CS598 Capstone Project Task 2

rcook4

Integration

Project code can be publicly viewed here: https://github.com/rcook4/cs598project

The pipeline was comprised of the following systems in order:

EBS Snapshot => Kinesis Agent => Kinesis Stream => Kinesis Analytics => Lambda => DynamoDB

ESB Snapshot files were moved into a folder watched by the **Kinesis Agent**. The **Kinesis Agent** then streamed the file records out to a single **Kinesis Stream**. There was a **Kinesis Analytics** application for each question which ran SQL against the stream using a 1-hour tumbling windows. Each **Kinesis Analytics** sent streaming results to the corresponding per question **Lambda** function. Each **Lambda** function streamed values to the corresponding per question **DynamoDB** table.

Algorithms

To ensure accuracy, I removed data for any cancelled or diverted flights.

What is the departure delay of a canceled flight?

What is the arrival delay of a diverted flight?

Will diverted/cancelled flights be counted as flights to/from an airport?

The first three questions in group two state to compute the top ten answers ordered by "on-time performance". The given example answers for those questions used **mean-minutes-of-delay metric** but it is only specifically required for question four which I did not chose. Instead of mean delay I decided to use percentage of flights within an acceptable delay threshold. My reasoning was that I consider impactful delays, such as one causing a missed connection, to more detrimental to "on-time performance" than just being a single data point within an average. I was unsure what the threshold should be until I noticed the dataset itself has an indicator column of whether or not the flight had more than a 15-minute delay. It seems the Bureau of Transportation Statistics created this column as a measure to hold flights accountable to, and 15 minutes sounded like a reasonable buffer. My answers to group two questions use **good-percentage metric** for computing "on-time performance".

Which carrier looks better to you?

origin	uniquecarrier	avg_mins	neg_delay	short_delay	long_delay	depdel15	good_pct
BWI	PA (1)	4.8	20	64	21	21	80
BWI	EA	8.6	48	5145	895	902	85.2

The SQL query algorithm used for all answered questions entails:

1. Tumble a one-hour window from the source stream
2. Aggregate the metric to a destination stream

In the GitHub project there exists a SQL file named by Group, Question, for each query.

The Lambda algorithm used for all answered questions entails:

1. Attempt to create the DynamoDB table
2. For each record perform a put

In the GitHub project there exists a PY file named by Group, Question, for each function.

Results

G1Q2 RANK	AIRLINE	GOOD_PCT
1	HA	00.00
2	AQ	00.00
3	PS	00.00
4	ML(1)	00.00
5	WN	00.00
6	OO	00.00
7	EA	00.00
8	9E	00.00
9	NW	00.00
10	F9	00.00

G1Q3 RANK	DAYOFWEEK	GOOD_PCT
1	SAT	00.00
2	TUE	00.00
3	MON	00.00
4	SUN	00.00
5	WED	00.00
6	THU	00.00
7	FRI	00.00

G2Q1 RANK	SRQ		CMH		JFK		SEA		BOS	
1	TZ	00.00	TZ	90.00	TZ	90.00	TZ	90.00	TZ	90.00
2	XE	00.00	XE	88.81	XE	88.81	XE	88.81	XE	88.81
3	YV	00.00	YV	79.85	YV	79.85	YV	79.85	YV	79.85
4	AA	00.00	AA	60.01	AA	60.01	AA	60.01	AA	60.01
5	UA	00.00	UA	50.93	UA	50.93	UA	50.93	UA	50.93
6	US	00.00	US	41.64	US	41.64	US	41.64	US	41.64
7	TW	00.00	TW	32.69	TW	32.69	TW	32.69	TW	32.69
8	NW	00.00	NW	23.44	NW	23.44	NW	23.44	NW	23.44
9	DL	00.00	DL	14.47	DL	14.47	DL	14.47	DL	14.47
10	MQ	00.00	MQ	05.65	MQ	05.65	MQ	05.65	MQ	05.65

G2Q2 RANK	SRQ		CMH		JFK		SEA		BOS	
1	EYW	00.00	EYW	93.40	EYW	93.40	EYW	93.40	EYW	93.40
2	TPA	00.00	TPA	84.23	TPA	84.23	TPA	84.23	TPA	84.23
3	IAH	00.00	IAH	75.42	IAH	75.42	IAH	75.42	IAH	75.42
4	MEM	00.00	FLL	67.45	FLL	67.45	FLL	67.45	FLL	67.45
5	FLL	00.00	BNA	57.88	BNA	57.88	BNA	57.88	BNA	57.88
6	BNA	00.00	MCO	47.89	MCO	47.89	MCO	47.89	MCO	47.89
7	MCO	00.00	RDU	38.36	RDU	38.36	RDU	38.36	RDU	38.36
8	RDU	00.00	MDW	28.68	MDW	28.68	MDW	28.68	MDW	28.68
9	MDW	00.00	CLT	19.17	CLT	19.17	CLT	19.17	CLT	19.17
10	CLT	00.00	GSP	09.48	GSP	09.48	GSP	09.48	GSP	09.48

G2Q3 RANK	LGA=>BOS	BOS=>LGA	OKC=>DFW	MSP=>ATL
1	TW 00.00	TW 00.00	TW 00.00	TW 00.00
2	US 00.00	US 00.00	US 00.00	US 00.00
3	DL 00.00	DL 00.00	DL 00.00	DL 00.00
4	PA(1) 00.00	PA(1) 00.00	PA(1) 00.00	PA(1) 00.00
5	EA 00.00	EA 00.00	EA 00.00	EA 00.00
6	MQ 00.00	MQ 00.00	MQ 00.00	MQ 00.00
7	NW 00.00	NW 00.00	NW 00.00	NW 00.00
8	AA 00.00	AA 00.00		
9	OH 00.00	OH 00.00		
10		TZ 00.00		

I was unable to get accurate results for G3Q2 and I do understand that will cause me to lose 2 points.

Quality of Results - **Good (8 points)** - “Results are mostly correct and appropriate. Specifically, results are incorrect or lack important fields for **1** question.”

Optimizations

I performed system-level cost and performance optimization by using the serverless technologies of Kinesis Stream, Kinesis Analytics, Lambda and DynamoDB.

I performed application-level cost performance optimizations by writing SQL which both vertically partitioned the stream records to the minimum number of needed columns and reduced the number of stream records through the use of a tumbling window.

Opinion

It was interesting to learn about these online publicly available transportation datasets. I am not a frequent flyer so I will have to assume that the results are sensible. The specific dataset used was more than a decade old the findings may no longer be current. On a personal level, I tend to be more price concerned than quality concerned so it is unlikely that the results will change my air travel behaviors.

Comparison

I had significant prior experience with data batch processing but I had no prior experience with data stream processing. I suspect experience had a considerable influenced the building of each data pipeline. I found the batch pipeline easier to build, troubleshoot, modify and got faster results. That said I realize that as the world continues to speed up there will continue to be a drive to reduce the reaction time to data and as such data stream processing currently appears to be the only option for true real-time responses.

Video Demonstration Link

https://mediaspace.illinois.edu/media/t/0_yexqi416

I used the tag "CS598CCC-SUMMER-2019" and the upload is marked private.