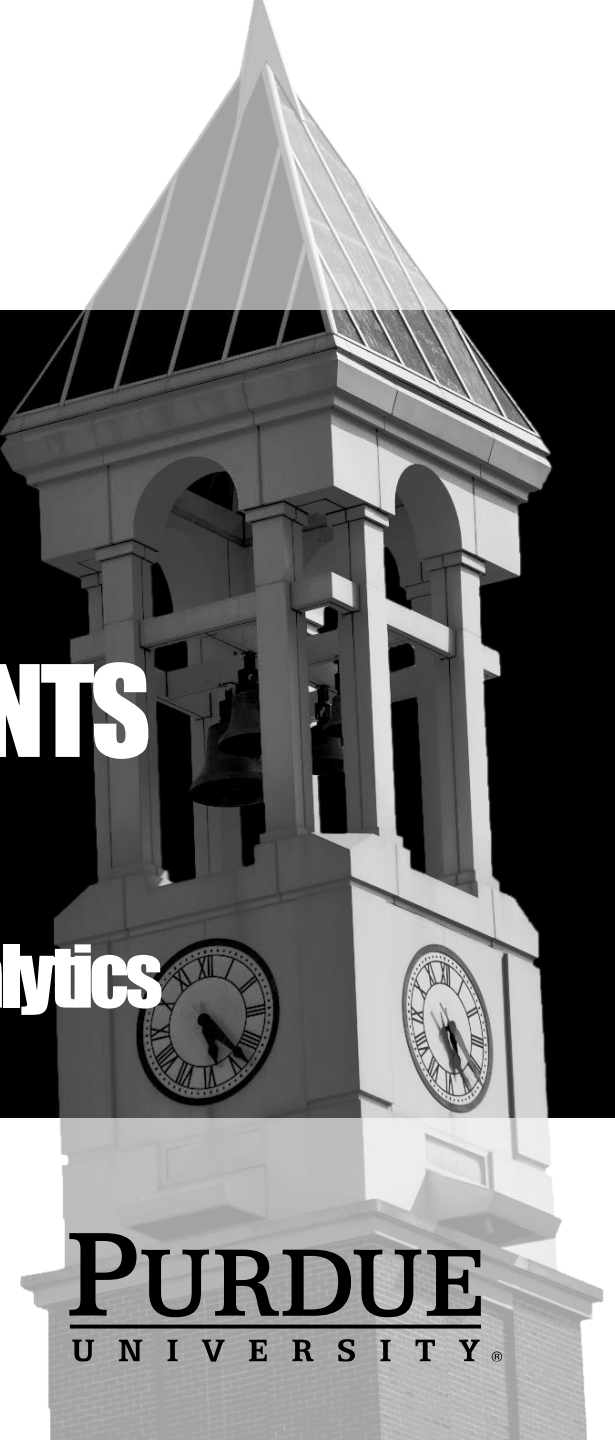
TEACHING ABCS TO BUSINESS STUDENTS

Undergraduate core and Masters in Business Analytics

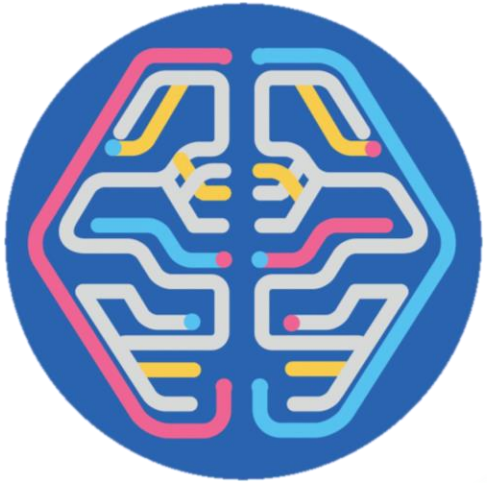


ALOK CHATURVEDI, ALOK@PURDUE.EDU

PURDUE UNIVERSITY®

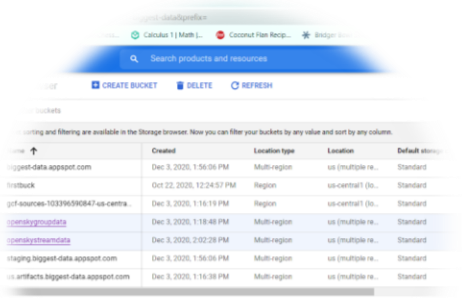
Teaching ABCs

Artificial Intelligence



Graduate BAIM

Big Data



Undergraduate Core

Cloud Computing

Undergraduate MIS core Course (MGMT 382)

- Objectives
  - Inspire students to be **fearless**
  - Overcome **Technophobia**
  - Develop **learning to learn** skills
  - Learn tangible skills that can improve job prospects
    - Google Cloud platform

Cloud Storage

BigQuery

Auto ML

Data Studio

ABC: ML in BigQuery

Create

```
#standardsql
CREATE OR REPLACE MODEL `bts_data.ontime`
OPTIONS
  (model_type='logistic_reg', input_label_cols=['on_time']) AS
SELECT
  IF(arr_delay < 15, 1, 0) AS on_time,
  carrier, origin, dest, dep_delay, taxi_out, distance
FROM
  `datapipe-1.bts_data.flights_unpart`
WHERE
  arr_delay IS NOT NULL
```

Evaluate

```
#standardsql
SELECT * FROM ML.EVALUATE(MODEL
`bts_data.ontime`,
(
SELECT
  IF(arr_delay < 15, 1, 0) AS on_time,
  carrier, origin, dest, dep_delay, taxi_out, distance
FROM
  `datapipe-1.bts_data.flights_unpart`
WHERE
  arr_delay IS NOT NULL
))
```

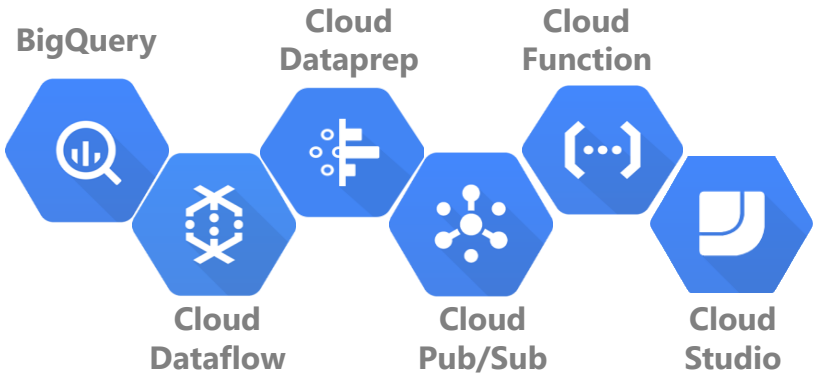
Create

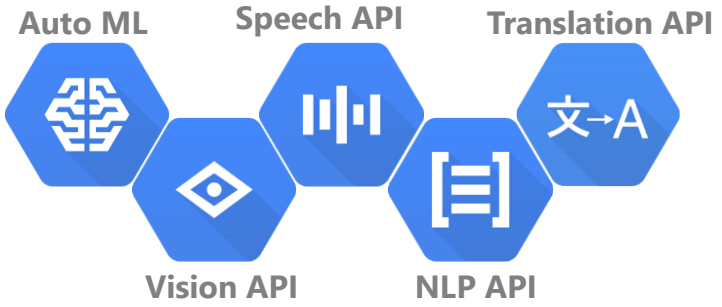
```
SELECT * FROM ml.PREDICT(MODEL `bts_data.ontime`, (
SELECT
  'AA' as carrier,  'DFW' as origin, 'LAX' as dest, dep_delay,
  18 as taxi_out, 1235 as distance
FROM
  UNNEST(GENERATE_ARRAY(-3, 10)) as dep_delay
))
```

Teaching ABCs to U/G core student

MS BAIM Big Data Course (MGMT 590)

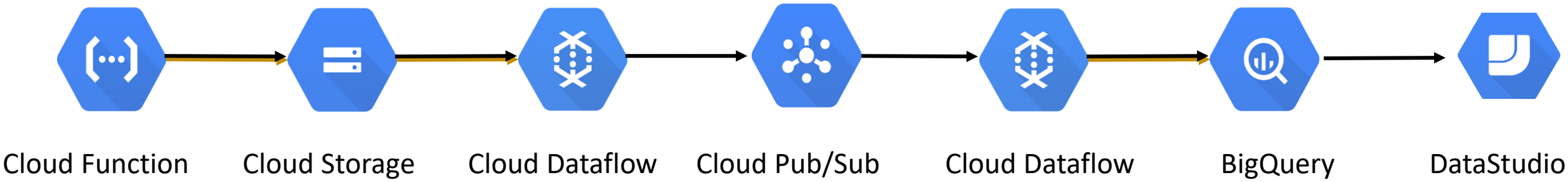
- Objectives
  - Inspire students to be **fearless**
  - Develop **learning to learn** skills
  - Learn to build **end-to-end complex data pipelines**





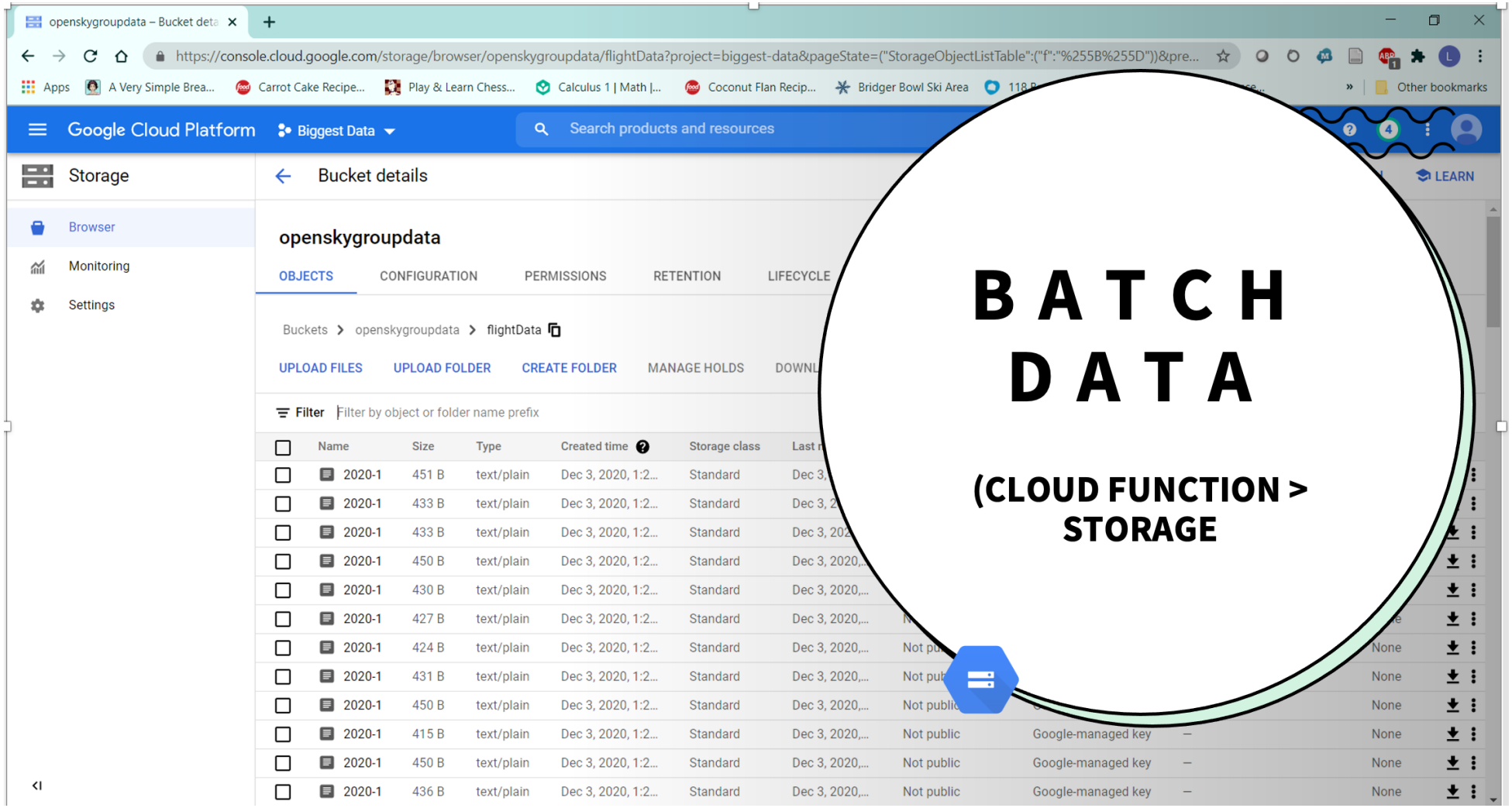
Building Data Pipeline

Batch Data Pipeline



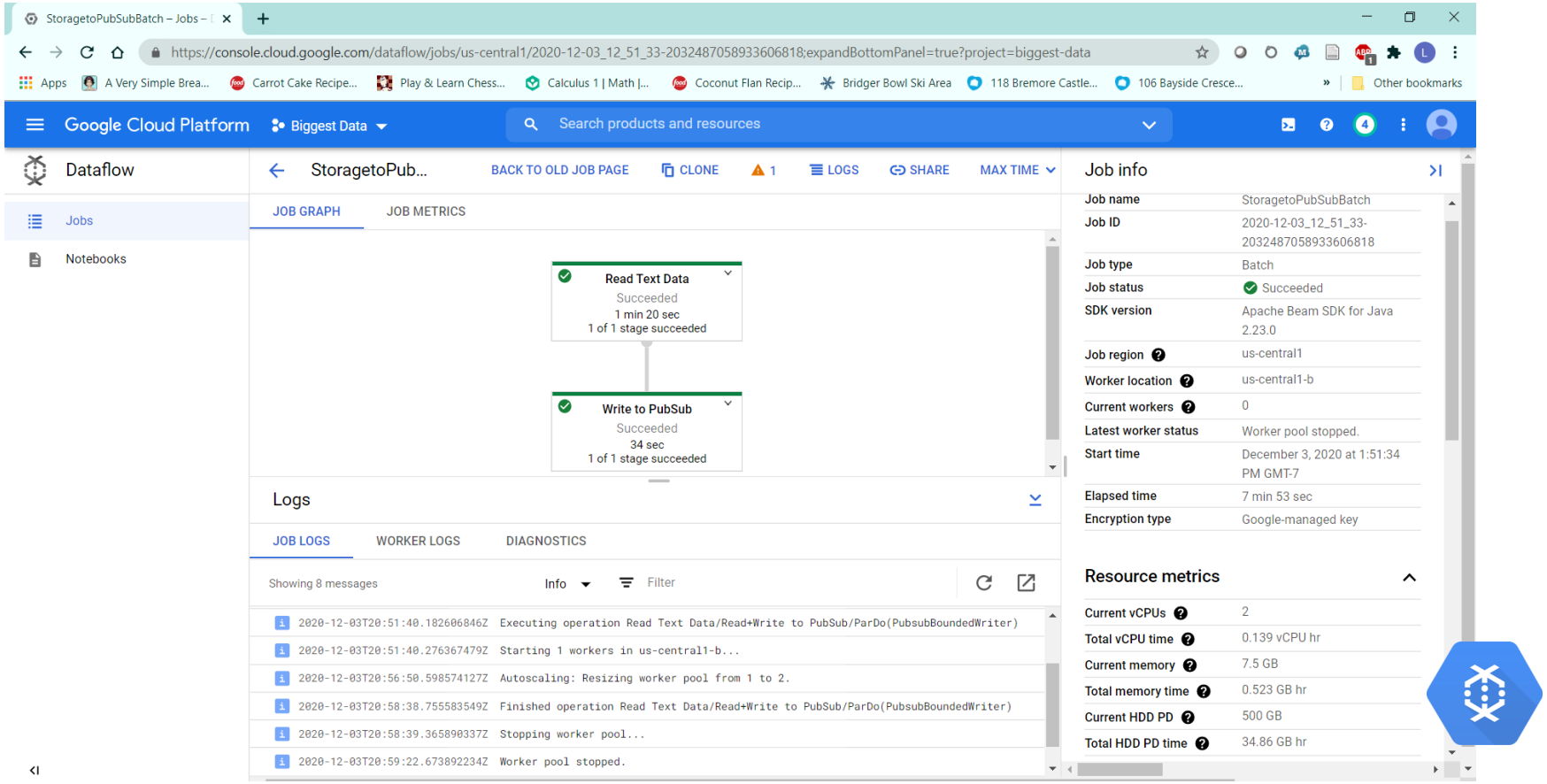
Building Data Pipeline

Batch Data Pipeline



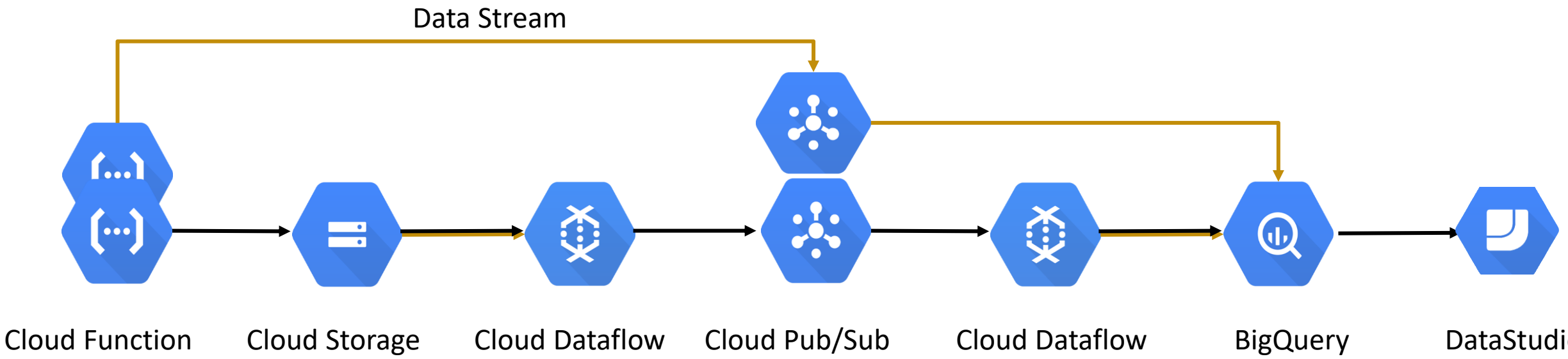
Building Data Pipeline

Batch Data Pipeline



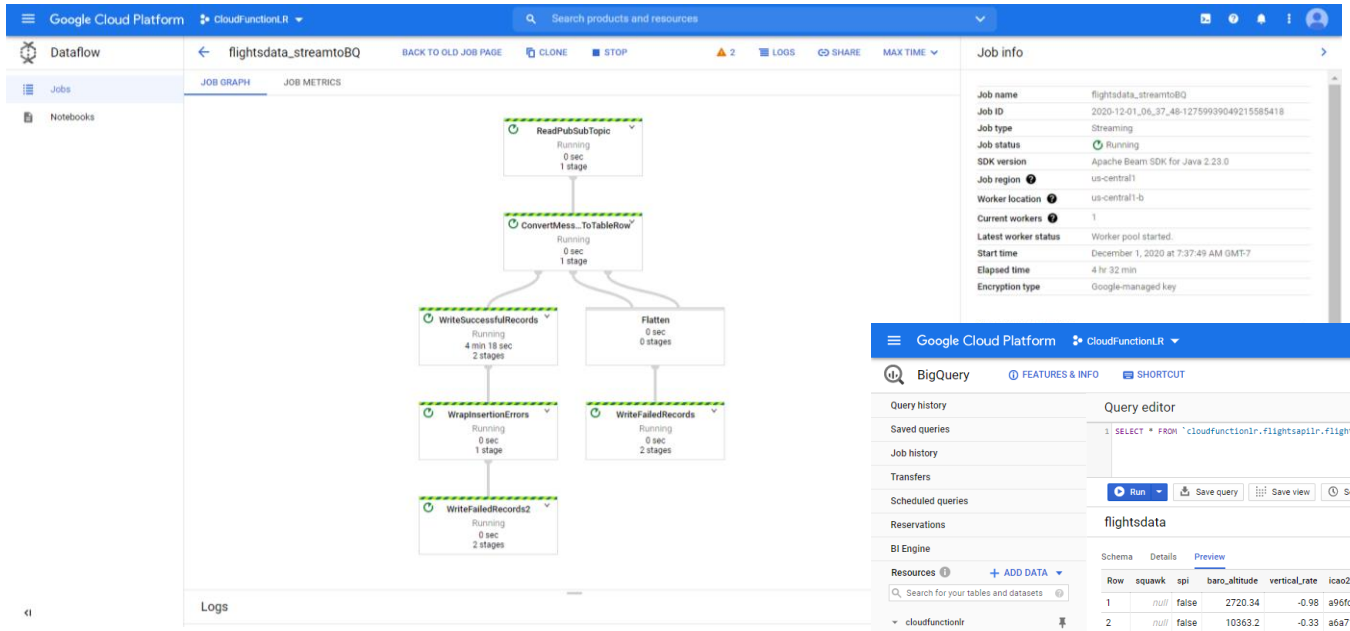
Building Data Pipeline

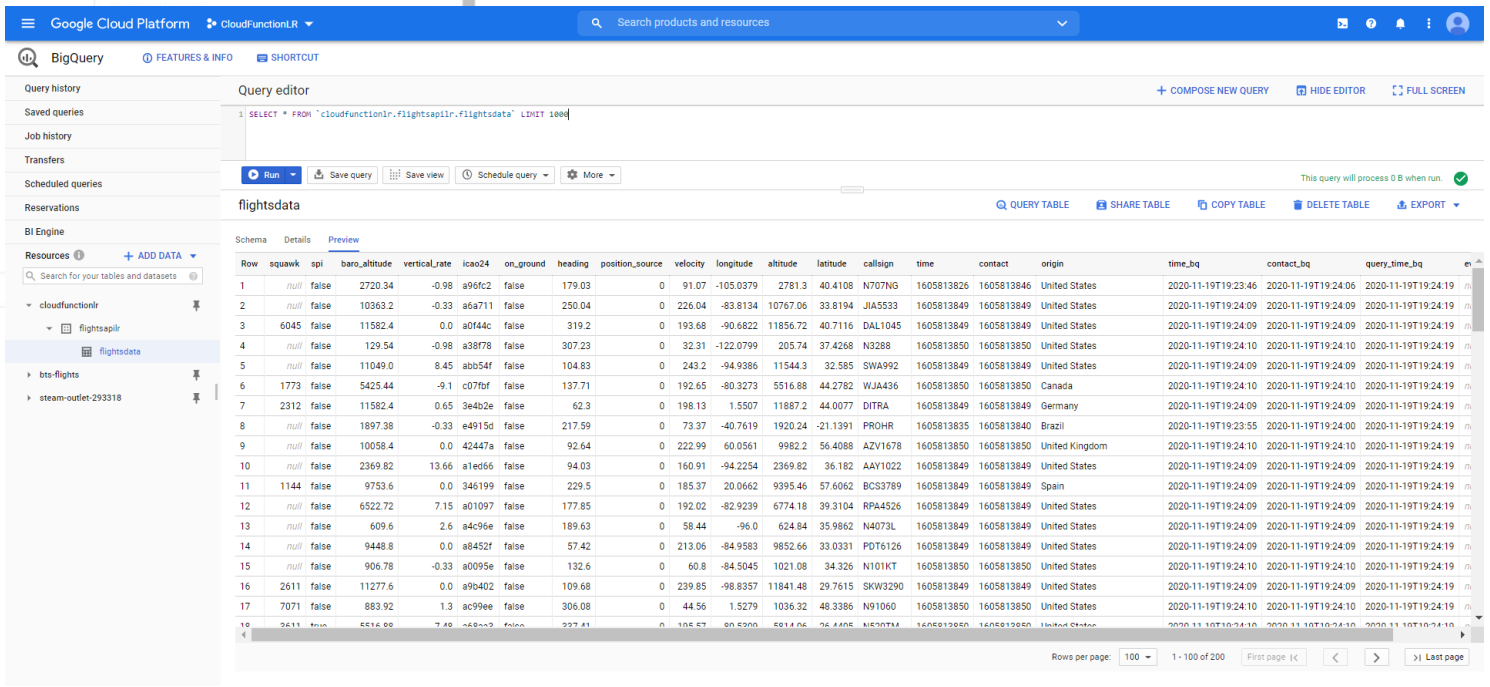
Streaming Data Pipeline



Building Data Pipeline

Streaming Data Pipeline





Row	squawk	spi	baro_altitude	vertical_rate	icao24	on_ground	heading	position_source	velocity	longitude	altitude	latitude	callsign	time	contact	origin	time_bq	contact_bq	query_time_bq
1	null	false	2720.34	-0.98	a96fc2	false	179.03	0	91.07	-105.0379	2781.3	40.4108	N707NG	1605813826	1605813846	United States	2020-11-19T19:23:46	2020-11-19T19:24:06	2020-11-19T19:24:19
2	null	false	10363.2	-0.33	a6a711	false	250.04	0	226.04	-83.8134	10767.06	33.8194	JIA5533	1605813849	1605813849	United States	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
3	6045	false	11582.4	0.0	a0f44c	false	319.2	0	193.68	-90.6822	11856.72	40.7116	DAL1045	1605813849	1605813849	United States	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
4	null	false	129.54	-0.98	a38f78	false	307.23	0	32.31	-122.0799	205.74	37.4268	N3288	1605813850	1605813850	United States	2020-11-19T19:24:10	2020-11-19T19:24:10	2020-11-19T19:24:19
5	null	false	11049.0	8.45	abb54f	false	104.83	0	243.2	-94.9386	11544.3	32.585	SWA992	1605813849	1605813849	United States	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
6	1773	false	5425.44	-9.1	c07fbf	false	137.71	0	192.65	-80.3273	5516.88	44.2782	WJA436	1605813850	1605813850	Canada	2020-11-19T19:24:10	2020-11-19T19:24:10	2020-11-19T19:24:19
7	2312	false	11582.4	0.65	3e4b2e	false	62.3	0	198.13	1.5507	11887.2	44.0077	DITRA	1605813849	1605813849	Germany	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
8	null	false	1897.38	-0.33	e4915d	false	217.59	0	73.37	-40.7619	1920.24	-21.1391	PROHR	1605813835	1605813840	Brazil	2020-11-19T19:23:55	2020-11-19T19:24:00	2020-11-19T19:24:19
9	null	false	10058.4	0.0	42447a	false	92.64	0	222.99	60.0561	9982.2	56.4088	AZV1678	1605813850	1605813850	United Kingdom	2020-11-19T19:24:10	2020-11-19T19:24:10	2020-11-19T19:24:19
10	null	false	2369.82	13.66	a1ed66	false	94.03	0	160.91	-94.2254	2369.82	36.182	AAY1022	1605813849	1605813849	United States	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
11	1144	false	9753.6	0.0	346199	false	229.5	0	185.37	20.0662	9395.46	57.6062	BCS3789	1605813849	1605813849	Spain	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
12	null	false	6522.72	7.15	a01097	false	177.85	0	192.02	-82.9239	6774.18	39.3104	RPA4526	1605813849	1605813849	United States	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
13	null	false	609.6	2.6	a4c96e	false	189.63	0	58.44	-96.0	624.84	35.9862	N4073L	1605813849	1605813849	United States	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
14	null	false	9448.8	0.0	a8452f	false	57.42	0	213.06	-84.9583	9852.66	33.0331	PDT6126	1605813849	1605813849	United States	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
15	null	false	906.78	-0.33	a0095e	false	132.6	0	60.8	-84.5045	1021.08	34.326	N101KT	1605813850	1605813850	United States	2020-11-19T19:24:10	2020-11-19T19:24:10	2020-11-19T19:24:19
16	2611	false	11277.6	0.0	a9b402	false	109.68	0	239.85	-98.8357	11841.48	29.7615	SKW3290	1605813849	1605813849	United States	2020-11-19T19:24:09	2020-11-19T19:24:09	2020-11-19T19:24:19
17	7071	false	883.92	1.3	ac99ee	false	306.08	0	44.56	1.5279	1036.32	48.3386	N91060	1605813850	1605813850	United States	2020-11-19T19:24:10	2020-11-19T19:24:10	2020-11-19T19:24:19

ML ANALYSIS ON BATCH AND STREAMING DATA

PURDUE UNIVERSITY

Krannert School of Management

• Big Query: ML Modeling

Question 1: Most major airlines operate across the globe and have agreements with other regional airlines that further extend their reach. Still, each airline has a geographic area that they operate within the most. But which airlines operate in similar spaces?

Question 2: How predictable are itineraries? Can we tell the origin of a flight from the airline, time and day of week?

Before modeling, let's add two fields, one for the Airline code and one for the day of the week, using the following SQL expression:

```
SELECT *, regexp_replace(left(callsign,3),"[0-9]","") AS airline, extract(DAYOFWEEK
from contact_bq) AS weekday
FROM  `flightsapilr.flightsdata`
```

Let's store this as query "fightsdata2"

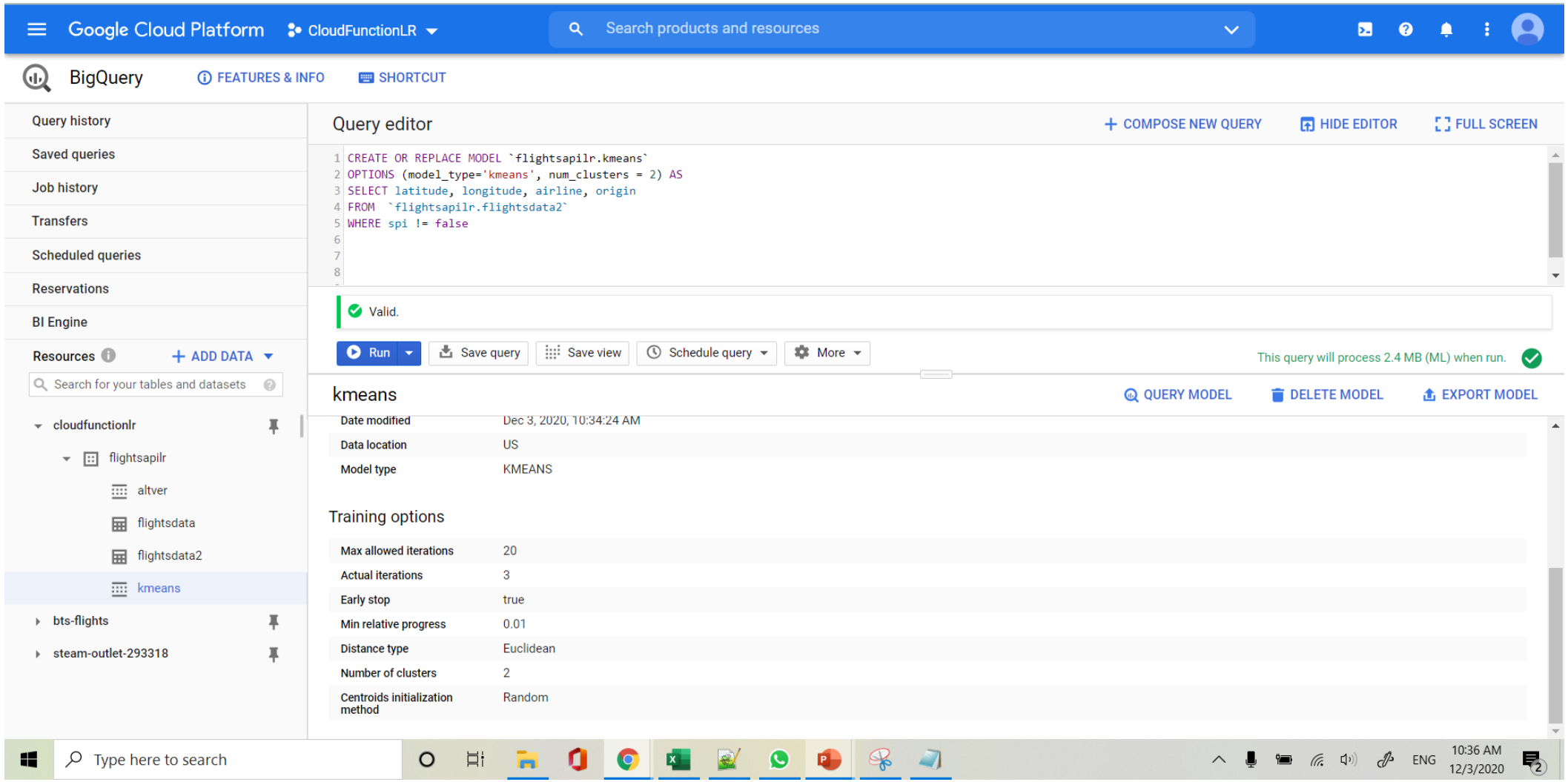
- Big Query: ML Clustering

We can build a k means clustering ML model to answer Question 1:

```
CREATE OR REPLACE MODEL `flightsapilr.kmeans`
OPTIONS (model_type='kmeans', num_clusters = 2) AS
SELECT latitude, longitude, airline, origin
FROM  `flightsapilr.flightsdata2`
WHERE spi != false
```

The number of clusters was selected through iterative trials to minimize the Davies-Bouldin index and MSD and maximize cluster interpretability. The values tried were integers 2 through 7.

Big Query: ML Clustering



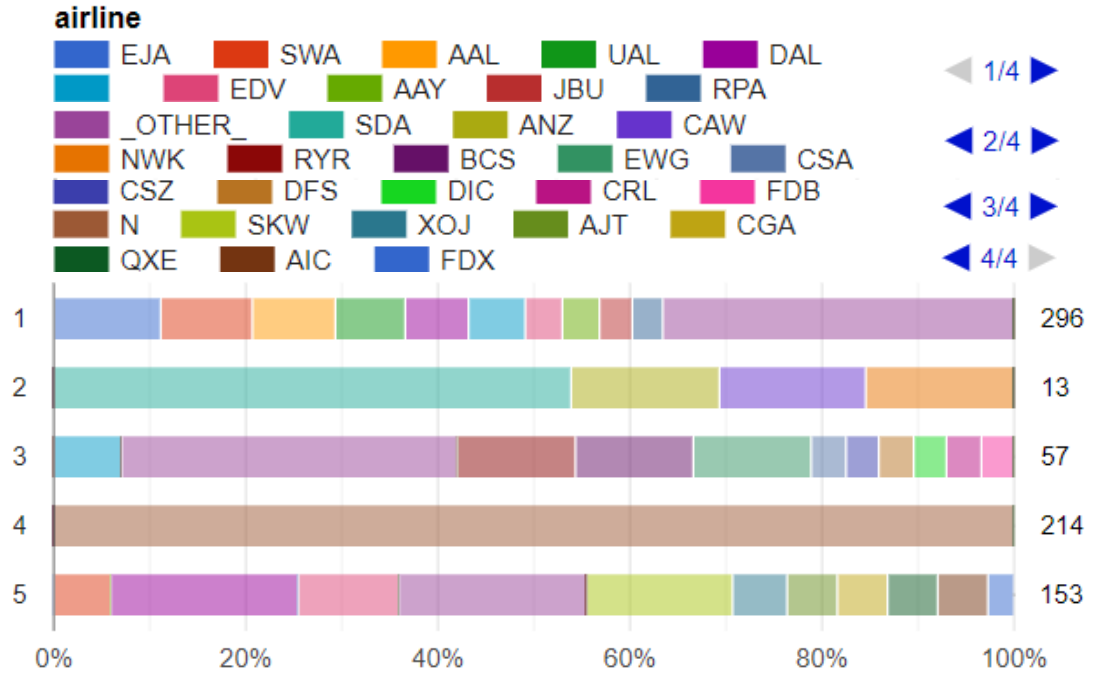
- Big Query: ML Clustering

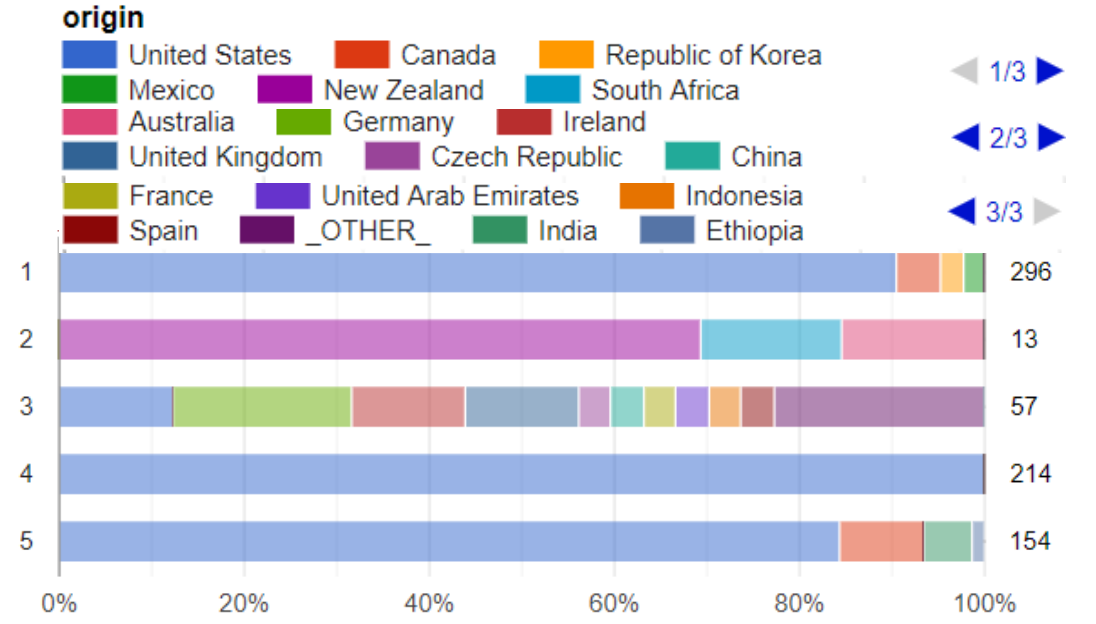
The results from building this model are:

Metrics

Davies–Bouldin index	1.7234
Mean squared distance	1.2103

Centroid Id	Count	latitude	longitude
1	296	35.1396	-92.3316
2	13	-36.9986	143.0028
3	57	41.4898	35.1548
4	214	37.8800	-92.9642
5	154	44.2159	-98.0174





Summary

- Challenge the students
- If you are fearless, students will be fearless
- Focus on learning to learn
- Zoom actually helps in peer-to-peer as you can ask the students to share their screens
- Working with large, real world data helps UG students grow cognitively
- Masters students appreciate what it takes to prepare data before you can use sexy ML models

References

- The Building Blocks of a Modern Data Platform
  - https://towardsdatascience.com/the-building-blocks-of-a-modern-data-platform-92e46061165
- Features in BigQuery's New UI for 2021
  - https://towardsdatascience.com/5-great-features-in-bigquerys-new-ui-for-2021-yes-it-has-tabs-c4bac66d66b
- BigQuery SQL Cheat Sheet
  - https://medium.com/data-school/the-best-bigquery-sql-cheat-sheet-for-beginners-81c762f72845
- BigQuery — Almost All You Need to Know
  - https://medium.com/swlh/bigquery-almost-all-you-need-to-know-f239e6b52279
- Working with Joins, Nested & Repeated Data
  - https://medium.com/google-cloud/bigquery-explained-working-with-joins-nested-repeated-data-1941646ccb5b
- Bigquery ML
  - https://towardsdatascience.com/search?q=Bigquery%20ML