Scoping the Project

Datasets used and Goal of data

I chose to do the Udacity provided project. The datasets used was the SAS dataset of the i94 forms, crossed with US City demographics data. Also used as supporting data sets was the i94 countries and i94 ports code descriptions. This supporting data sets and the demographic datasets would be used as dimensions in a star schema data warehouse with the i94 form dataset as the star.

Not all columns where used from the demographics and i94 form datasets. I was mostly interested in the data that will help analyze, where people of a certain age where ending up. Where people born in 1977 more likely to show up in “so and so” state and where people born in 1994 likely to visit “so and so” state more. If you look at the data dictionary provided with the project you will see a column with explanation as to why each column was kept.

Example below is a query I made and then opened the results in tableau. I actually found that most states have a normal distribution of age. All well guess my answer is not very interesting. However, I did get my answer. SUCCESS!

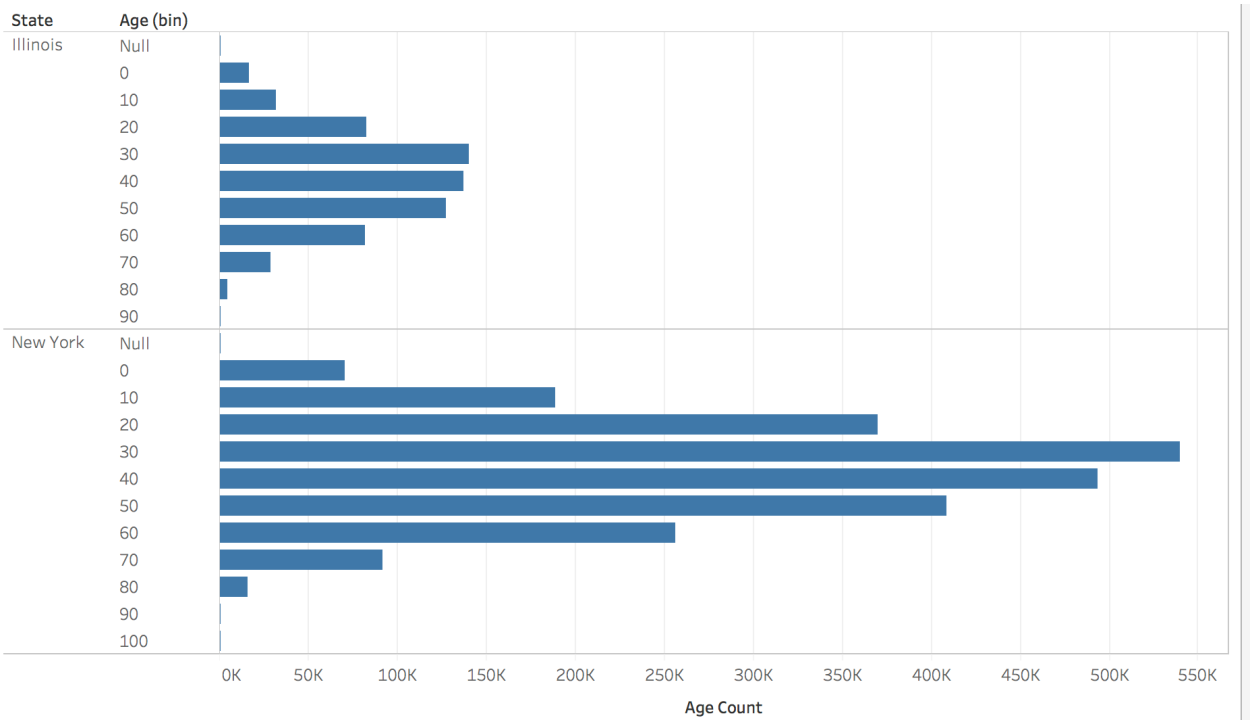
select d.state, i.age, COUNT(i.age) as AGE_COUNT, i.gender, COUNT(i.gender) as Gender_COUNT

from i94 i, demographics d

where i.port_id = d.port_id

group by d.state, i.age, i.gender

order by state, age



Technology used and Goal

For this project I wanted to incorporate Spark, Airflow, and AWS (S3, Redshift). Although its not completely necessary to use these technologies, I used them for two reasons. One reason is I wanted to spend more time using and troubleshooting with these tools and the second reason is that these tools do help even if its overkill.

Spark was run on a stand-alone cluster locally on my machine along with Hadoop installed and sharing its class path with Spark. This was necessary to leverage reading and writing files directly from s3 with spark. Spark did help partition the very large i94 form dataset by age.

Airflow was definitely useful to organize my pipeline's dependencies. Subdags, where used and Custom Operators where designed to group and repetitive tasks easier to handle.

Most importantly S3 and Redshift were like the glue in all this. It is was really easy to copy from s3 to my redshift cluster and the redshift cluster allowed me to distribute the large i94 form data set by age.

Addressing Other Scenarios

- The data was increased by 100x.
- The pipelines would be run on a daily basis by 7 am every day.
- The database needed to be accessed by 100+ people.

If the data was increase by 100x I would definitely have to leverage Spark better than I am now. For now I'm probably not partitioning my dataset in the best way I can. I did some data exploration on the i94 form and It seems that age might be a good column to partition on. I found that the average age of an i94 applicant was 38 and that the standard deviation of age was 17 years and the percentiles looked like so **25%:** 26, **50%:** 39 , **75%:** 52.

If the pipelines need to be run on a daily basis by 7am every day. I would take advantage airflow dag's schedule_interval config value.

If the database needed to be accessed by 100+ people I think again It would nice to leverage distributed computing and use redshift as a database and using i94 age column as a distribute key for the same reasons as above.

Defending Modeling Choices

Choice of Star Schema Design

I think going for the data warehouse star schema approach was a good idea. The i94 form was essentially already a fact table data dump. It used so many codes to reference other data dimension like ports, countries, visa codes etc. Then it includes measure fields about the i94 applicant like age,

application date etc. The dimensions then became the country names for the codes, the visa types for the visa codes, the us demographic data tied together by port_id.

Using age as partitioning and distribution key on spark and redshift

Using age as a distkey for the very large fact table and then distributing the other data on all nodes of my redshift cluster is defensible because I found that the average age of an i94 applicant was 38 and that the standard deviation of age was 17 years and the percentiles looked like so **25%:** 26, **50%:** 39 , **75%:** 52. The demographics data is not very large and also the other tables were made from supporting data that was not very large either.