School of Information Studies
Syracuse University

IST 615 – Cloud Management

Project Progress Report

Car Retail Website

Due Date: 11/14/2022
Submission Date: 11/14/2022
Number of Pages: 06

Team members:

- Mikhail Pravin Pinto (mpinto01@syr.edu)
- Kirat Saran (ksaran@syr.edu)

Contents

Introduction:

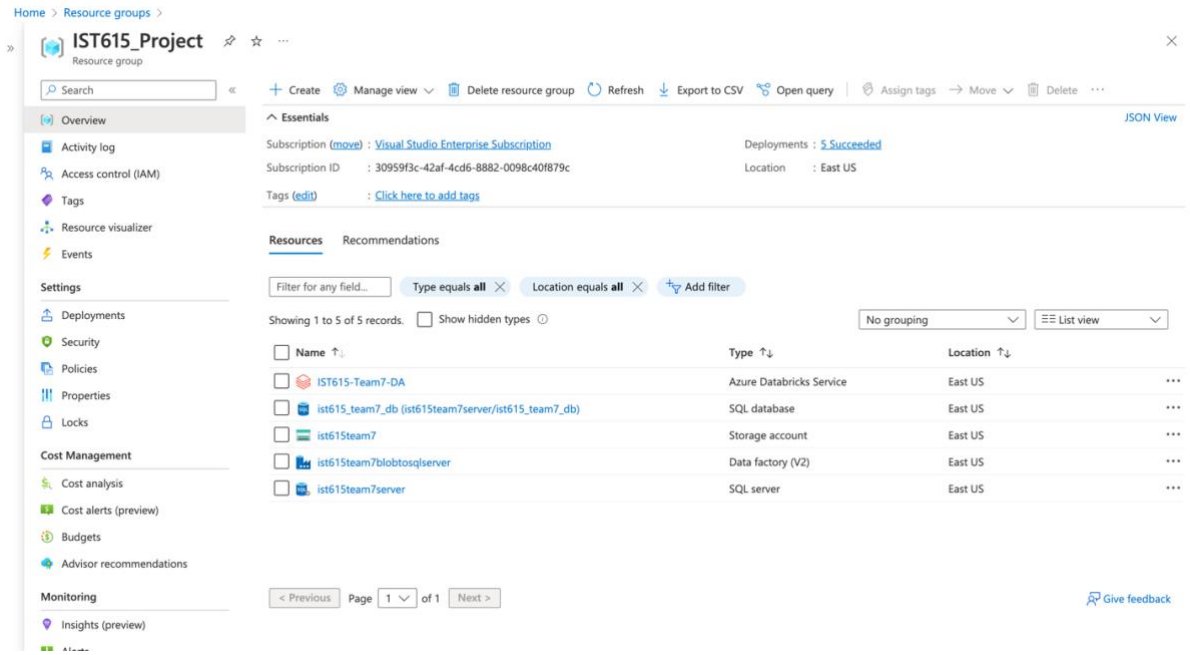
We are building a car retail website using cloud where buyers can buy cars based on manufacturers, mileage, distance travelled, etc since the demand to buy vehicles has increased post pandemic.
The dataset is from the site CarDekho.com which was uploaded to Kaggle.com.
Link to the dataset:
https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data?datasetId=62920
We'll use Microsoft Azure Cloud Services to implement this.

Progress:

1. *Data Cleaning*

CarDekho.com dataset was downloaded in the form of a csv file. We cleaned the data by removing rows with null values in columns like condition, manufacturer, model, fuel, odometer, and transmission.
Column with County name was also dropped as it had more than 50% null values. Similarly, the size and region URL was also removed due to a lot of missing values. The description was also dropped as the column contained identical descriptions for more than 50% of the records.

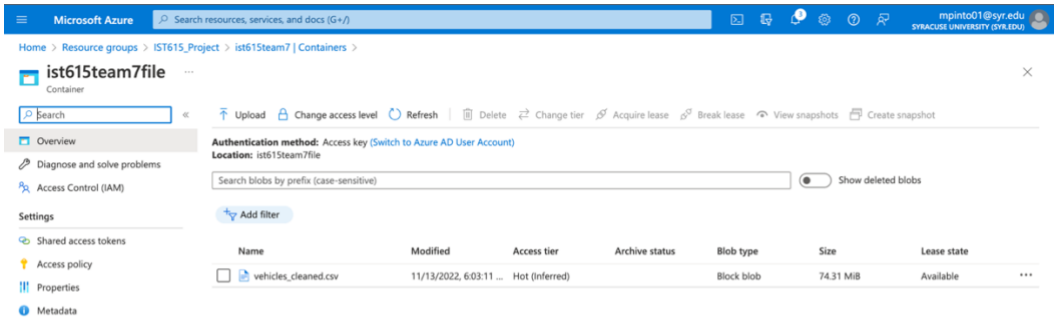
We created a resource group for the project to manage and associate all the multiple resources that would be used for this project.



Screenshot 1: Resource group created for the project.

2. *Azure Blob Storage*

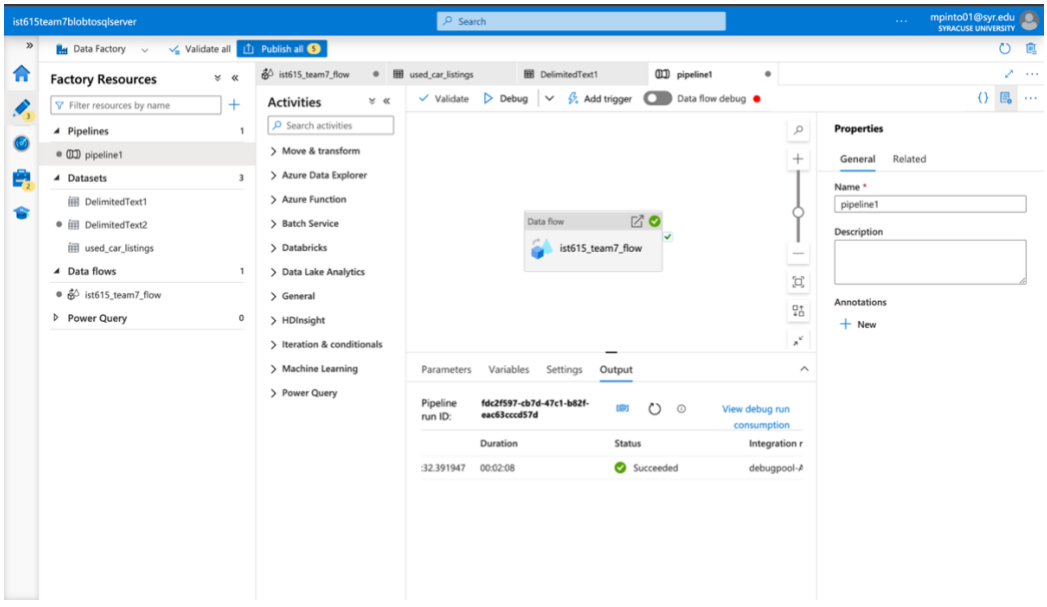
Our cleaned dataset file ‘vehicles_cleaned.csv’ was uploaded to Azure blob service and stored in a container.



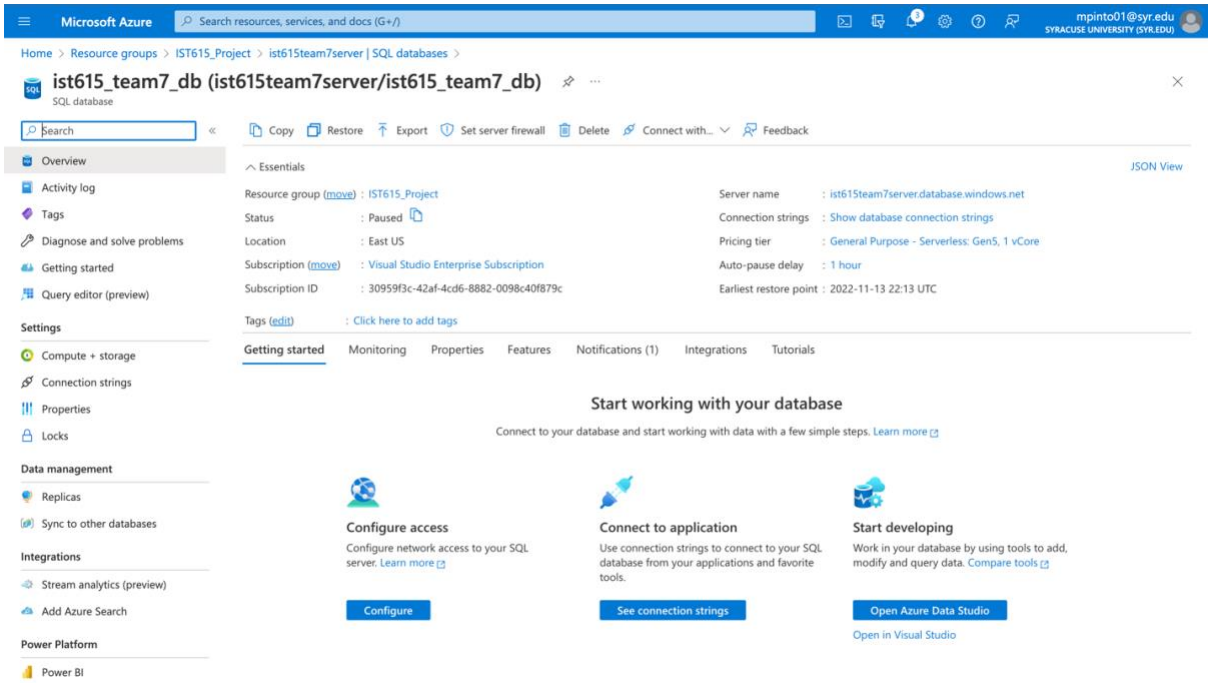
Screenshot 2: A container created in Azure blob storage to upload and store our cleaned csv data file.

3. *Azure Data Factory*

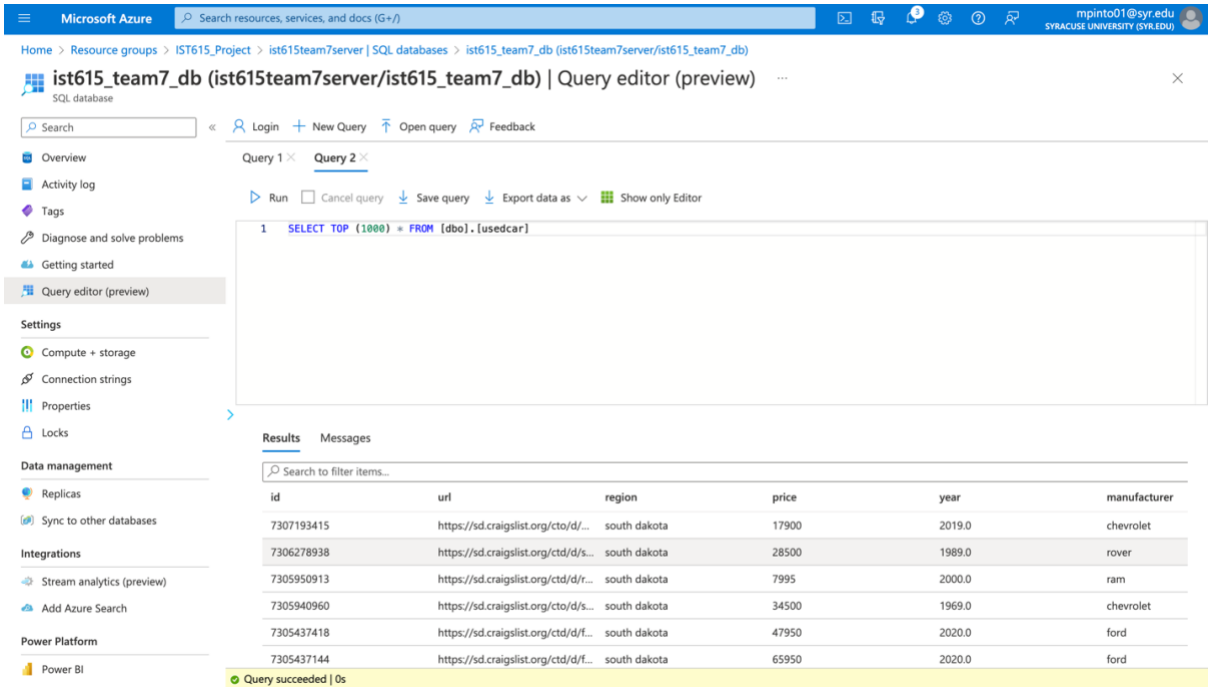
Azure Data Factory has been used to facilitate data integration and data transformation. A data flow was created to get data from the Azure Blob storage container to Azure SQL Server DB. This was executed using a pipeline to get data from Azure blob storage to SQL Server.



Screenshot 3: A pipeline created and executed in Azure data factory to get data from Azure blob storage to SQL Server.



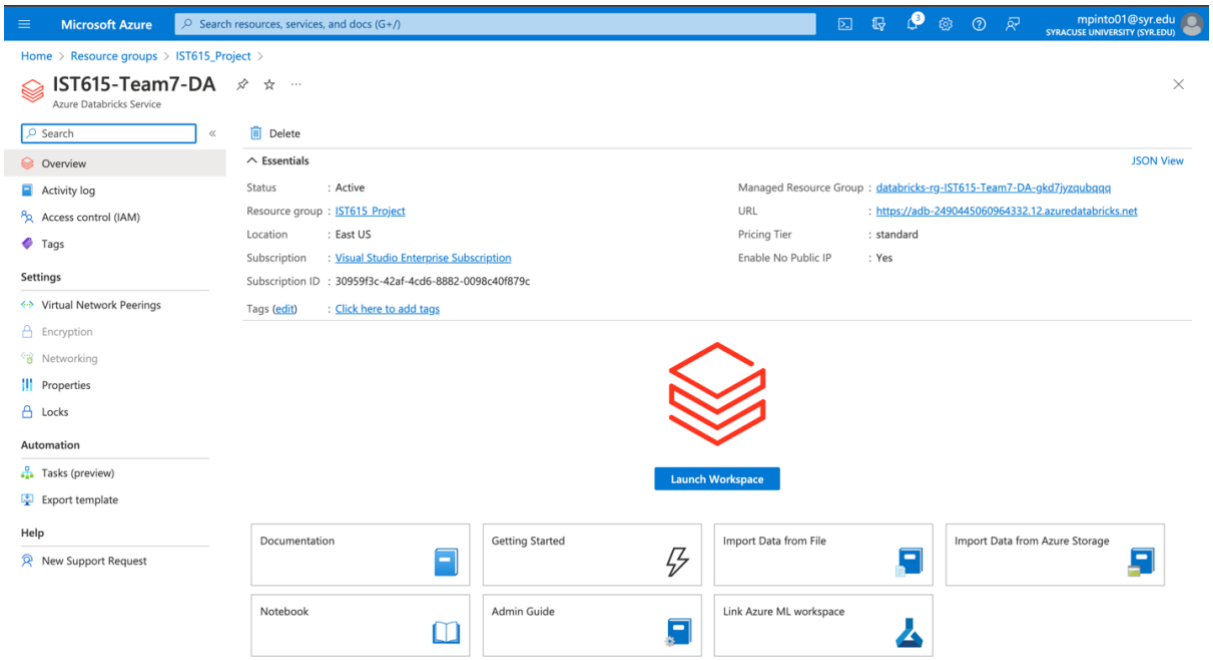
Screenshot 4: Azure SQL server and DB created for the purpose of data handling and data transformation.



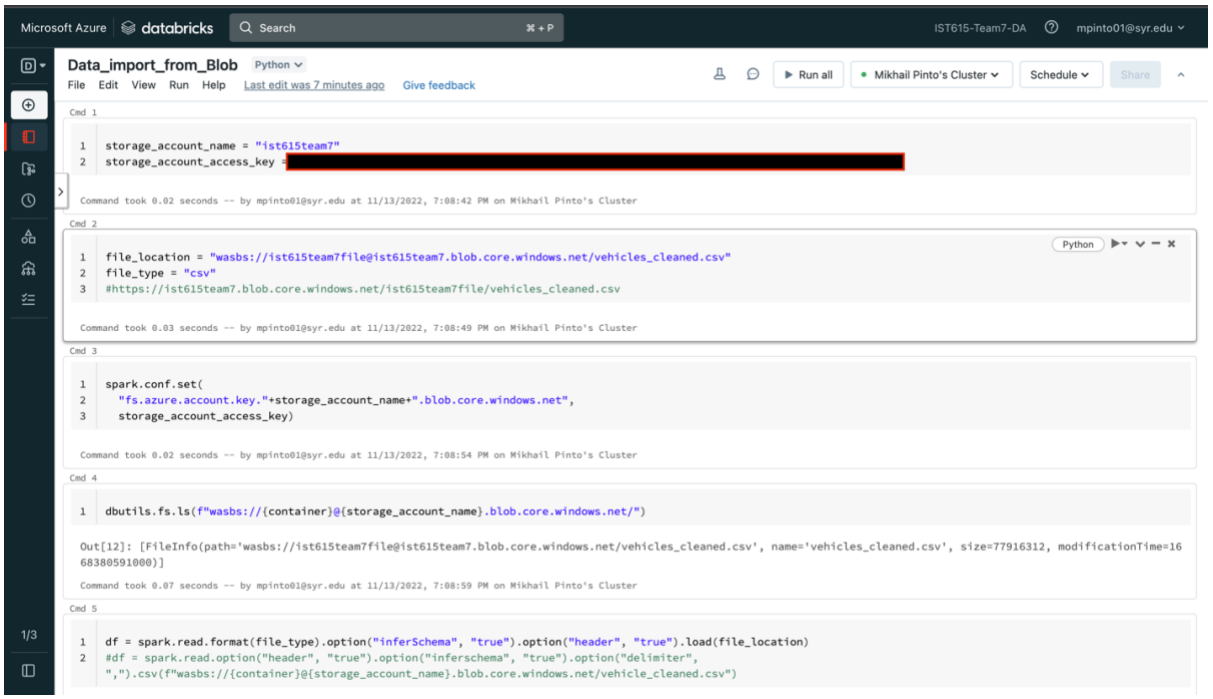
Screenshot 5: An SQL query to display top 1000 rows out of 234k total records.

4. *Azure Databricks*

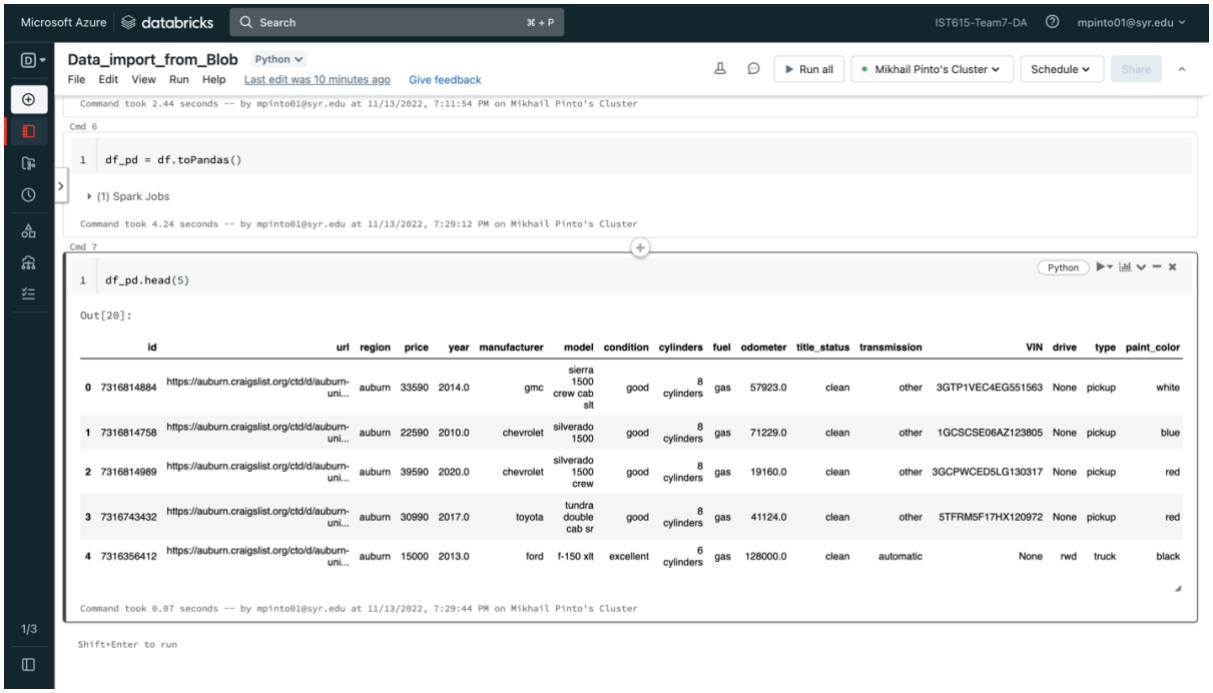
This would be used for the purpose of python scripting, analysis and model creation in our project. We wrote a script to integrate Azure Blob storage and Azure Databricks. Using this script, we can read data from our container in blob storage to do further analysis on our dataset.



Screenshot 6: Setting up Azure Databricks service for our project.



Screenshot 7: Script written to integrate Azure Blob storage and Azure Databricks



Screenshot 7: Converting our data to a pandas dataframe using databricks service.

Future Implementations:

S.No.	Implementation	Expected Completion
1.	Implemenation of basic front end • Search Page This page will have user inputs like the brand of the car, the transmission of the car, the type of car, the color of the car, price of the car. This will retrieve car data for that brand as well as cars with similar characteristics • Post Page According to the price of the cars already present in the dataset, we will estimate the price of user's car if a user wants to know how much their car is valued in the market. If the user likes the estimated price, they can list their car on our website. This would add the data to the database in the SQL server.	November 21st – 30th, 2022.
2.	Create PowerBI Dashboards based on our analysis and modelling	December 1st, 2022.
3.	Ready with our final project and presentation	December 5th, 2022.