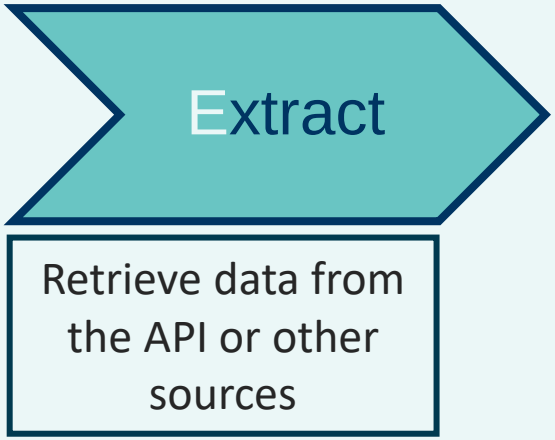
deda.next

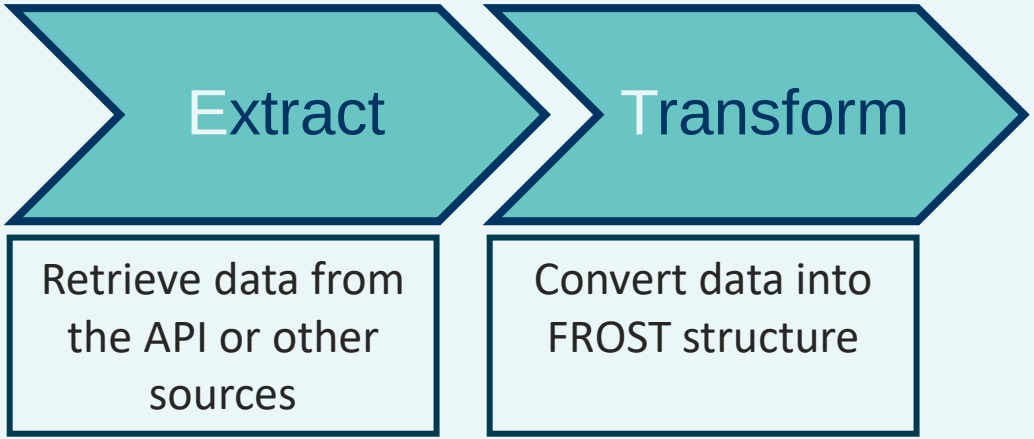
05 – Data Ingestion in FROST via Apache NiFi

Beatrice Olivari

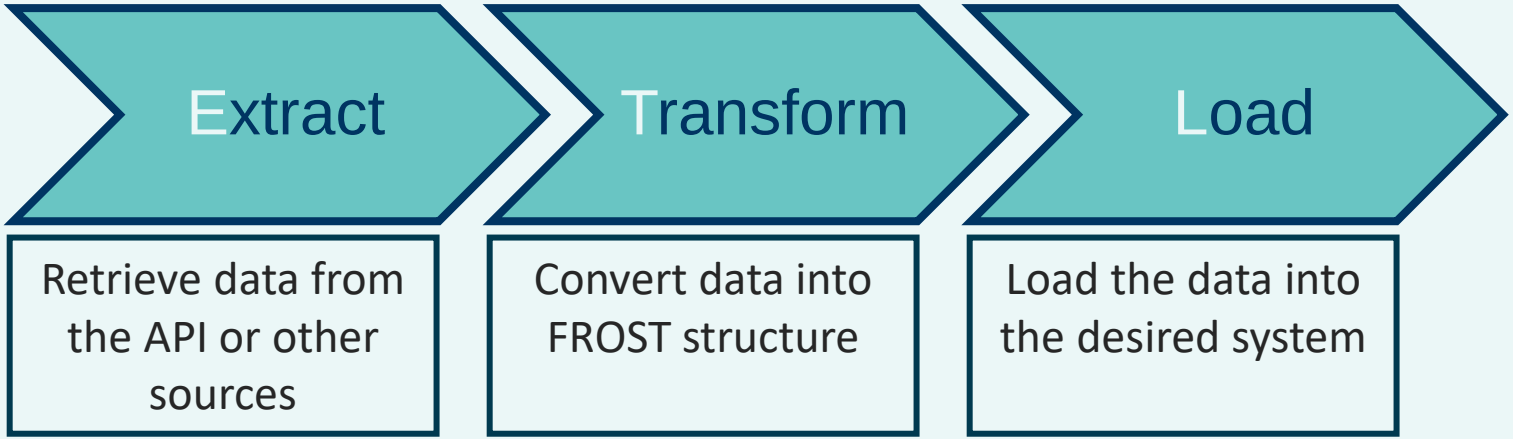
Go with the (ETL) FLOW



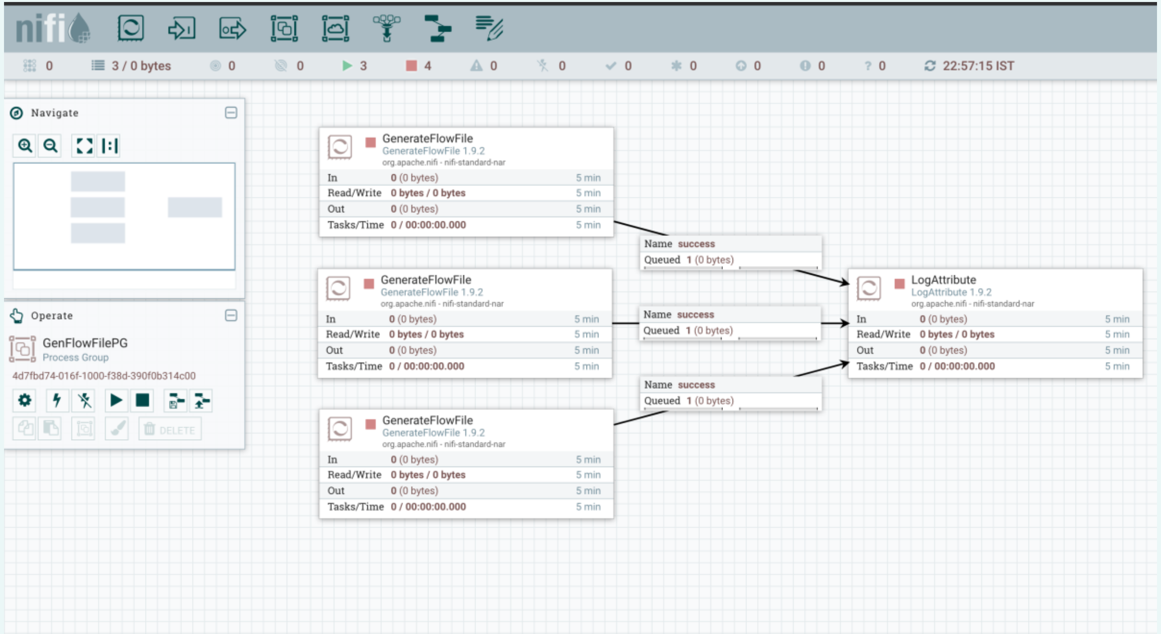
Go with the (ETL) FLOW



Go with the (ETL) FLOW



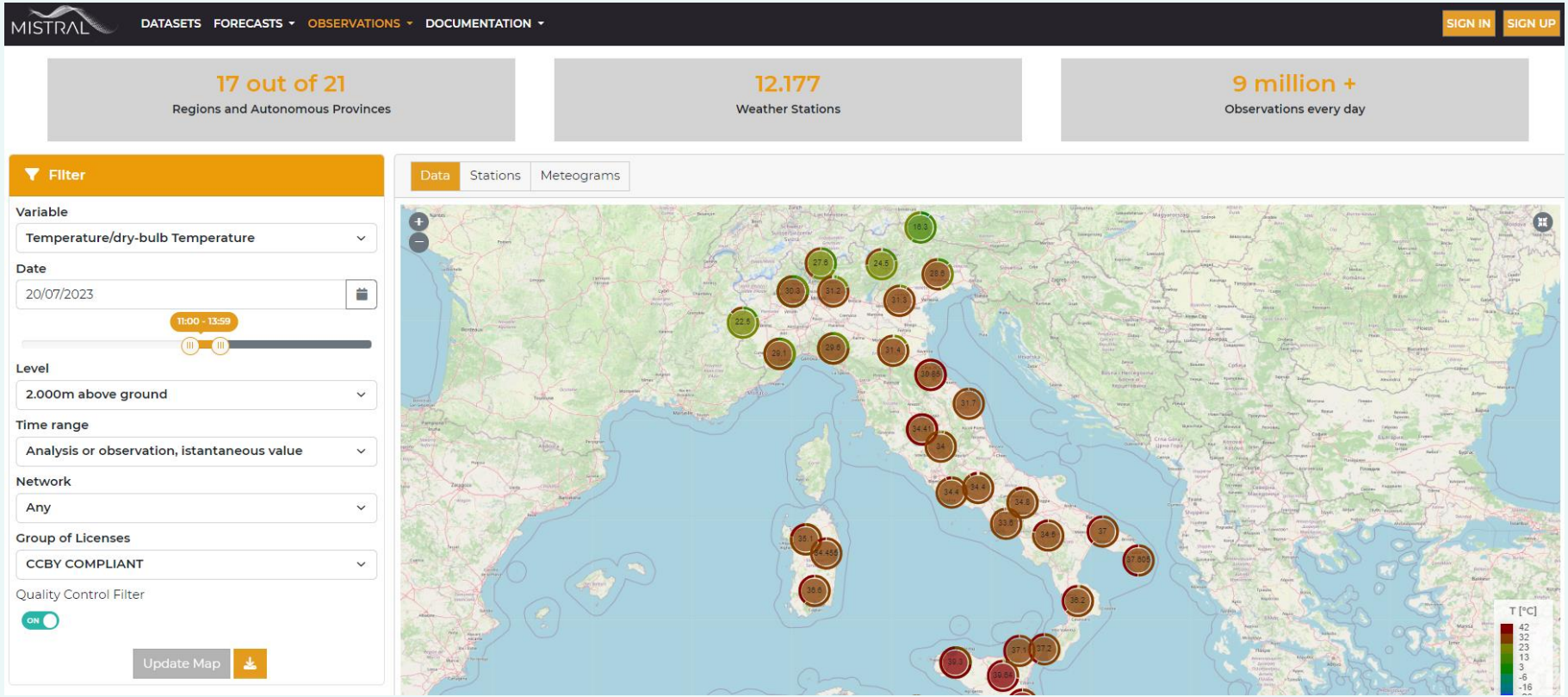
The tool: Apache NiFi



Example flow: Meteorological data

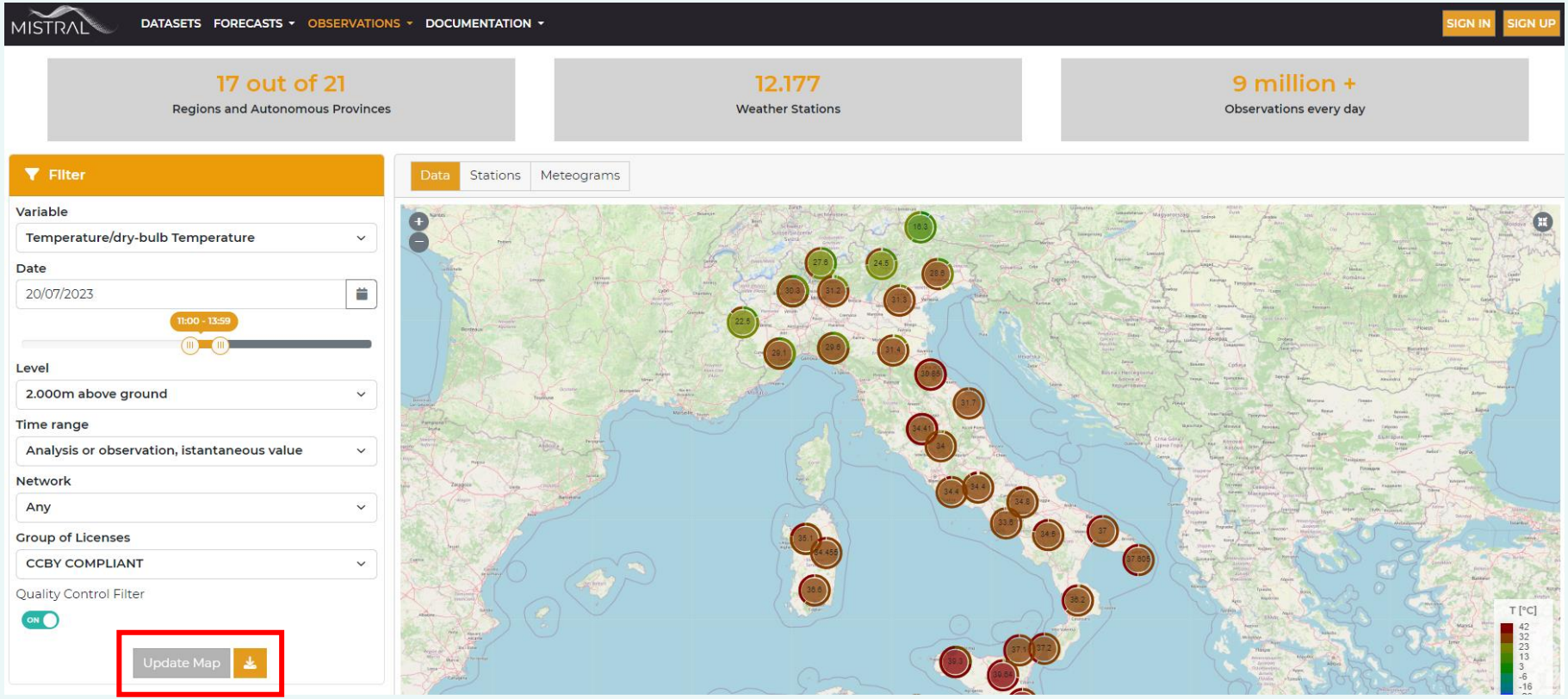
- MISTRAL is an **EU financed** project **started in 2018**
- The **goal** of the MISTRAL portal is to facilitate and foster the re-use of the datasets by the weather community, as well as by its cross-area communities

Mistral: User Interface



https://meteohub.mistralportal.it/app/maps/observations

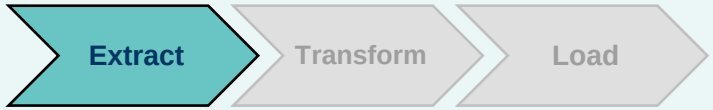
Mistral: User Interface



https://meteohub.mistralportal.it/app/maps/observations

Mistral: Data Extraction

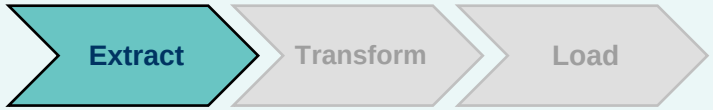
How do we extract data in a continued, up to date and automated way?



Mistral: Data Extraction

How do we extract data in a continued and automated way?

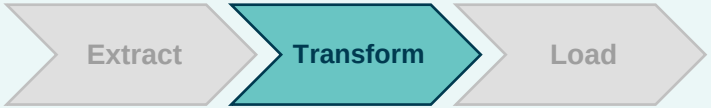
Mistral API



Mistral: Data Transformation

deda.next x EDIAQI

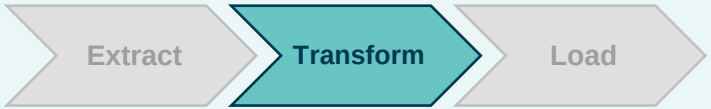
How do we make data STA compliant?



Mistral: Data Transformation

How do we make data STA compliant?

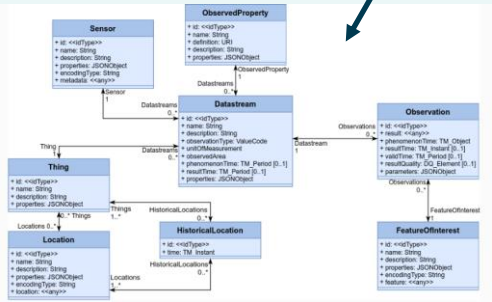
Mapping

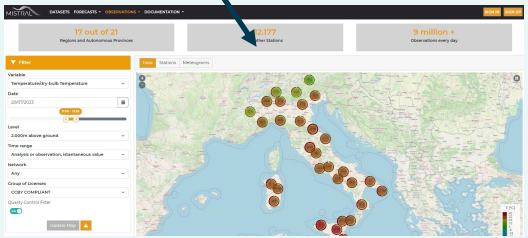


Mistral: Mapping

STA entity

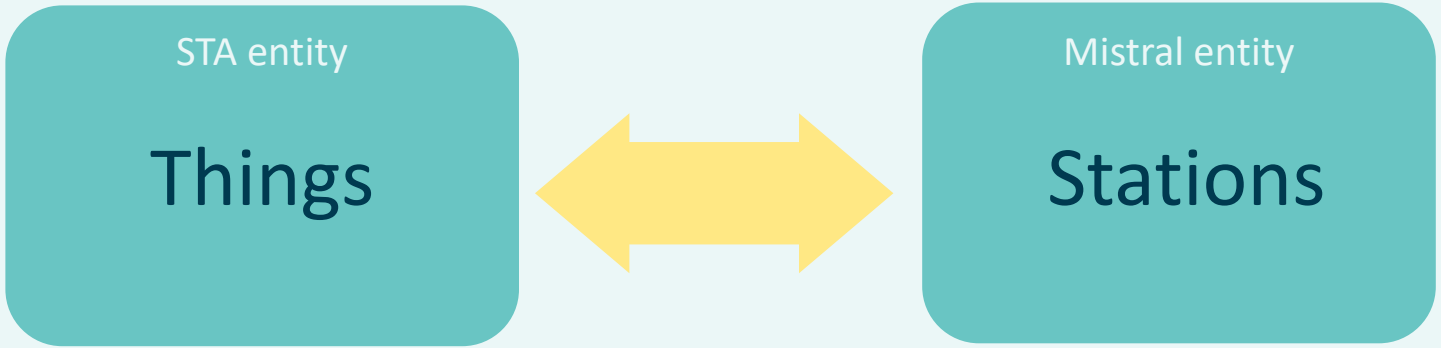
Mistral entity

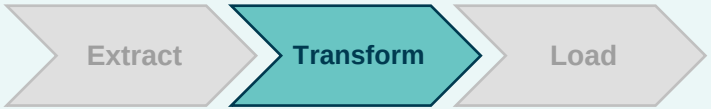




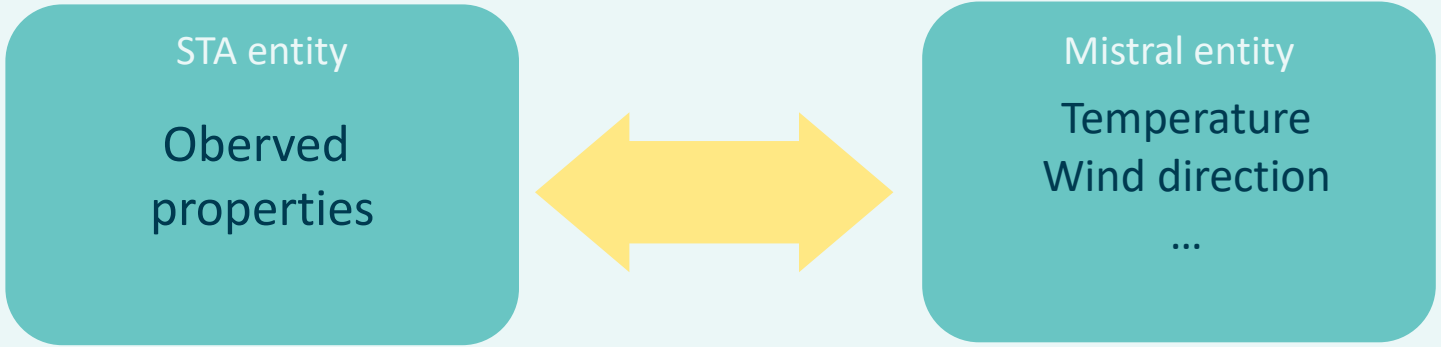
Extract → Transform → Load

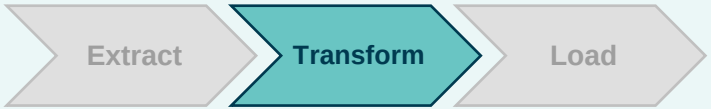
Mistral: Mapping



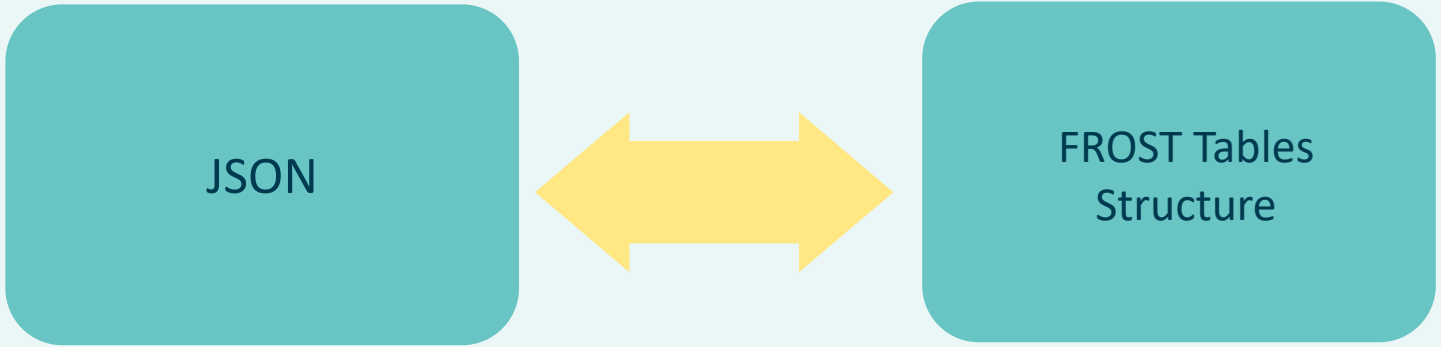


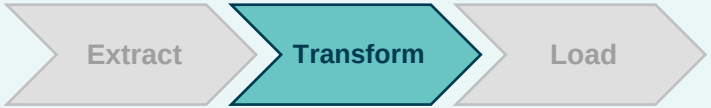
Mistral: Mapping



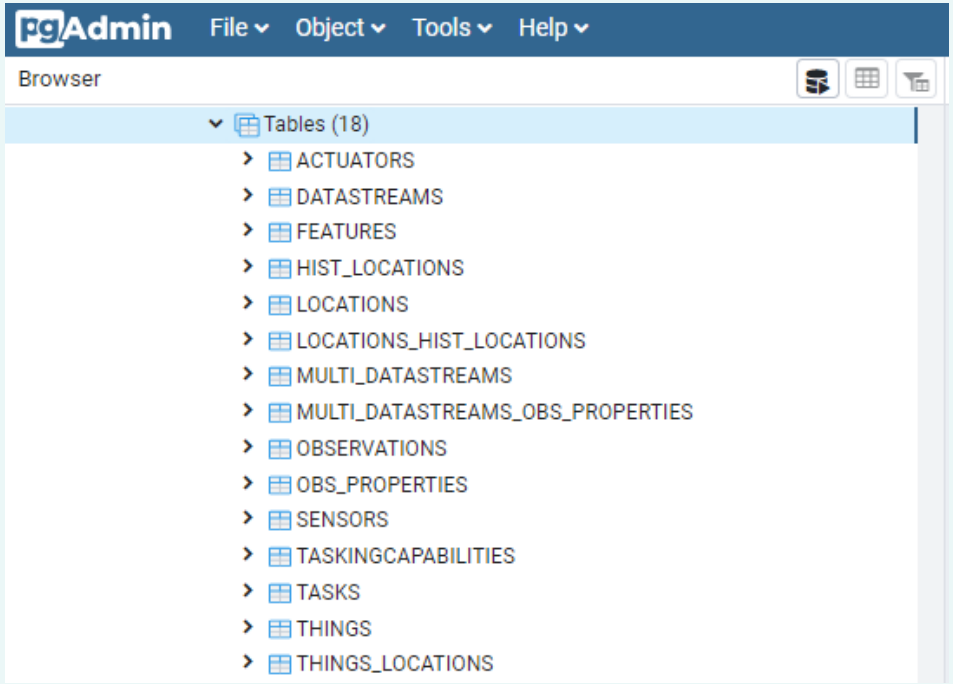


Mistral: Transformation



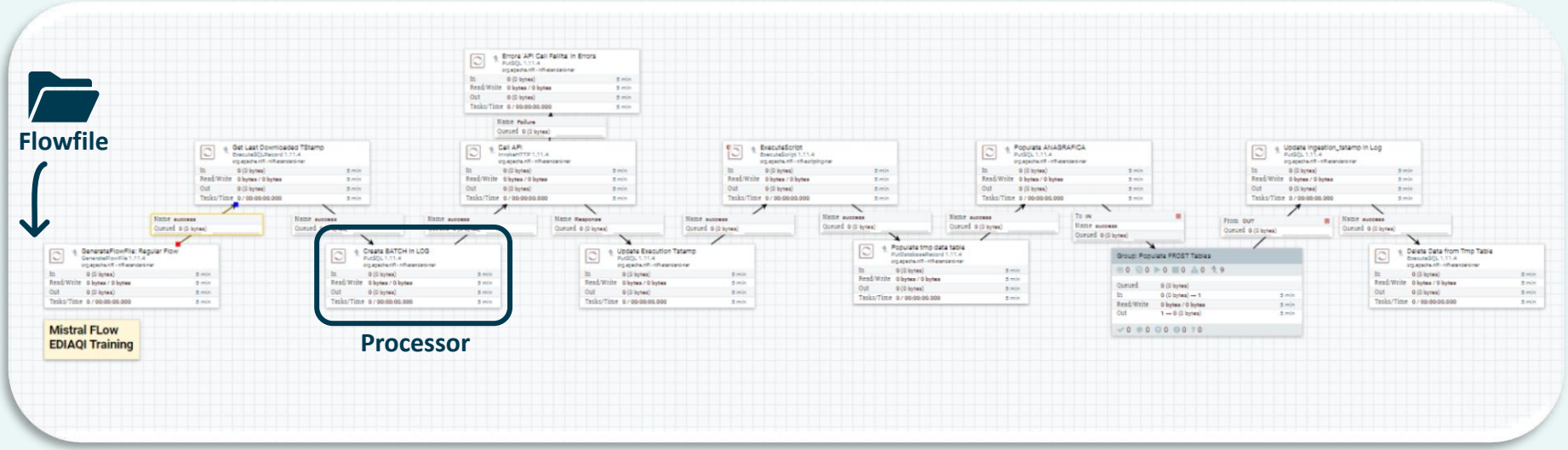


Mistral: Load

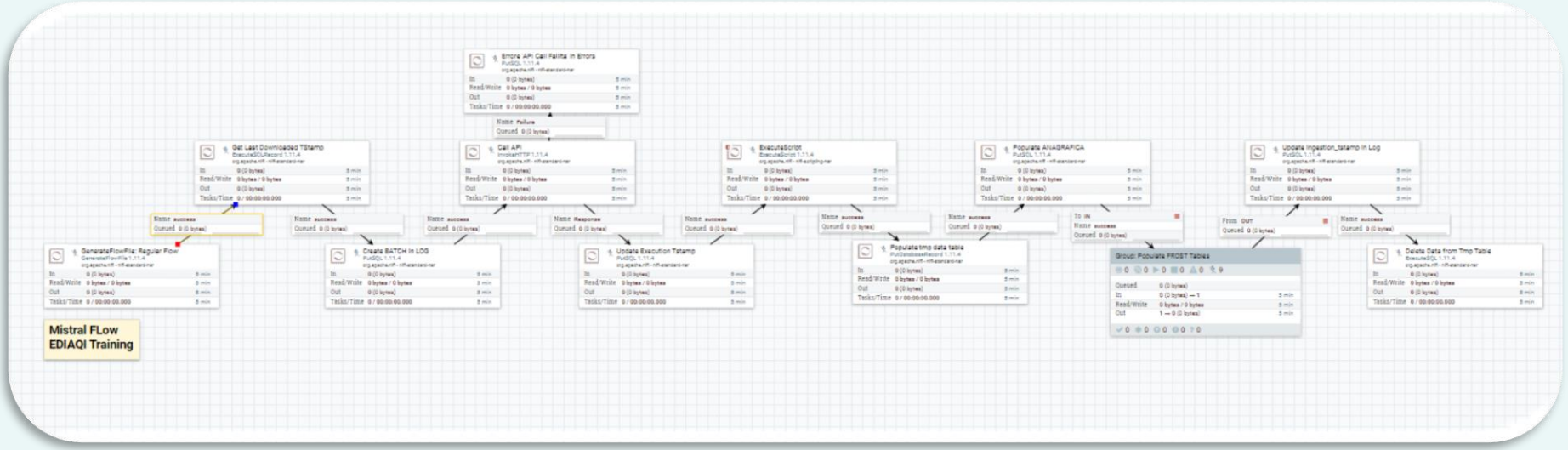


Extract → Transform → **Load**

Mistral: NiFi Flow

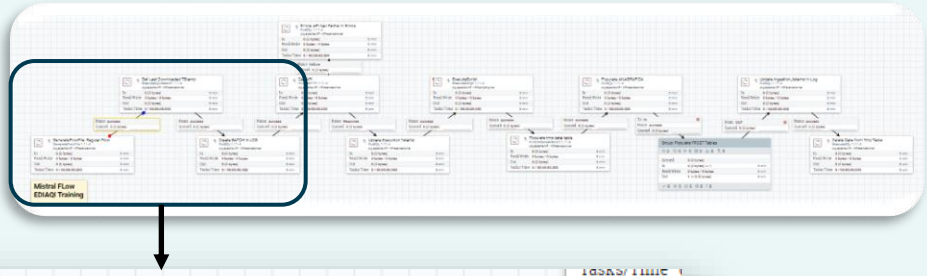


Mistral: NiFi Flow



1 flow trip = 1 day of data

Mistral: NiFi Flow



Get Last Downloaded TStamp
ExecuteSQLRecord 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

GenerateFlowFile: Regular Flow
GenerateFlowFile 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Create BATCH in LOG
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

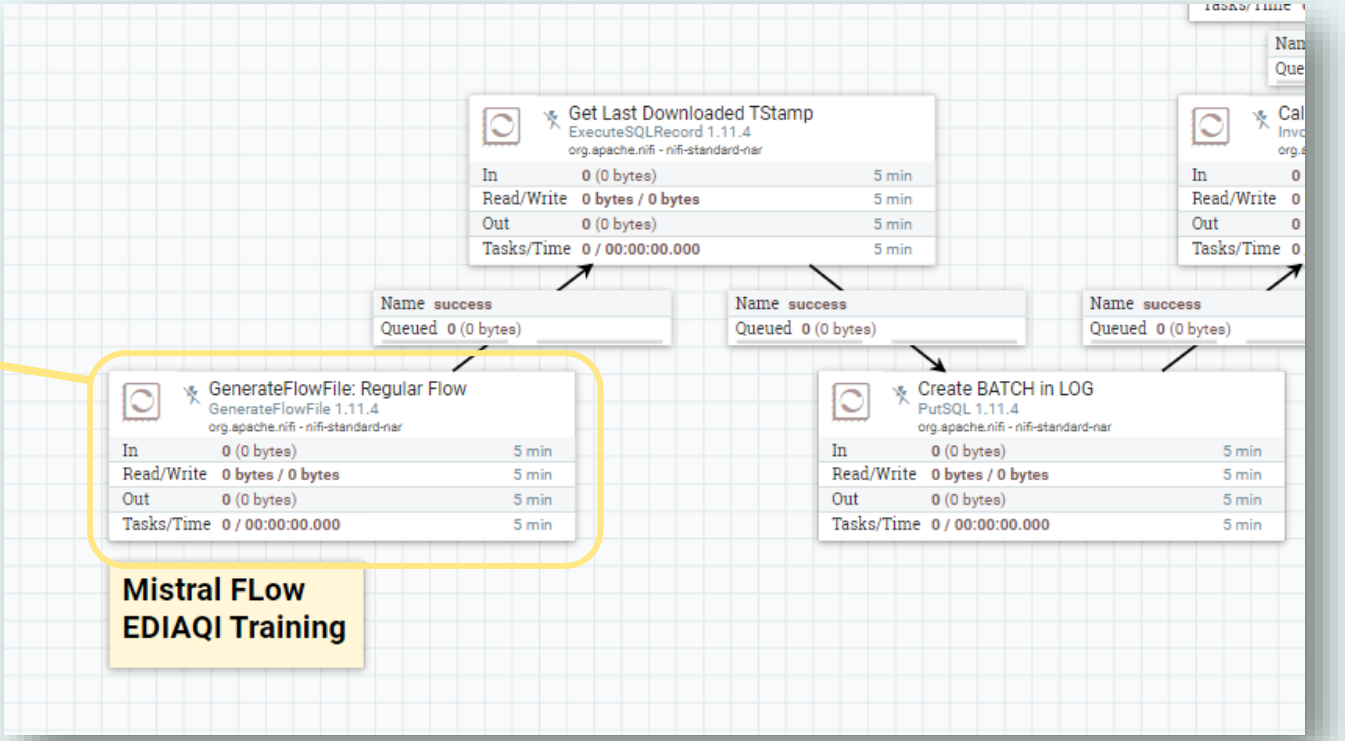
Mistral FLow
EDIAQI Training

deda.next x EDIAQI

Mistral: NiFi Flow

Initial processor

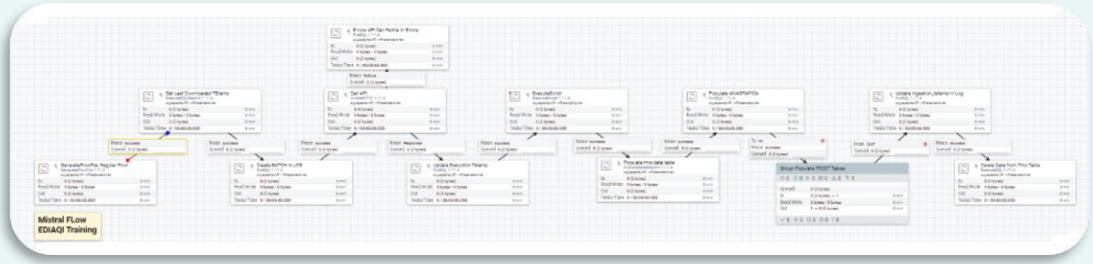
1st of January



deda.next x EDIAQI

Mistral: NiFi Flow

In the next rounds,
how do we **keep track** of what day of data
has already been downloaded?



deda.next
x
EDIAQI

Mistral: NiFi Flow

In the next rounds,
how do we **keep track** of what day of data
has already been downloaded?

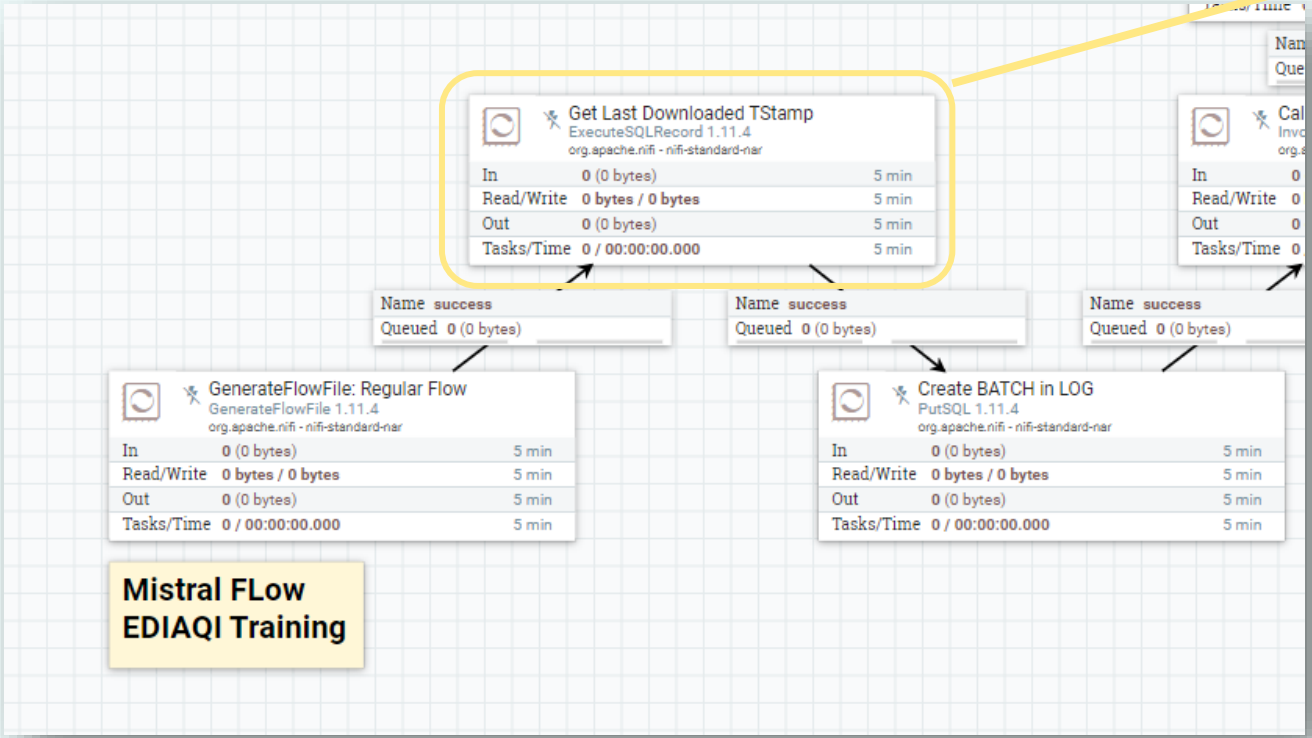
Log table

Batch_ID	Start_Date	Stop_Date	API_Excecution_Tstamp	FROST_Ingestion_Tstamp

deda.next x EDIAQI

Mistral: NiFi Flow

Keeping track of ingested data



Log table

deda.next x EDIAQI

Mistral: NiFi Flow

In the next rounds,
how do we **keep track** of what day of data
has already been downloaded?

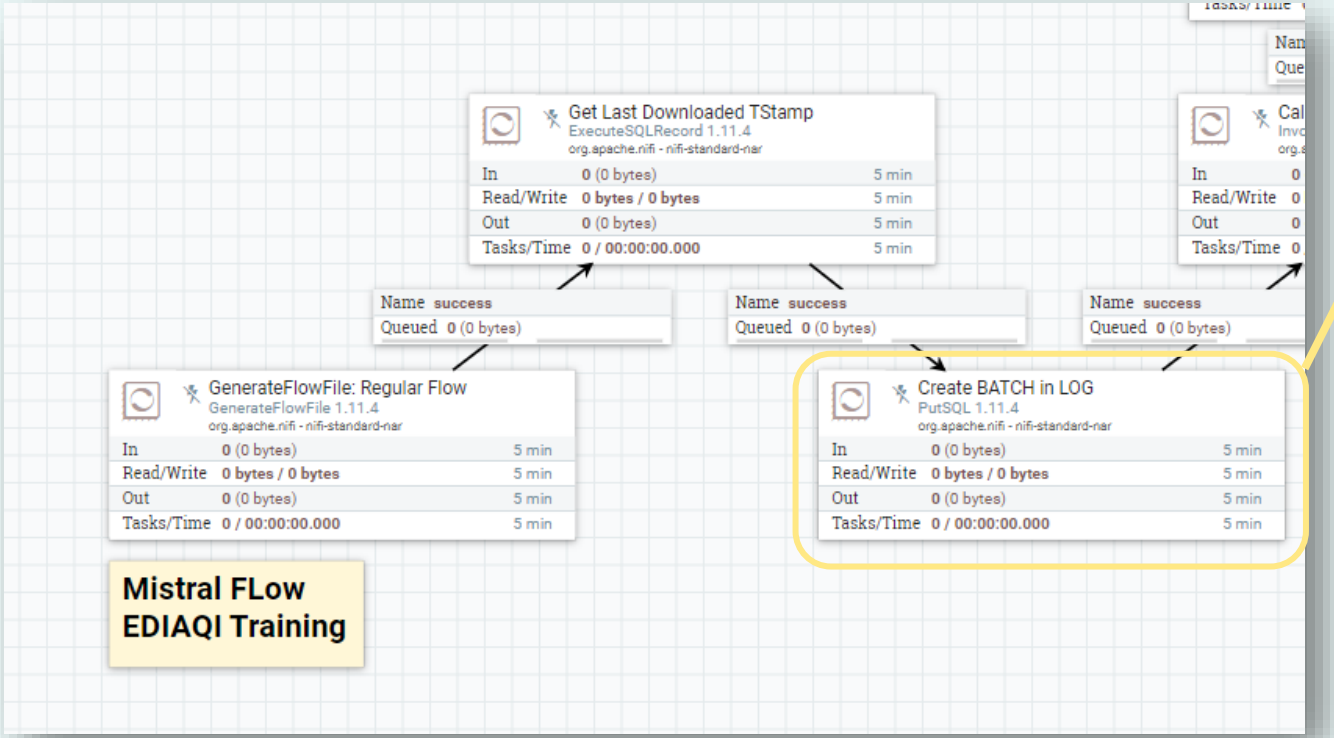
Log table

Batch_ID	Start_Date	Stop_Date	API_Excecution_Tstamp	FROST_Ingestion_Tstamp
1	01/01/2023 00:00	01/01/2023 23:59	07/21/2023 10:00	07/21/2023 10:01

deda.next x EDIAQI

Mistral: NiFi Flow

How do we keep track of what data have already been downloaded?



Log table

deda.next x EDIAQI

Mistral: NiFi Flow

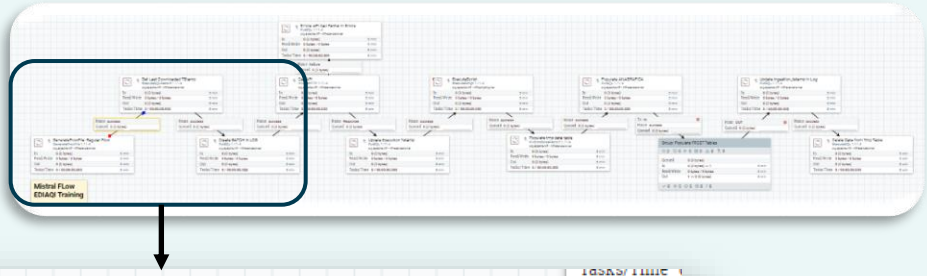
In the next rounds,
how do we **keep track** of what day of data
has already been downloaded?

Log table

Batch_ID	Start_Date	Stop_Date	API_Excecution_Tstamp	FROST_Ingestion_Tstamp
1	01/01/2023 00:00	01/01/2023 23:59	07/21/2023 10:00	07/21/2023 10:01
2	02/01/2023 00:00	02/01/2023 23:59		

deda.next x EDIAQI

Mistral: NiFi Flow



Get Last Downloaded TStamp
ExecuteSQLRecord 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

GenerateFlowFile: Regular Flow
GenerateFlowFile 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

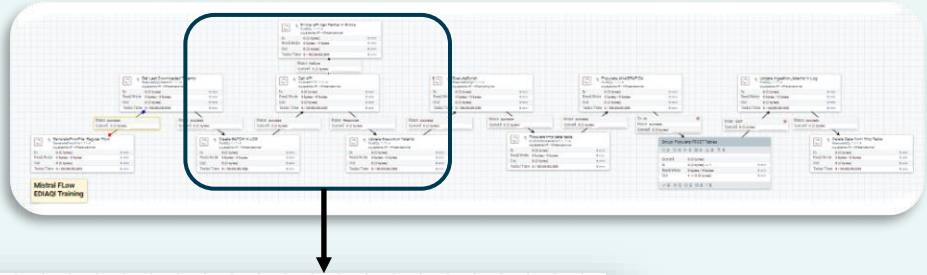
Create BATCH in LOG
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Mistral FLow
EDIAQI Training

deda.next x EDIAQI

Mistral: NiFi Flow



Errore 'API Call Fallita' in Errors
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name Failure
Queued 0 (0 bytes)

Call API
InvokeHTTP 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

Name Response
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

Create BATCH in LOG
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Update Execution Tstamp
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

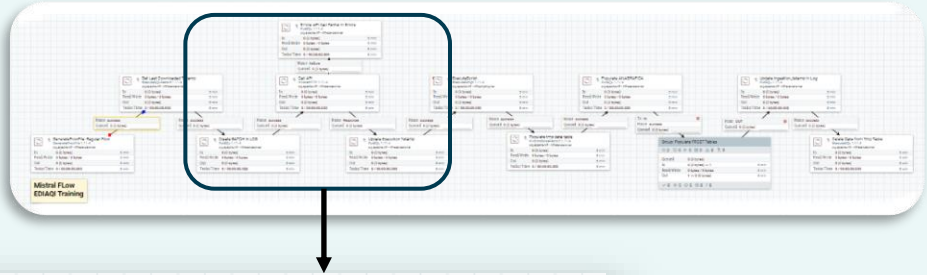
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Endpoint
+ start date
+ stop date
+ variable to measure

JSON

deda.next x EDIAQI

Mistral: NiFi Flow



Errore 'API Call Fallita' in Errors
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name Failure
Queued 0 (0 bytes)

Call API
InvokeHTTP 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

mp
5 min
5 min
5 min
5 min

success
0 (0 bytes)

Name success
Queued 0 (0 bytes)

Name Response
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

Create BATCH in LOG
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Update Execution Tstamp
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Log table

deda.next x EDIAQI

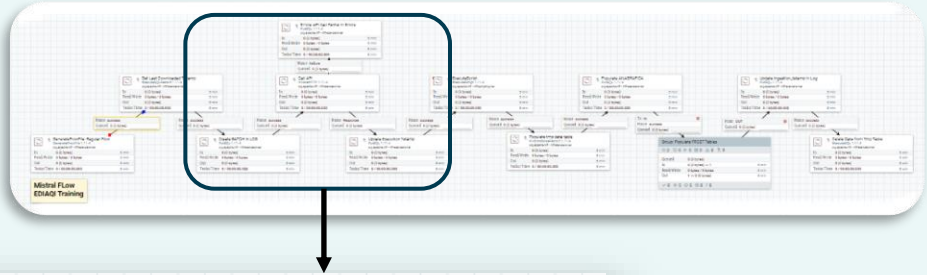
Mistral: NiFi Flow

Log table

Batch_ID	Start_Date	Stop_Date	API_Excecution_Tstamp	FROST_Ingestion_Tstamp
1	01/01/2023 00:00	01/01/2023 23:59	07/21/2023 10:00	07/21/2023 10:01
2	02/01/2023 00:00	02/01/2023 23:59	07/21/2023 10:30	

deda.next x EDIAQI

Mistral: NiFi Flow



Errore 'API Call Fallita' in Errors
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name Failure
Queued 0 (0 bytes)

Call API
InvokeHTTP 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

mp
5 min
5 min
5 min
5 min

success
0 (0 bytes)

Name success
Queued 0 (0 bytes)

Name Response
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

Create BATCH in LOG
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Update Execution Tstamp
PutSQL 1.11.4
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

How do we deal with errors?

Error table

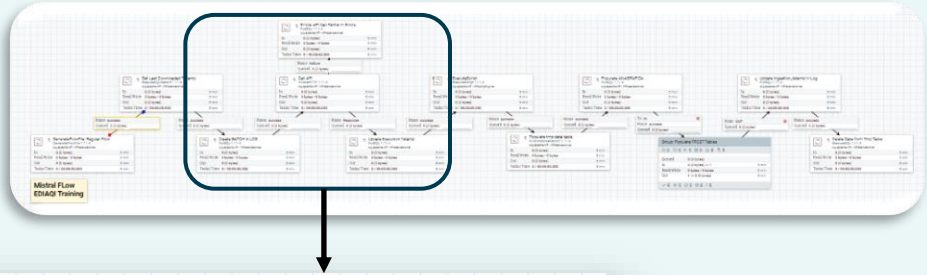
deda.next x EDIAQI

Mistral: NiFi Flow

How do you actually go from JSON data to FROST data?

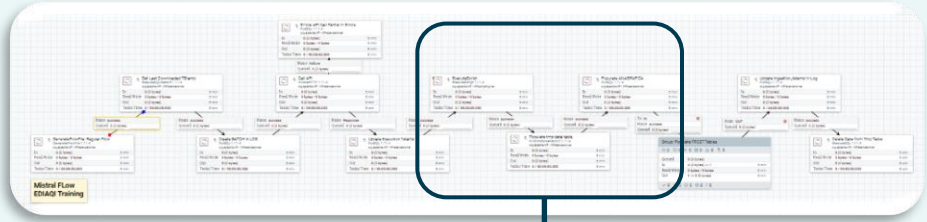
deda.next x EDIAQI

Mistral: NiFi Flow

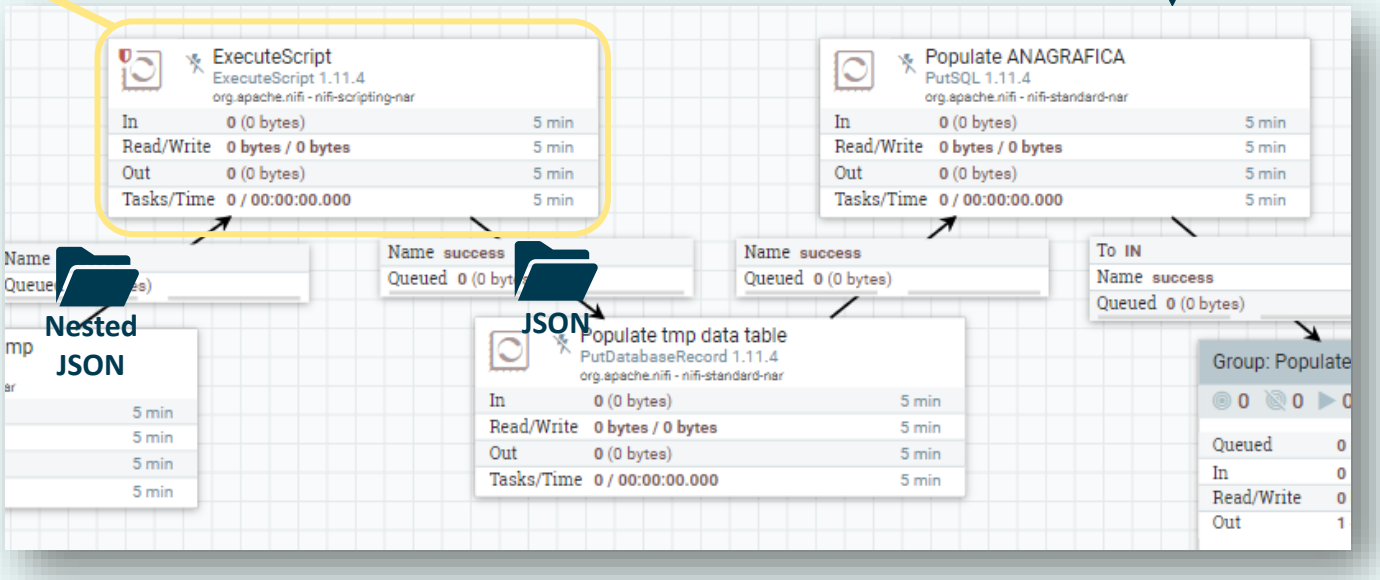


deda.next x EDIAQI

Mistral: NiFi Flow

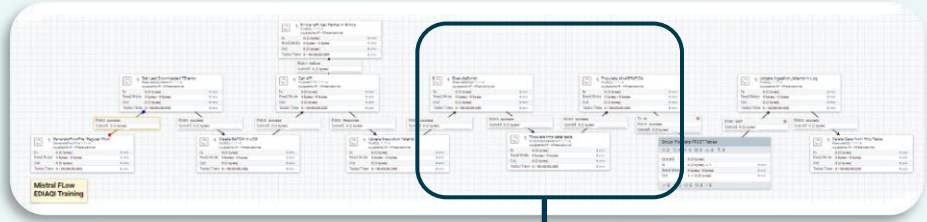


Jython scritp

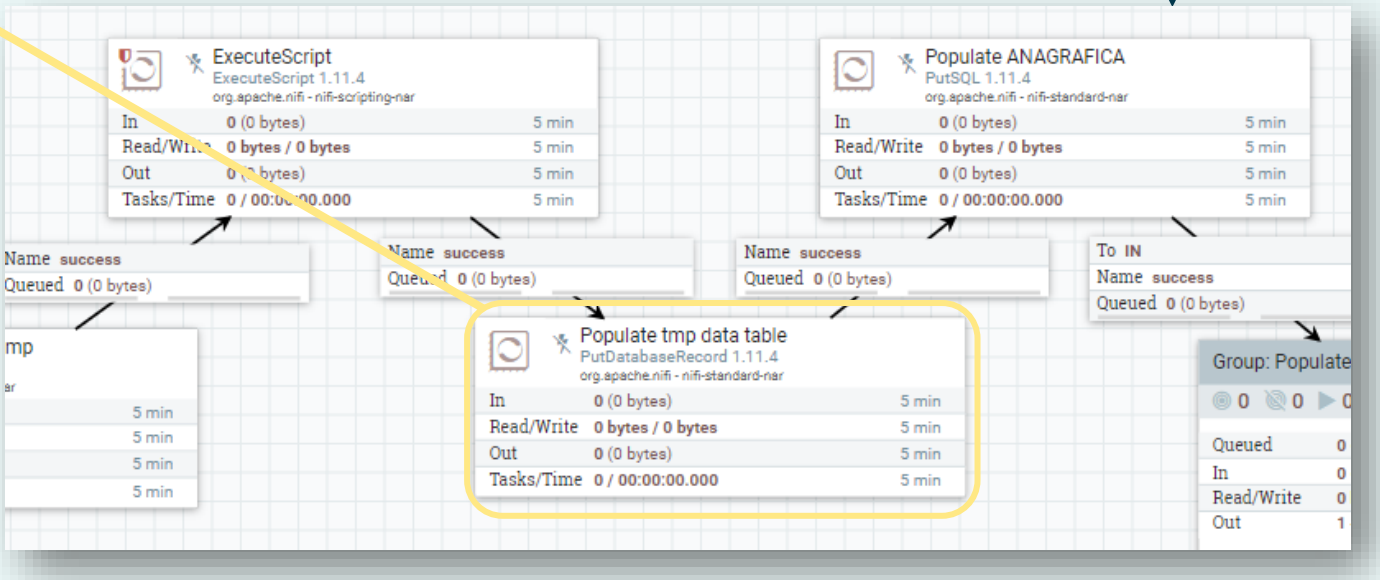


deda.next x EDIAQI

Mistral: NiFi Flow



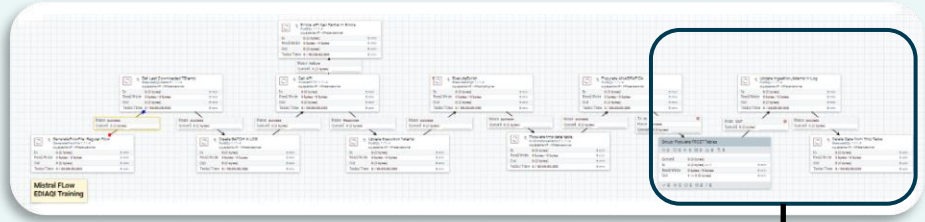
Support DB table to temporarily store the data

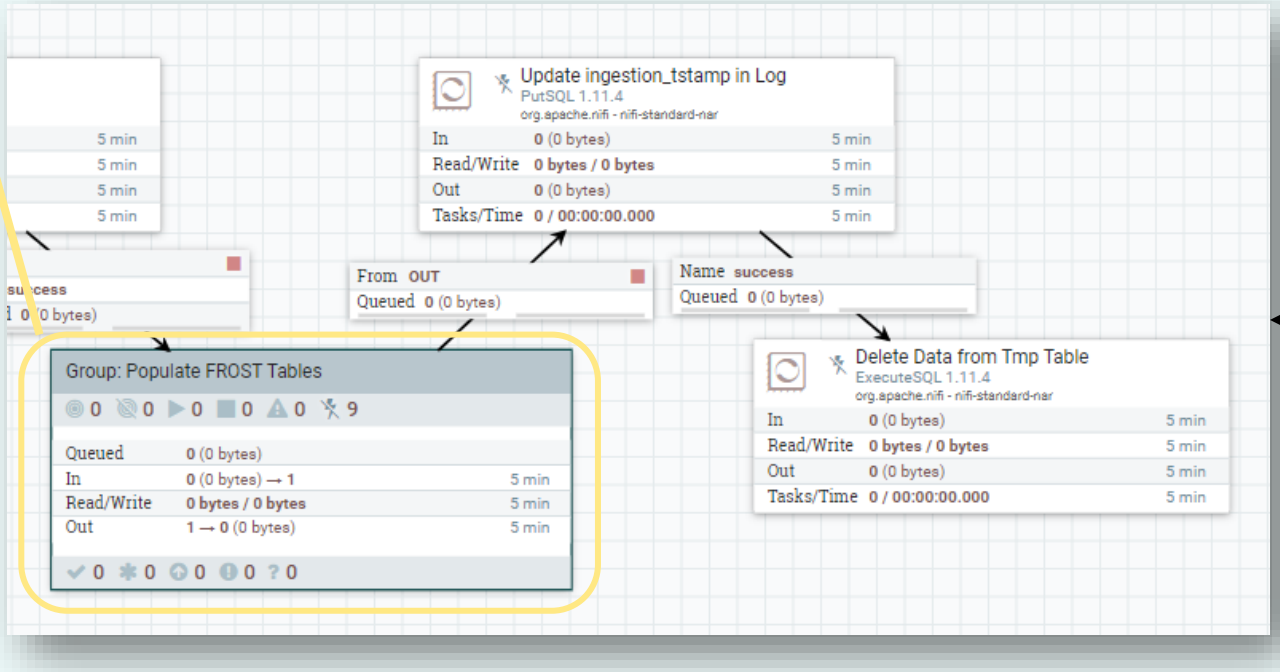


deda.next x EDIAQI

Mistral: NiFi Flow

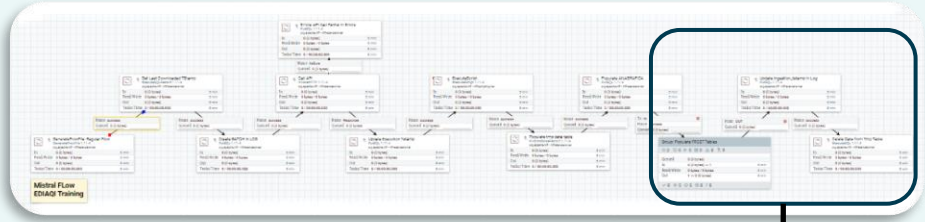
SQL queries to populate FROST tables



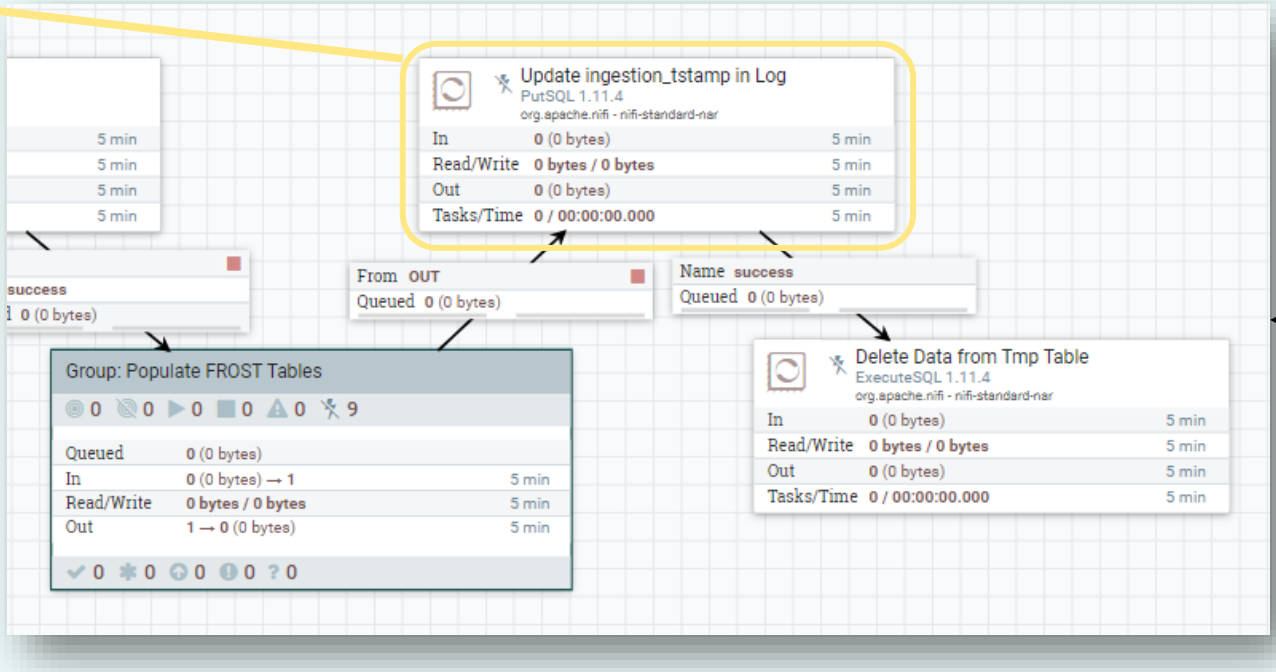


deda.next x EDIAQI

Mistral: NiFi Flow



Log table



deda.next x EDIAQI

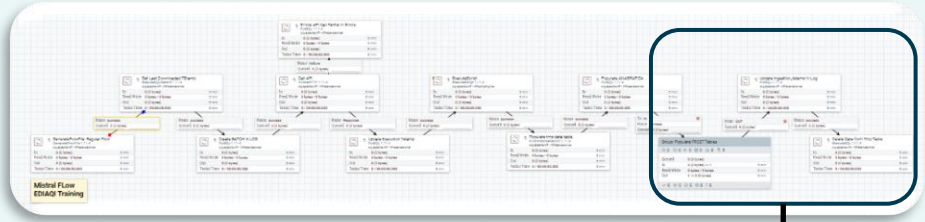
Mistral: NiFi Flow

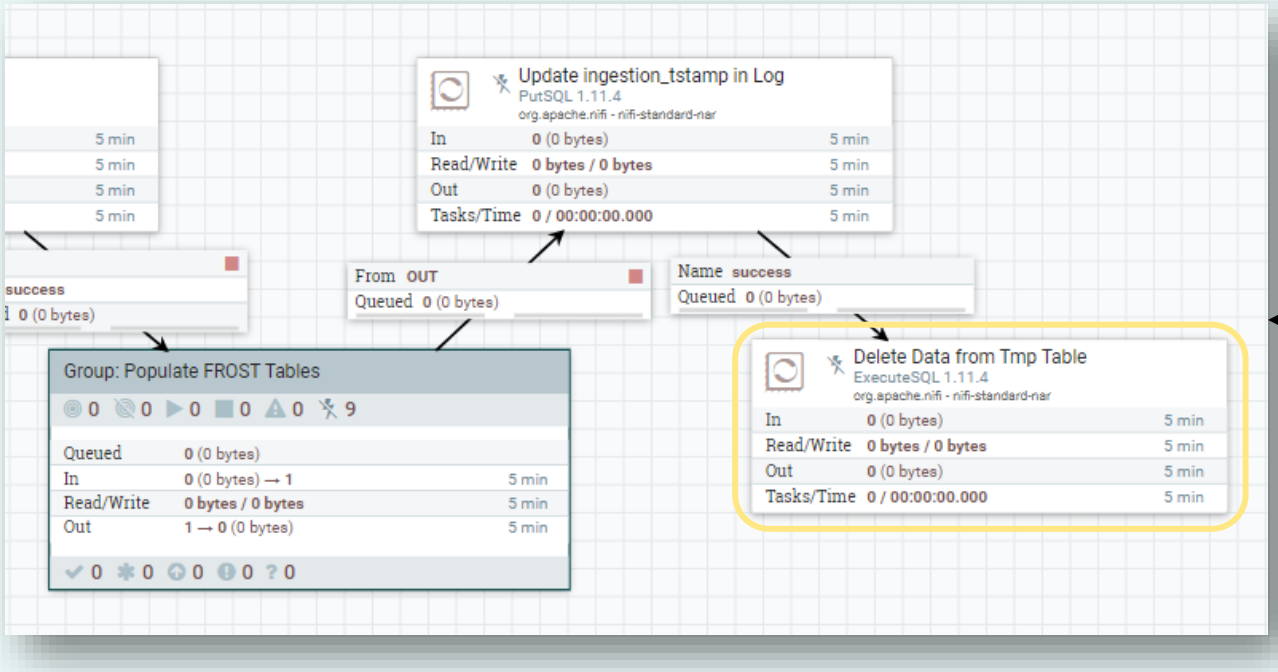
Log table

Batch_ID	Start_Date	Stop_Date	API_Excecution_Tstamp	FROST_Ingestion_Tstamp
1	01/01/2023 00:00	01/01/2023 23:59	07/21/2023 10:00	07/21/2023 10:01
2	02/01/2023 00:00	02/01/2023 23:59	07/21/2023 10:30	07/21/2023 10:32

deda.next x EDIAQI

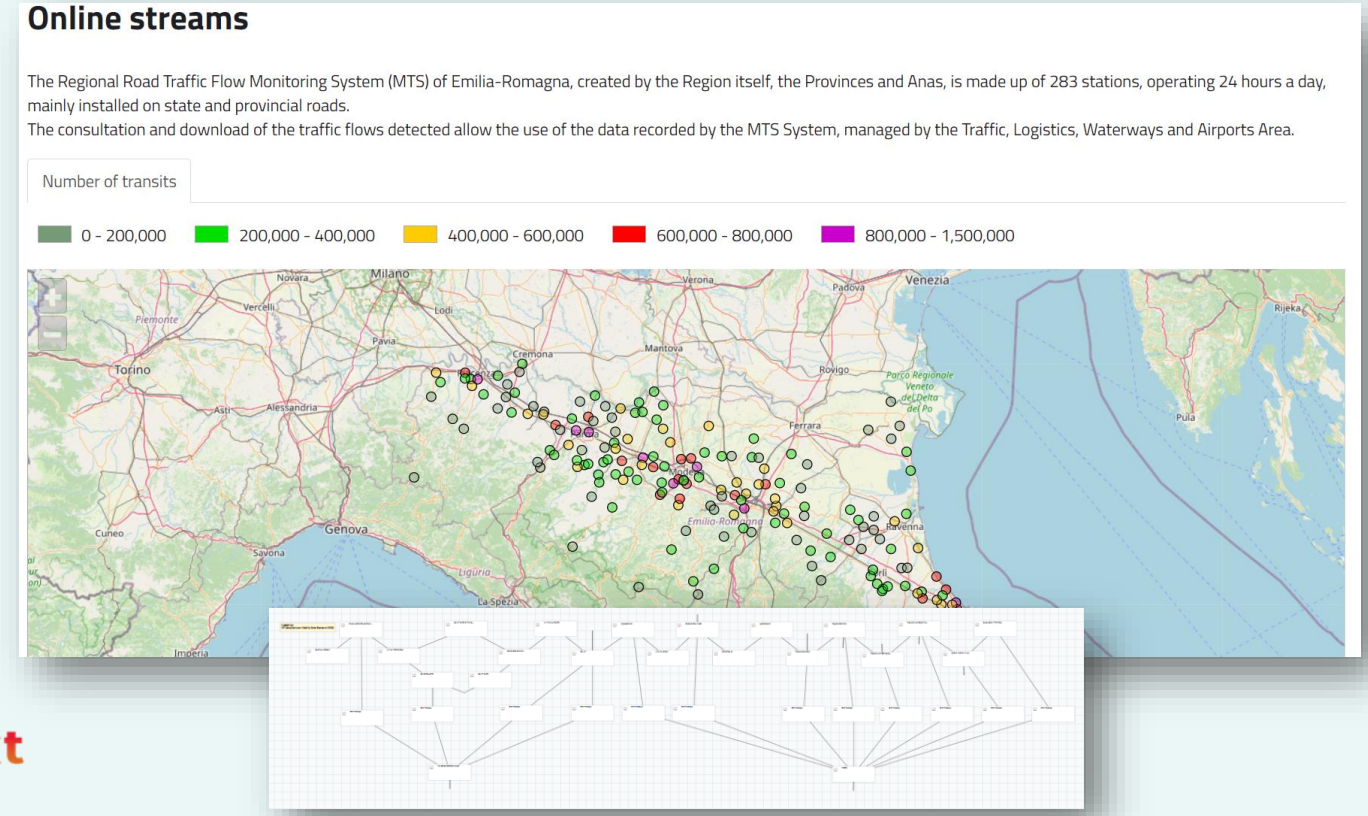
Mistral: NiFi Flow





deda.next x EDIAQI

Other examples: road traffic data



deda.next x EDIAQI

https://servizissiir.regione.emilia-romagna.it/FlussiMTS/

Other examples: outdoor AQ data

https://airbreakferrara.net/che-aria-tira/