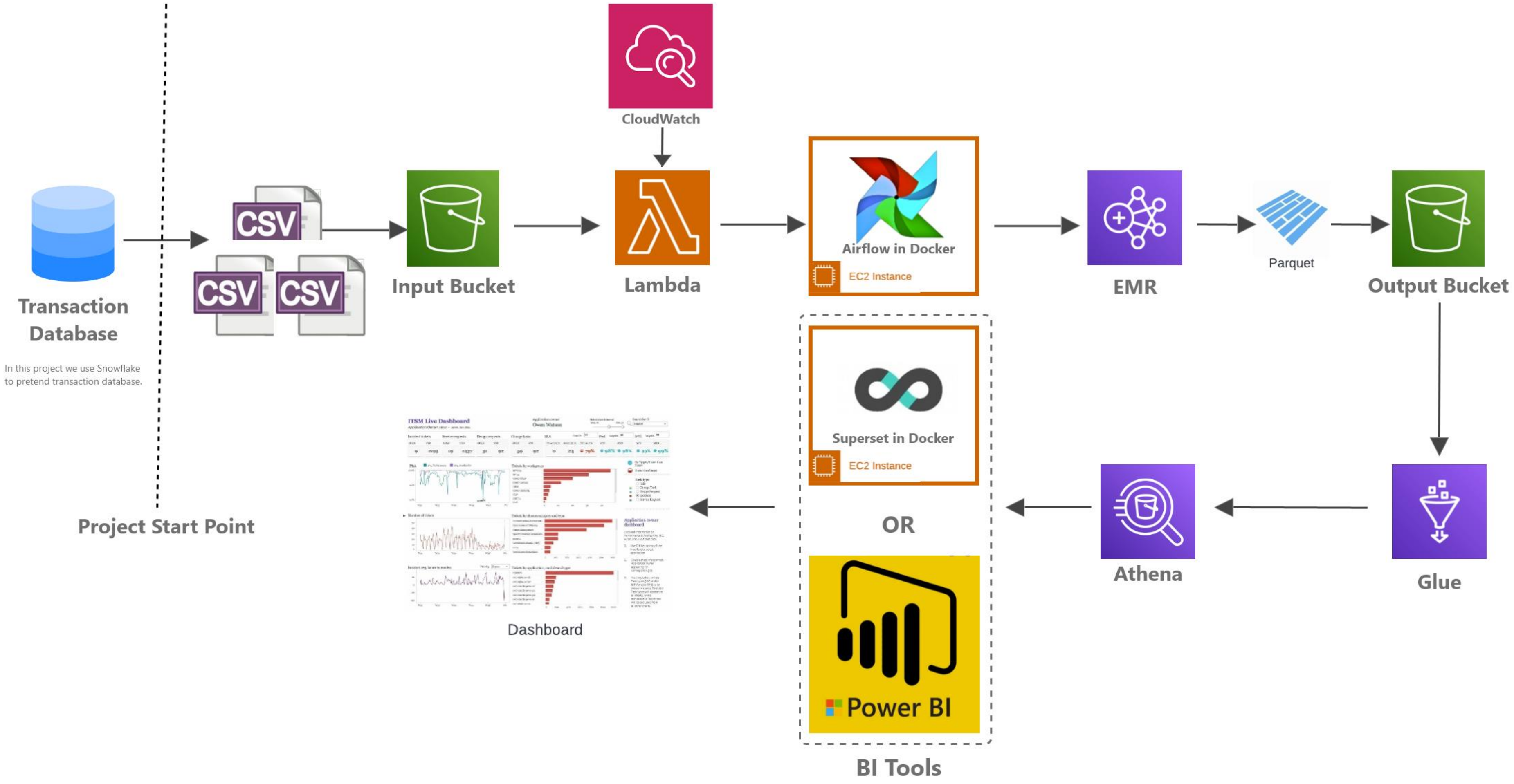
WeCloudData

Data Engineering Bootcamp

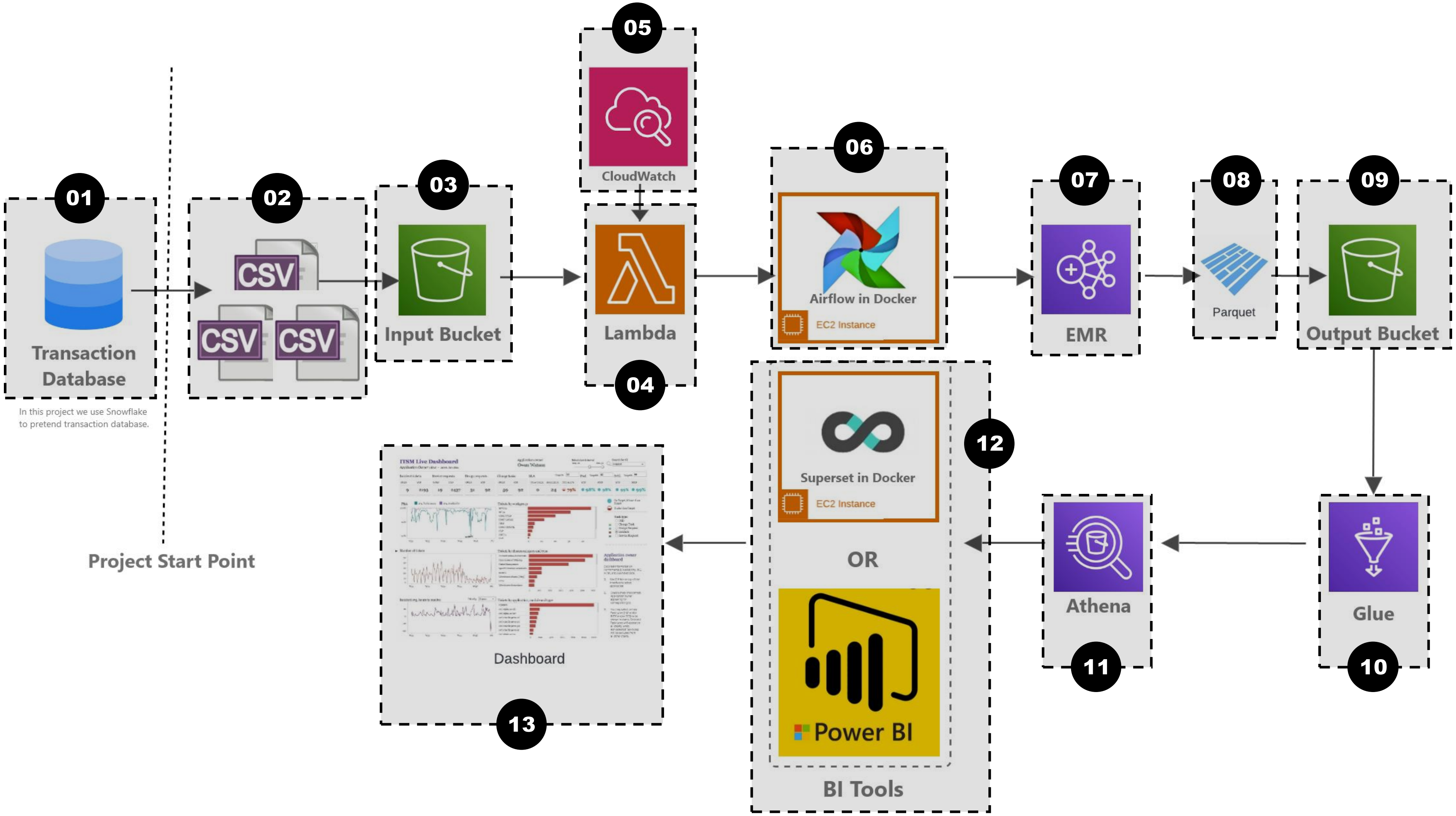
Chichi Aguoji

Architecture 1



In this project we use Snowflake to pretend transaction database.

Architecture 1



In this project we use Snowflake to pretend transaction database.

01

Transaction Database

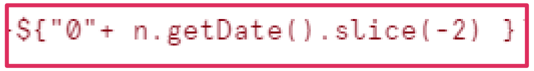
Transaction Database

Challenges

- Zero padded dates in Javascript

Modifications

- Adjusted the stored procedure to include:



```
${"0"+ n.getDate().slice(-2) }
```

  - This adds a “0” in the front of the date then uses the slice function to get a substring of the last 2 digits in the date.

02

Data Pull

Challenges

- None

Modifications

- None

03

Input Bucket

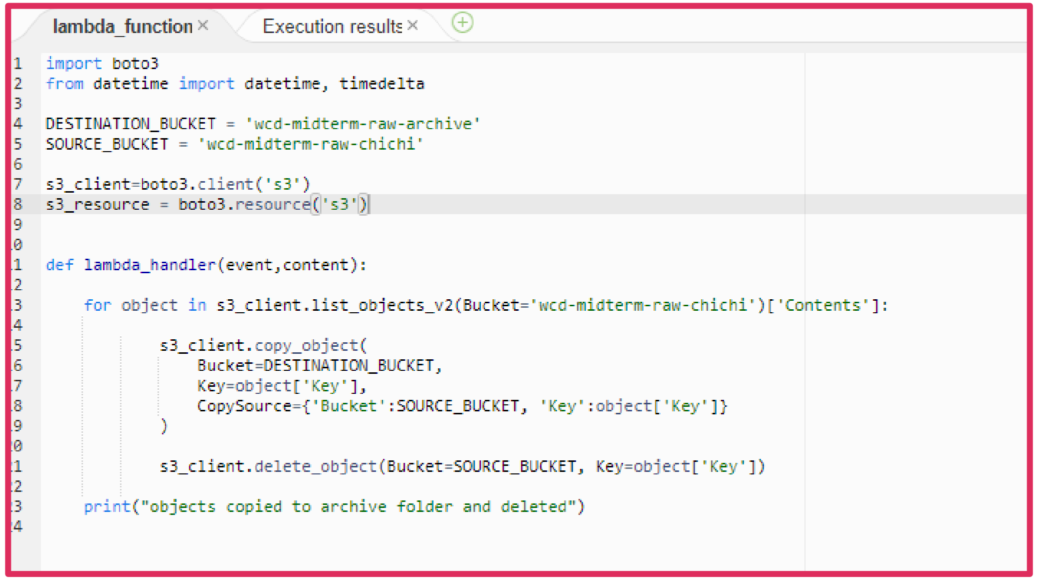
Input Bucket

Challenges

- Daily data being appended into S3 bucket

Modifications

- Created another S3 bucket to store historical data.
- Used Lambda function to move data into this separate bucket.



```python
import boto3
from datetime import datetime, timedelta

DESTINATION_BUCKET = 'wcd-midterm-raw-archive'
SOURCE_BUCKET = 'wcd-midterm-raw-chichi'

s3_client=boto3.client('s3')
s3_resource = boto3.resource('s3')


def lambda_handler(event,content):

    for object in s3_client.list_objects_v2(Bucket='wcd-midterm-raw-chichi')['Contents']:

            s3_client.copy_object(
                Bucket=DESTINATION_BUCKET,
                Key=object['Key'],
                CopySource={'Bucket':SOURCE_BUCKET, 'Key':object['Key']}
            )

            s3_client.delete_object(Bucket=SOURCE_BUCKET, Key=object['Key'])

    print("objects copied to archive folder and deleted")

```

04

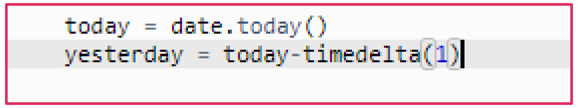
Lambda

Challenges

- Issues with the run date being 1 day after the file date
- Call to subprocess kept failing

Modifications

- Attempted to use datetime library to use timedelta to get previous day’s date, but it wasn’t working so ended up using substrings. Substrings logic broke during testing when new month began so used the following:



```
today = date.today()
yesterday = today-timedelta(1)
```

- Didn’t realize the content-type was entered incorrectly, entered ‘application/json/’ instead of ‘application/json’

05

CloudWatch

Cloud Watch

Challenges

- None

Modifications

- None

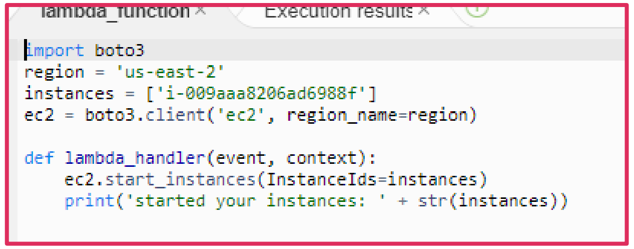
Airflow in Docker

Challenges

- Updating EC2 IP address after every start
- Unhelpful Logs
- Having to leave EC2 instance running in order to test

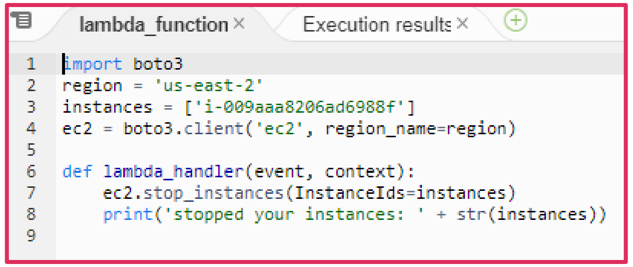
Modifications

- Set up Elastic IP Address for EC2 instance
- Manually Timed Errors
- Set up 2 additional Lambda functions to start and stop EC2 instance each night to test the data drop



```
import boto3
region = 'us-east-2'
instances = ['i-009aaa8206ad6988f']
ec2 = boto3.client('ec2', region_name=region)

def lambda_handler(event, context):
    ec2.start_instances(InstanceIds=instances)
    print('started your instances: ' + str(instances))
```



```
import boto3
region = 'us-east-2'
instances = ['i-009aaa8206ad6988f']
ec2 = boto3.client('ec2', region_name=region)

def lambda_handler(event, context):
    ec2.stop_instances(InstanceIds=instances)
    print('stopped your instances: ' + str(instances))

```

EMR

EMR

Challenges

- Manual Start
- Configuring access to Airflow
- Testing PySpark code in Databricks w/o a local run

Modifications

- Added a step to start EMR programmatically
- Trial/Error, Google, Rewatching Lectures
- Ended up having to run code on EMR in chunks until found errors

08

Parquet

Challenges

- Understanding partitions labeling
- Partition amount and number of partitions not aligning

Modifications

- None
- None

09

Output Bucket

Output Bucket

Challenges

- Using folder structure to facilitate BI analysis

Modifications

- None

10

Glue

Glue

Challenges

- None

Modifications

- None

Athena

Athena

Challenges

- None

Modifications

- None

BI Tools

Challenges

- Connecting PowerBI to Athena

Modifications

- Leveraged ODBC Connection and AWS Secret Access Key

Dashboard

Dashboard

Challenges

- Having questions about the data that the data couldn’t answer
- No Automatic Refresh of data
- Wanting to use visualizations not supported by the data

Modifications

- Only visualized questions the data could answer
- None
- Created supporting dummy data