Altinity

Connect Python to ClickHouse Data!

R. Hodges -- SF Python Meetup

What's this ClickHouse data warehouse you keep talking about?

Understands SQL

Runs on bare metal to cloud

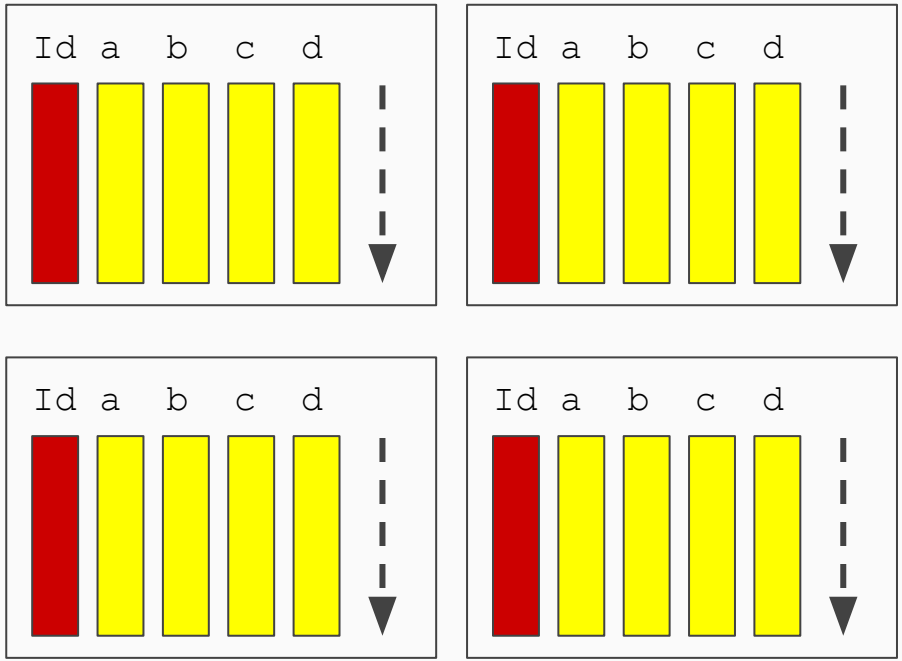
Stores data in columns

Runs queries in parallel

Scales to many petabytes

Is open source (Apache 2.0)

Is WAY fast!



What do we mean by “WAY fast”?

```
SELECT toYear(FlightDate) t,
    sum(Cancelled)/count(*) cancelled,
    sum(DepDel15)/count(*) delayed
FROM airline.ontime GROUP BY t ORDER BY t
```

```
┌────t─┬─────cancelled─────────┬─delayed──────────────┐
│ 1987 │ 0.01500580107422783 1 │   0.1584768101867168 3 │
│ 1988 │ 0.00964284396135711 5 │   0.1308220763323091 3 │
...
│ 2017 │  0.013998818968705 09 │   0.1947870137354604 8 │
└──────┴──────────────────────┴──────────────────────┘
```

31 rows in set. Elapsed: 0.402 sec. Processed 173.82 million rows, 1.74 GB (432.76 million rows/s., 4.33 GB/s.)

(Amazon md5.2xlarge: Xeon(R) Platinum 8175M, 8vCPU, 30GB RAM, NVMe SSD)

Let’s get that data into a Jupyter Notebook!

```
from sqlalchemy import create_engine
%load_ext sql
```

```
%%sql clickhouse://default:@localhost/airline
SELECT toYear(FlightDate) t,
    sum(Cancelled)/count(*) cancelled,
    sum(DepDel15)/count(*) delayed
FROM airline.ontime GROUP BY t ORDER BY t
```

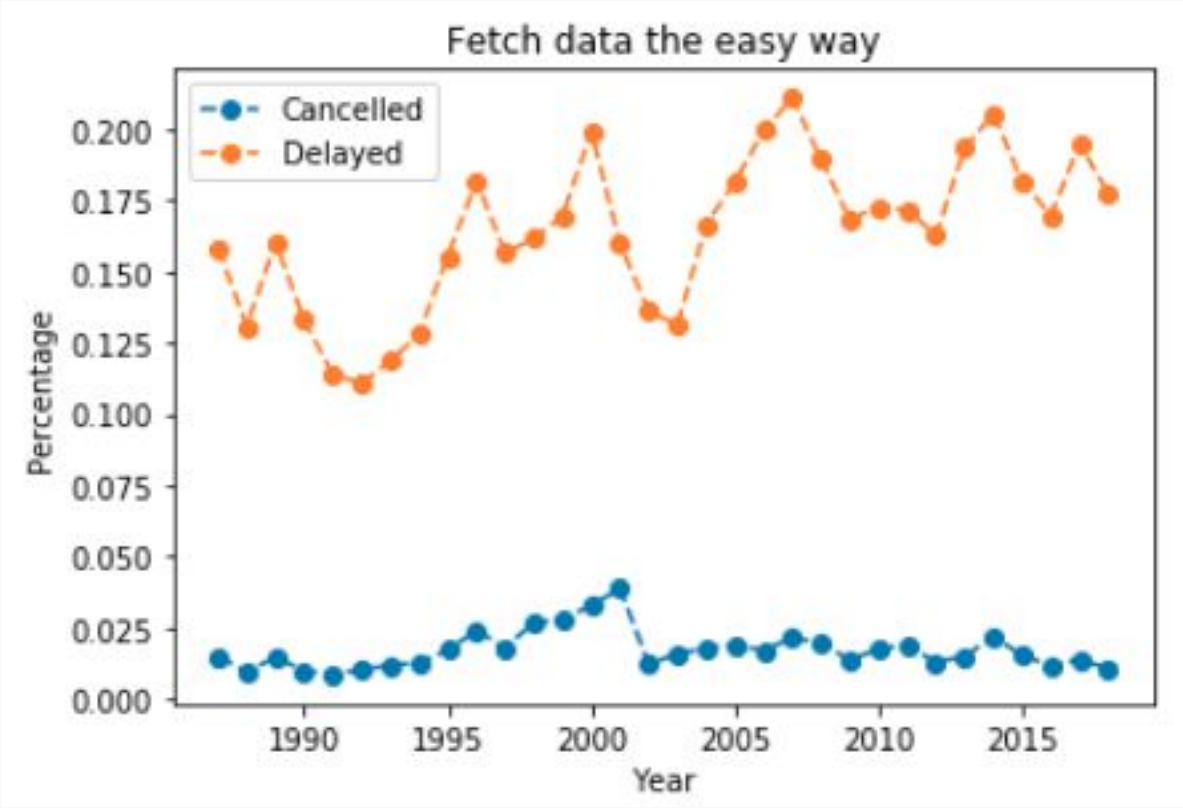
```
result = _
df = result.DataFrame()
df.tail()
```

Now we can make a nice graph

```
import matplotlib.pyplot as plt
%matplotlib inline

plt.plot('t', 'cancelled', data=df, linestyle='--',
       marker='o', label='Cancelled')
plt.plot('t', 'delayed', data=df, linestyle='--',
       marker='o', label='Delayed')
plt.xlabel('Year')
plt.ylabel('Percentage')
plt.legend(loc='upper left')
plt.title('Fetch data the easy way')
plt.show()
```

And here it is!



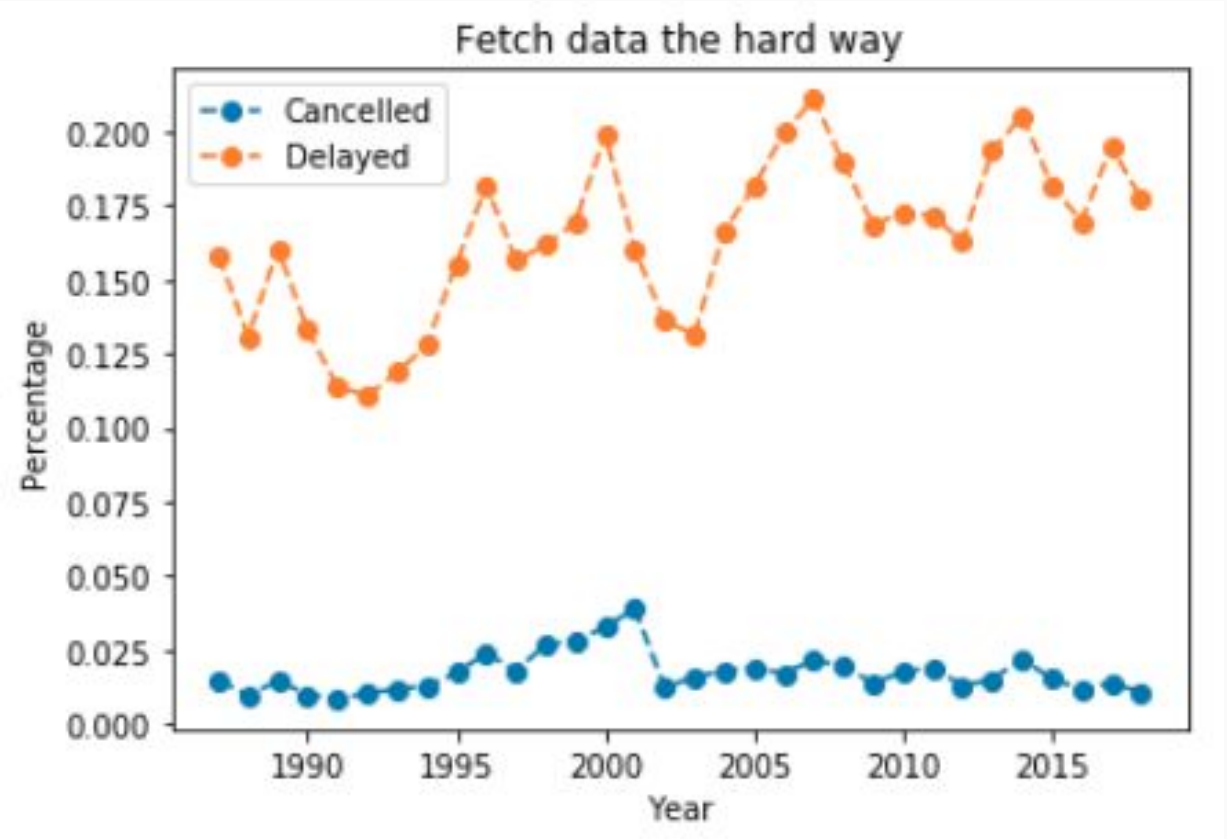
We can also make direct calls to the SQL API and pandas...

```
import pandas
from clickhouse_driver import Client

client = Client('localhost', database='airline')
result, columns = client.execute(
    'SELECT toYear(FlightDate) t,'
    'sum(Cancelled)/count(*) cancelled,'
    'sum(DepDel15)/count(*) delayed '
    'FROM airline.ontime GROUP BY t ORDER BY t',
    with_column_types=True)

df2 = pandas.DataFrame(result,
     columns=[tuple[0] for tuple in columns])
```

Now graph it again using df2 and a different title



More Information about Python and Clickhouse

ClickHouse Github Project -- https://github.com/yandex/ClickHouse

ClickHouse and Python: Jupyter Notebooks -- https://www.altinity.com/blog/2019/2/25/clickhouse-and-python-jupyter-notebooks

Python Code Samples -- https://github.com/Altinity/clickhouse-python-examples

Thank you!

Contacts:
info@altinity.com
rhodges@altinity.com

Visit us at:
https://www.altinity.com

Read Our Blog:
https://www.altinity.com/blog