COVID-19 CLOUD DATA WAREHOUSE AND ANALYTICS

Team 1:

Kunal Jaiswal, Sharvari Karnik, Keval Vankudre

Project Dashboard URL: https://bit.ly/CovidDataIntegration

Objectives:

Create a Data Warehouse by integrating COVID public datasets from different sources and design dashboards to get insights.

- Perform source systems analysis on data sources (& compare)
- Ingest & Integrate from/to various sources to generate BI Dashboards
- Gain experience with cloud databases & cloud DW
- Gain experience with data integration & data pipeline tools using data from both cloud and on
- premise (notebook) sources

Deliverables:

- Dashboards visualizing COVID data tracking US states and counties
    - Microsoft PowerBI, Tableau or Google Data Studio
- Data should be loaded into:
    - Google BigQuery
    - Azure SQL (note: likely limitations on student credits)
    - SQL Server or MySQL (as a fallback if team exhausts their free student credits)
- Review & compare data integration tools

Tools Used:

- Talend Data Quality Open Studio - For Data Profiling
- Stitch - Used to load COVID public dataset and flat files to Big Query
- Talend Pipeline Designer - For creating staging area and for Data Integration
- Talend Big Data - For Data Integration

Cloud Technologies Used:

- Google Big Query
- Azure SQL

BI tool used:

- Power BI

List of COVID dataset sources used:

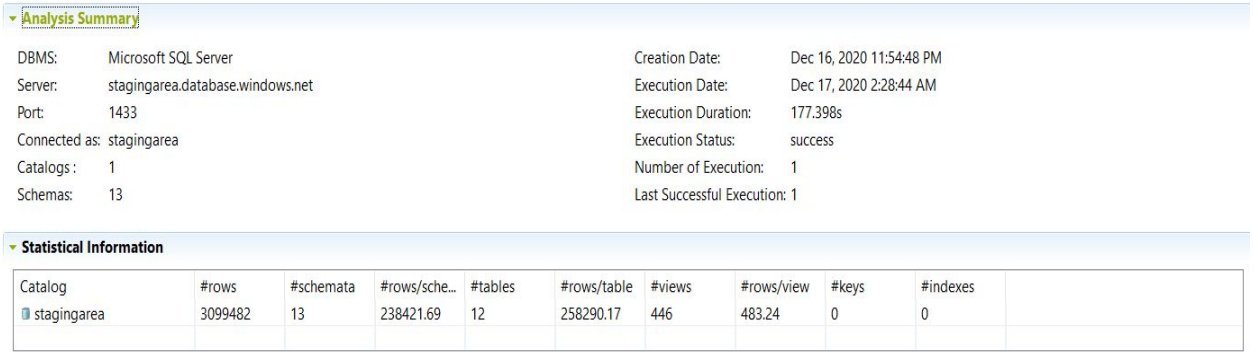
- Open Data - By state and county
- New York Times Data - By state and county
- The COVID Tracking Project - By state
- Oxford Policy Tracker - By State
- American Heart Association (AHA) - By county
- Flat files (from the open source github) - Demographics

Source System Analysis :

Source system analysis has been performed in Talend Data Profiling. Below snippets give an overview about the type of data that we have on our hands.

Connection Overview Analysis:

This analysis returns an overview of the content of your database. It computes the number of tables and the number of rows per table for each catalog and/or schema, etc. It also counts the number of indexes and primary keys, etc.



From the above analysis we get a gist of data that we have our hands.

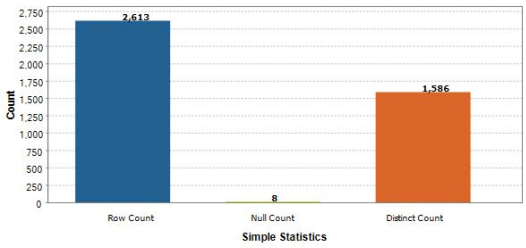
Column Analysis:

This type helps to analyze and manually assign the indicators on each column, such as number of nulls, frequency table, summary statistics, pattern matching indicators, etc.

Column: aha_hospital_beds.county_name

Simple Statistics

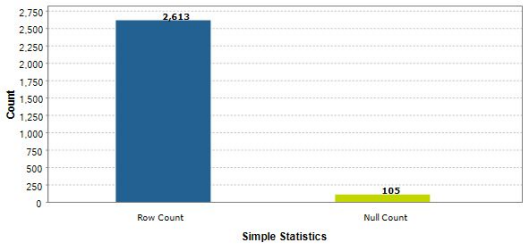
Label	Count	%
Row Count	2613	100.00%
Null Count	8	0.31%
Distinct Count	1586	60.70%



Column: aha_hospital_beds.total_hospital_beds

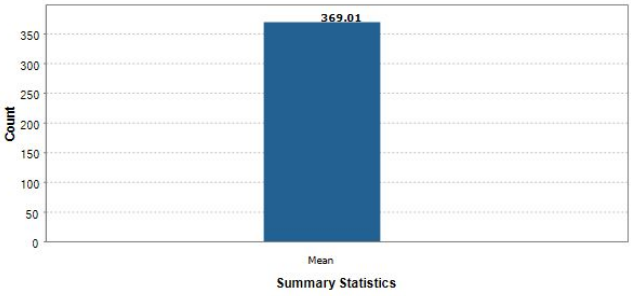
Simple Statistics

Label	Count	%
Row Count	2613	100.00%
Null Count	105	4.02%



Summary Statistics

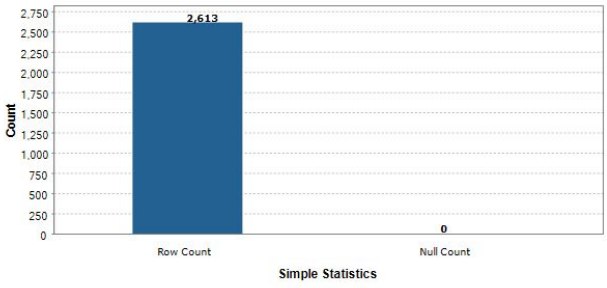
Label	Value
Mean	369.00518341307816



Column: aha_hospital_beds.state_name

Simple Statistics

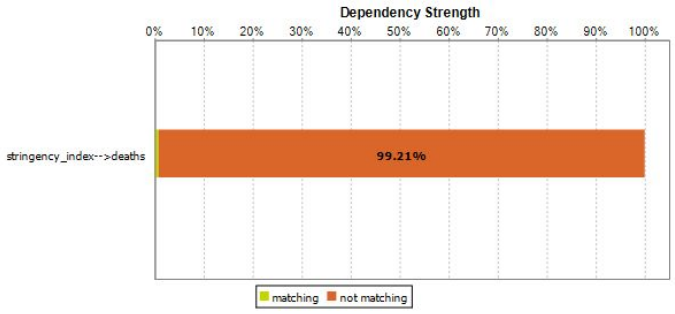
Label	Count	%
Row Count	2613	100.00%
Null Count	0	0.00%



Functional Dependency Analysis:

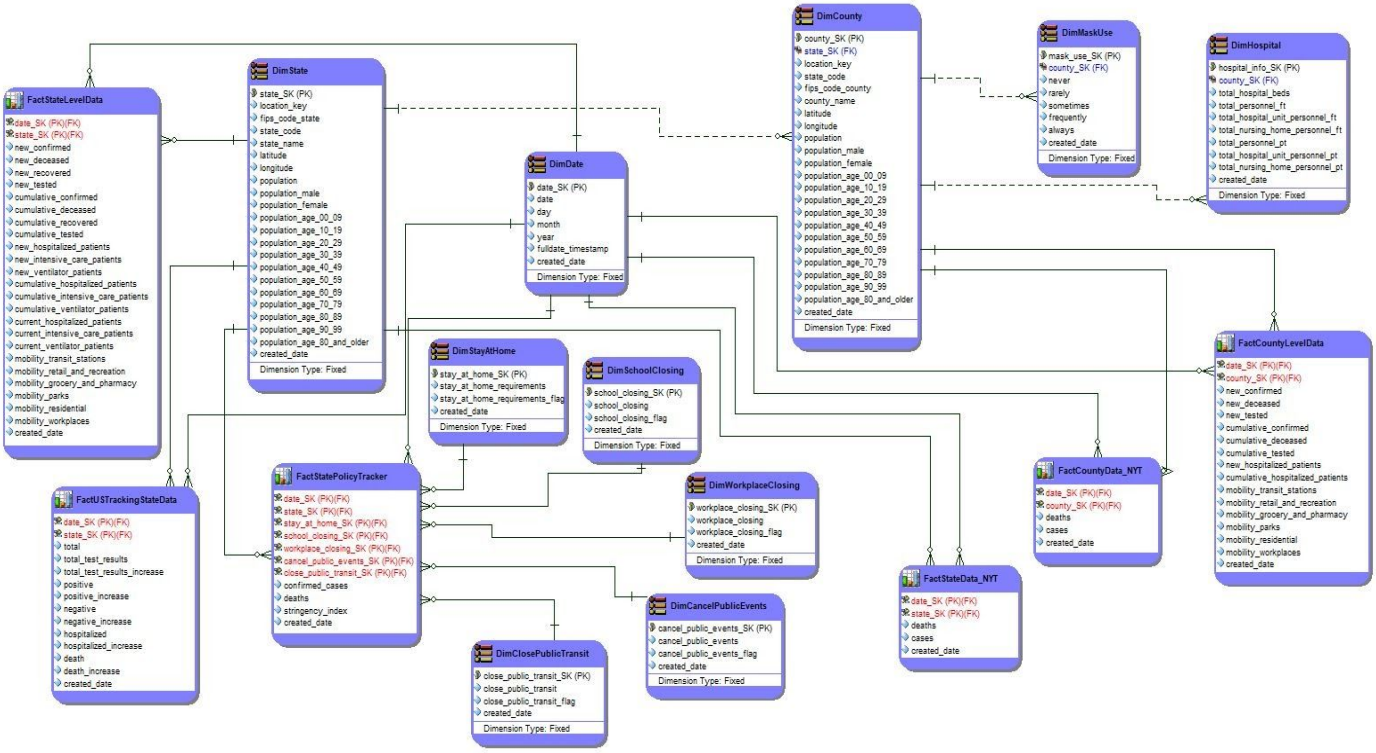
This analysis enables you to detect anomalies in your column dependencies. It determines to which extent the values of a column A determines the values of a column B. This relationship is noted as A->B.

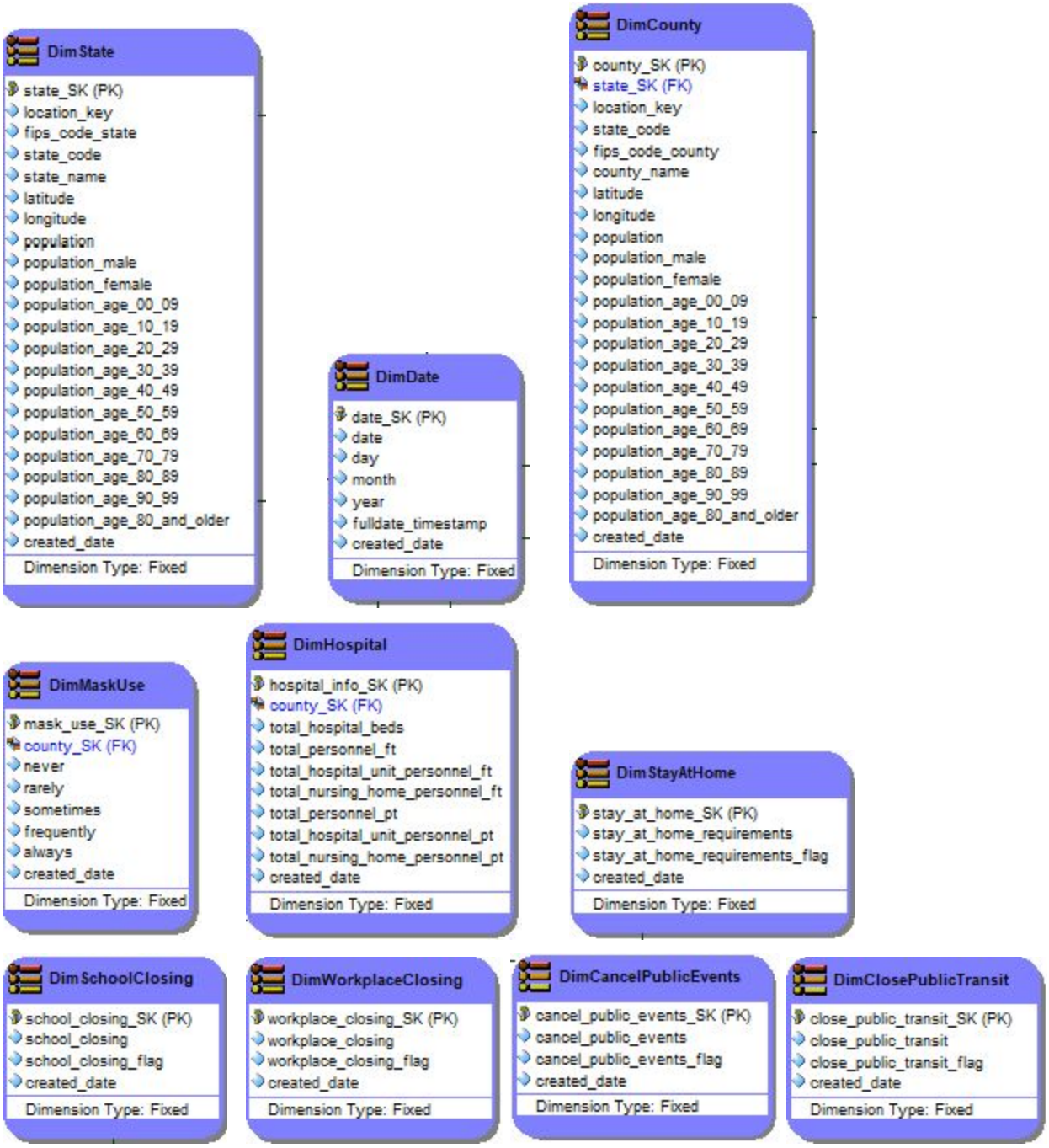
Dependency	#Match	%Match	#row
stringency_index-->deaths	85	0.79%	10791

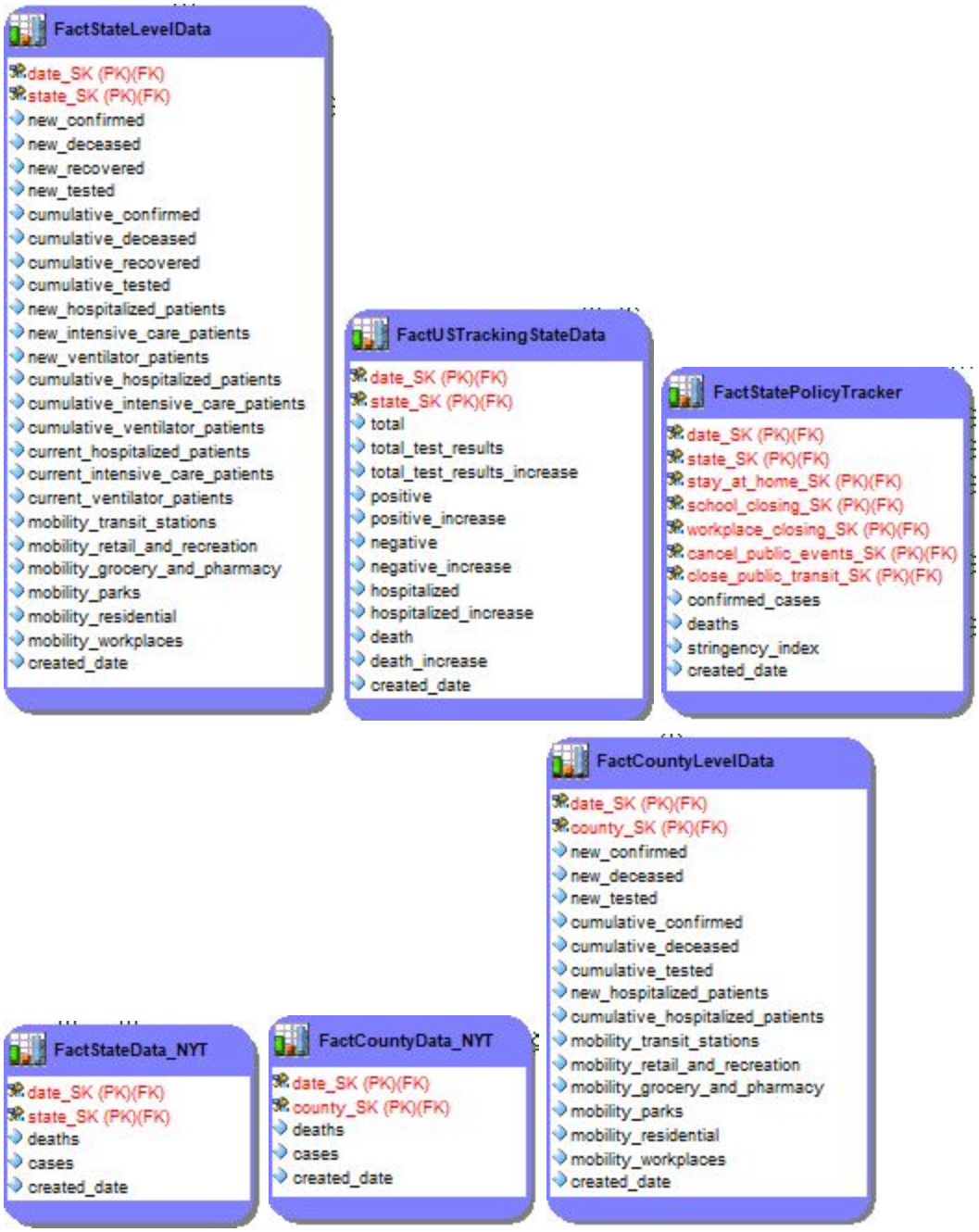


From this we can say that, stringency_index determines only 0.79% of deaths. From this we can also infer that to determine the deaths more detailed analysis apart from stringency_index is needed.

Data Model




DimState
state_SK (PK)
location_key
fips_code_state
state_code
state_name
latitude
longitude
population
population_male
population_female
population_age_00_09
population_age_10_19
population_age_20_29
population_age_30_39
population_age_40_49
population_age_50_59
population_age_60_69
population_age_70_79
population_age_80_89
population_age_90_99
population_age_80_and_older
created_date
Dimension Type: Fixed
DimDate
date_SK (PK)
date
day
month
year
fulldate_timestamp
created_date
Dimension Type: Fixed
DimCounty
county_SK (PK)
state_SK (FK)
location_key
state_code
fips_code_county
county_name
latitude
longitude
population
population_male
population_female
population_age_00_09
population_age_10_19
population_age_20_29
population_age_30_39
population_age_40_49
population_age_50_59
population_age_60_69
population_age_70_79
population_age_80_89
population_age_90_99
population_age_80_and_older
created_date
Dimension Type: Fixed
DimMaskUse
mask_use_SK (PK)
county_SK (FK)
never
rarely
sometimes
frequently
always
created_date
Dimension Type: Fixed
DimHospital
hospital_info_SK (PK)
county_SK (FK)
total_hospital_beds
total_personnel_ft
total_hospital_unit_personnel_ft
total_nursing_home_personnel_ft
total_personnel_pt
total_hospital_unit_personnel_pt
total_nursing_home_personnel_pt
created_date
Dimension Type: Fixed
DimStayAtHome
stay_at_home_SK (PK)
stay_at_home_requirements
stay_at_home_requirements_flag
created_date
Dimension Type: Fixed
DimSchoolClosing
school_closing_SK (PK)
school_closing
school_closing_flag
created_date
Dimension Type: Fixed
DimWorkplaceClosing
workplace_closing_SK (PK)
workplace_closing
workplace_closing_flag
created_date
Dimension Type: Fixed
DimCancelPublicEvents
cancel_public_events_SK (PK)
cancel_public_events
cancel_public_events_flag
created_date
Dimension Type: Fixed
DimClosePublicTransit
close_public_transit_SK (PK)
close_public_transit
close_public_transit_flag
created_date
Dimension Type: Fixed



Review and compare data integration tools:

Talend Big Data

- Significantly easier to create data pipelines in Talend Big Data than Talend pipeline designer
- Easy to access database and create connections

- Can change and compare data types easily
- Can perform join operations easily
- Easy to understand and interpret the errors
- Takes a bit longer to process, integrate and ingest the data in the data warehouse than Talend Pipeline Designer

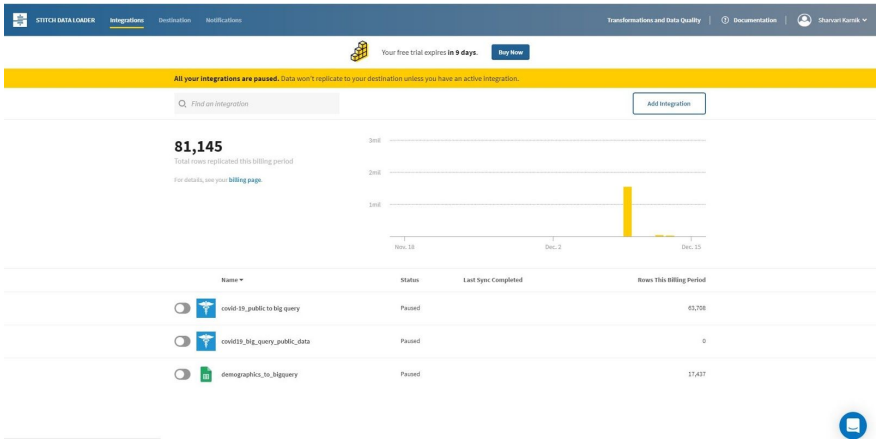
Stitch

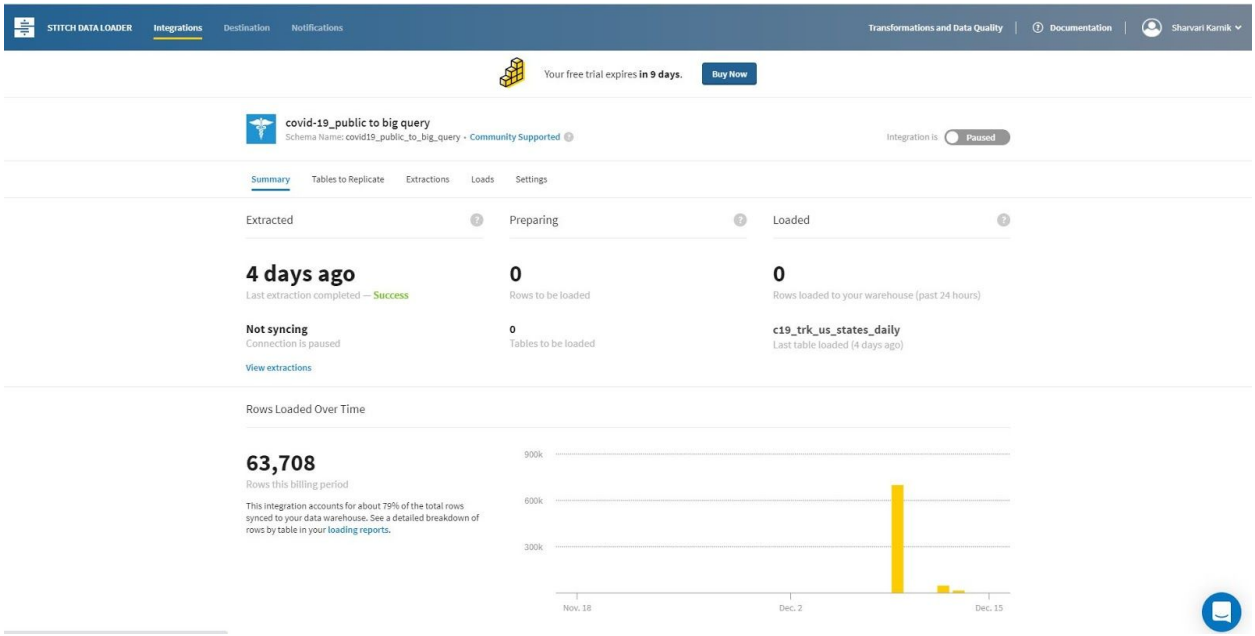
- Easy to use
- Comes with plethora of connectors for building connections with different sources which helped in seamless data ETL
- Fast and efficient

Talend Pipeline Designer

- Easy drag and drop ETL tool
- Can use different processor to play with the data
- Can preview the input data and output data after using a processor
- Takes considerably less time to process, integrate and ingest the data in the data warehouse than Talend Big Data
- Difficult to change the data types of the attributes and perform join operations with other source
- It would be better to access a particular data source if it had a proper dataset organization
- No truncation option before running a pipeline
- A bit difficult to understand the errors
- Might cause timeout errors if the data is too large to process

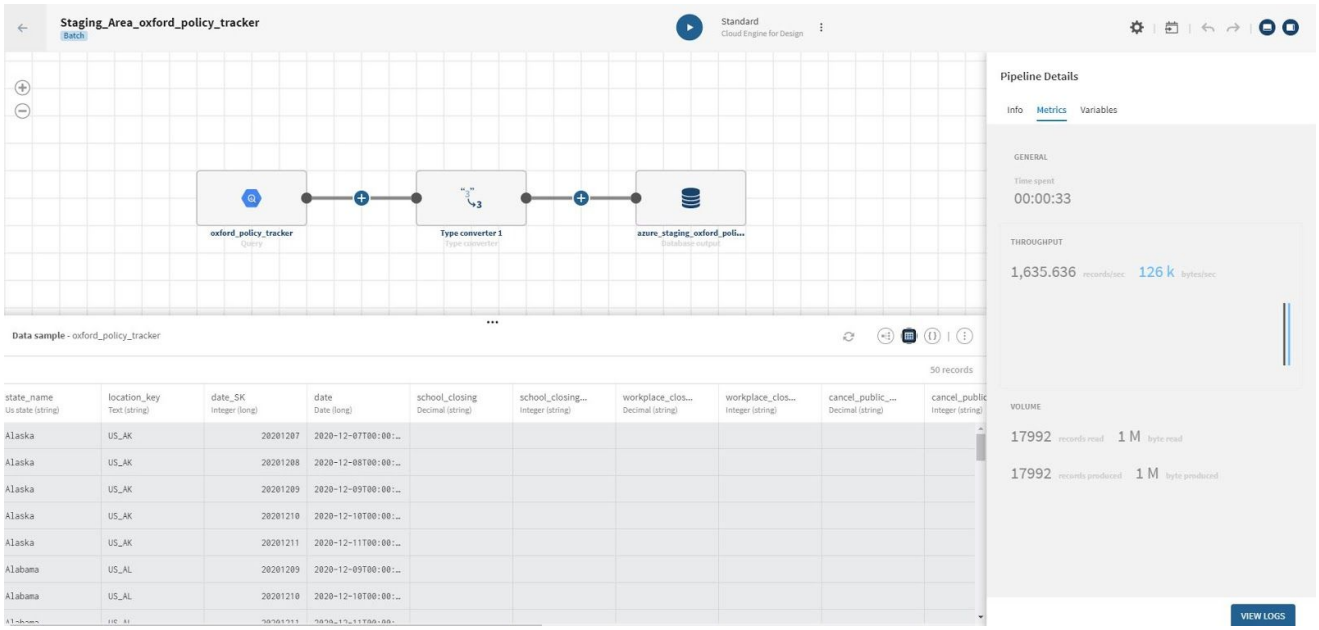
Screenshots:

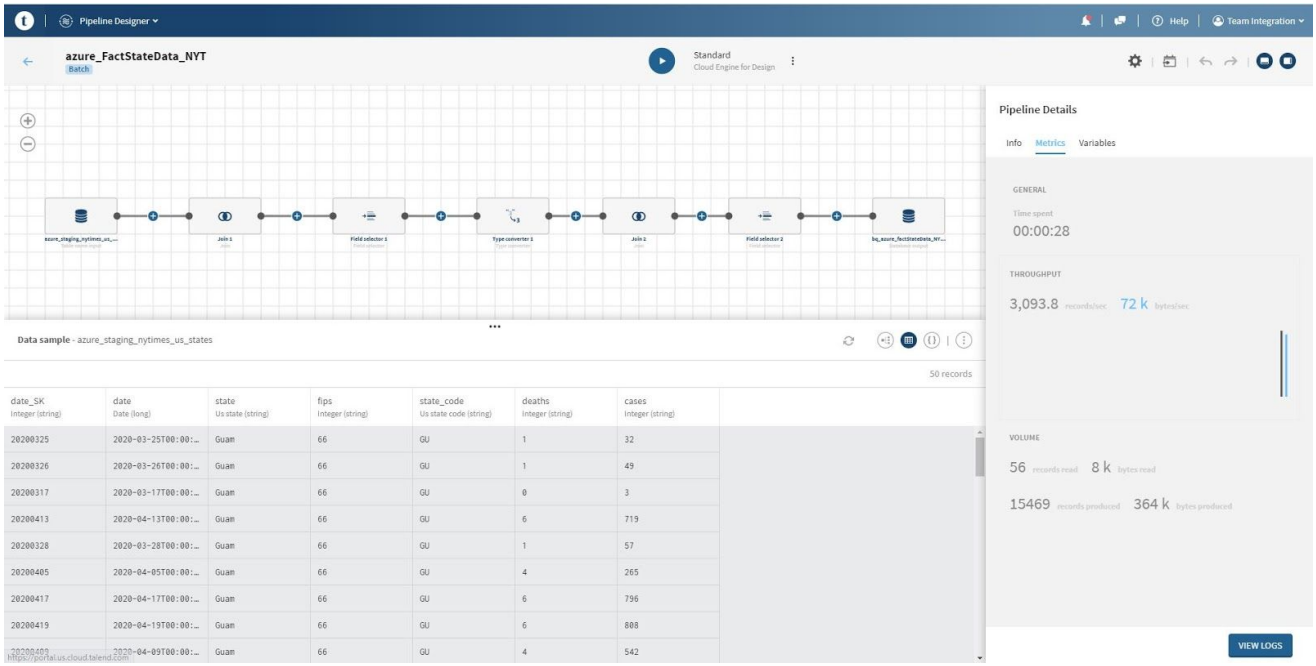


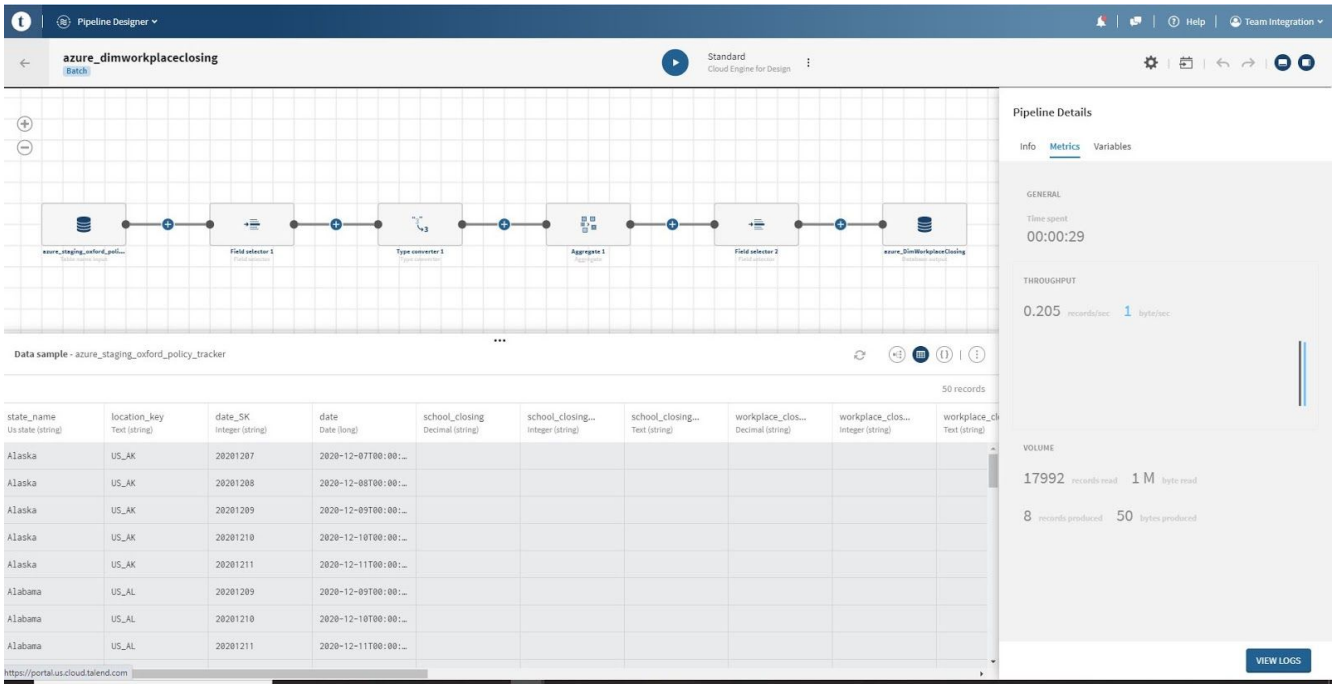


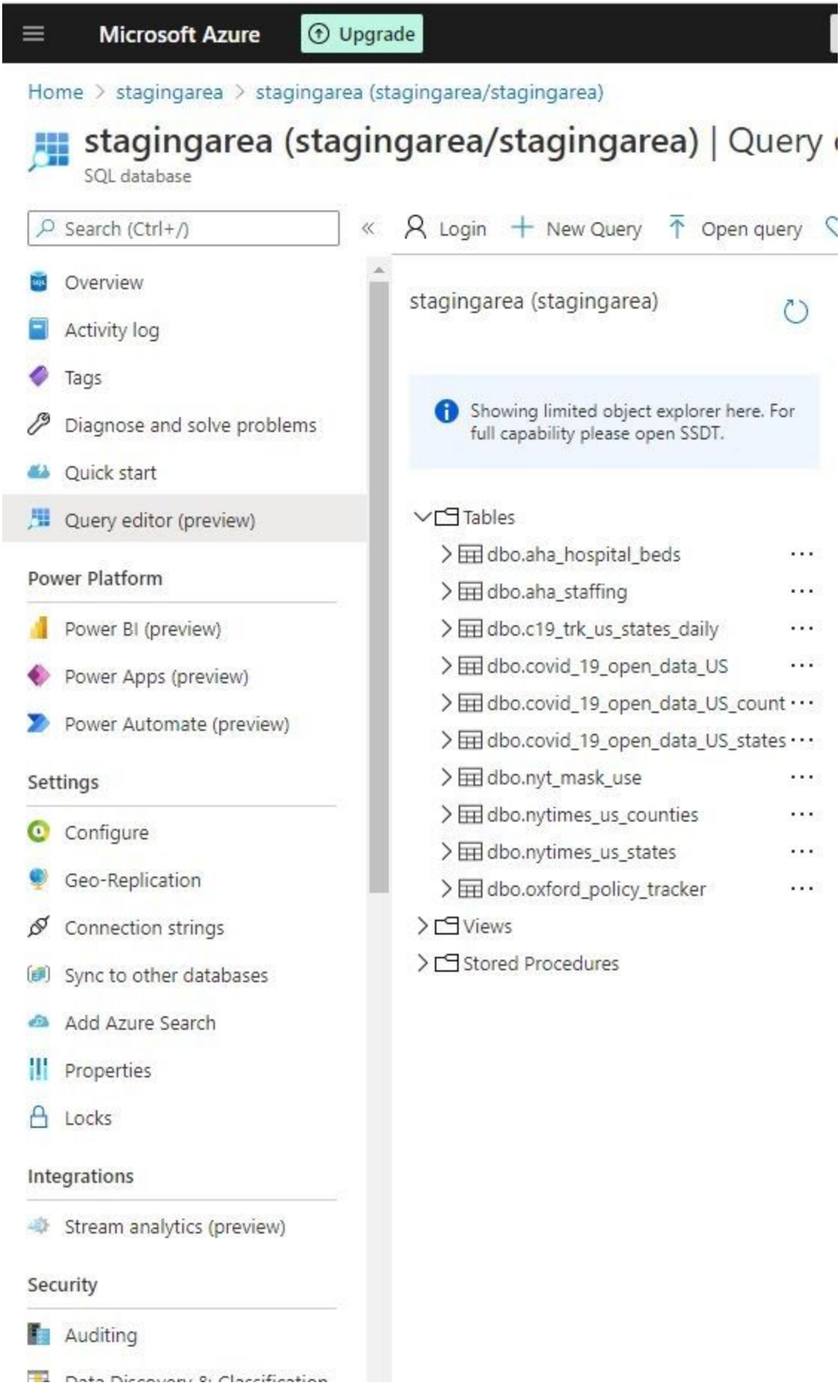
azure_dimcancelpublicevents		Batch	Successful	about 2 hours ago
azure_dimclosepublictransit		Batch	Successful	about 2 hours ago
azure_dimcounty		Batch	Successful	about 20 hours ago
azure_dimDate		Batch	Successful	2 minutes ago
azure_dimschoolclosing		Batch	Successful	about 2 hours ago
azure_dimstate		Batch	Successful	about 20 hours ago
azure_dimstayathome		Batch	Successful	about 2 hours ago
azure_dimworkplaceclosing		Batch	Successful	about 2 hours ago
azure_FactStateData_NYT		Batch	Successful	about 18 hours ago
azure_maskuse		Batch	Successful	about 19 hours ago

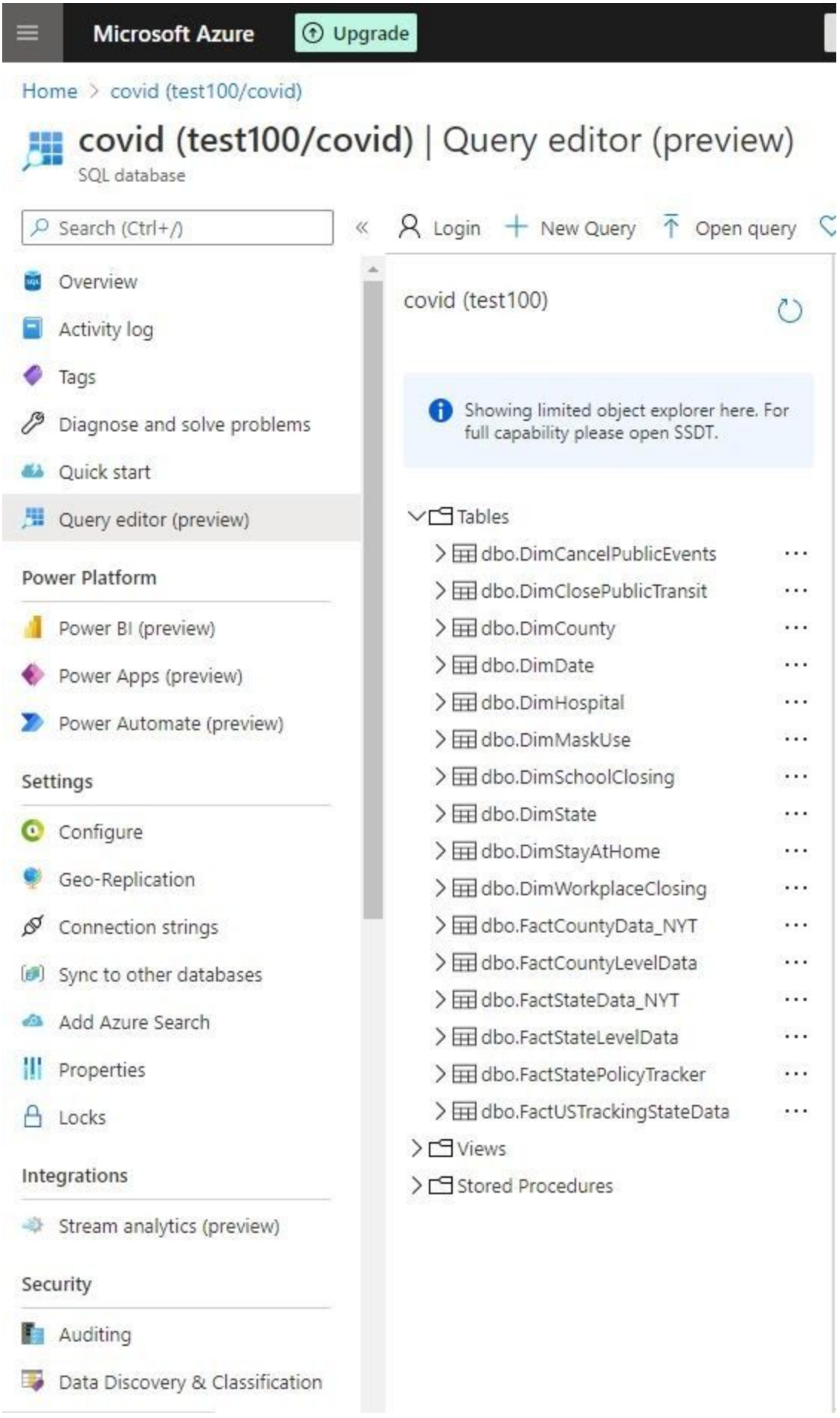
Staging_Area_nyt_mask_use		Batch	Successful	about 19 hours ago
Staging_Area_open_data_US_county		Batch	Successful	about 19 hours ago
Staging_Area_open_data_US_states		Batch	Successful	about 19 hours ago
Staging_Area_oxford_policy_tracker		Batch	Successful	about 12 hours ago
Staging_Area_trk_us_states_daily		Batch	Successful	about 22 hours ago
StagingArea_hospital_beds		Batch	Successful	about 23 hours ago
StagingArea_nytimes_us_counties		Batch	Successful	about 23 hours ago
StagingArea_nytimes_us_states		Batch	Successful	about 23 hours ago
StagingArea_Open_data_US		Batch	Successful	about 23 hours ago
StagingArea_staffing		Batch	Successful	about 23 hours ago


Staging_Area_oxford_policy_tracker
Batch
Standard
Cloud Engine for Design
oxford_policy_tracker
Query
Type converter 1
Type converter
azure_staging_oxford_poli...
Database output
Pipeline Details
Info
Metrics
Variables
GENERAL
Time spent
00:00:33
THROUGHPUT
1,635.636 records/sec
126 k bytes/sec
VOLUME
17992 records read
1 M byte read
17992 records produced
1 M byte produced
VIEW LOGS
Data sample - oxford_policy_tracker
50 records
state_name
Us state (string)
location_key
Text (string)
date_SK
Integer (long)
date
Date (long)
school_closing
Decimal (string)
school_closing...
Integer (string)
workplace_clos...
Decimal (string)
workplace_clos...
Integer (string)
cancel_public_...
Decimal (string)
cancel_public...
Integer (string)
Alaska US_AK 20201207 2020-12-07T00:00:...
Alaska US_AK 20201208 2020-12-08T00:00:...
Alaska US_AK 20201209 2020-12-09T00:00:...
Alaska US_AK 20201210 2020-12-10T00:00:...
Alaska US_AK 20201211 2020-12-11T00:00:...
Alabama US_AL 20201209 2020-12-09T00:00:...
Alabama US_AL 20201210 2020-12-10T00:00:...


Pipeline Designer
Help
Team Integration
azure_FactStateData_NYT
Batch
Standard
Cloud Engine for Design
Pipeline Details
Info
Metrics
Variables
GENERAL
Time spent
00:00:28
THROUGHPUT
3,093.8 records/sec
72 k bytes/sec
VOLUME
56 records read
8 k bytes read
15469 records produced
364 k bytes produced
VIEW LOGS
Data sample - azure_staging_nytimes_us_states
50 records
date_SK
Integer (string)
date
Date (long)
state
Us state (string)
fips
Integer (string)
state_code
Us state code (string)
deaths
Integer (string)
cases
Integer (string)
20200325 2020-03-25T00:00:... Guam 66 GU 1 32
20200326 2020-03-26T00:00:... Guam 66 GU 1 49
20200317 2020-03-17T00:00:... Guam 66 GU 0 3
20200413 2020-04-13T00:00:... Guam 66 GU 6 719
20200328 2020-03-28T00:00:... Guam 66 GU 1 57
20200405 2020-04-05T00:00:... Guam 66 GU 4 265
20200417 2020-04-17T00:00:... Guam 66 GU 6 796
20200419 2020-04-19T00:00:... Guam 66 GU 6 808
20200409 2020-04-09T00:00:... Guam 66 GU 4 542
https://portal.us.cloud.talend.com


Pipeline Designer
Help
Team Integration
azure_dimworkplaceclosing
Batch
Standard
Cloud Engine for Design
Pipeline Details
Info
Metrics
Variables
GENERAL
Time spent
00:00:29
THROUGHPUT
0.205 records/sec
1 byte/sec
VOLUME
17992 records read
1 M byte read
8 records produced
50 bytes produced
VIEW LOGS
azure_staging_oxford_poli...
Field selector 1
Type converter 1
Aggregate 1
Field selector 2
azure_DimWorkplaceClosing
Data sample - azure_staging_oxford_policy_tracker
50 records
state_name
Us state (string)
location_key
Text (string)
date_SK
Integer (string)
date
Date (long)
school_closing
Decimal (string)
school_closing...
Integer (string)
school_closing...
Text (string)
workplace_clos...
Decimal (string)
workplace_clos...
Integer (string)
Alaska US_AK 20201207 2020-12-07T00:00:...
Alaska US_AK 20201208 2020-12-08T00:00:...
Alaska US_AK 20201209 2020-12-09T00:00:...
Alaska US_AK 20201210 2020-12-10T00:00:...
Alaska US_AK 20201211 2020-12-11T00:00:...
Alabama US_AL 20201209 2020-12-09T00:00:...
Alabama US_AL 20201210 2020-12-10T00:00:...
Alabama US_AL 20201211 2020-12-11T00:00:...
https://portal.us.cloud.talend.com


Microsoft Azure
Upgrade
Home > stagingarea > stagingarea (stagingarea/stagingarea)
stagingarea (stagingarea/stagingarea) | Query
SQL database
Search (Ctrl+/)
Login
New Query
Open query
Overview
Activity log
Tags
Diagnose and solve problems
Quick start
Query editor (preview)
Power Platform
Power BI (preview)
Power Apps (preview)
Power Automate (preview)
Settings
Configure
Geo-Replication
Connection strings
Sync to other databases
Add Azure Search
Properties
Locks
Integrations
Stream analytics (preview)
Security
Auditing
stagingarea (stagingarea)
Showing limited object explorer here. For full capability please open SSDT.
Tables
dbo.aha_hospital_beds
dbo.aha_staffing
dbo.c19_trk_us_states_daily
dbo.covid_19_open_data_US
dbo.covid_19_open_data_US_count
dbo.covid_19_open_data_US_states
dbo.nyt_mask_use
dbo.nytimes_us_counties
dbo.nytimes_us_states
dbo.oxford_policy_tracker
Views
Stored Procedures


Microsoft Azure
Upgrade
Home > covid (test100/covid)
covid (test100/covid) | Query editor (preview)
SQL database
Search (Ctrl+/)
Login
New Query
Open query
Overview
Activity log
Tags
Diagnose and solve problems
Quick start
Query editor (preview)
Power Platform
Power BI (preview)
Power Apps (preview)
Power Automate (preview)
Settings
Configure
Geo-Replication
Connection strings
Sync to other databases
Add Azure Search
Properties
Locks
Integrations
Stream analytics (preview)
Security
Auditing
Data Discovery & Classification
covid (test100)
Showing limited object explorer here. For full capability please open SSDT.
Tables
dbo.DimCancelPublicEvents
dbo.DimClosePublicTransit
dbo.DimCounty
dbo.DimDate
dbo.DimHospital
dbo.DimMaskUse
dbo.DimSchoolClosing
dbo.DimState
dbo.DimStayAtHome
dbo.DimWorkplaceClosing
dbo.FactCountyData_NYT
dbo.FactCountyLevelData
dbo.FactStateData_NYT
dbo.FactStateLevelData
dbo.FactStatePolicyTracker
dbo.FactUSTrackingStateData
Views
Stored Procedures

Run | Cancel query | Save query | Export data as | Show only Editor

```
1  SELECT TOP (1000) * FROM [dbo].[FactCountyLevelData]
```

Results | Messages

Search to filter items...

date_SK	county_SK	new_confirmed	new_deceased	new_tested	cumulative_confirmed	cumulative_deceased	cumulative_tested
20200208	1621	0	0	0	0	0	0
20200208	1622	0	0	0	0	0	0
20200208	1623	0	0	0	0	0	0
20200208	1624	0	0	0	0	0	0
20200208	1625	0	0	0	0	0	0
20200208	1626	0	0	0	0	0	0
20200208	1627	0	0	0	0	0	0
20200208	1628	0	0	0	0	0	0

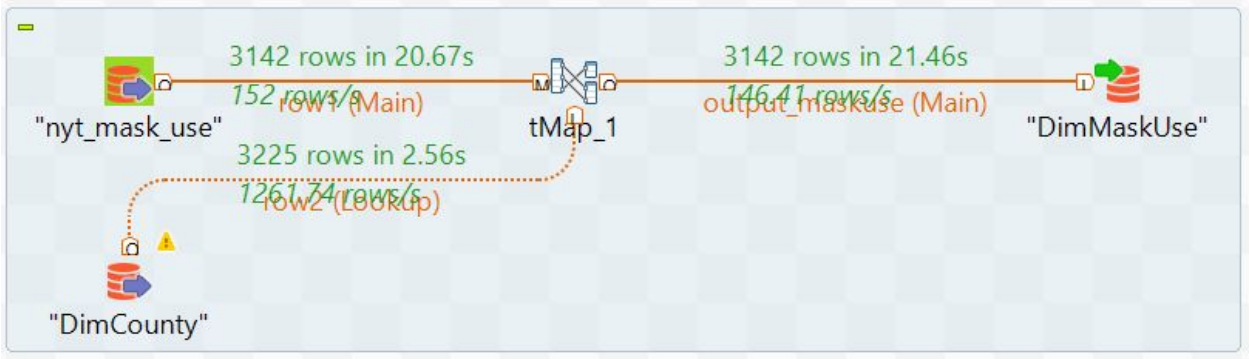
Run | Cancel query | Save query | Export data as | Show only Editor

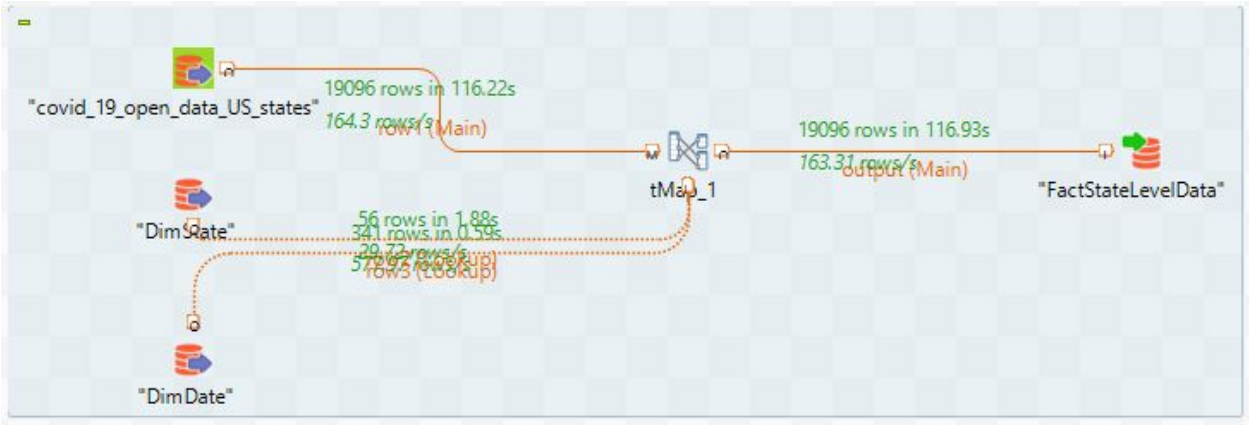
```
1  SELECT TOP (1000) * FROM [dbo].[DimMaskUse]
```

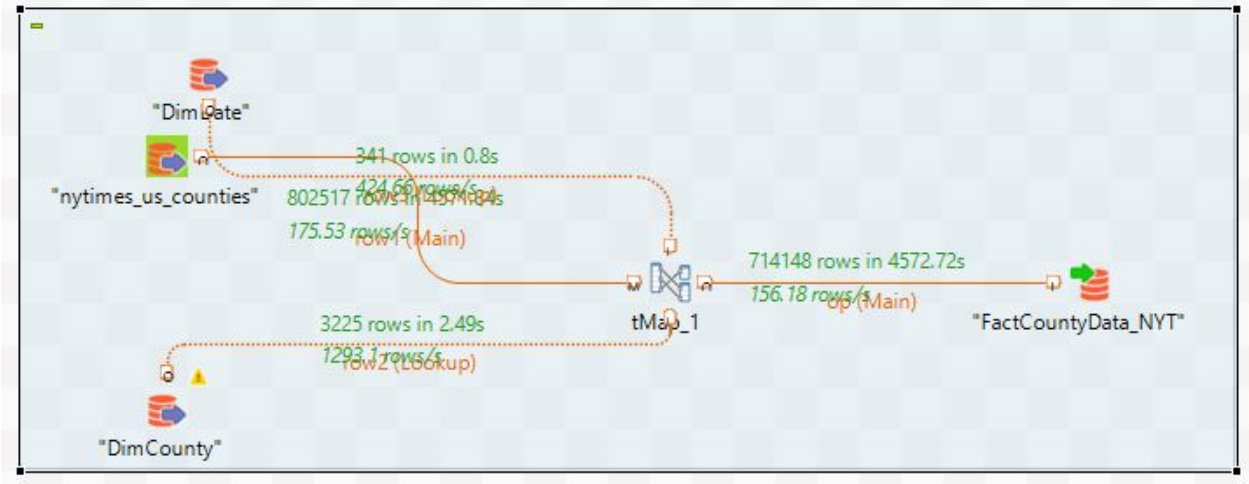
Results | Messages

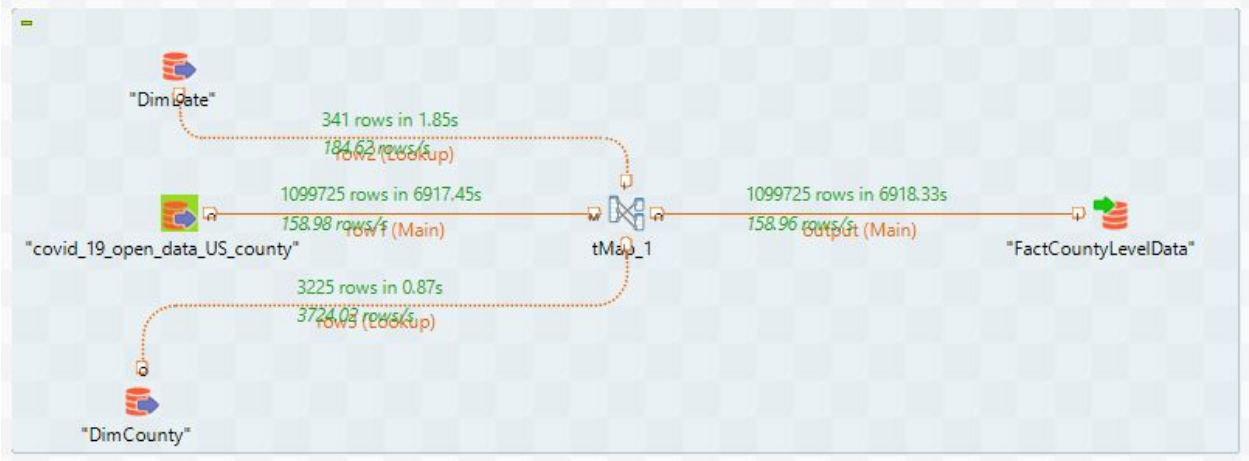
Search to filter items...

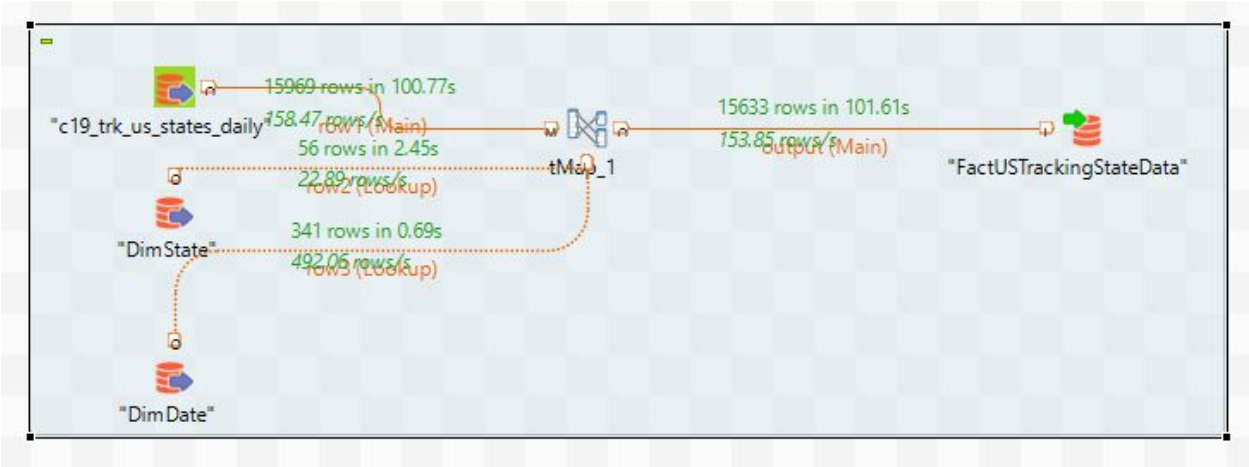
mask_use_SK	county_SK	never	rarely	sometimes	frequently	always	created_date
1	1	0.053	0.074	0.134	0.295	0.444	2020-12-16T21:04:48.913000(
2	2	0.083	0.059	0.098	0.323	0.436	2020-12-16T21:04:48.913000(
3	3	0.067	0.121	0.12	0.201	0.491	2020-12-16T21:04:48.913000(
4	4	0.02	0.034	0.096	0.278	0.572	2020-12-16T21:04:48.913000(
5	5	0.053	0.114	0.18	0.194	0.459	2020-12-16T21:04:48.913000(
6	6	0.031	0.04	0.144	0.286	0.5	2020-12-16T21:04:48.913000(
7	7	0.102	0.053	0.257	0.137	0.451	2020-12-16T21:04:48.913000(
8	8	0.152	0.108	0.13	0.167	0.442	2020-12-16T21:04:48.913000(

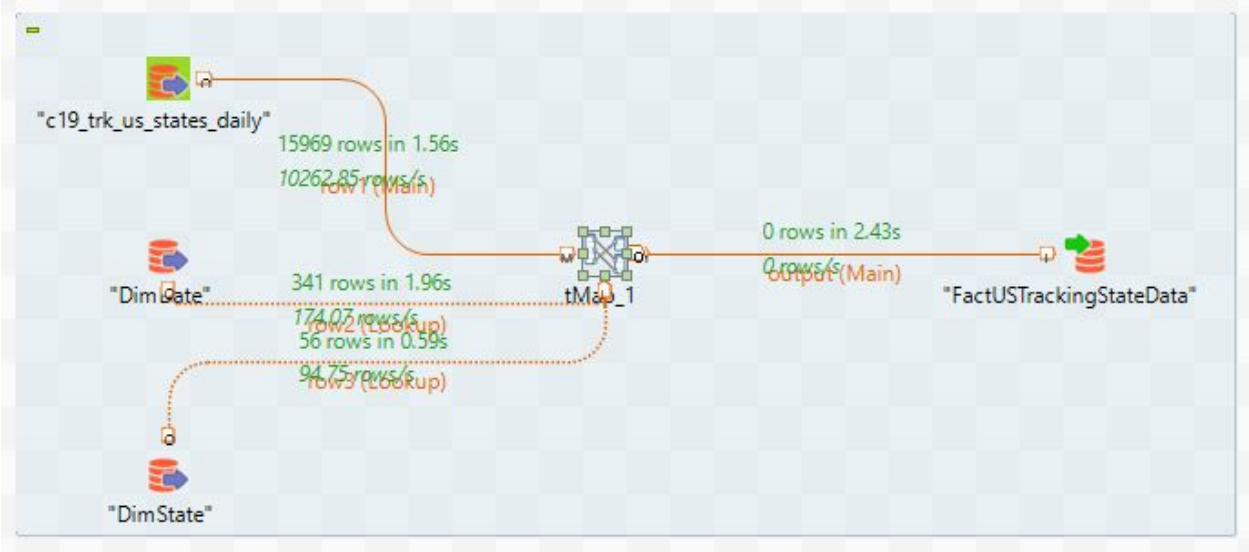


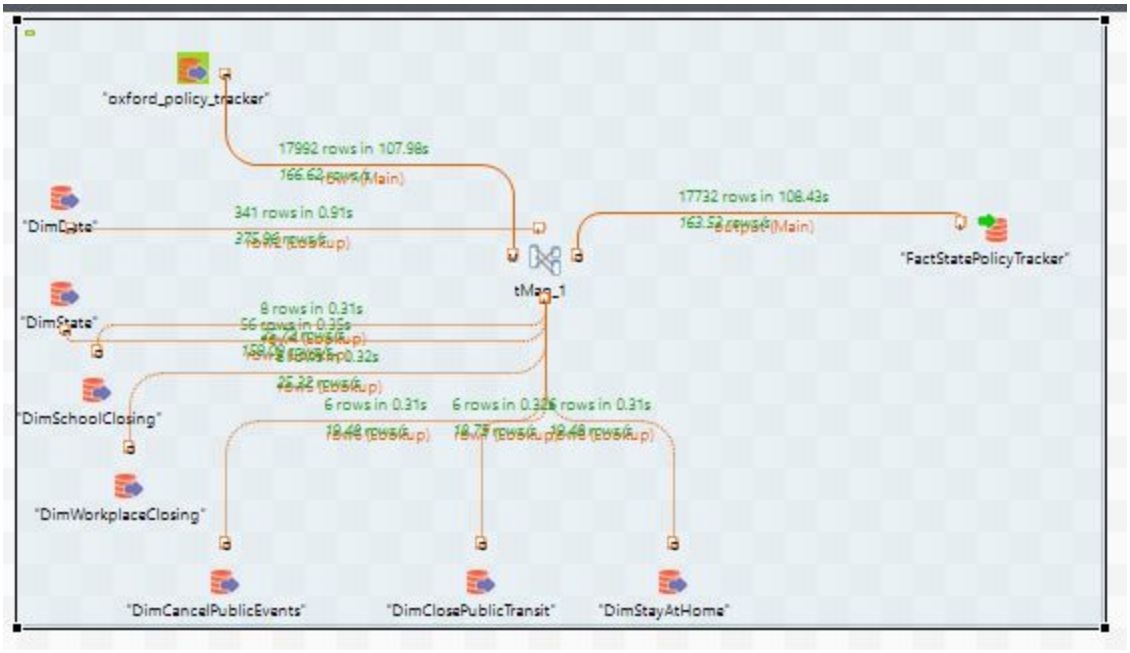

"covid_19_open_data_US_states"
19096 rows in 116.22s
164.3 rows/s
row1 (Main)
19096 rows in 116.93s
163.31 rows/s
output (Main)
tMap_1
"FactStateLevelData"
"DimState"
56 rows in 1.88s
341 rows in 0.59s
29.72 rows/s
576.62 rows/s
row3 (Lookup)
"DimDate"


"DimDate"
"nytimes_us_counties"
341 rows in 0.8s
424.66 rows/s
802517 rows in 4574.84s
175.53 rows/s
row1 (Main)
714148 rows in 4572.72s
156.18 rows/s
output (Main)
tMap_1
"FactCountyData_NYT"
3225 rows in 2.49s
1293.1 rows/s
row2 (Lookup)
"DimCounty"


"DimDate"
341 rows in 1.85s
184.62 rows/s
row2 (Lookup)
"covid_19_open_data_US_county"
1099725 rows in 6917.45s
158.98 rows/s
row1 (Main)
tMap_1
1099725 rows in 6918.33s
158.96 rows/s
output (Main)
"FactCountyLevelData"
3225 rows in 0.87s
3724.03 rows/s
row3 (Lookup)
"DimCounty"


15969 rows in 100.77s
158.47 rows/s
row1 (Main)
"c19_trk_us_states_daily"
56 rows in 2.45s
22.89 rows/s
row2 (Lookup)
"DimState"
341 rows in 0.69s
492.06 rows/s
row3 (Lookup)
"DimDate"
tMap_1
15633 rows in 101.61s
153.85 rows/s
output (Main)
"FactUSTrackingStateData"


"c19_trk_us_states_daily"
15969 rows in 1.56s
10262.85 rows/s
row1 (Main)
"DimDate"
341 rows in 1.96s
174.07 rows/s
row2 (Lookup)
56 rows in 0.59s
94.75 rows/s
row3 (Lookup)
"DimState"
tMap_1
0 rows in 2.43s
0 rows/s
output (Main)
"FactUSTrackingStateData"



PowerBI Dashboard：

URL: https://bit.ly/CovidDataIntegration