Rideshare and Weather

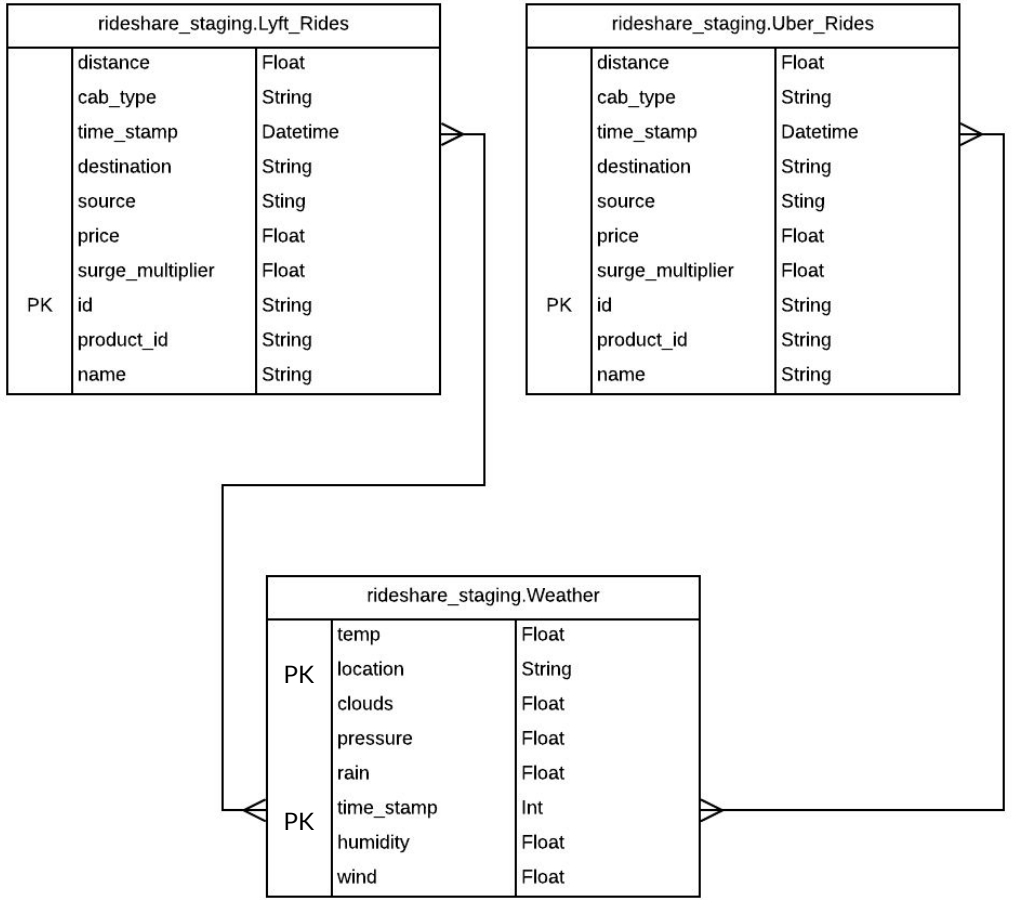
Jerry and Fernando, Uber_Fair

Problem Statement

Is there a relationship between weather and the price for the ride?

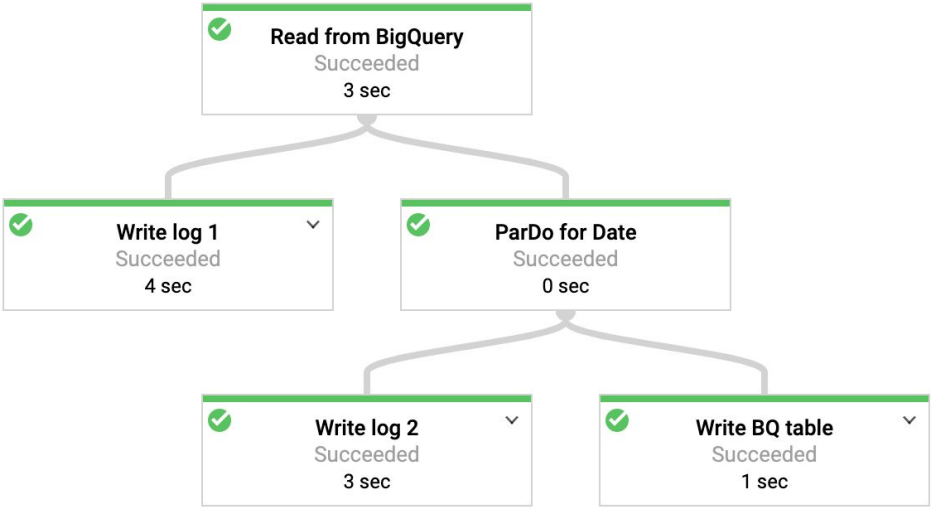
Dateset

- Ride share data from Uber and Lyft
- Weather data from NOAA
- Add IDs
- Connect the IDs
- Fixed time_stamp



Beam Pipeline

- Timestamp - Int
- Timestamp - datetime
- Fixed the timestamp inconsistent

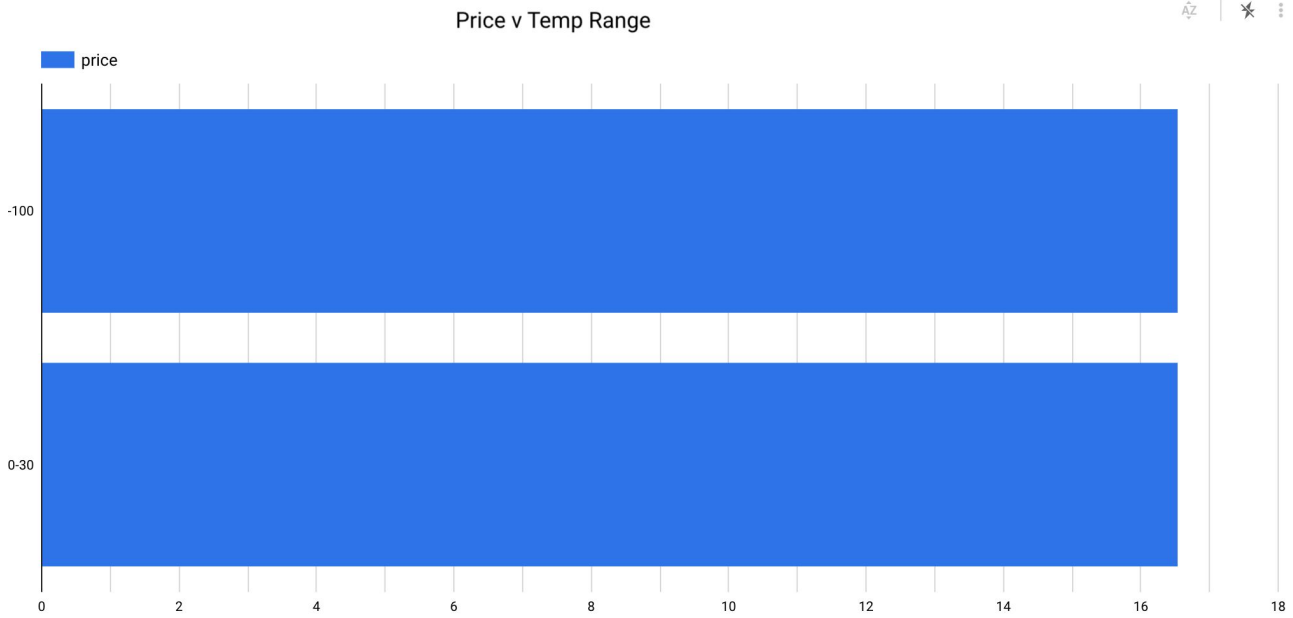


Airflow - DAG

- Uset Airflow DAG to create our second dataset modeled and staging datasets
- Loaded in the raw data
- Applied date transforms
- Applied ID generation
- Generated the modeled table with fixed values and present ID's

```
import datetime
from airflow import models
from airflow.operators.bash_operator import BashOperator
from airflow.operators.dummy_operator import DummyOperator

default_dag_args = {
    # https://airflow.apache.org/faq.html#what-s-the-deal-with-start-date
    'start_date': datetime.datetime(2019, 12, 07)
}

staging_dataset = 'weather_workflow_staging'
modeled_dataset = 'weather_workflow_modeled'
id_dataset = 'weather_modeled'

bq_query_start = 'bq query --use_legacy_sql=false '

create_modeled_sql = 'create or replace table ' + modeled_dataset + '''.weather as
                    SELECT T1.temp, T1.location, T1.clouds, T1.pressure, T1.rain, T1.time_stamp, T1.humidity, T1.wind, T2.id
                    FROM ''' + staging_dataset + '''.weather T1
                    JOIN ''' + id_dataset + '''.weather T2
                    ON T1.location = T2.location and T1.time_stamp = T2.time_stamp
                    ORDER BY 1, 2'''

with models.DAG(
        'weather_uberfair_workflow',
        schedule_interval=None,
        default_args=default_dag_args) as dag:

    create_staging_dataset = BashOperator(

            task_id='create_staging_dataset',
            bash_command='bq --location=US mk --dataset ' + staging_dataset)

    create_modeled_dataset = BashOperator(
            task_id='create_modeled_dataset',
            bash_command='bq --location=US mk --dataset ' + modeled_dataset)

    load_weather = BashOperator(
            task_id='load_weather',
            bash_command='bq --location=US load --autodetect --skip_leading_rows=1 \
                        --source_format=CSV ' + staging_dataset + '.weather \
                        "gs://uber_fair_data/dataset1/weather.csv"',
            trigger_rule='one_success')

    create_modeled = BashOperator(
            task_id='create_modeled',
            bash_command=bq_query_start + "'" + create_modeled_sql + "'",
            trigger_rule='one_success')


    create_staging_dataset >> create_modeled_dataset >> load_weather >> create_modeled
```

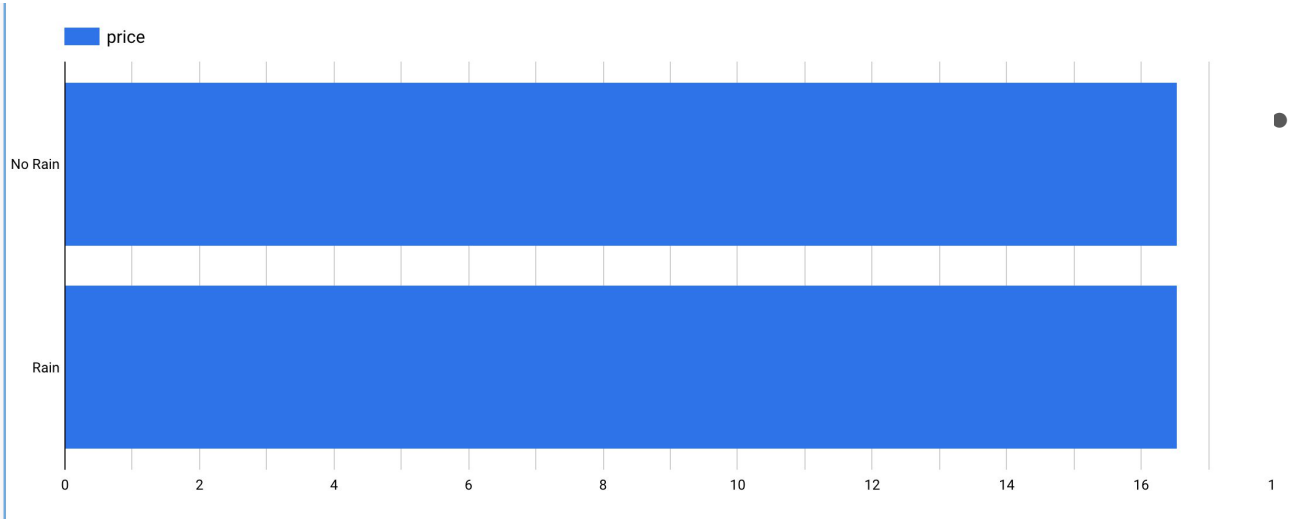
Temperature Comparison



- Initially did a groupby
- Switch to subquery, case solution

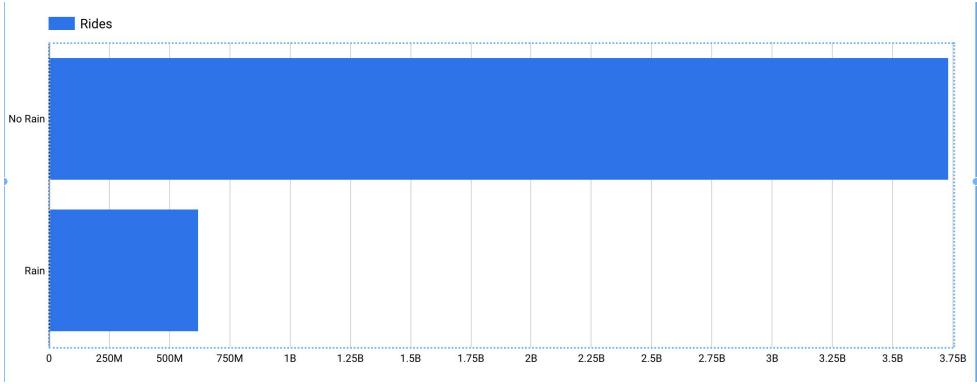
```
select t.temp_level , avg(r.price) price
from (select case
    when w.temp between 0 and 30 then
"0-10"
    when w.temp between 30 and 100 then
"30-100"
    else "End Range"
 end as temp_level
 from `uber-fair.weather_modeled.weather`
w)t, `uber-fair.rideshare_modeled.Rider` r
 group by t.temp_level
 order by t.temp_level
```

Rain Price Comparison



- No major price difference

Raining VS Not Raining



	No Rain	Rain
Total Rides	3730108122	619605474
Days	745	121
Average Rides	5006856	5120706

Future improves

- Wider range of data for the rides
  - Different cities
  - Track over time
- More detailed weather data
- Better Beam/Airflow transformation that allow us to use more data

Thank you Q&A