Visualizing Real Time Data Using ClickHouse and Superset

Srini Kadamati -- Preset
Robert Hodges -- Altinity

Presenter Bios

Robert Hodges - CEO at Altinity

30+ years on DBMS plus virtualization and security. ClickHouse is DBMS #20

Srini Kadamati - Senior Developer Advocate at Preset.io

6+ years in data science & data science education. Committer to Apache Superset.

Introducing ClickHouse and Superset

ClickHouse SQL Data Warehouse

- C++ Binary
- Apache 2.0 License
- Active community
- Column storage with compression
- Efficient parallel processing
- Sharding and replication

Superset Data Visualization and Exploration Platform

- Python + Javascript (Browser)
- Apache 2.0 License
- Active community
- Supports nearly any SQL database
- Dozens of chart types
- Preset Cloud managed service

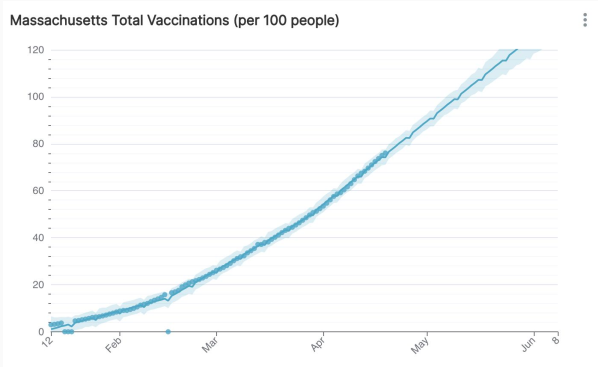
Visualization and data analysis

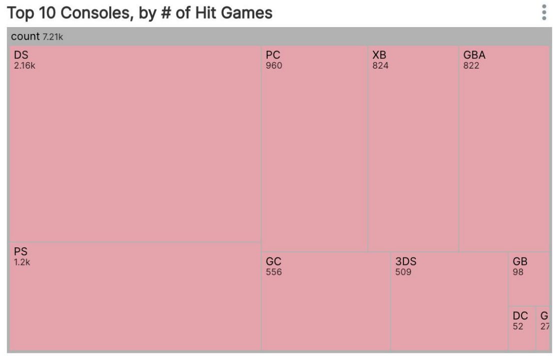
The data revolution

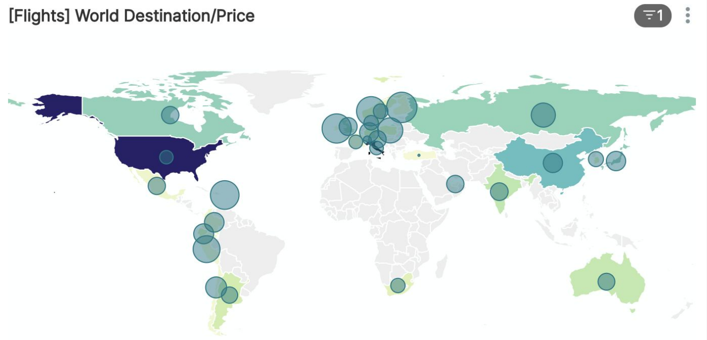
The main purpose of data science is insight

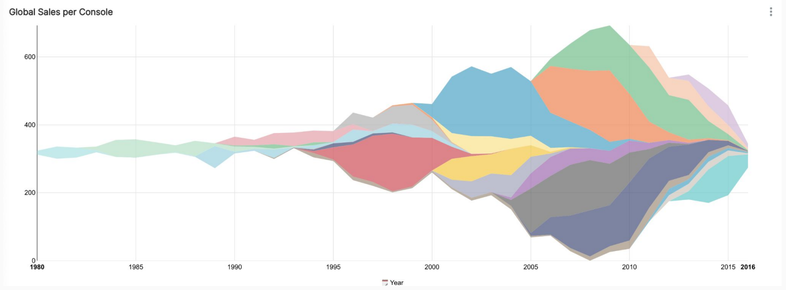
Data + Visualization!

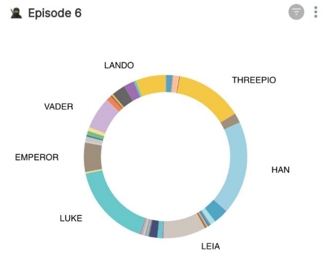
Effective data visualization is one of the best ways to get to **insight** from large, complex amounts of data







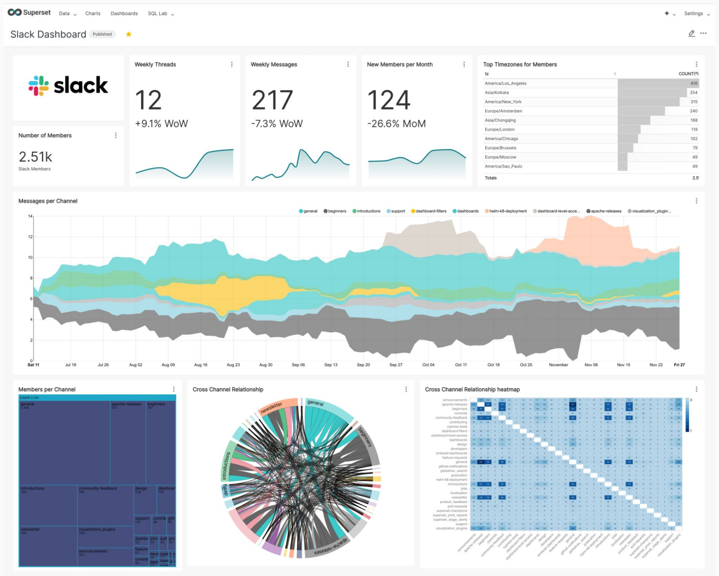




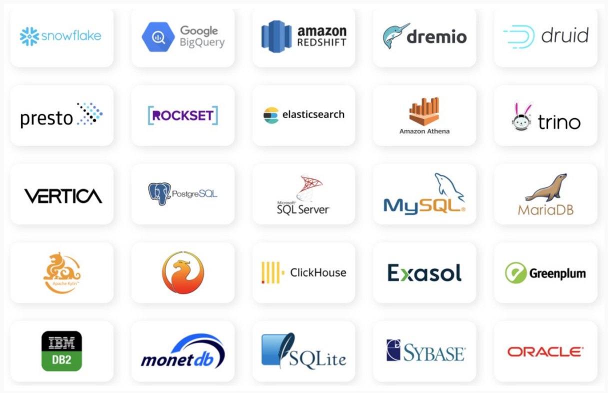
What is Superset & how does it work?

Altinity preset

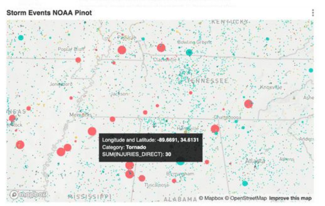
What is Apache Superset?

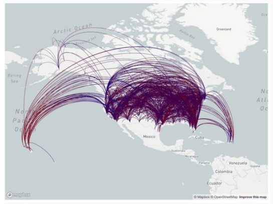


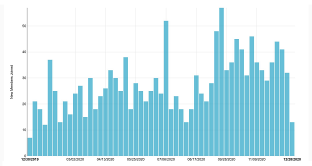
Modern open source BI platform

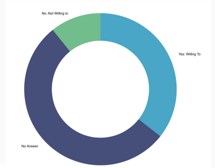


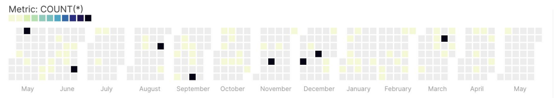
Works with nearly any SQL speaking data engine











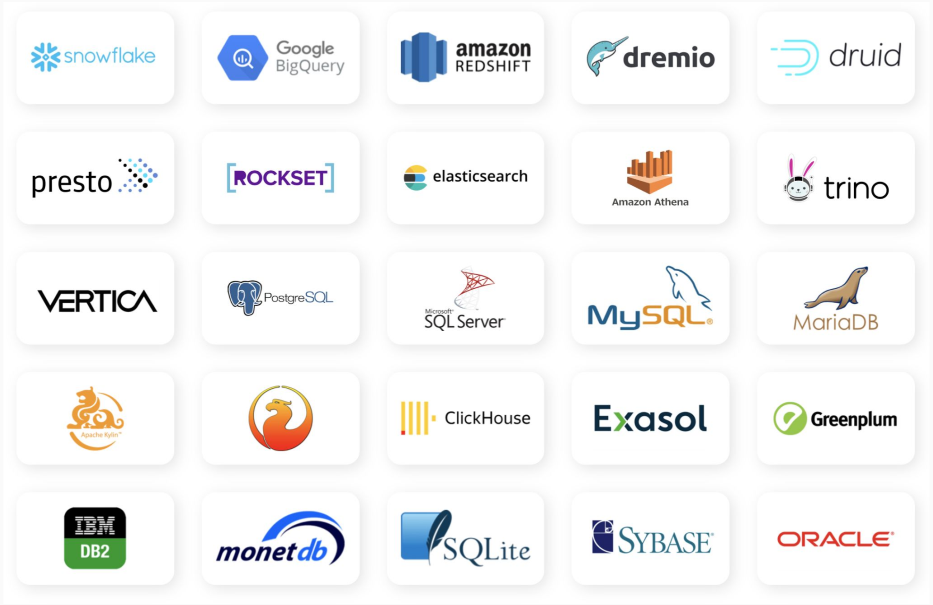
Large diversity of charts

Support for SQL Speaking Databases

Superset can query data from nearly any SQL speaking database

Requirements:

- Python DB-API 2.0 driver
  - https://www.python.org/dev/peps/pep-0249/
- SQLAlchemy Dialect
  - Incomplete list here: https://docs.sqlalchemy.org/en/14/dialects/

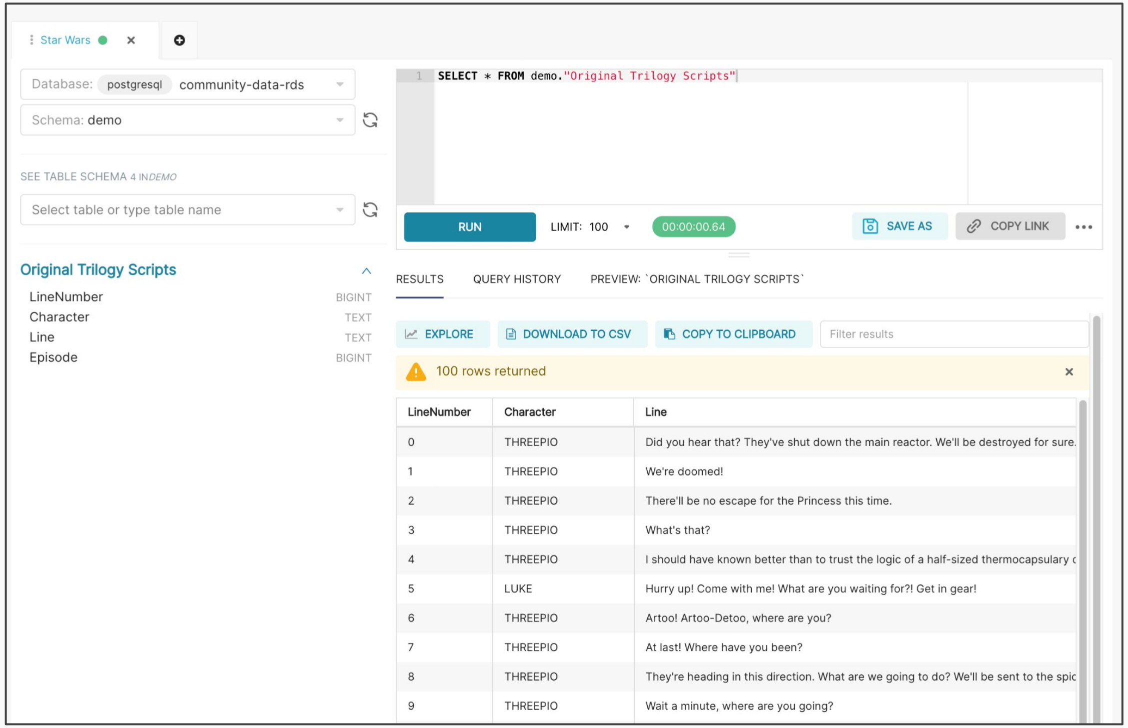


Blog Post: Building New Database Connectors for Superset

SQL Lab

Browser based SQL IDE (focused on writing analytical queries)

- Explore data
- Sculpt data for visualization
- Save transformed data as virtual datasets, db VIEWs, or db TABLEs



```sql
SELECT * FROM demo."Original Trilogy Scripts"
```

Original Trilogy Scripts

LineNumber	BIGINT
Character	TEXT
Line	TEXT
Episode	BIGINT

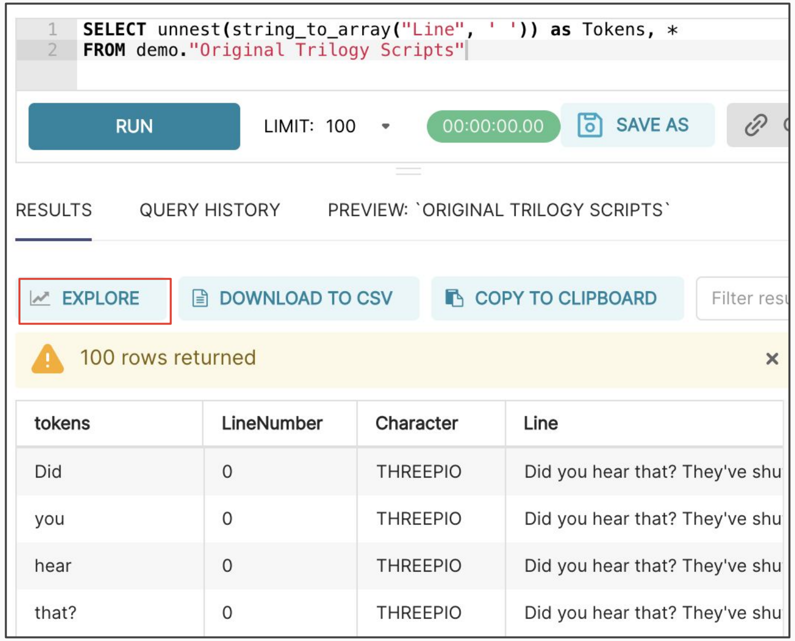
100 rows returned

LineNumber	Character	Line
0	THREEPIO	Did you hear that? They've shut down the main reactor. We'll be destroyed for sure.
1	THREEPIO	We're doomed!
2	THREEPIO	There'll be no escape for the Princess this time.
3	THREEPIO	What's that?
4	THREEPIO	I should have known better than to trust the logic of a half-sized thermocapsulary c
5	LUKE	Hurry up! Come with me! What are you waiting for?! Get in gear!
6	THREEPIO	Artoo! Artoo-Detoo, where are you?
7	THREEPIO	At last! Where have you been?
8	THREEPIO	They're heading in this direction. What are we going to do? We'll be sent to the spic
9	THREEPIO	Wait a minute, where are you going?

Altinity

preset

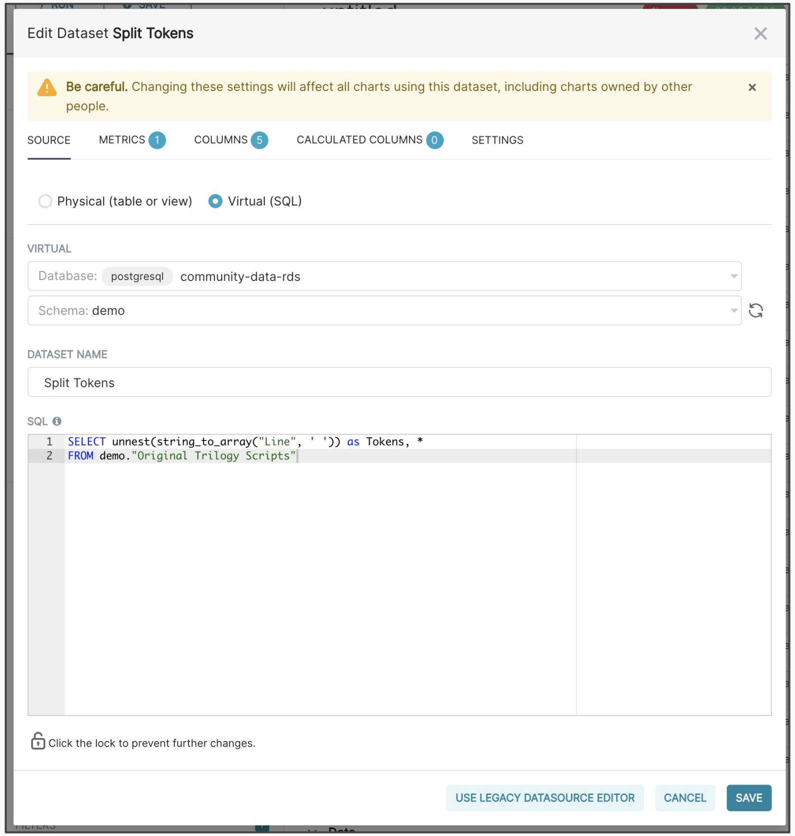
Superset Semantic Layer



```
SELECT unnest(string_to_array("Line", ' ')) as Tokens, *
FROM demo."Original Trilogy Scripts"
```

100 rows returned

tokens	LineNumber	Character	Line
Did	0	THREEPIO	Did you hear that? They've shu
you	0	THREEPIO	Did you hear that? They've shu
hear	0	THREEPIO	Did you hear that? They've shu
that?	0	THREEPIO	Did you hear that? They've shu



Edit Dataset **Split Tokens**

Be careful. Changing these settings will affect all charts using this dataset, including charts owned by other people.

Physical (table or view) [✓ Virtual (SQL)]

VIRTUAL

Database: postgresql community-data-rds

Schema: demo

DATASET NAME

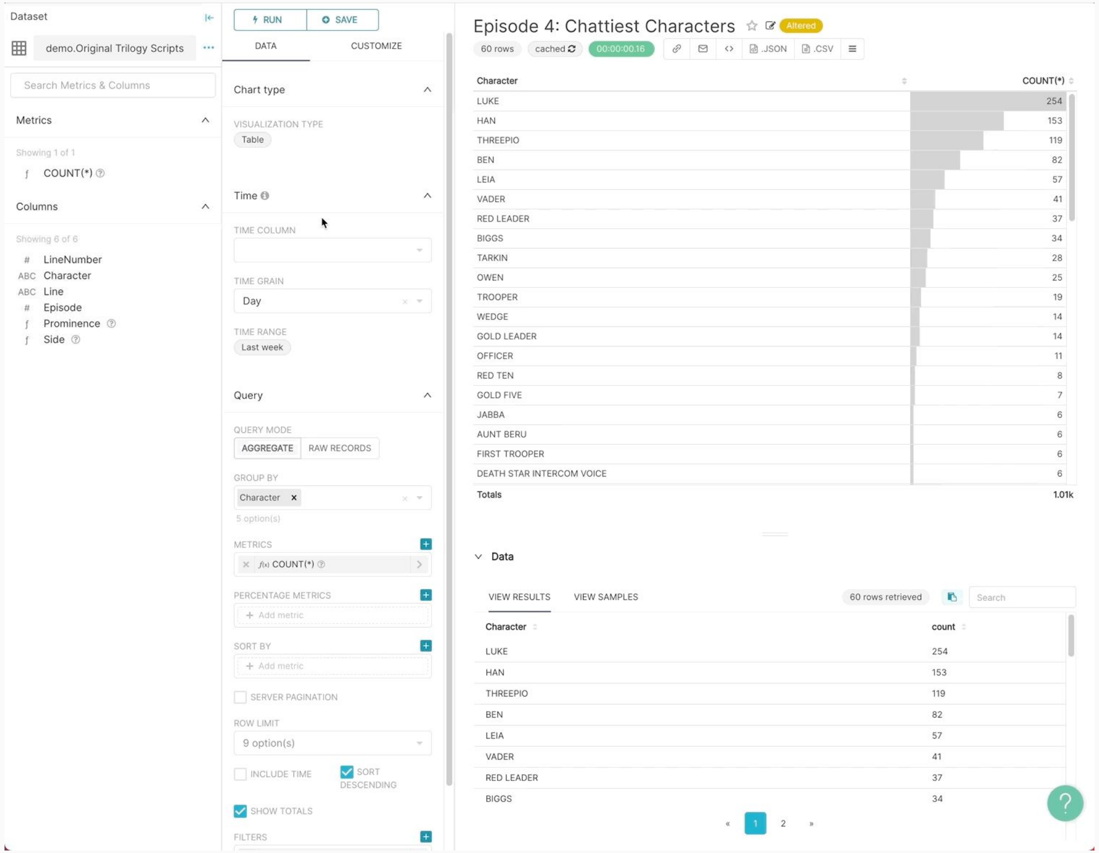
Split Tokens

SQL

```
SELECT unnest(string_to_array("Line", ' ')) as Tokens, *
FROM demo."Original Trilogy Scripts"
```

Click the lock to prevent further changes.

No Code Chart Builder (Explore)



Dataset: demo.Original Trilogy Scripts

Metrics (Showing 1 of 1): COUNT(*)

Columns (Showing 6 of 6): LineNumber, Character, Line, Episode, Prominence, Side

RUN · SAVE · DATA · CUSTOMIZE

Chart type

VISUALIZATION TYPE: Table

Time

TIME COLUMN:

TIME GRAIN: Day

TIME RANGE: Last week

Query

QUERY MODE: [✓ AGGREGATE] RAW RECORDS

GROUP BY: Character (5 option(s))

METRICS: COUNT(*)

PERCENTAGE METRICS: + Add metric

SORT BY: + Add metric

☐ SERVER PAGINATION

ROW LIMIT: 9 option(s)

☐ INCLUDE TIME ☑ SORT DESCENDING

☑ SHOW TOTALS

FILTERS

Episode 4: Chattiest Characters (Altered)

60 rows · cached · 00:00:00.16 · .JSON · .CSV

Character	COUNT(*)
LUKE	254
HAN	153
THREEPIO	119
BEN	82
LEIA	57
VADER	41
RED LEADER	37
BIGGS	34
TARKIN	28
OWEN	25
TROOPER	19
WEDGE	14
GOLD LEADER	14
OFFICER	11
RED TEN	8
GOLD FIVE	7
JABBA	6
AUNT BERU	6
FIRST TROOPER	6
DEATH STAR INTERCOM VOICE	6
Totals	1.01k

Data

VIEW RESULTS · VIEW SAMPLES · 60 rows retrieved

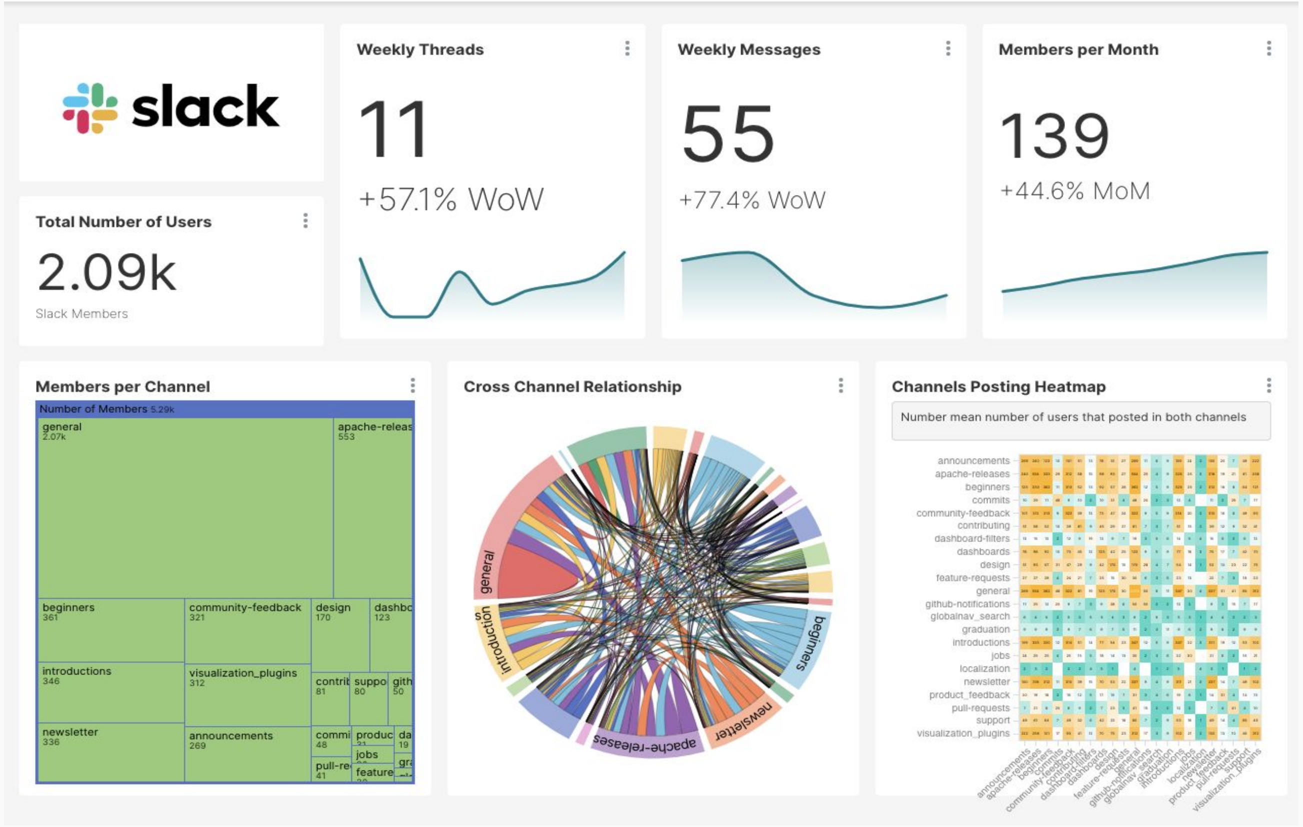
Character	count
LUKE	254
HAN	153
THREEPIO	119
BEN	82
LEIA	57
VADER	41
RED LEADER	37
BIGGS	34

« 1 2 »

Altinity

preset

Slack Community Dashboard



[Build your own Slack dashboard]

Altinity

preset

Introducing ClickHouse and Superset

ClickHouse is a new option for analytic services

An industrial strength, Apache 2.0 SQL data warehouse

Installs on a laptop in 60 seconds

Runs in cloud, containers, and bare metal

Answers queries in milliseconds

ClickHouse performance meets or exceeds proprietary SQL data warehouses

Portable C++ binary

Advanced SQL implementation

Column storage with high compression

Distributed, vectorized queries

Built-in sharding and replication

Scales from laptop to 100s of nodes

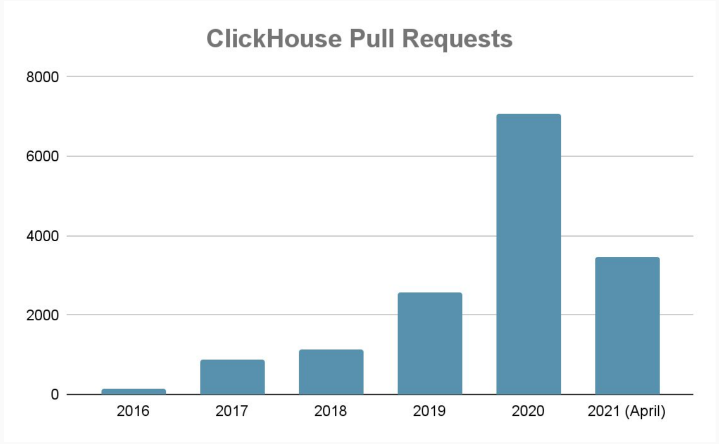
Popular Apache 2.0 licensing

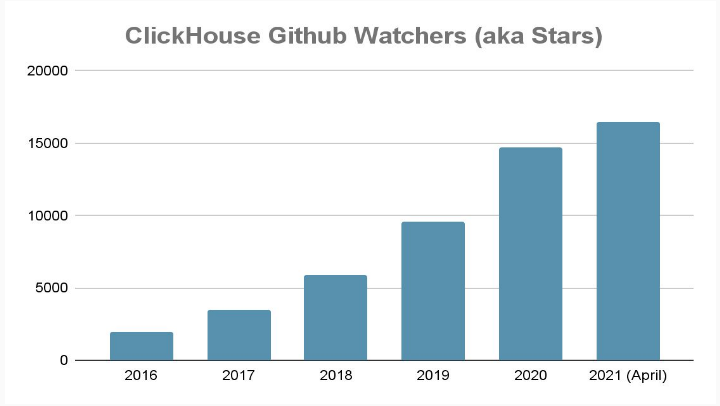
We use ClickHouse extensively and it's been great.

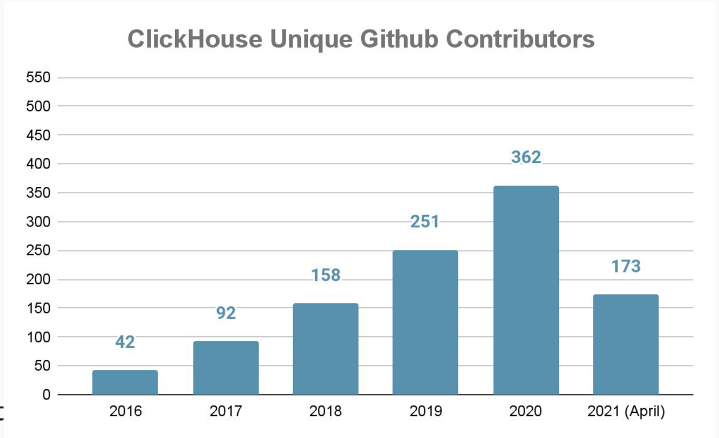
John Graham-Cumming, CloudFlare CTO
Hacker News, 2020

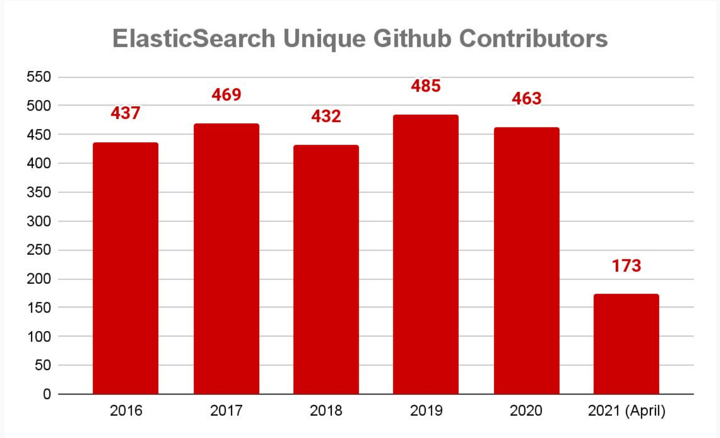
Bloomberg Cisco Flipkart Comcast criteo Instana Cloudflare Contentsquare Lifestreet Nuna Semrush Spotify Yandex Sentry Walmart eBay Uber Deutsche Bank

ClickHouse grew from an in-house Yandex project to an international community that rivals ElasticSearch





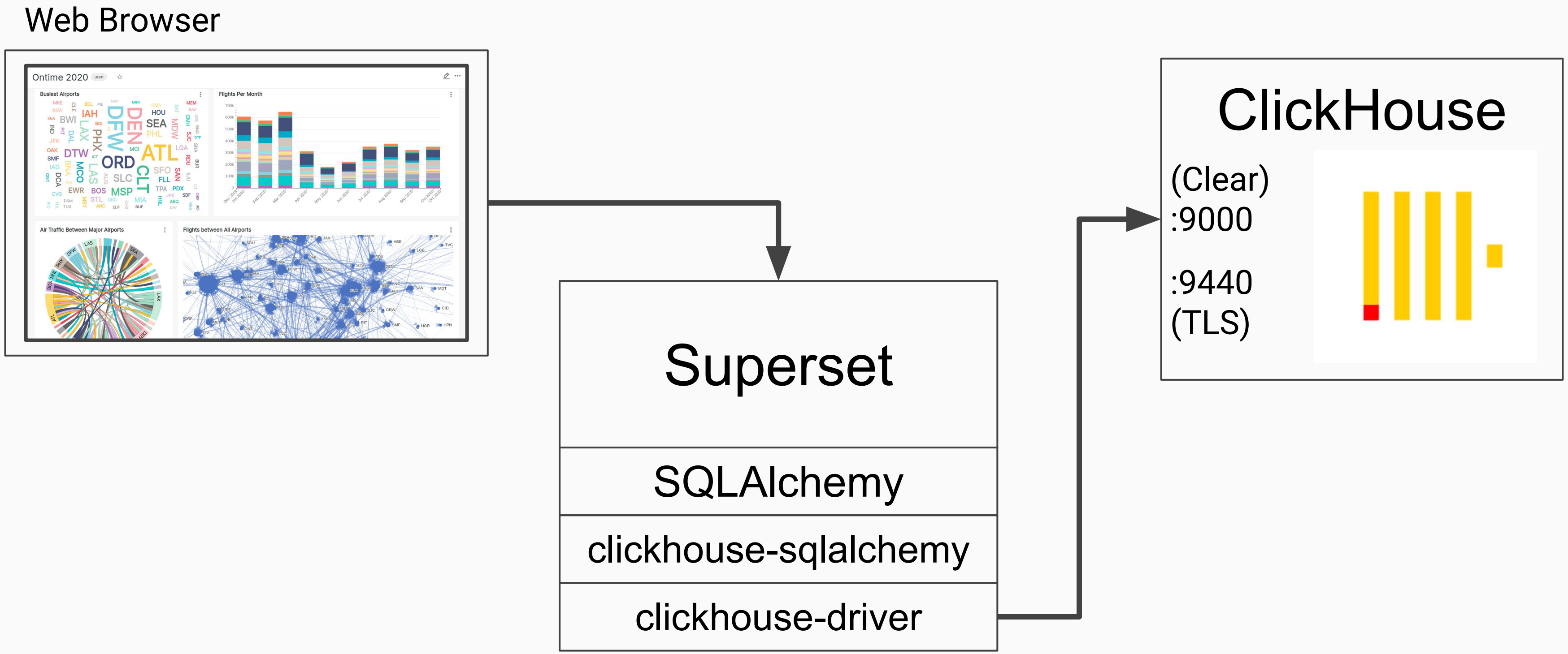




Altinity

preset

Superset connection to ClickHouse



ClickHouse SQLAlchemy Drivers (An aside)

★

sqlalchemy-clickhouse
Apache 2.0
Developed by Marek Vavrusa

- Currently documented in Superset
- Uses ClickHouse HTTP Interface
- No TLS support
- Current pypi.org release: 0.1.5.post0, Aug 9 2018

★ ★

clickhouse-sqlalchemy
Apache 2.0
Developed by Konstantin Lebedev

- Supported by Altinity
- Uses ClickHouse Native TCP
- TLS support
- Bug fixes for Superset
- Current pypi.org release: 0.1.6, Mar 15 2021

Setting up Superset (Ubuntu)

```
# Install Superset
export FLASK_APP=superset
pip install apache-superset
superset db upgrade
superset fab create-admin
superset load_examples
superset init

# Add ClickHouse driver
pip install clickhouse-sqlalchemy

# Start Superset
superset run -p 8088 --with-threads --reload --debugger
```

Ensure version >= 0.1.6!

ClickHouse connection strings

SQLAlchemy URL format:

```
clickhouse+native://[user:pw]@host[:port]/database[?options…]
```

ClickHouse on localhost (e.g., your laptop)

```
clickhouse+native://localhost/default
```

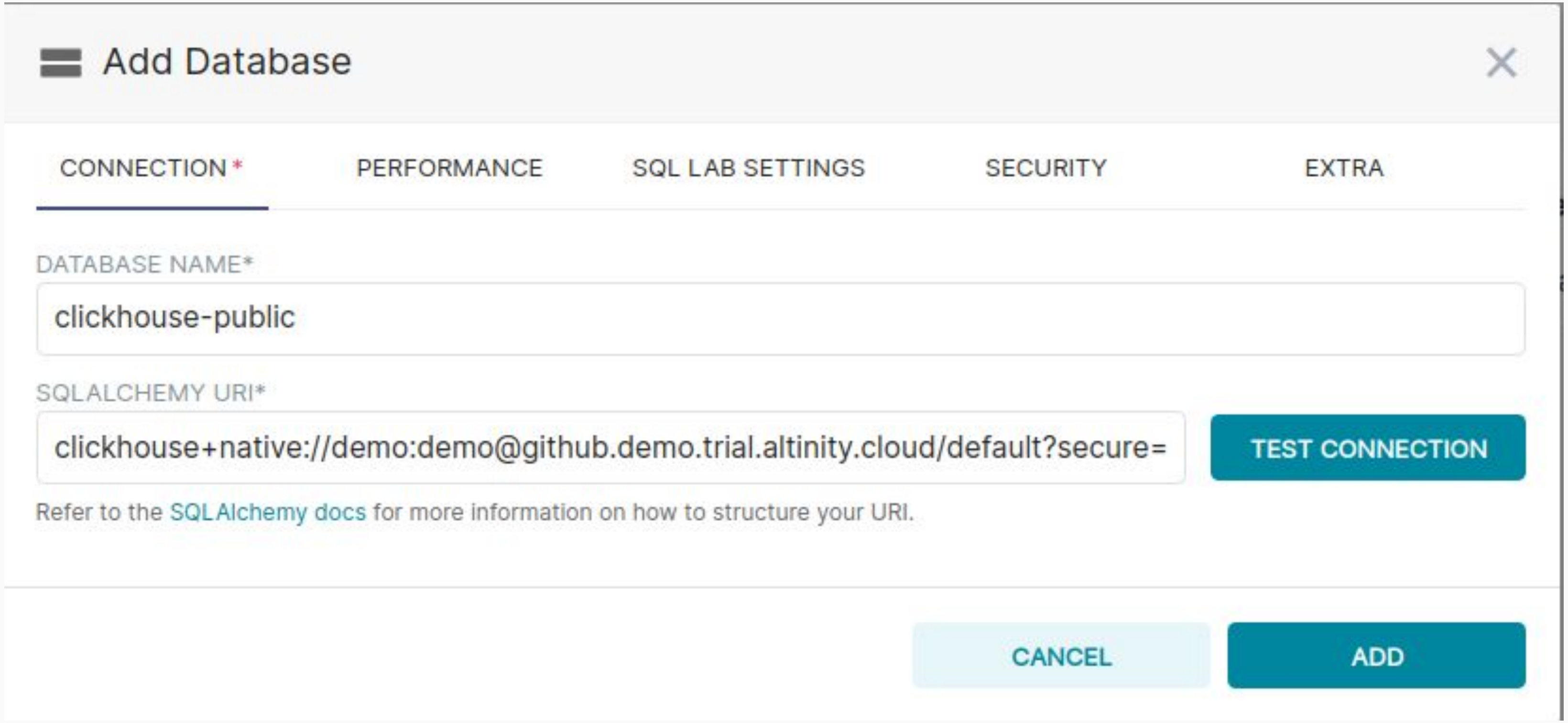
ClickHouse public endpoint:

```
clickhouse+native://demo:demo@github.demo.trial.altinity.clou
d/default?secure=true
```

Altinity

preset

Database connection page



Add Database

DATABASE NAME*

clickhouse-public

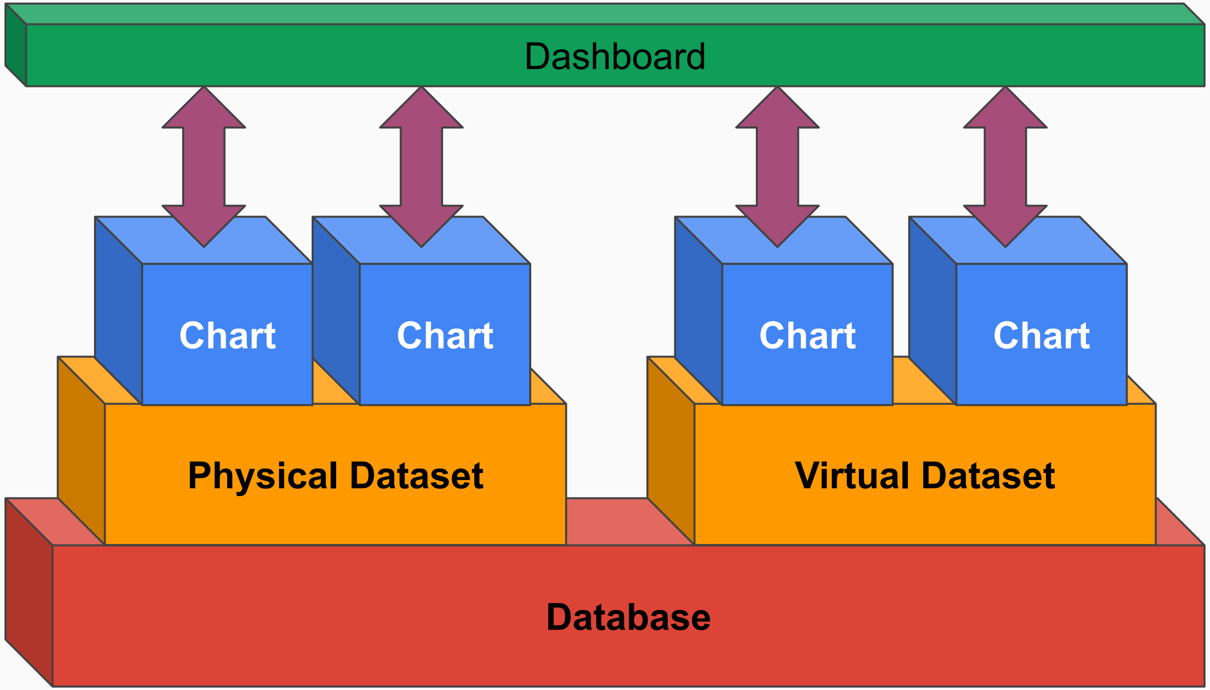
SQLALCHEMY URI*

clickhouse+native://demo:demo@github.demo.trial.altinity.cloud/default?secure=

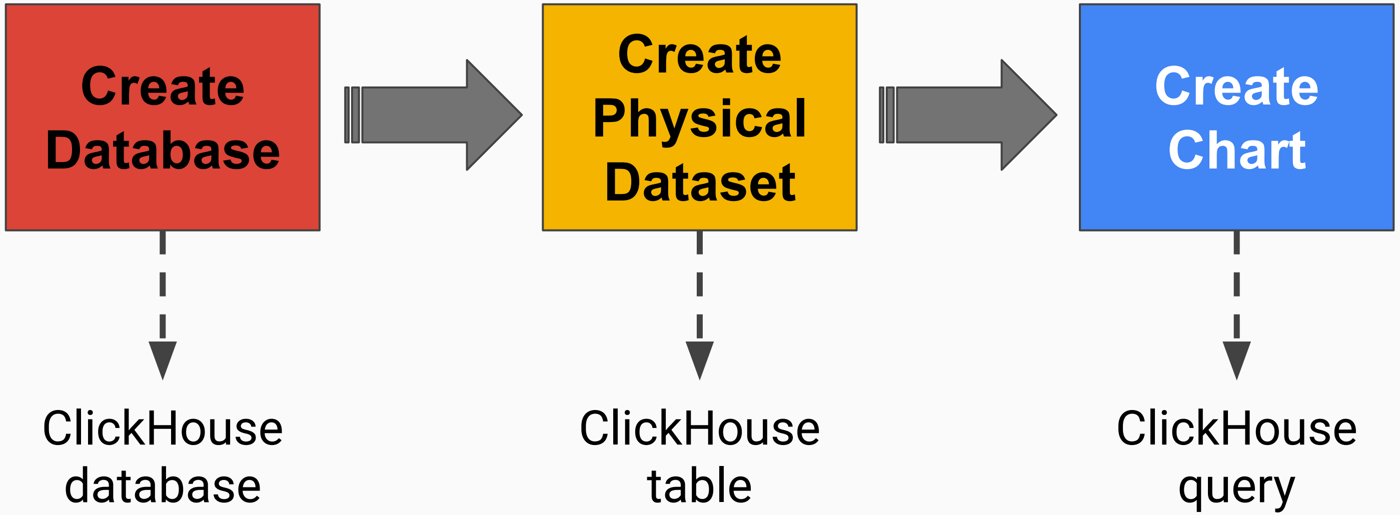
Refer to the SQLAlchemy docs for more information on how to structure your URI.

Fun with ClickHouse Data

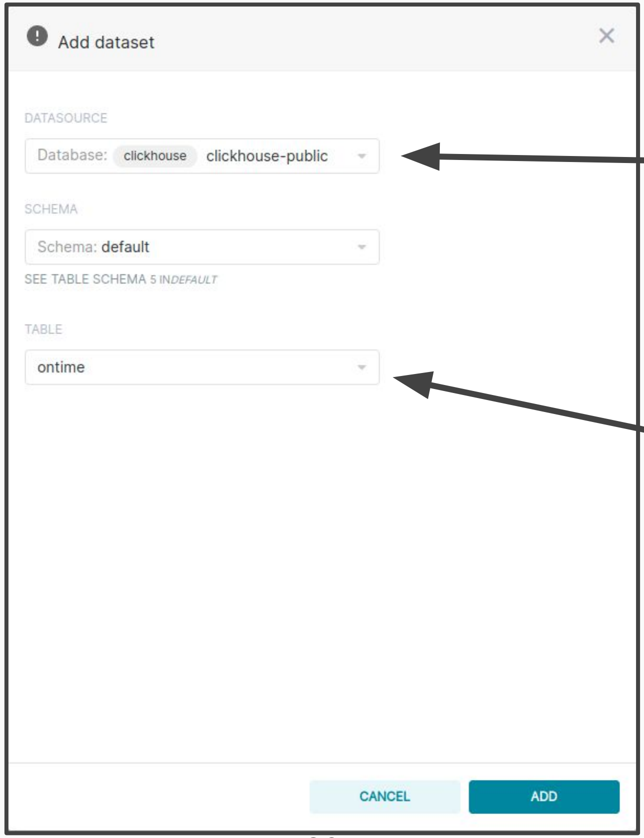
Superset dashboard organization



Creating a time series chart



Physical dataset creation page



Add dataset

DATASOURCE

Database: clickhouse clickhouse-public

SCHEMA

Schema: default

SEE TABLE SCHEMA 5 IN*DEFAULT*

TABLE

ontime

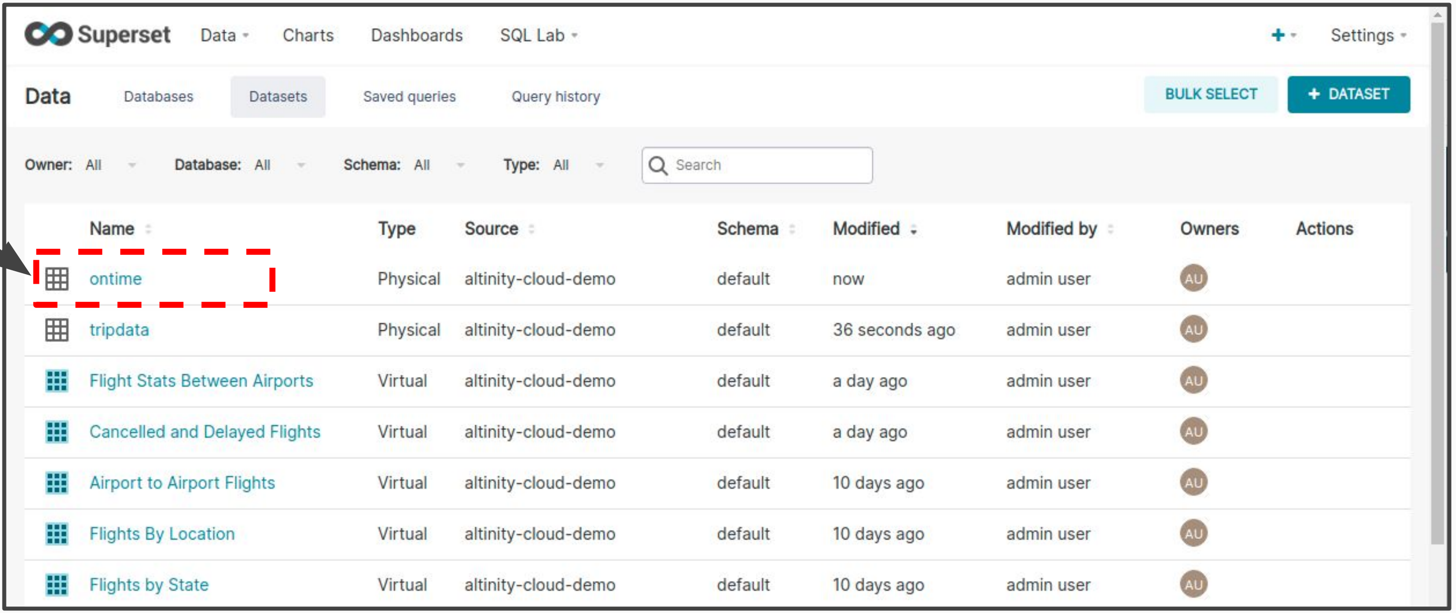
Database

ClickHouse Connection

Table

Starting a chart from a dataset

Dataset



Name	Type	Source	Schema	Modified	Modified by	Owners	Actions
ontime	Physical	altinity-cloud-demo	default	now	admin user	AU	
tripdata	Physical	altinity-cloud-demo	default	36 seconds ago	admin user	AU	
Flight Stats Between Airports	Virtual	altinity-cloud-demo	default	a day ago	admin user	AU	
Cancelled and Delayed Flights	Virtual	altinity-cloud-demo	default	a day ago	admin user	AU	
Airport to Airport Flights	Virtual	altinity-cloud-demo	default	10 days ago	admin user	AU	
Flights By Location	Virtual	altinity-cloud-demo	default	10 days ago	admin user	AU	
Flights by State	Virtual	altinity-cloud-demo	default	10 days ago	admin user	AU	

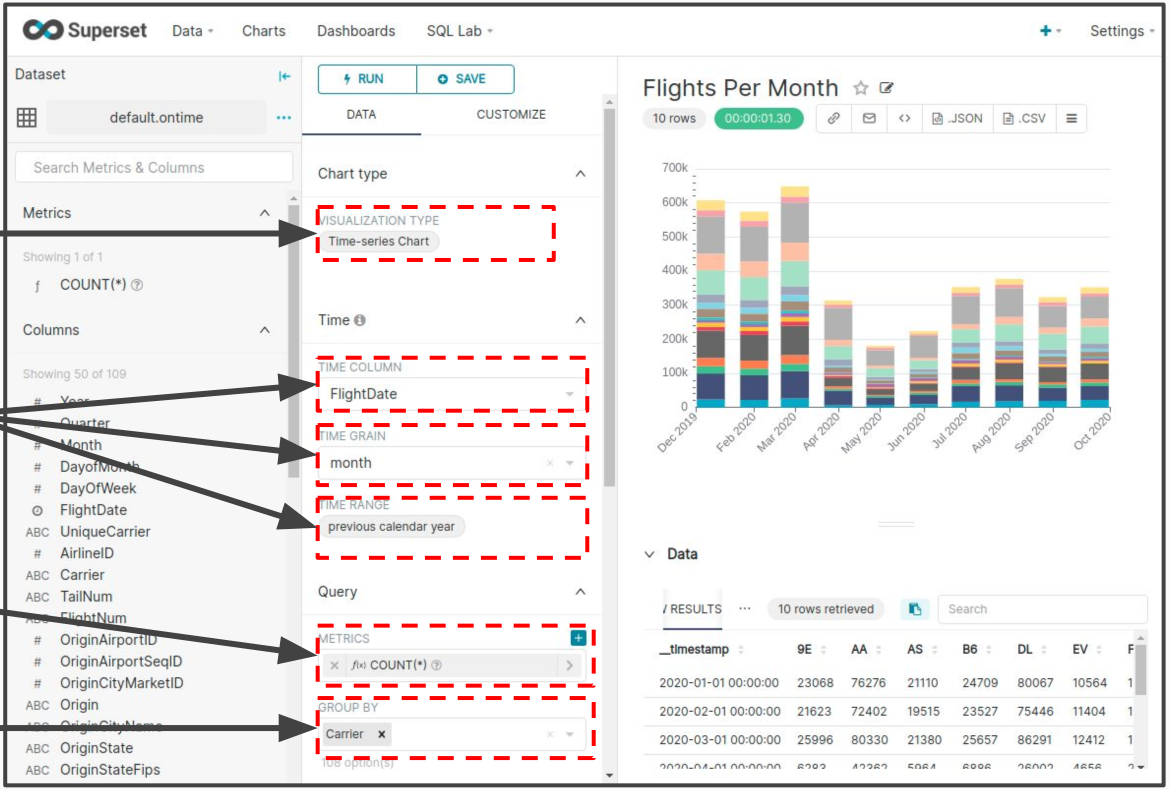
Time series chart from physical dataset

Chart Type

Time Dimension

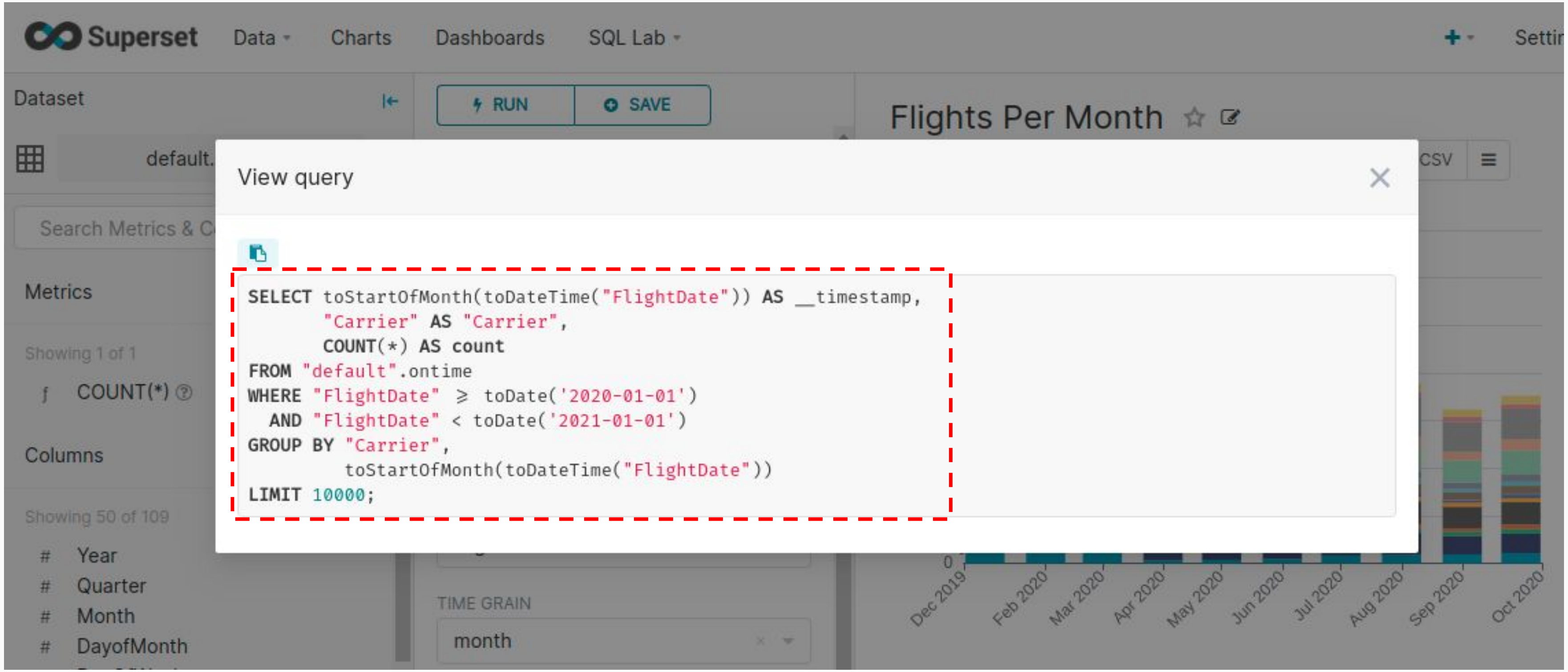
Metric

Group



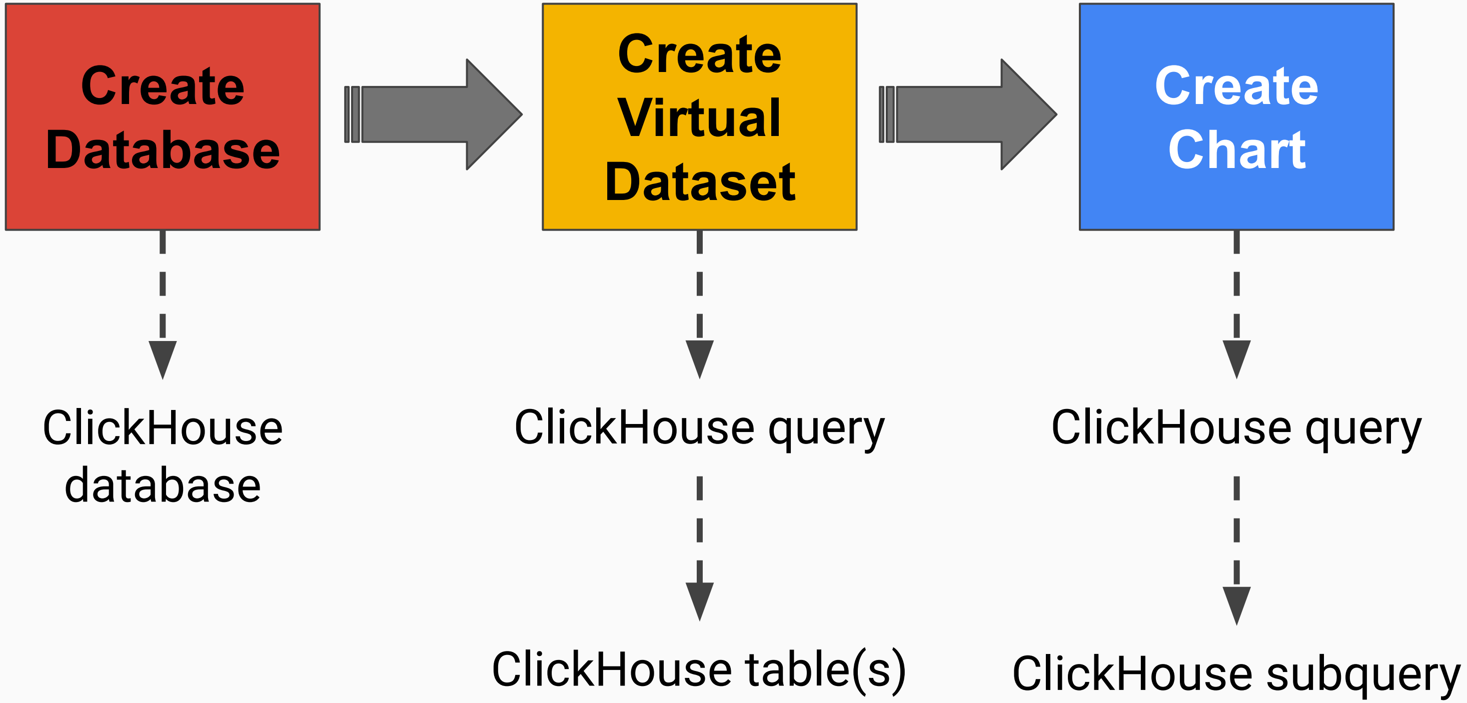
__timestamp	9E	AA	AS	B6	DL	EV	F
2020-01-01 00:00:00	23068	76276	21110	24709	80067	10564	1
2020-02-01 00:00:00	21623	72402	19515	23527	75446	11404	1
2020-03-01 00:00:00	25996	80330	21380	25657	86291	12412	1
2020-04-01 00:00:00	6283	42362	5964	6886	26002	4656	2

Time series query is generated



```
SELECT toStartOfMonth(toDateTime("FlightDate")) AS __timestamp,
       "Carrier" AS "Carrier",
       COUNT(*) AS count
FROM "default".ontime
WHERE "FlightDate" >= toDate('2020-01-01')
  AND "FlightDate" < toDate('2021-01-01')
GROUP BY "Carrier",
         toStartOfMonth(toDateTime("FlightDate"))
LIMIT 10000;
```

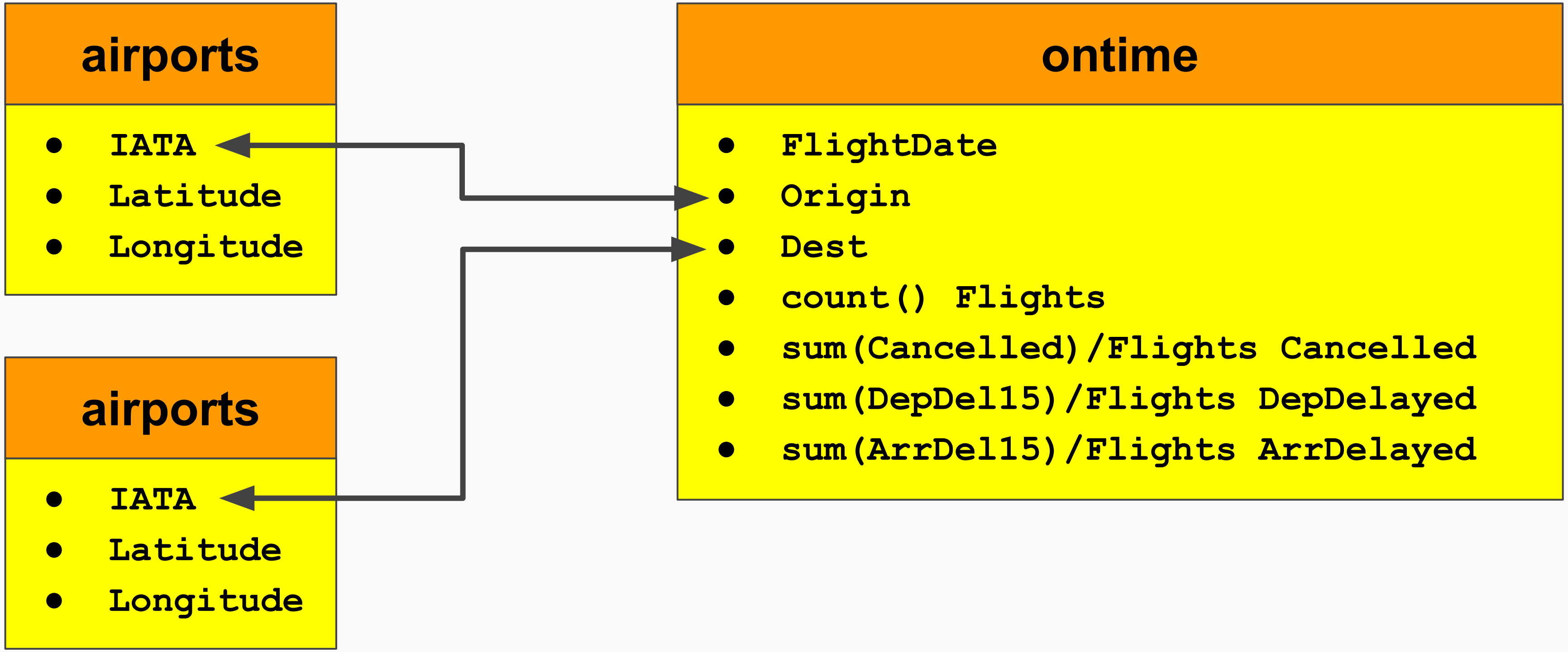
Creating a chart on a virtual dataset



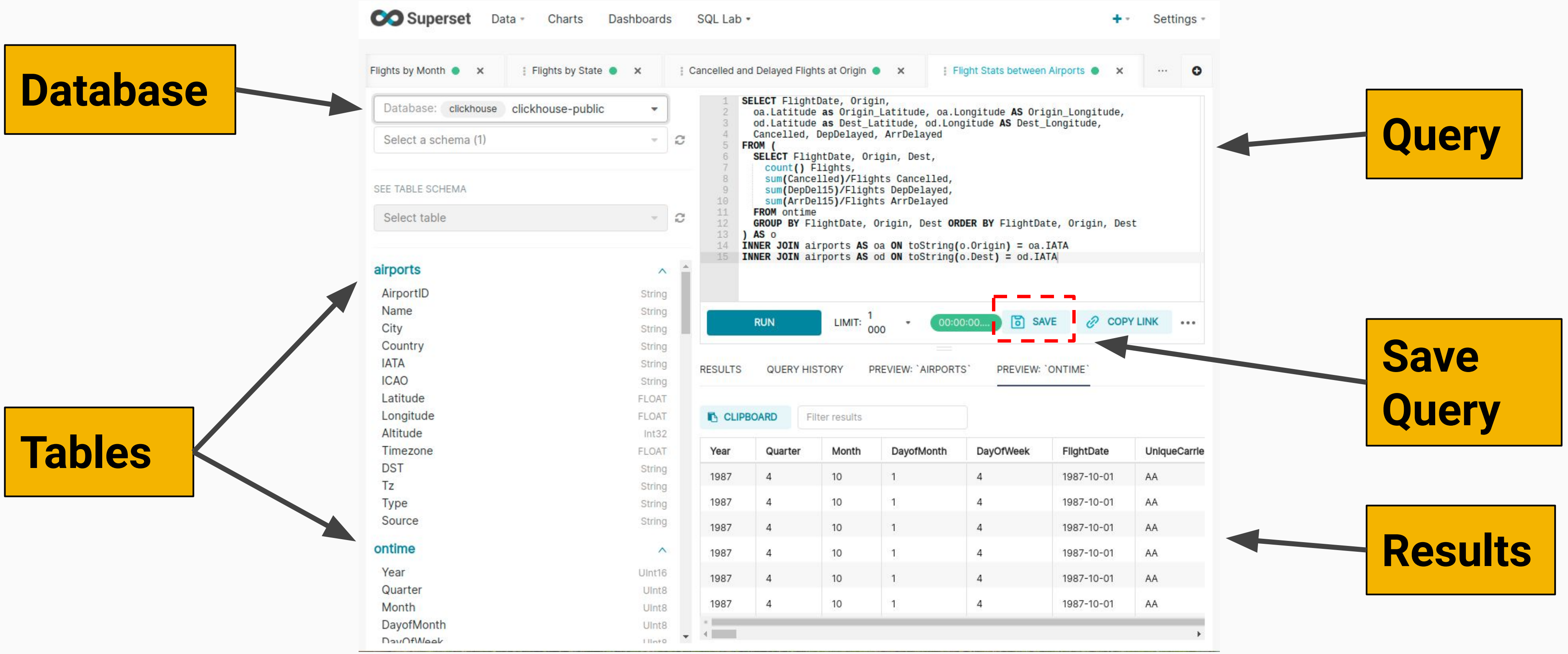
A query that answers multiple questions

```
SELECT FlightDate, Origin, Dest,
  oa.Latitude as Origin_Latitude, oa.Longitude AS Origin_Longitude,
  od.Latitude as Dest_Latitude, od.Longitude AS Dest_Longitude,
  Flights, Cancelled, DepDelayed, ArrDelayed
FROM (
  SELECT FlightDate, Origin, Dest, count() Flights,
    sum(Cancelled)/Flights Cancelled, sum(DepDel15)/Flights DepDelayed,
    sum(ArrDel15)/Flights ArrDelayed
  FROM ontime
  GROUP BY FlightDate, Origin, Dest ORDER BY FlightDate, Origin, Dest
) AS o
INNER JOIN airports AS oa ON toString(o.Origin) = oa.IATA
INNER JOIN airports AS od ON toString(o.Dest) = od.IATA
```

Using JOIN to add airport LAT/LONG



Build, run, and save query in SQL Lab



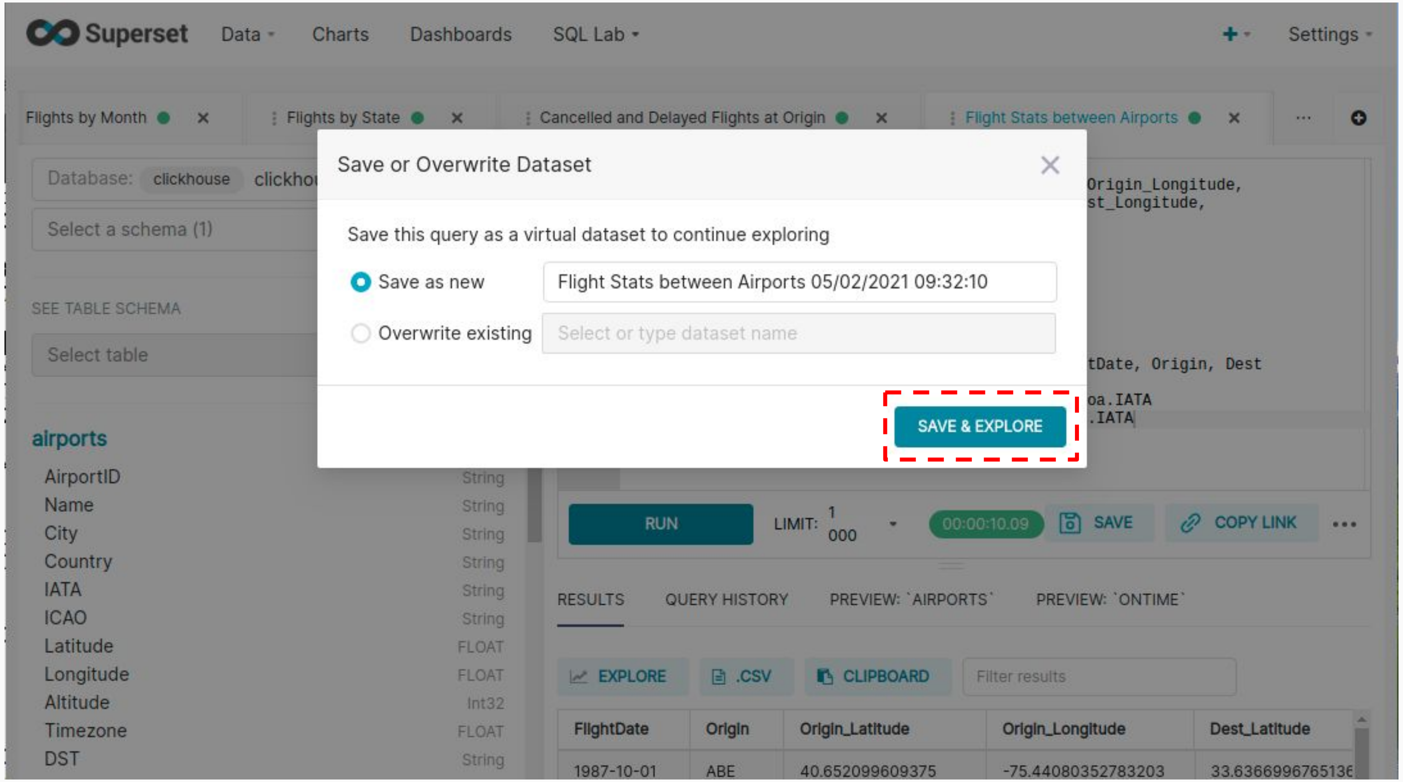
airports	
AirportID	String
Name	String
City	String
Country	String
IATA	String
ICAO	String
Latitude	FLOAT
Longitude	FLOAT
Altitude	Int32
Timezone	FLOAT
DST	String
Tz	String
Type	String
Source	String

ontime	
Year	UInt16
Quarter	UInt8
Month	UInt8
DayofMonth	UInt8

```sql
SELECT FlightDate, Origin,
  oa.Latitude as Origin_Latitude, oa.Longitude AS Origin_Longitude,
  od.Latitude as Dest_Latitude, od.Longitude AS Dest_Longitude,
  Cancelled, DepDelayed, ArrDelayed
FROM (
  SELECT FlightDate, Origin, Dest,
    count() Flights,
    sum(Cancelled)/Flights Cancelled,
    sum(DepDel15)/Flights DepDelayed,
    sum(ArrDel15)/Flights ArrDelayed
  FROM ontime
  GROUP BY FlightDate, Origin, Dest ORDER BY FlightDate, Origin, Dest
) AS o
INNER JOIN airports AS oa ON toString(o.Origin) = oa.IATA
INNER JOIN airports AS od ON toString(o.Dest) = od.IATA
```

Year	Quarter	Month	DayofMonth	DayOfWeek	FlightDate	UniqueCarrie
1987	4	10	1	4	1987-10-01	AA
1987	4	10	1	4	1987-10-01	AA
1987	4	10	1	4	1987-10-01	AA
1987	4	10	1	4	1987-10-01	AA
1987	4	10	1	4	1987-10-01	AA
1987	4	10	1	4	1987-10-01	AA

Use EXPLORE to save as dataset



Save or Overwrite Dataset

Save this query as a virtual dataset to continue exploring

[✓ Save as new] Flight Stats between Airports 05/02/2021 09:32:10

Overwrite existing — Select or type dataset name

airports

AirportID	String
Name	String
City	String
Country	String
IATA	String
ICAO	String
Latitude	FLOAT
Longitude	FLOAT
Altitude	Int32
Timezone	FLOAT
DST	String

```
Origin_Longitude,
st_Longitude,

tDate, Origin, Dest

oa.IATA
.IATA
```

FlightDate	Origin	Origin_Latitude	Origin_Longitude	Dest_Latitude
1987-10-01	ABE	40.652099609375	-75.44080352783203	33.636699676513

Creating a deck.gl Arc chart

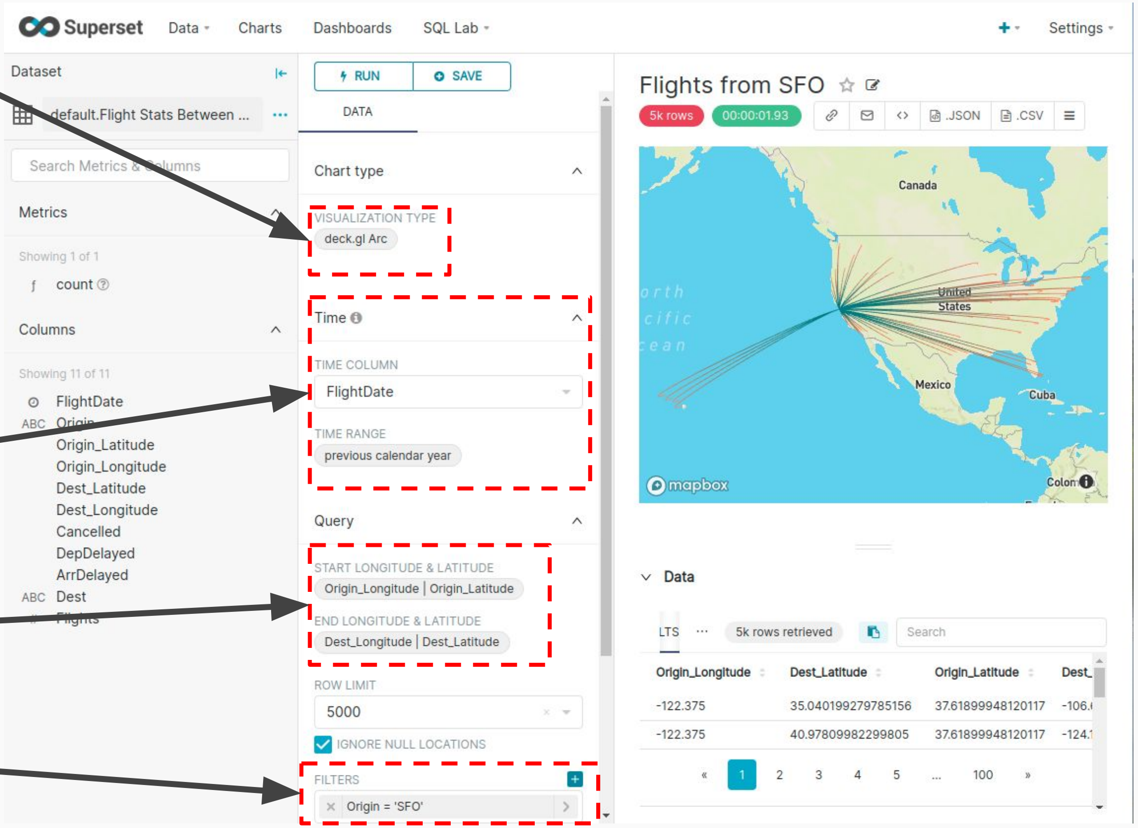
Chart Type*

* Requires a Mapbox token -- See docs

Time Dimension

Lat/Long

Filter



Dataset: default.Flight Stats Between ...

Metrics — Showing 1 of 1: count

Columns — Showing 11 of 11: FlightDate, Origin, Origin_Latitude, Origin_Longitude, Dest_Latitude, Dest_Longitude, Cancelled, DepDelayed, ArrDelayed, Dest, Flights

Chart type

VISUALIZATION TYPE: deck.gl Arc

Time

TIME COLUMN: FlightDate

TIME RANGE: previous calendar year

Query

START LONGITUDE & LATITUDE: Origin_Longitude | Origin_Latitude

END LONGITUDE & LATITUDE: Dest_Longitude | Dest_Latitude

ROW LIMIT: 5000

✓ IGNORE NULL LOCATIONS

FILTERS: Origin = 'SFO'

Flights from SFO

5k rows · 00:00:01.93 · .JSON · .CSV

Data

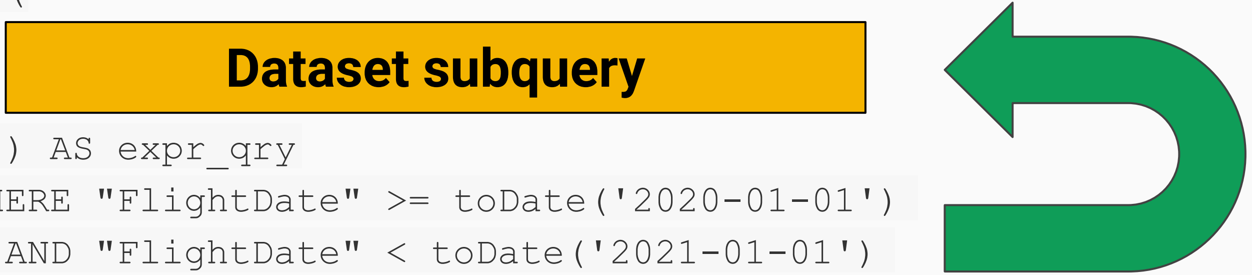
LTS · 5k rows retrieved · Search

Origin_Longitude	Dest_Latitude	Origin_Latitude	Dest_
-122.375	35.040199279785156	37.61899948120117	-106.
-122.375	40.97809982299805	37.61899948120117	-124.1

« 1 2 3 4 5 ... 100 »

How Superset queries virtual datasets

```
SELECT "Origin_Longitude" AS "Origin_Longitude",
       "Dest_Latitude" AS "Dest_Latitude",
       "Origin_Latitude" AS "Origin_Latitude",
       "Dest_Longitude" AS "Dest_Longitude"
FROM
  (
```



```
  ) AS expr_qry
WHERE "FlightDate" >= toDate('2020-01-01')
  AND "FlightDate" < toDate('2021-01-01')
  AND "Origin" = 'SFO'
  AND "Dest_Latitude" IS NOT NULL AND "Dest_Longitude" IS NOT NULL
  AND "Origin_Latitude" IS NOT NULL AND "Origin_Longitude" IS NOT NULL
LIMIT 5000;
```

Filters pushed down to base table

Creating Word Cloud chart

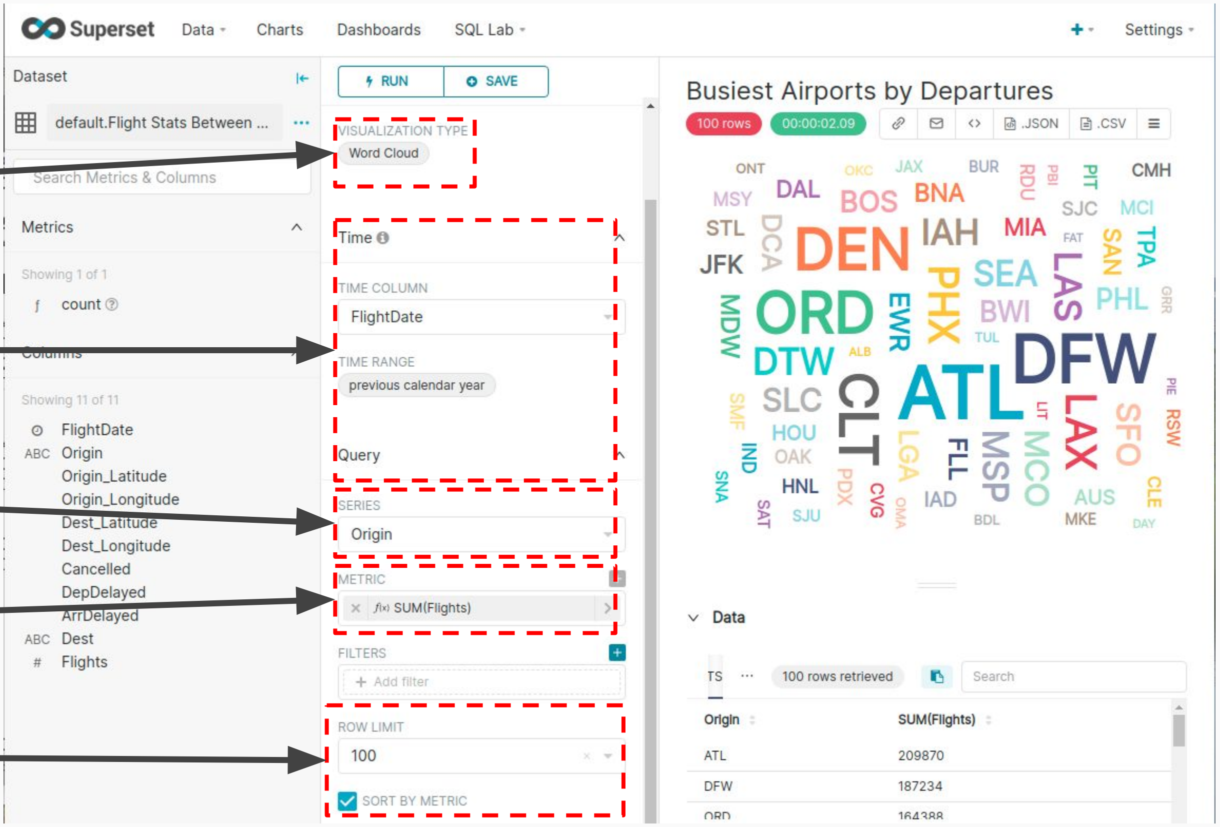
Chart Type

Time Dimension

Series

Metric

Limit & Sort Order



Columns: FlightDate, Origin, Origin_Latitude, Origin_Longitude, Dest_Latitude, Dest_Longitude, Cancelled, DepDelayed, ArrDelayed, Dest, Flights

VISUALIZATION TYPE: Word Cloud

Time

TIME COLUMN: FlightDate

TIME RANGE: previous calendar year

Query

SERIES: Origin

METRIC: SUM(Flights)

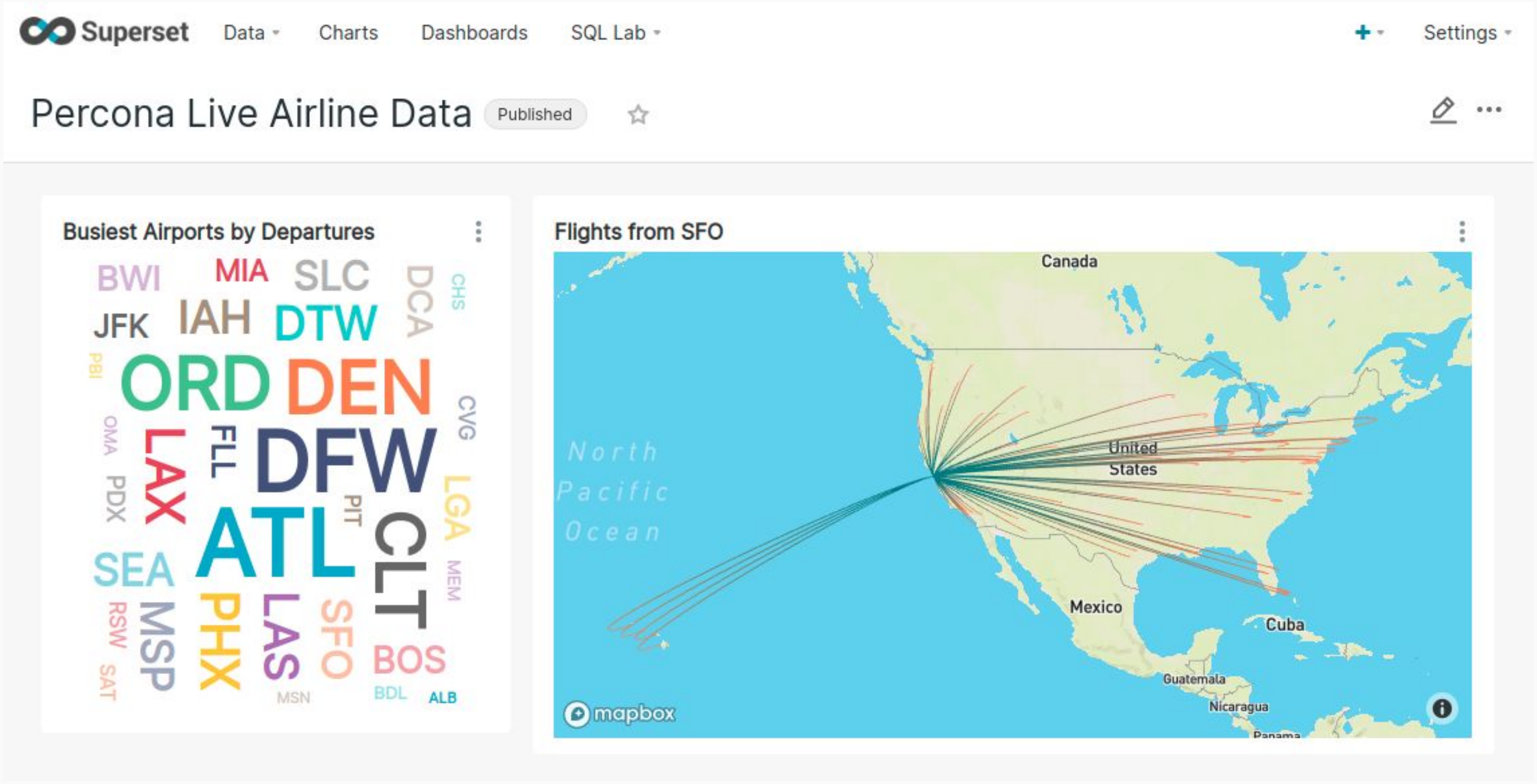
FILTERS: + Add filter

ROW LIMIT: 100

✓ SORT BY METRIC

Origin	SUM(Flights)
ATL	209870
DFW	187234
ORD	164388

Putting charts in a dashboard



Building from here

- There are *many* more charts available in Superset
- You can build complex dashboards
- Superset caching and clustering can amortize query overhead

Plans for the future

Roadmap for Superset and ClickHouse

- Improve clickhouse-sqlalchemy driver
  - There are still a few bugs :/
- Build out and document usage examples for users
  - Blog Post on Clickhouse <> Superset
- Full integration between Preset Cloud and Altinity.Cloud
  - Ensure both services available in common set of AWS regions
  - Automatic setup of AWS PrivateConnect between services

Questions?

Thank you!!

Srini Kadamati
Preset
https://preset.io

Robert Hodges
Altinity
https://altinity.com

Altinity preset