LAB1

Challenge questions

Now that you know how to load files to Amazon S3, try the following exercises to test your knowledge. Your lab instructor has model solutions. However, some of the exercises have more than one way to solve them.

Challenge one

The taxi data includes data for two different vendors. The **vendorid** field has two possible values: *1* and 2. Write a query to count the number of rides for vendor 1.

Challenge two

The taxi data includes data for payment type. Payment type *1* is for payments that are made by credit card. Write a query to total the number of trips that were paid for by credit card.

exclude first line 구문을 키는 것을 늦게 확인해서 시간을 많이 소비했다.

Challenge one

SQL query

Amazon S3 Select supports only the SELECT SQL command. Using the S3 console, you can extract up to 40 MB of recor
from an object that is up to 128 MB in size. To work with larger files or more records, use the AWS CLI, AWS SDK, or Am
S3 REST API. For more complex SQL queries, use **Amazon Athena**

Add SQL from templates | Run SQL query

```
1  /* To create reference point for writing SQL queries, you can display the first 5 records of ir
     data by running the following SQL query: SELECT * FROM s3object s LIMIT 5 */
2  SELECT count(*) FROM s3object WHERE VendorID='1'
```

Query results

Query results are not available after you choose **Close** or navigate away. Choose **Download results** to download a copy
the following query results.

Download results

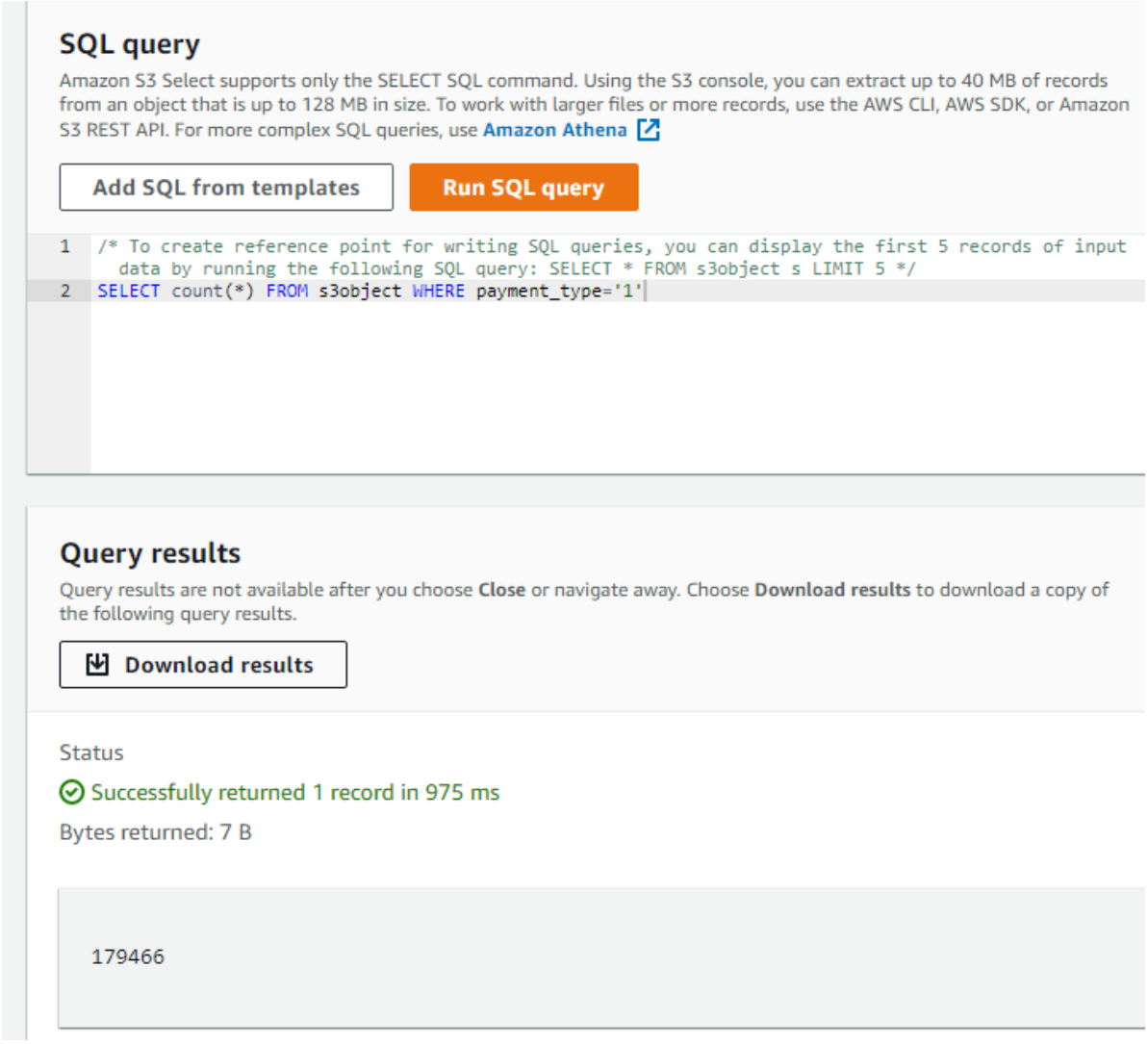
Status

Successfully returned 1 record in 1042 ms

Bytes returned: 7 B

```
123047
```

Chanllenge two



LAB2

Challenge one

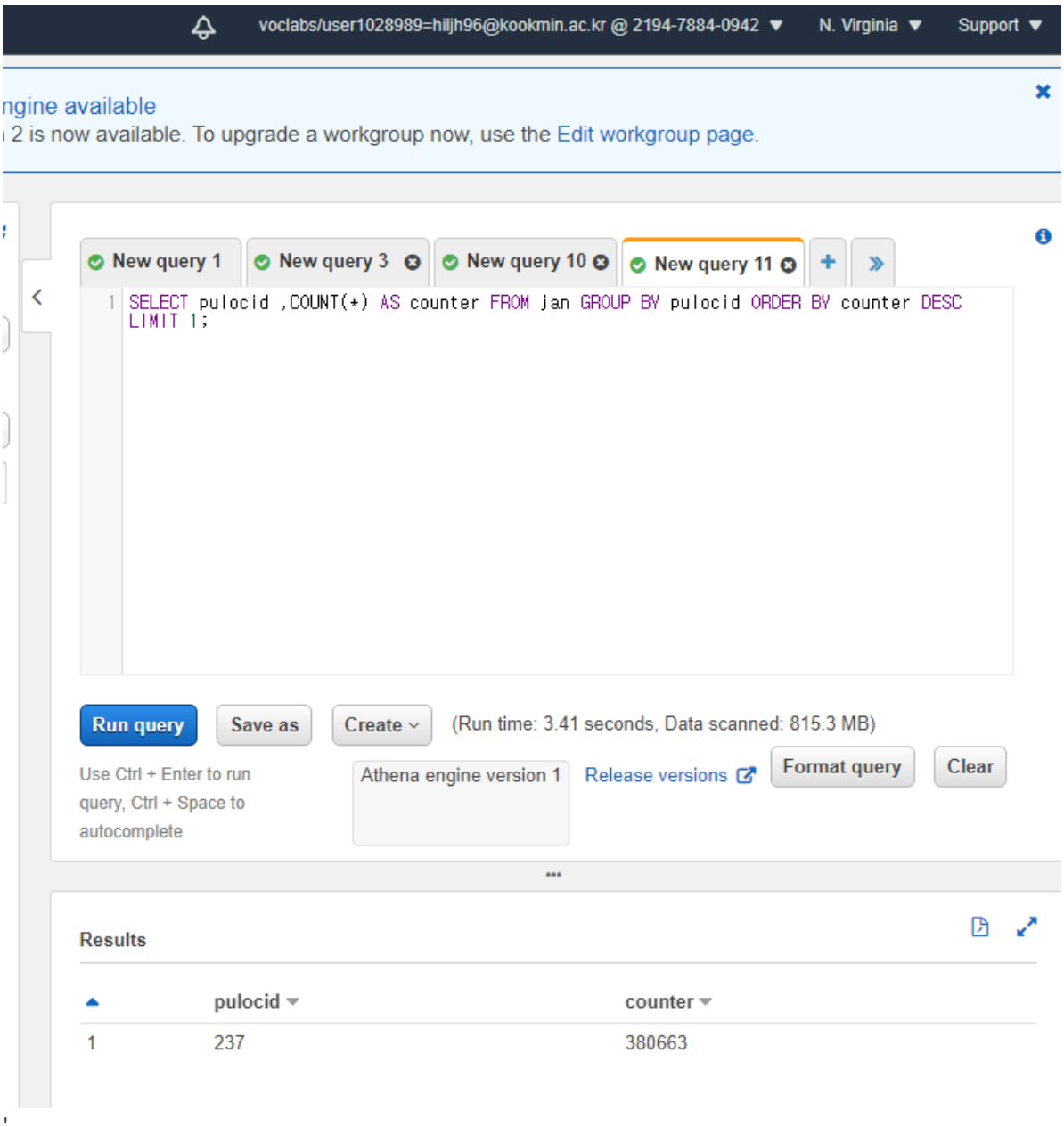
Write a query that identifies the most common pickup location in January 2017.

Challenge two

Write a query to compare the average distance for trips that were paid with credit cards and the average distance for trips that were paid with cash in January 2017.

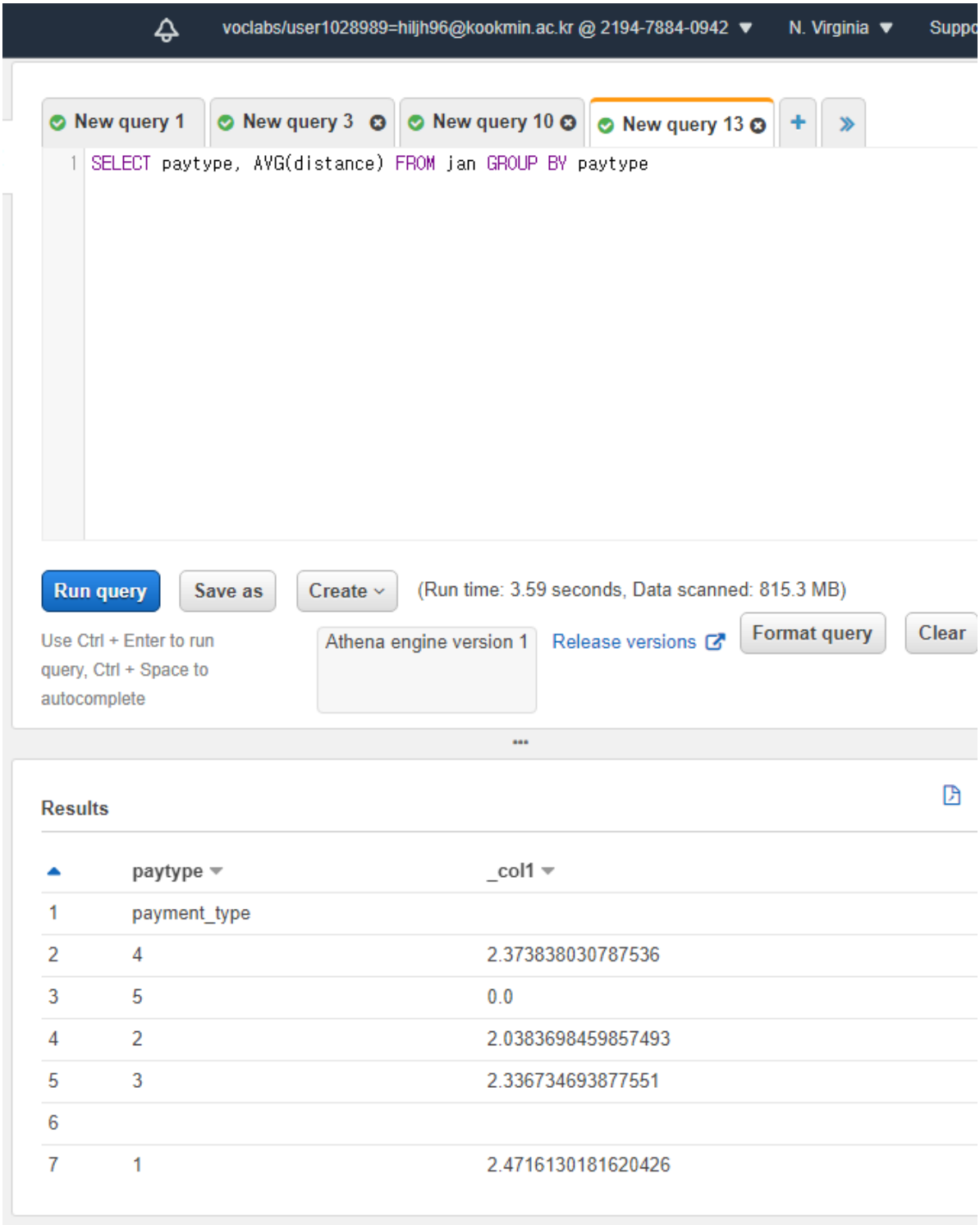
Chanllenge one

GROUP BY를 써서 그런지 view로 만들어지지 않아서 일단 출력했다.



Chanllenge two

TEMP 테이블을 만들어서 저장해서 2개 SELECT로 크레딧 카드와 캐시 컬럼만 따올려 했는데 권한 문제인지 테이블을 그릴 수 없었다.



1이 creditcard이며 2가 cash이다.

LAB3

Challenge question

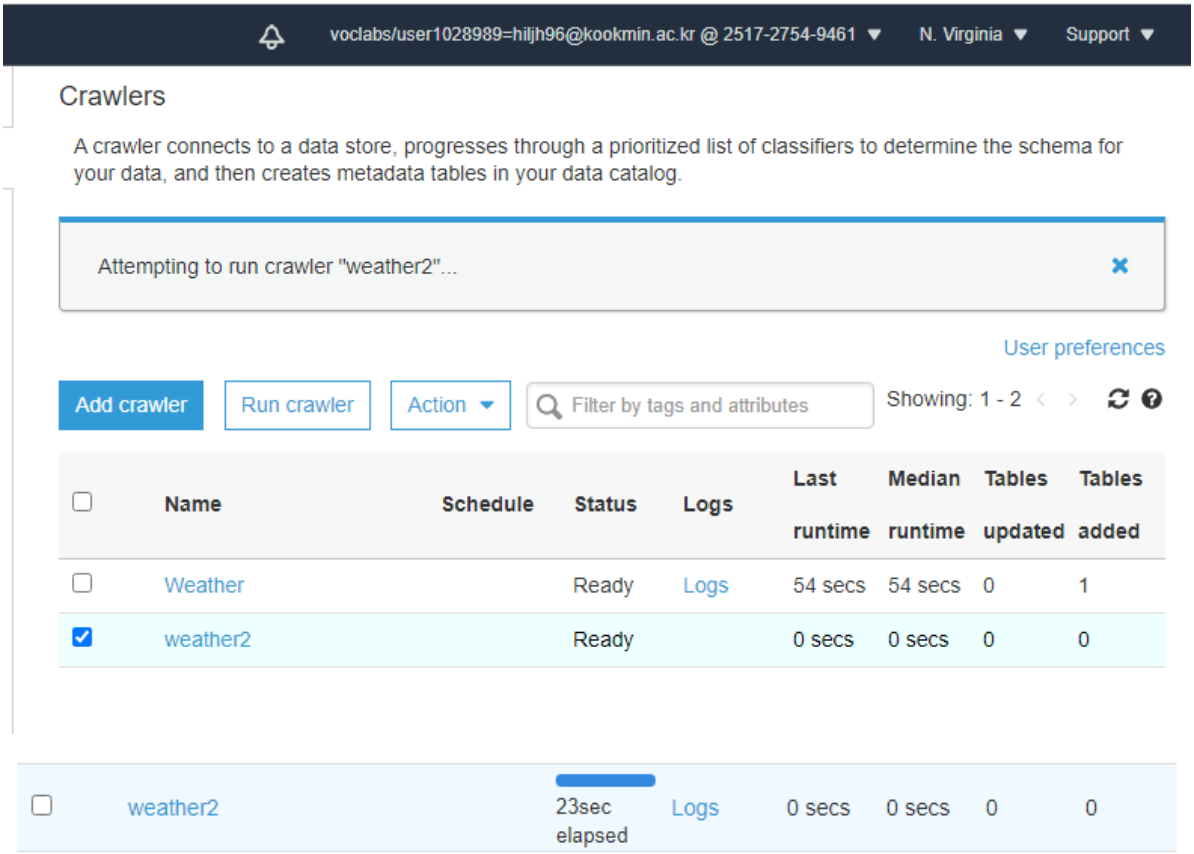
Now that you know how AWS Glue and Athena work together, try the following exercise to test your knowledge. Your lab instructor has a model solution. However, there is more than one way to solve the challenge.

The Global Historical Climatology Network Daily receives data from around the world. You can download data that describes these stations at the following location: ghcnd-stations.txt. The data dictionary for the observation and stations data is available at the following location: Readme.

Note The ghcnd-stations.txt file is a fixed-width text file. Before you use it with AWS Glue, you must convert it to a comma-separated values (CSV) file, or one of the other file formats that AWS Glue supports. One easy way to convert a text file to a .csv file is to open the text file with a spreadsheet program and then save the file in .csv format. You can also find free utility programs on the internet that can help with this process.

For this challenge, do the following tasks:

- Use AWS Glue to create a table for the weather stations.
- Write a query in Athena to count the number of stations that are not in the US or Canada. The first two characters of the station ID field indicate the country where the station is located. The country codes for the United States is *US* and the country code for Canada is *CA*.



weather2를 크롤링 중이다. AWS Athena는 Query 이상하게 돼지 않았다.

Lab 5

Challenge question

Now that you can use the Bokeh Python package to create data visualizations, try the following challenge to apply your skills to a real-world case.

AnyCompany Airlines has collected sample data for flight departures. They asked you to analyze the data to determine if there is an association between departure times and flight delays. You can access the sample data from Amazon S3. Develop a visualization that will describe this association.

```
Error: invalid argument type
sh-4.2$ aws s3 cp s3://aws-tc-largeobjects/CUR-TF-200-ACBDFO-1/Lab5/flightdata.csv data.csv
download: s3://aws-tc-largeobjects/CUR-TF-200-ACBDFO-1/Lab5/flightdata.csv to ./data.csv
sh-4.2$
```

jupyter flight Last Checkpoint: 3분 전 (unsaved changes)

Logout

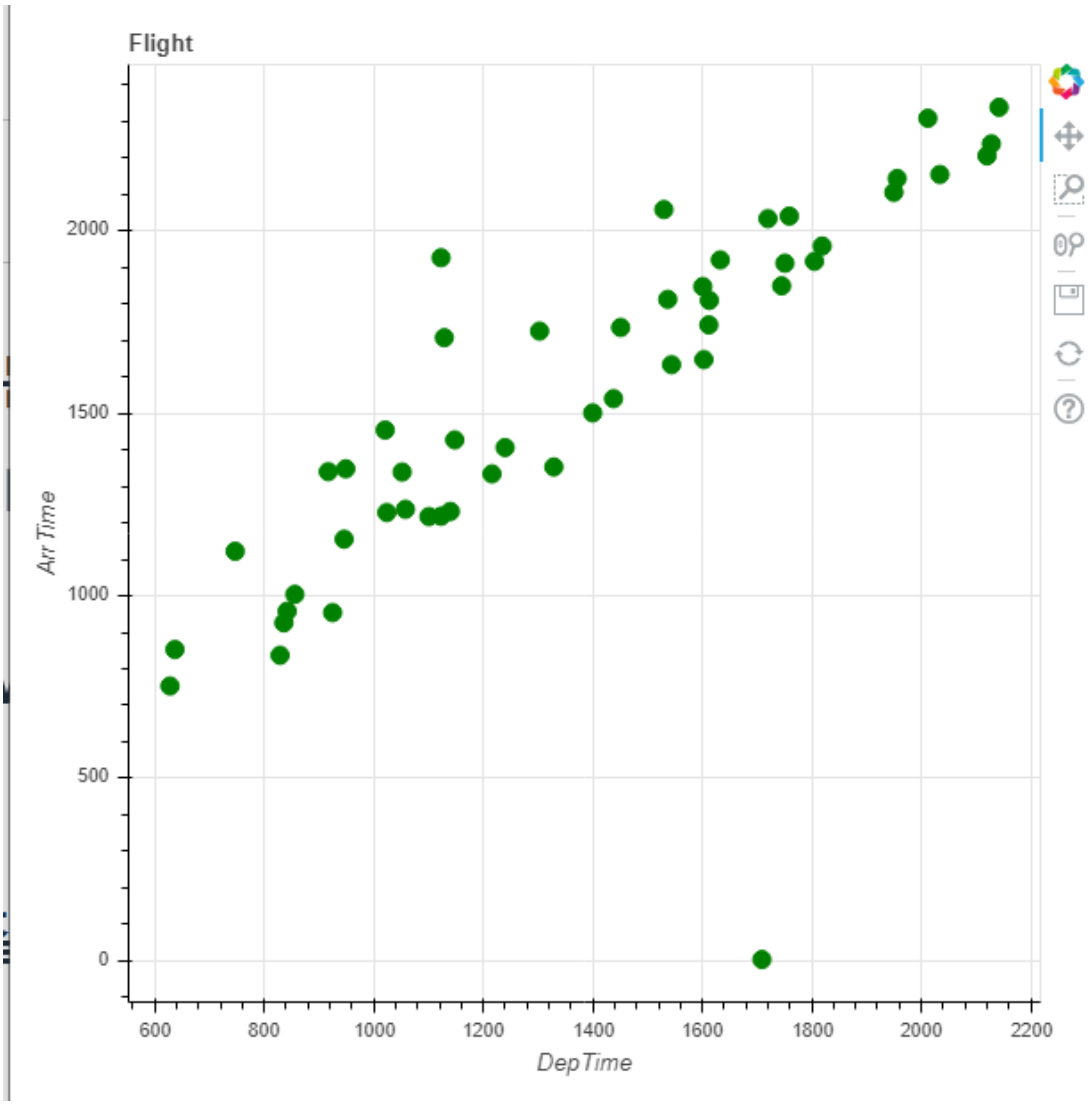
File Edit View Insert Cell Kernel Widgets Help Trusted conda_python3

Run Code nbdiff

In [1]:
```
import pandas as pd
from bokeh.plotting import figure, output_file, show
from bokeh.models import ColumnDataSource

output_file('flightdata.html')
df = pd.read_csv('/home/ec2-user/data.csv')
sample = df.sample (50)
source = ColumnDataSource(sample)
p=figure()
p.circle(x='DepTime', y='ArrTime', source=source, size=10, color='green')
p.title.text = 'Flight'
p.xaxis.axis_label = "DepTime"
p.yaxis.axis_label = "ArrTime"
show(p)
```

In [ ]:



가로 컬럼은 Dep Time 세로 컬럼은 Arr Time으로 수정해서 그렸다.

Lab 6

Challenge question

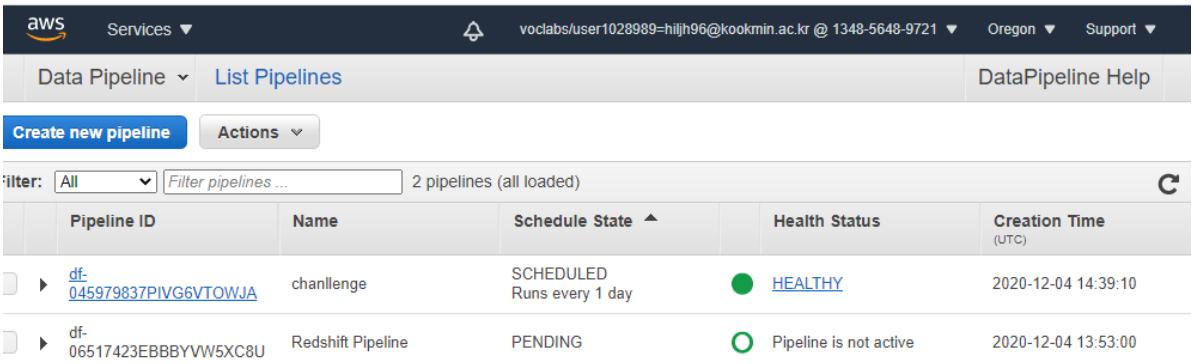
Now that you can automate loading data by using AWS Data Pipeline, try the following challenge to apply your skills to a real-world case.

Your manager is pleased that you automated the process of loading data to Amazon Redshift. He would now like you to do two additional tasks:

- Create a pipeline that will load a second month of data.
- Determine the most common pickup locations for each of the two months.

The February data is in the following Amazon S3 location:

```
s3://aws-tc-largeobjects/CUR-TF-200-ACBDFO-1/Lab6/February
```



클러스터 생성 시 VPC 설정하지 않고 실행하려 하니 VPC 설정 없이 생성이 안되는 문제가 있었다. 그래서 default값으로 설정하고 실행하니 redshift pipeline이 IAM 오류 또는 생성이 안되는 문제가 생겼다.