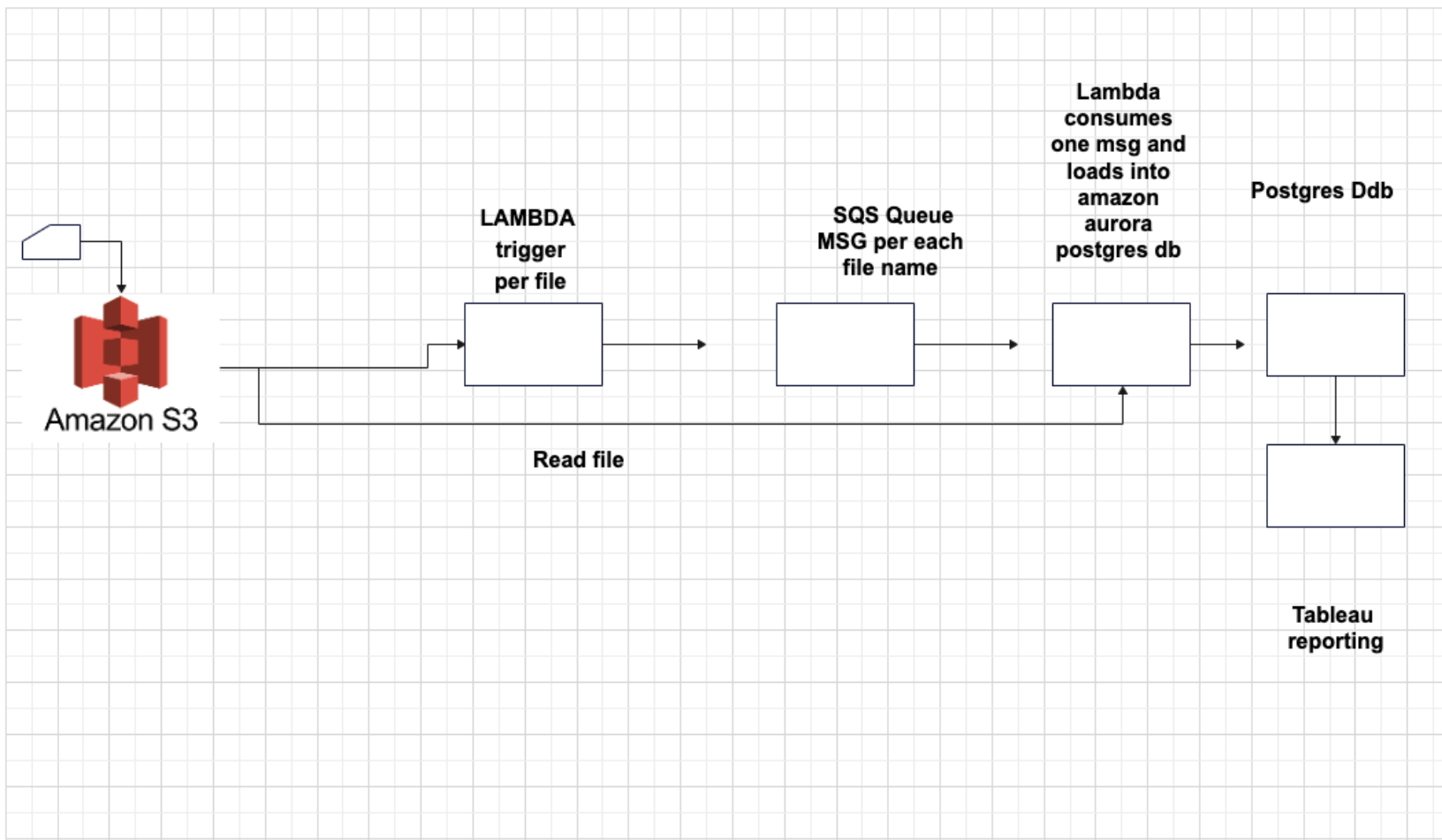
Contraints: Operations are peroformed on Data from the year 2009 due to resource contraints on the local machine

In [14]:
```
from IPython.display import Image
Image(filename='/Users/arunkumarkokkula/Desktop/Architecture.png')
```

Out[14]:



Loading to database

In [1]:
```
file_path = "/Users/arunkumarkokkula/Desktop/archive"
```

In [2]:
```
import pandas as pd
from sqlalchemy import create_engine
import psycopg2
import io
```

In [3]:
```
df = pd.read_csv("/Users/arunkumarkokkula/Desktop/archive/2009.csv")
```

In [4]:
```
engine = create_engine('postgresql+psycopg2://postgres:deep@localhost:5433/flights')
```

In [5]:
```
df['FL_DATE'] = pd.to_datetime(df['FL_DATE'])
```

In [6]:
```
df.head(0).to_sql('flights', engine, if_exists='replace',index=False) #drops old table and creates new empty ta
```

Out[6]:
```
0
```

In [7]:
```
conn = engine.raw_connection()
cur = conn.cursor()
output = io.StringIO()
df.to_csv(output, sep='\t', header=False, index=False)
output.seek(0)
contents = output.getvalue()
cur.copy_from(output, 'flights', null="") # null values become ''
conn.commit()
```

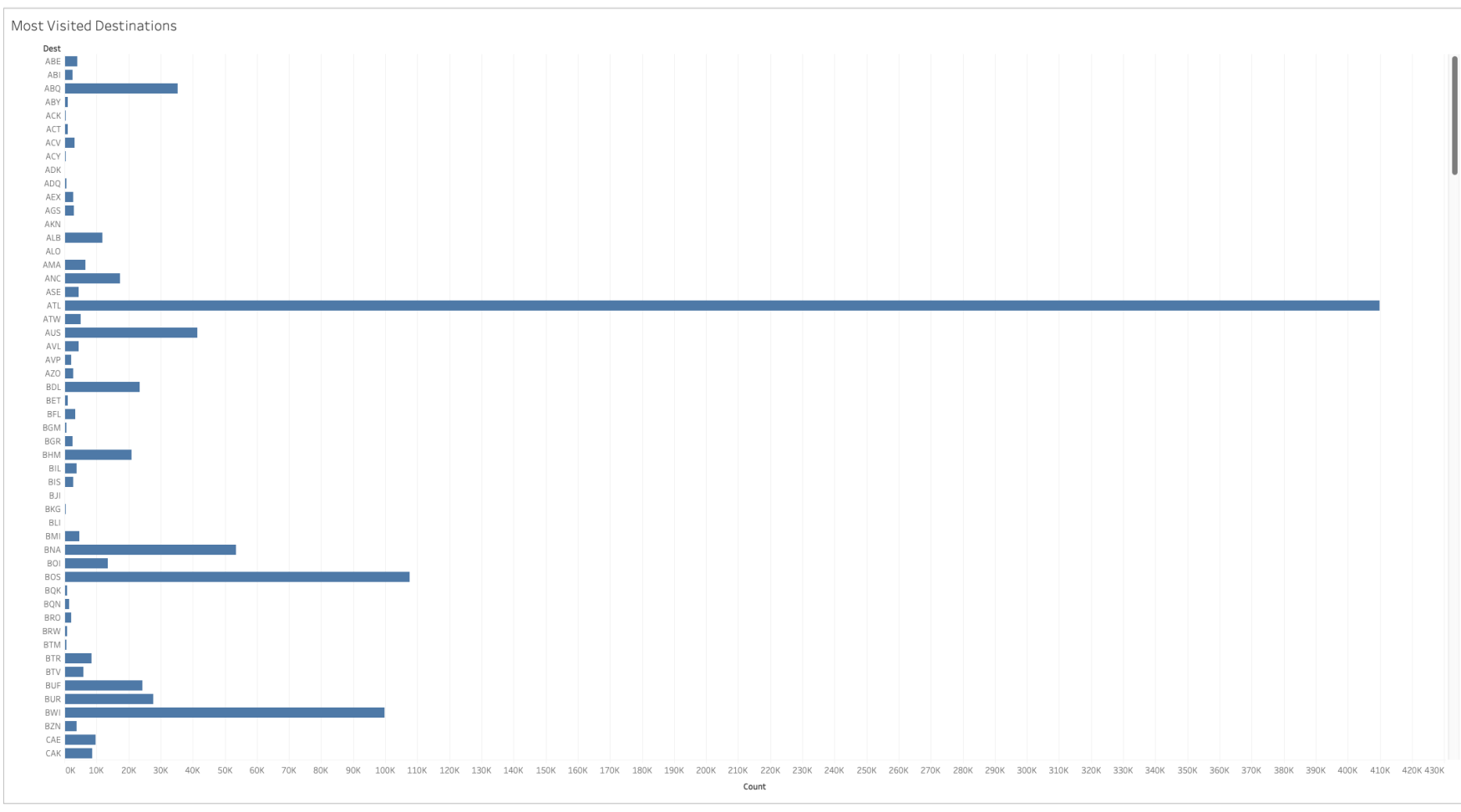
SQL queries

1. Most visited Destinations

select "DEST", count(*) from public.flights where "CANCELLED"=0 and "DIVERTED"=0 group by "DEST" order by count DESC

In [18]:
```
from IPython.display import Image
Image(filename='/Users/arunkumarkokkula/Desktop/1mostvisited.png')
```

Out[18]:



2. *Month with most Cancellations*

select to_char("FL_DATE", 'YYYY-MM'), count(*) from public.flights where "CANCELLED">0 group by to_char("FL_DATE", 'YYYY-MM') order by count DESC limit 1

3. *Airports that have the highest departure delay*

-- assuming that we need to pull the highest departure delays of every airport

select "ORIGIN", max("DEP_DELAY") from public.flights where "DEP_DELAY">0 group by "ORIGIN" order by max desc

-- we are pulling the airport that has the highest departure delay

with cte as (select "ORIGIN", max("DEP_DELAY") max_delay from public.flights where "DEP_DELAY">0 group by "ORIGIN"), cte2 as (select "ORIGIN", max_delay, rank() over(order by max_delay desc) from cte) select "ORIGIN", max_delay from cte2 where rank=1

4. *Routes with most diversions*

with cte as (select "ORIGIN", "DEST", count(*) route_max from public.flights where "DIVERTED">0 group by "ORIGIN", "DEST"), cte2 as (select "ORIGIN", "DEST", "route_max", rank() over(order by route_max desc) from cte) select "ORIGIN", "DEST", "route_max" from cte2 where rank=1

5. the most connected airport

with cte1 as (select distinct "ORIGIN", "DEST" from public.flights), cte2 as (select "ORIGIN", count(*) connections from cte1 group by "ORIGIN"), cte3 as (select "ORIGIN","connections", rank() over(order by connections desc) from cte2) select "ORIGIN", "connections" from cte3 where rank=1

In [ ]:
```
```