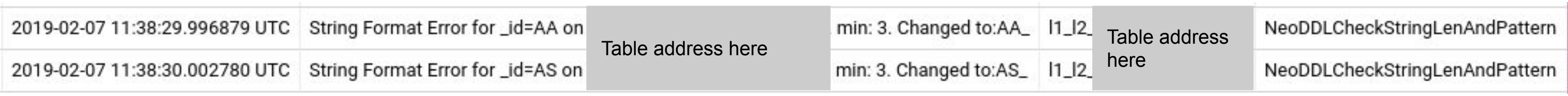
dp_inserted_at	dp_error_msg.message	dp_error_msg.etl_name	dp_error_msg.class
2019-07-09 18:03:38.372274 UTC	payment_method is null. Changed to: UNDEFINED	l1_l2_ Table address here	NeoDDLCheckNull
2019-07-08 17:50:08.404565 UTC	payment_method is null. Changed to: UNDEFINED	l1_l2_ Table address here	NeoDDLCheckNull



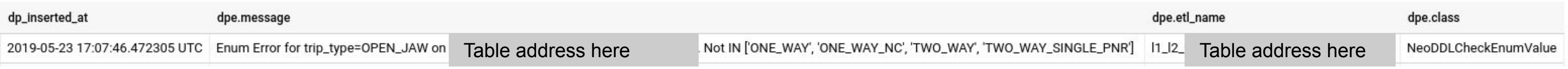
	dpe.message
TC	Enum Error for user_payment_status=OVERPAID on Table address here Not IN ['NOT_VERIFIED', 'VERIFIED']

Well. Thank you.. I guess?

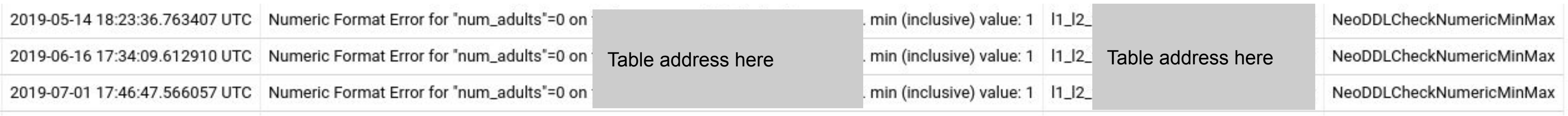
Sample Records with Error Tagging



2019-02-07 11:38:29.996879 UTC	String Format Error for _id=AA on	Table address here	min: 3. Changed to:AA_	l1_l2_	Table address here	NeoDDLCheckStringLenAndPattern
2019-02-07 11:38:30.002780 UTC	String Format Error for _id=AS on		min: 3. Changed to:AS_	l1_l2_		NeoDDLCheckStringLenAndPattern

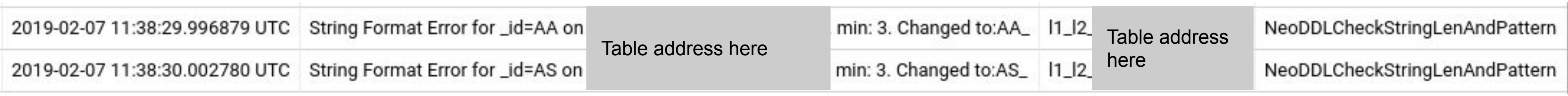


dp_inserted_at	dpe.message			dpe.etl_name		dpe.class
2019-05-23 17:07:46.472305 UTC	Enum Error for trip_type=OPEN_JAW on	Table address here	. Not IN ['ONE_WAY', 'ONE_WAY_NC', 'TWO_WAY', 'TWO_WAY_SINGLE_PNR']	l1_l2_	Table address here	NeoDDLCheckEnumValue

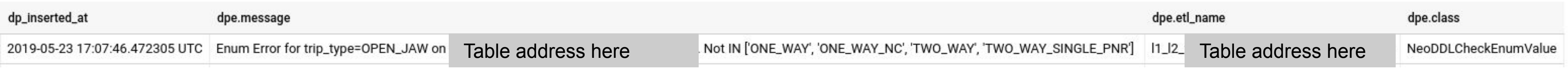


2019-05-14 18:23:36.763407 UTC	Numeric Format Error for "num_adults"=0 on	Table address here	. min (inclusive) value: 1	l1_l2_	Table address here	NeoDDLCheckNumericMinMax
2019-06-16 17:34:09.612910 UTC	Numeric Format Error for "num_adults"=0 on		. min (inclusive) value: 1	l1_l2_		NeoDDLCheckNumericMinMax
2019-07-01 17:46:47.566057 UTC	Numeric Format Error for "num_adults"=0 on		. min (inclusive) value: 1	l1_l2_		NeoDDLCheckNumericMinMax

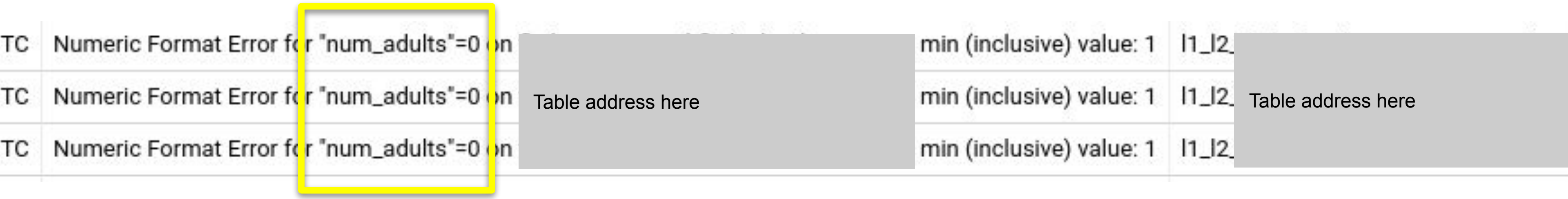
Sample Records with Error Tagging



2019-02-07 11:38:29.996879 UTC	String Format Error for _id=AA on	Table address here	min: 3. Changed to:AA_	l1_l2_	Table address here	NeoDDLCheckStringLenAndPattern
2019-02-07 11:38:30.002780 UTC	String Format Error for _id=AS on		min: 3. Changed to:AS_	l1_l2_		NeoDDLCheckStringLenAndPattern



dp_inserted_at	dpe.message			dpe.etl_name		dpe.class
2019-05-23 17:07:46.472305 UTC	Enum Error for trip_type=OPEN_JAW on	Table address here	Not IN ['ONE_WAY', 'ONE_WAY_NC', 'TWO_WAY', 'TWO_WAY_SINGLE_PNR']	l1_l2_	Table address here	NeoDDLCheckEnumValue



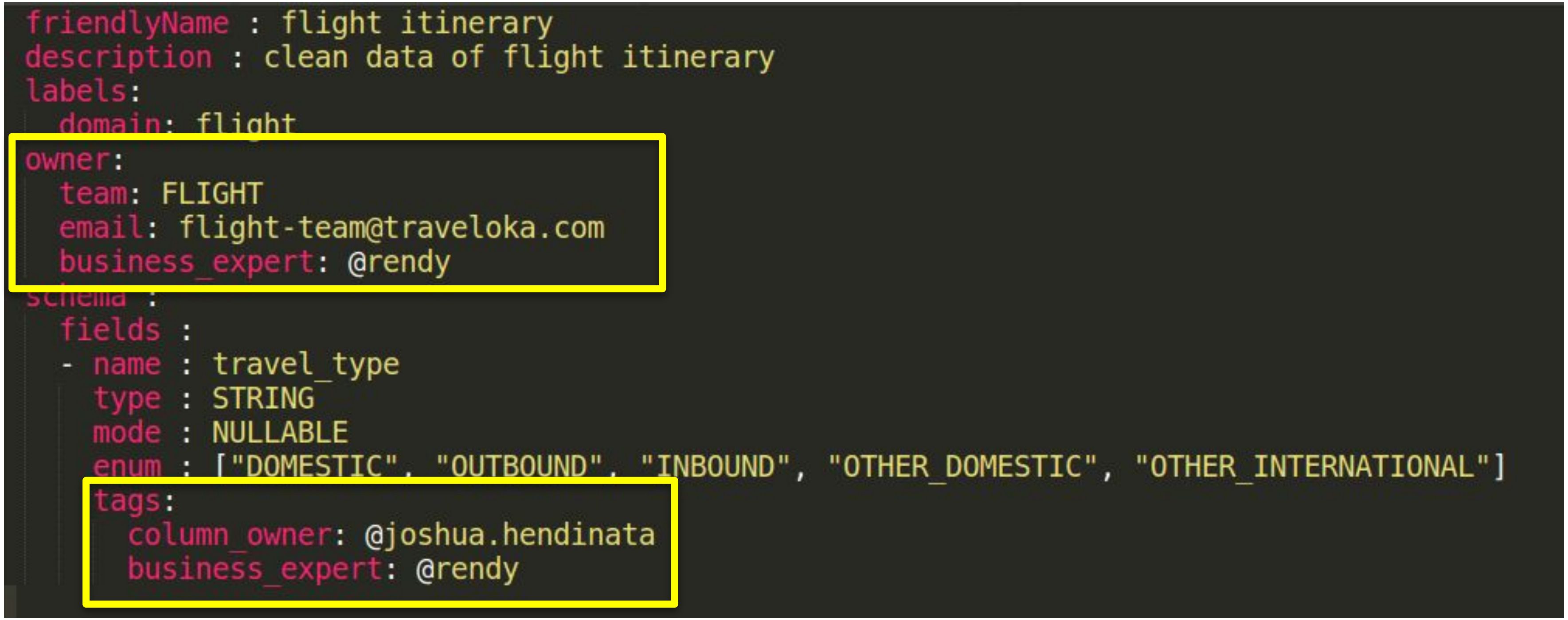
TC	Numeric Format Error for "num_adults"=0 on		min (inclusive) value: 1	l1_l2_	
TC	Numeric Format Error for "num_adults"=0 on	Table address here	min (inclusive) value: 1	l1_l2_	Table address here
TC	Numeric Format Error for "num_adults"=0 on		min (inclusive) value: 1	l1_l2_	

What a brave young soul..

Future Plan

traveloka

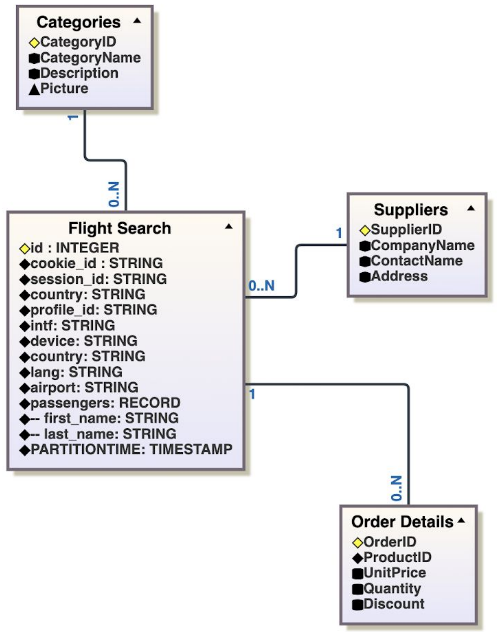
Add More Business Metadata for Data Cataloging



```
friendlyName : flight itinerary
description : clean data of flight itinerary
labels:
  domain: flight
owner:
  team: FLIGHT
  email: flight-team@traveloka.com
  business_expert: @rendy
schema :
  fields :
  - name : travel_type
    type : STRING
    mode : NULLABLE
    enum : ["DOMESTIC", "OUTBOUND", "INBOUND", "OTHER_DOMESTIC", "OTHER_INTERNATIONAL"]
    tags:
      column_owner: @joshua.hendinata
      business_expert: @rendy
```

Add Metadata on Data Model Relationship

- Foreign key and target table
- Enable automatic star schema diagram generation



traveloka

Thank You!

rendy@traveloka.com
joshua.hendinata@traveloka.com