



**Unext**

# EV Carbon Emission Analysis

B3\_Rhea\_Mehta  
B3\_Harsh\_Thakur  
B3\_Pragun\_Sood  
B3\_Akshat\_Gupta



# Introduction

- In light of the urgent need for a sustainable and environmentally conscious transportation ecosystem, the Canada Climate Department and Transportation has gathered a substantial dataset comprising millions of observations and over 20 diverse variables.
- This dataset forms the cornerstone of a crucial initiative aimed at conducting an in-depth analysis of carbon emissions from both Electric Vehicles (EVs) and conventional vehicles.

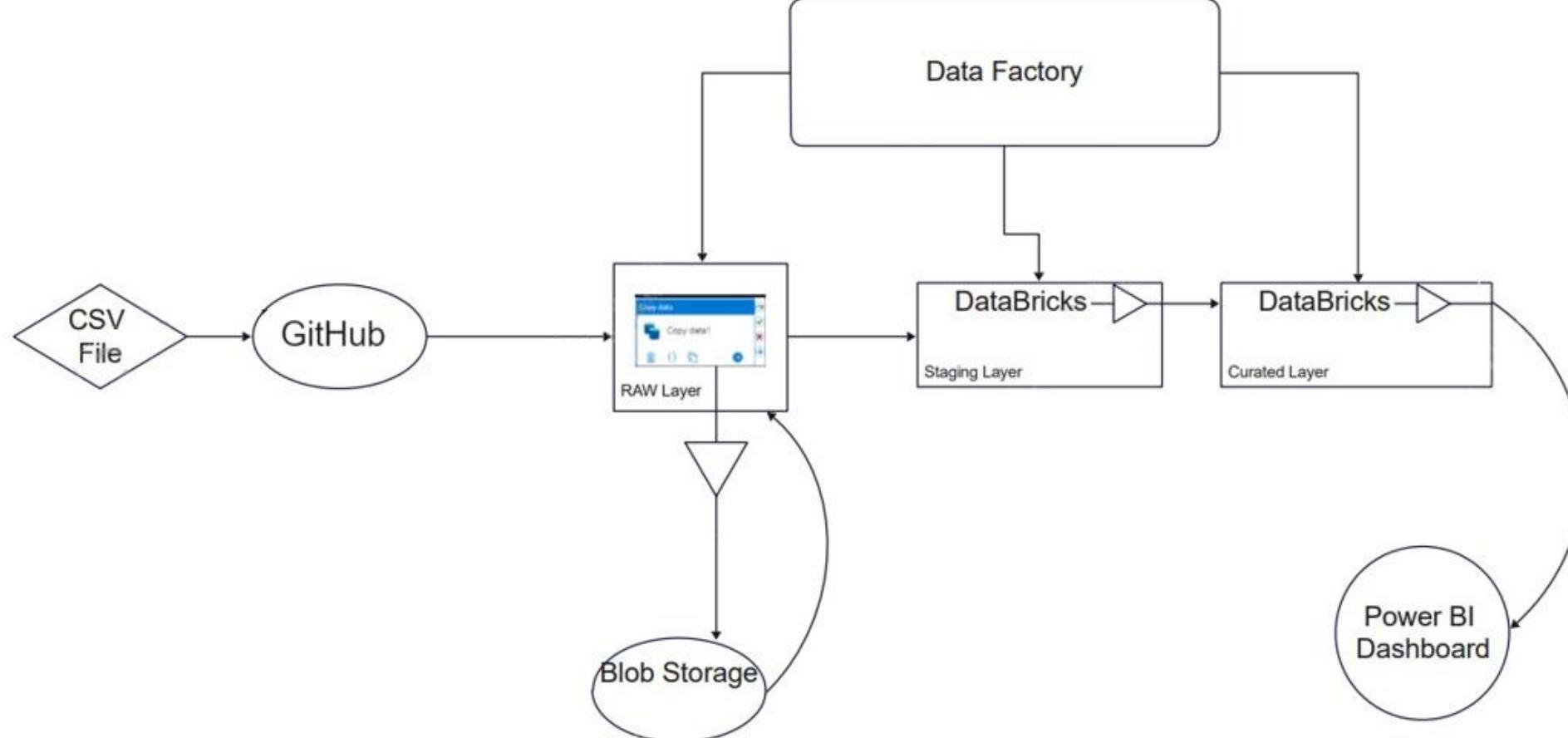


# Problem Statement

- We have to harness the power of Azure cloud services and the Spark framework to construct a robust data engineering application. The application should be capable of efficiently processing, transforming, and analyzing this extensive dataset to extract meaningful insights regarding carbon emissions.
- The successful completion of this project will significantly contribute to the understanding of carbon emissions in the context of Electric Vehicles and conventional vehicles, ultimately guiding policy decisions towards a more sustainable transportation ecosystem



# Data Flow Diagram





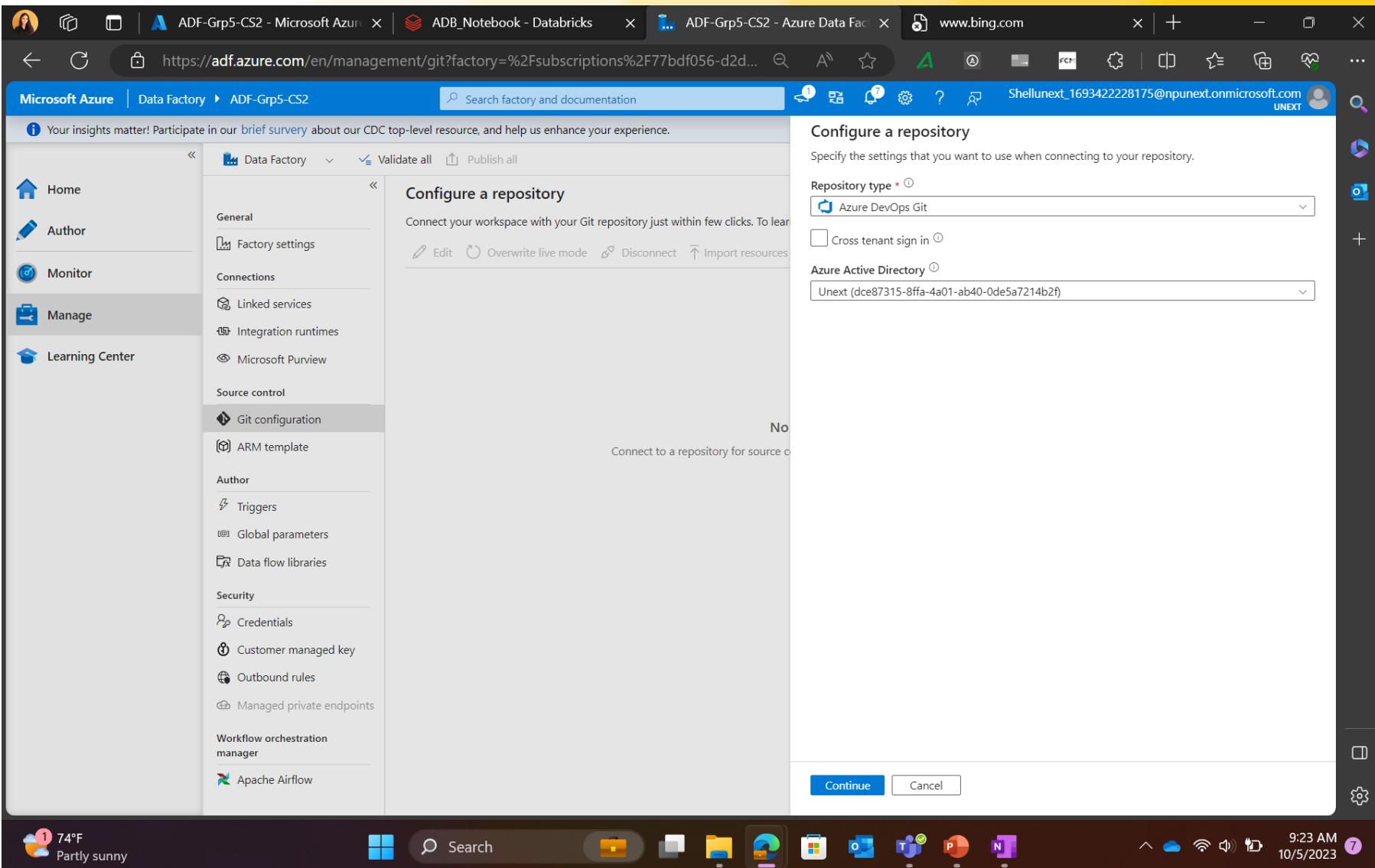
# Services Used

- Azure Data Factory
- Azure Data Bricks
- Power BI
- Azure DevOps
- GitHub
- Azure Storage Account



# CI/CD Workflow

## 1. Git Configuration for ADF



The screenshot shows the Microsoft Azure Data Factory management interface. The left sidebar has 'Manage' selected. The main area is titled 'Configure a repository' under 'Source control'. It shows a list of items: General, Factory settings, Connections, Integration runtimes, Microsoft Purview, Source control, Git configuration, ARM template, Author, Triggers, Global parameters, Data flow libraries, Security, Credentials, Customer managed key, Outbound rules, Managed private endpoints, Workflow orchestration manager, and Apache Airflow. On the right, there's a 'Configure a repository' form with fields for 'Repository type' (set to 'Azure DevOps Git') and 'Azure Active Directory' (set to 'Unext (dce87315-8ffa-4a01-ab40-0de5a7214b2f)'). Below the form are 'Continue' and 'Cancel' buttons.



ADF-Grp5-CS2 - Microsoft Azure | New query | Account Console - Databricks | ADF-Grp5-CS2 - Azure Data Fact... + - □ X

https://adf.azure.com/en/management/git?factory=%2Fsubscriptions%2F77bd056-d2d... Search factory and documentation

Microsoft Azure | Data Factory > ADF-Grp5-CS2

Your insights matter! Participate in our brief survey about our CDC top-level resource, and help us enhance your experience.

Home / main branch Validate all Save all Publish Preview experience Off

Author

Monitor

Manage

Learning Center

Configure a repository

Connect your workspace with your Git repository just within few clicks. To learn more about best practices about CI/CD please view document here. [CI/CD best practices](#)

Edit Overwrite live mode Disconnect Import resources

Repository type Azure DevOps Git

Azure DevOps Account Shellunext1693422212529

Project name EV Carbon Emission Analysis

Repository name EV Carbon Emission Analysis

Collaboration branch main

Publish branch adf\_publish

Root folder /

Last published commit d913ca1389e798809592451b4a15e7294f447c60

Tenant dce87315-8ffa-4a01-ab40-0de5a7214b2f

Publish (from ADF Studio) Enabled

Custom comment Enabled

Git configuration

ARM template

Author

Triggers

Global parameters

Data flow libraries

Security

Credentials

Customer managed key

Outbound rules

Managed private endpoints

Workflow orchestration manager

Apache Airflow

76°F Partly sunny

Search

9:45 AM 10/5/2023



A screenshot of a Microsoft Edge browser window showing the Azure DevOps interface for a project named "EV Carbon Emission Analysis".

The browser tab bar shows multiple open tabs, including "Subscription", "Data factory", "akshatgupta1", "EV Carbon En", "CS2\_Shell\_ID", "CS2 - Edraw", "www.bing.co", and a blank tab.

The main content area displays the Azure DevOps repository structure:

- EV Carbon Emission Analysis** (repository name)
- dataset**
- factory**
- linkedService**
- pipeline**
- azure-pipelines.yml**
- publish\_config.json**
- readme.md**

The "Files" tab is selected in the left sidebar. A message at the bottom of the file list says "Initialized by Azure Data Factory!"

The right panel shows a table of files with the following data:

Name ↑	Last change	Commits
dataset	6h ago	e7d40a59 Adding database...
factory	6h ago	e7d40a59 Adding database...
linkedService	6h ago	e7d40a59 Adding database...
pipeline	6h ago	e7d40a59 Adding database...
azure-pipelines.yml	1h ago	cdac09d9 Update azure-...
publish_config.json	6h ago	cfdb8c70d Update publis...
readme.md	6h ago	d59d9e16 Initial commit. ...

At the top right of the file list, there is a red "failed" status indicator and a "Clone" button.



Search



4:00 PM  
ENG IN  
10/5/2023



# YAML to push the changes back to GitHub

```
trigger:
- '*'

pool:
  vmImage: 'windows-latest'

steps:
- task: UsePythonVersion@0
  inputs:
    versionSpec: '3.x'
    addPath: true

- script:
  - git config --global user.email "guptakshat02@gmail.com"
  - git config --global user.name "akshatgupta18"
  displayName: 'Configure Git'

- checkout: self
  persistCredentials: true

- script:
  - git clone https://ghp_dunko4RccV4z40RbAEn7bqavxQOhN1E50JA@github.com/Vishwanatha87/testdf.git
  - cd testdf
  - cp -r https://bia12t4legrdxpsri2o4zndvyjw7wqmi77ihhebbf5nvhlxfwba@dev.azure.com/Shellunext16934221259/CS2_Ed/_apis/pipelines/16934221259/runs/16934221259/outputs/1/contents/* .
```

This branch is 14 commits ahead, 1 commit behind main.

Commit	Author	Date	Message
678cbcc	RheasCode	52 minutes ago	Add files via upload
		1 hour ago	dataset Add files via upload
		1 hour ago	factory Adding pipeline: PL_Grp5_CS2
		1 hour ago	linkedService Adding pipeline: PL_Grp5_CS2
		1 hour ago	pipeline Adding pipeline: PL_Grp5_CS2
		52 minutes ago	ADB_Notebook (1).ipynb Add files via upload
		52 minutes ago	Aggregation_Notebook.ipynb Add files via upload
		53 minutes ago	CS2_Grp5_Report.pbix Add files via upload
		1 hour ago	publish_config.json Update publish_config.json
		1 hour ago	readme.md Initial commit.



# Azure Data Factory

Home >

Microsoft.DataFactory-20231004151104 | Overview

Deployment

Search

Delete

Cancel

Redeploy

Download

Refresh

Overview

Inputs

Outputs

Template

✓ Your deployment is complete

Deployment name : Microsoft.DataFactory-202... Start time : 10/4/2023, 3:13:22 PM  
Subscription : npunext-1673505276902 Correlation ID : 9ce1fbcd-299d-4e13-9592...  
Resource group : RG\_Grp5\_CS2

> Deployment details

✓ Next steps

Go to resource

Give feedback

Tell us about your experience with deployment

## ADF Deployed

## Storage account created

Microsoft Azure Search resources, services, and docs (G+/)

Home > storagegrp5cs2\_1696412825099 | Overview >

storagegrp5cs2 Storage account

Search Upload Open in Explorer Delete Move Refresh Open in mobile CLI / PS Feedback JSON View

Overview Activity log Tags Diagnose and solve problems Access Control (IAM) Data migration Events Storage browser Storage Mover

Resource group (move) RG\_Grp5\_CS2 Location East US Subscription (move) npunext-1673505276902 Subscription ID 77bdf056-d2d7-4d5f-85dd-48ed6d6eac2b Disk state Available Tags (edit) Add tags

Properties Monitoring Capabilities (7) Recommendations (0) Tutorials Tools + SDKs

Blob service Hierarchical namespace Disabled Security + networking Default access tier Enabled

Require secure transfer for REST API operations



Data Factory   Validate all   Publish all

Factory Resources < >

DS\_github DS\_Sink PL\_Grp5\_CS2

Validate Debug Add trigger

Pipelines 1

PL\_Grp5\_CS2

Change Data Capture (preview) 0

Datasets 2

DS\_Sink DS.github

Data flows 0

Power Query 0

Filter resources by name +

Copy data ✓ Github to Blob

Parameters Variables Settings Output

Pipeline run ID: b68208b7-7ba0-479c-8877-4b97392600c1

Pipeline status Succeeded

All status

Showing 1 - 1 of 1 items

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User
Github to Blob	Succeeded	Copy data	10/4/2023, 3:51:39 PM	14s	AutoResolveIntegrator	

## Copy Data Activity

Pipeline Success

Microsoft Azure | Data Factory > ADF-Grp5-CS2

Search factory and documentation

Home Author Monitor Manage Learning Center

Activities < >

DS\_github DS\_Sink PL\_Grp5\_CS2

Validate Validate copy runtime Debug Add trigger

Copy data ✓ Github to Blob

Source dataset \* DS.github

Request method \* GET

Additional headers

Request body

General Source Sink Mapping Settings User properties

Source dataset \* DS.github

Request method \* GET

Additional headers

Request body



## Parent Folder

Azure Storage Container Overview for 'storagegrp5cs2 | Containers'. The 'parent' container is listed in the table:

Name	Last modified	Anonymous access level	Lease state
Slogs	10/4/2023, 3:17:51 PM	Private	Available
parent	10/4/2023, 3:20:03 PM	Private	Available

## File Added to Raw

Azure Storage Blob Overview for 'parent / raw'. The 'raw' blob container is listed in the table:

Name	Modified	Access tier	Archive status	Blob type
[...]	10/4/2023, 3:51:51 PM	Hot (Inferred)		Block blob
ADF_RawData				

## Raw Folder Created

Azure Storage Container Overview for 'parent'. A new blob named 'raw' is listed in the table:

Name	Modified	Access tier	Archive status	Blob type
raw				

## Raw data

Azure Storage Blob Preview for 'raw/ADF\_RawData'. The blob contains the following text:

```
1 Make,Model,Vehicle Class,Engine Size(L),Cylinders,Transmission,Fuel Type,Fuel Consumption City (L/100 km),Fuel Consumption Hwy (L/100 km),Fuel Co  
2 ACURA,ILX,COMPACT,2,4,AS5,Z,9,9,6,7,8,5,33,196  
3 ACURA,ILX,COMPACT,2,4,4,M6,Z,11,2,7,7,9,6,29,221  
4 ACURA,ILX HYBRID,COMPACT,1,5,4,AV7,Z,1,6,5,8,5,9,48,136  
5 ACURA,MDX 4WD,SUV - SMALL,3,5,6,AS6,Z,12,7,9,1,11,1,25,255  
6 ACURA,RDX AWD,SUV - SMALL,3,5,6,AS6,Z,12,1,8,7,10,6,27,244  
7 ACURA,RLX,MID-SIZE,3,5,6,AS6,Z,11,9,7,7,10,28,230  
8 ACURA,TL,MID-SIZE,3,5,6,AS6,Z,11,8,8,1,10,1,28,232  
9 ACURA,TL AWD,MID-SIZE,3,7,6,AS6,Z,12,8,9,11,1,25,255  
10 ACURA,TL AWD,MID-SIZE,3,7,6,M6,Z,13,4,9,5,11,6,24,267  
11 ACURA,TSX,COMPACT,2,4,4,AS5,Z,18,6,7,5,9,2,31,212  
12 ACURA,TSX,COMPACT,2,4,4,M6,Z,11,2,8,1,9,8,29,221  
13 ACURA,TSX,COMPACT,3,5,6,AS5,Z,12,1,8,3,10,4,27,239  
14 ALFA ROMEO,4C,TWO-SEATER,1,8,4,AM6,Z,9,7,6,9,8,4,34,193  
15 ASTON MARTIN,B99,MINICOMPACT,5,9,12,A6,Z,18,12,6,15,6,18,359  
16 ASTON MARTIN,RAPIDE,SUBCOMPACT,5,9,12,A6,Z,18,12,6,15,6,18,359  
17 ASTON MARTIN,V8 VANTAGE,TWO-SEATER,4,7,8,AM7,Z,17,4,11,3,14,7,19,338
```



# Azure Data Bricks

The screenshot shows the Microsoft Azure portal's "RG\_Grp5\_CS2\_ADB\_Grp5\_CS2 | Overview" page. A prominent green success message box at the top right states "Deployment succeeded" for deployment "RG\_Grp5\_CS2\_ADB\_Grp5\_CS2" to resource group "RG\_Grp5\_CS2". Below this, there are sections for "Cost management", "Microsoft Defender for Cloud", and "Work with an expert". The left sidebar lists "Overview", "Inputs", "Outputs", and "Template". The bottom of the screen shows a Windows taskbar with various icons.

## ADB Deployment

The screenshot shows the Azure Data Bricks Compute blade. It displays a table of clusters, with one entry for "Shell76 Unext's Cluster" which has a state of "13.3", a runtime of "42 GB", and 12 cores. The blade also includes sections for "Compute", "SQL", "Data Engineering", and "Machine Learning". The bottom of the screen shows a Windows taskbar.

## ADB Compute

The screenshot shows an Azure Data Bricks Notebook titled "ADB\_Notebook - Python". The notebook contains a single cell with Python code for mounting a blob storage account. The code imports `dbutils` and `storageAccountName` and `storageAccountAccessKey`. It then attempts to mount the storage account using `dbutils.fs.mount` and prints the result. The output shows the command took 18.12 seconds and was successful. The left sidebar of the notebook interface is visible.

## ADB Mount Success

The screenshot shows an Azure Data Bricks Notebook titled "ADB\_Notebook - Python". The notebook contains four cells of Python code for raw mounting a storage account. The first cell uses `dbutils.fs.mounts` to list existing mounts. The second cell uses `dbutils.fs.mount` to mount a blob storage account. The third cell uses `dbutils.fs.ls` to list files in the mounted directory. The fourth cell uses `dbutils.fs.unmount` to unmount the storage account. The output of the second cell shows the command took 0.41 seconds. The left sidebar of the notebook interface is visible.

## ADB Raw Mount



ADB\_Notebook Python

```
df = spark.read.option("inferSchema", True).option("header", True).csv("/mnt/staging/raw/ADF_RawData")
```

df.display()

Make	Model	Vehicle Class	Engine Size(L)	Cylinders
ACURA	ILX	COMPACT	2	4
ACURA	ILX	COMPACT	2.4	4
ACURA	ILX HYBRID	COMPACT	1.5	4
ACURA	MDX 4WD	SUV - SMALL	3.5	6
ACURA	RDX AWD	SUV - SMALL	3.5	6
ACURA	RLX	MID-SIZE	3.5	6
ACURA	Tl	MID-SIZE	3.5	6

7,385 rows | 0.76 seconds runtime

## Reading File

ADB\_Notebook Python

```
df.distinct() == df.dropDuplicates()
```

print(df.count() - df\_distinct.count())

1103

1103 Command took 0.81 seconds -- by shellunext\_1693422228175@pnunext.onmicrosoft.com at 10/5/2023, 9:55:49 AM on Shell76 Unext's Cluster

## Duplicates Dropped

ADB\_Notebook Python

```
df_distinct.count()
```

df\_distinct.dropna().count()

6282

6282 Command took 0.51 seconds -- by shellunext\_1693422228175@pnunext.onmicrosoft.com at 10/5/2023, 9:58:22 AM on Shell76 Unext's Cluster

1

Shift+Enter to run  
Shift+Ctrl+Enter to run selected text

## Dropping NA values

ADB\_Notebook Python

```
from functools import reduce
```

```
oldColumns = df_distinct.schema.names
```

```
newColumns = ["Make", "Model", "Vehicle_class", "Engine_size(L)", "Cylinder", "Transmission", "Fuel_type", "FC_city", "FC_fwy", "FC_comb", "FC_comb_mp", "CO2_emission(g/km)"]
```

```
df_rename = reduce(lambda data, idx: data.withColumnRenamed(oldColumns[idx], newColumns[idx]), range(len(oldColumns)), df_distinct)
```

```
df_rename = pyspark.sql.DataFrameDataframe = [Make string, Model string ... 10 more fields]
```

df\_rename

df\_rename.display()

Make	Model	Vehicle_class	Engine_size(L)	Cylinder
CHEVROLET	SONIC RS	COMPACT	1.4	4
FORD	TRANSIT CONNECT WAGON	SPECIAL PURPOSE VEHICLE	2.5	4

## Renaming columns



```
1 from pyspark.sql.functions import *
2 df_new = df_rename.withColumn("Fuel_new", when((df_rename.Fuel_type == "N"),("Petro!")).when((df_rename.Fuel_type == "X"),("Hybrid!")).when((df_rename.Fuel_type == "Z"),("CNG!")).when((df_rename.Fuel_type == "D"),("Diesel!")).when((df_rename.Fuel_type == "E"), ("Electric!")))

> df_new: pyspark.sql.dataframe.DataFrame = [Company: string, Model: string ... 11 more fields]

Command took 0.15 seconds -- by shellunext_169342228175@pnunext.onmicrosoft.com at 10/5/2023, 11:25:50 AM on ShellT6 Unix's Cluster
```

```
Cmd 16
1 cols = ["Fuel_type", "Fuel_new"]
2 df_new.select(*cols).show()

+---+---+
|Fuel_type|Fuel_new|
+---+---+
|X| Hybrid|
|X| Hybrid|
|E| Electric|
|Z| CNG|
|X| Hybrid|
|X| Hybrid|
|Z| CNG|
|X| Hybrid|
|Z| CNG|
+---+
```

## Fuel New Type

```
1 columnsdrop = ["Fuel_type", "FC_comb_mpg"]
2 df_final = df_new.drop(*columnsdrop)

> df_final: pyspark.sql.dataframe.DataFrame = [Company: string, Model: string ... 9 more fields]

Command took 0.10 seconds -- by shellunext_169342228175@pnunext.onmicrosoft.com at 10/5/2023, 11:29:13 AM on ShellT6 Unix's Cluster
```

```
Cmd 18
1 df_final.schema

StructType([StructField('Company', StringType(), True), StructField('Model', StringType(), True), StructField('Vehicle_class', StringType(), True), StructField('Engine_size(L)', DoubleType(), True), StructField('Cylinder', IntegerType(), True), StructField('Transmission', StringType(), True), StructField('FC_city', DoubleType(), True), StructField('FC_highway', DoubleType(), True), StructField('FC_comb', DoubleType(), True), StructField('CO2_emission(g/km)', IntegerType(), True), StructField('Fuel', StringType(), True)])
```

```
Cmd 19
1
```

## Drop Column

```
1 df_final.display()

+---+---+---+---+---+
|Company|Model|Vehicle_class|Engine_size(L)|Cylinder|
+---+---+---+---+---+
|1| CHEVROLET| SONIC RS| COMPACT| 1.4| 4|
|2| FORD| TRANSIT CONNECT WAGON| SPECIAL PURPOSE VEHICLE| 2.5| 4|
|3| GMC| SAVANA 1500 CARGO| VAN - CARGO| 5.3| 8|
|4| LEXUS| GS 450h| MID-SIZE| 3.5| 6|
|5| NISSAN| SENTRA| MID-SIZE| 1.8| 4|
|6| CHEVROLET| CAMARO 2LS| COMPACT| 3.6| 6|
|7| ASTON MARTIN| VR VANTAGE S| TWO-SEATER| 4.7| 8|
+---+---+---+---+---+
6,282 rows | 0.60 seconds runtime
Refreshed 1 hour ago
```

```
Command took 0.60 seconds -- by shellunext_169342228175@pnunext.onmicrosoft.com at 10/5/2023, 1:39:12 PM on ShellT6 Unix's Cluster
```

```
Cmd 18
1 df_final.write.parquet("/mnt/parent/staging/stg_co2emissions")
```

```
> (2) Spark Jobs
```

## Parquet Upload



```
dbutils.fs.ls("/mnt/parent/staging/")

[FileInfo(path='dbfs:/mnt/parent/staging/raw/', name='raw', size=0, modificationTime=0),
 FileInfo(path='dbfs:/mnt/parent/staging/stg_co2emissions/', name='stg_co2emissions', size=0, modificationTime=169493369000)]

Command took 0.12 seconds -- by shellunext_1693422228175@pnunext.onmicrosoft.com at 10/5/2023, 1:42:14 PM on ShellT6 Unext's Cluster
```

```
df = spark.read.option("inferSchema", True).option("header", True).parquet("/mnt/parent/staging/stg_co2emissions/")

(1) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Company: string, Model: string ... 9 more fields]

Command took 1.59 seconds -- by shellunext_1693422228175@pnunext.onmicrosoft.com at 10/5/2023, 1:45:36 PM on ShellT6 Unext's Cluster
```

## Agg Mount

```
df = df.drop("Transmission")

(1) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Company: string, Model: string ... 8 more fields]

Command took 0.07 seconds -- by shellunext_1693422228175@pnunext.onmicrosoft.com at 10/5/2023, 1:45:47 PM on ShellT6 Unext's Cluster
```

```
from pyspark.sql.functions import col
df_aggregate = df.withColumn("FC_c60h40", ((col("FC_city") * 0.6) + (col("FC_hwy") * 0.4)))

(1) Spark Jobs
df_aggregate: pyspark.sql.dataframe.DataFrame = [Company: string, Model: string ... 9 more fields]

Command took 0.13 seconds -- by shellunext_1693422228175@pnunext.onmicrosoft.com at 10/5/2023, 2:00:37 PM on ShellT6 Unext's Cluster
```

## Weighted Columns Created

```
df_aggregate.display()

(1) Spark Jobs
Table + New result table: OFF

Company Model Vehicle_class Engine_size(L) Cylinder
1 CHEVROLET SONIC RS COMPACT 1.4 4
2 FORD TRANSIT CONNECT WAGON SPECIAL PURPOSE VEHICLE 2.5 4
3 GMC SAVANA 1500 CARGO VAN - CARGO 5.3 8
4 LEXUS GS 450h MID-SIZE 3.5 6
5 NISSAN SENTRA MID-SIZE 1.8 4
6 CHEVROLET CAMARO ZLS COMPACT 3.6 6
7 DXTON MARTIN VR VANTAGE C TWIN-POWER 4.7 8

6,282 rows | 0.44 seconds runtime
Refreshed 59 minutes ago

Command took 0.44 seconds -- by shellunext_1693422228175@pnunext.onmicrosoft.com at 10/5/2023, 2:01:41 PM on ShellT6 Unext's Cluster
```

```
df_aggregate.write.csv("/mnt/parent/staging/curated/final_agg_canada")

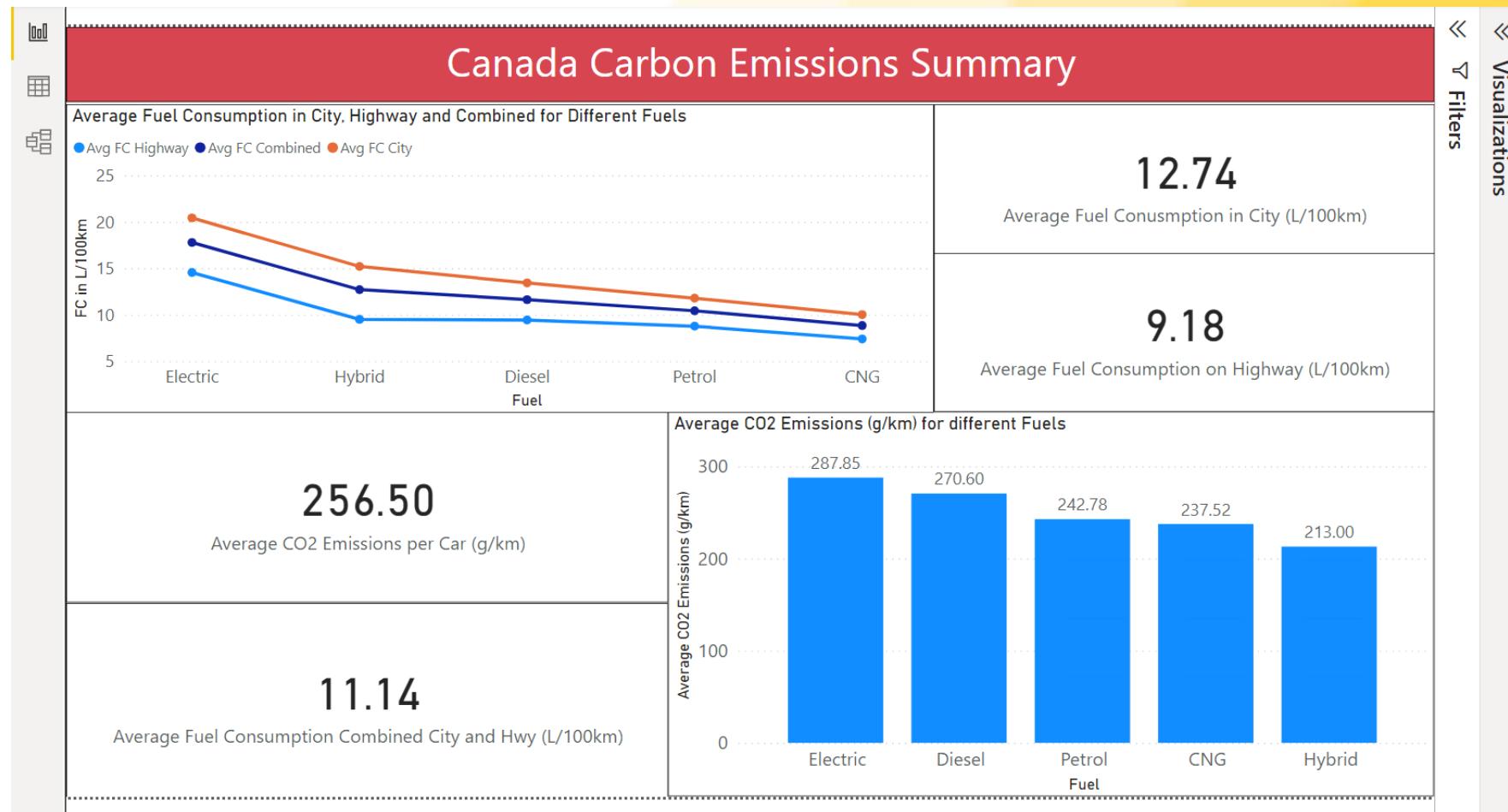
(1) Spark Jobs
```

## Agg Uploading to Blob



# Power BI

Q. What are the most influencing features that affect the CO2 emission the most?





Q. Determine or test the influence of different variables on the emission of CO<sub>2</sub>.

Get the latest data by refreshing all visuals in this report.

## Canada Carbon Emissions Comprehensive Report

Filters

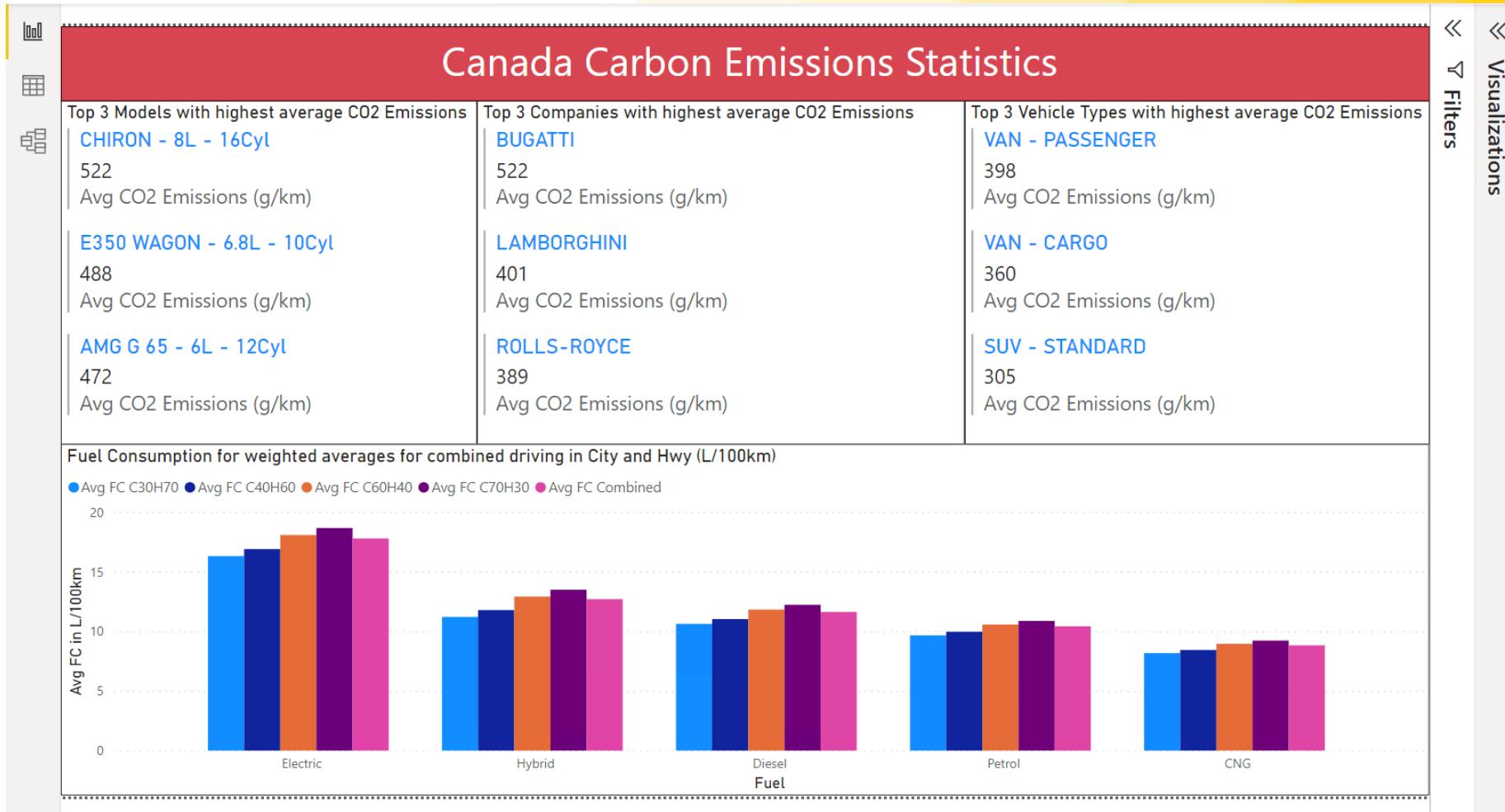
Visualizations

Company	Fuel	Model	Vehicle_class	
All	All	All	All	
<b>Fuel Consumption (FC) and Carbon Emissions for all Car Models</b>				
Company	Avg FC in City (L/100km)	Avg FC in Highway (L/100km)	Avg FC Combined (L/100km)	Avg CO <sub>2</sub> (g/km)
ACURA	11.03	8.19	9.75	
COMPACT	10.48	7.35	9.08	
ILX - 2.4L - 4Cyl	10.12	7.10	8.76	
ILX - 2L - 4Cyl	9.80	6.70	8.40	
ILX HYBRID - 1.5L - 4Cyl	6.00	5.95	6.00	
TLX - 2.4L - 4Cyl	9.70	6.73	8.40	
TLX - 3.5L - 6Cyl	11.20	6.90	9.20	
TLX A-SPEC - 2.4L - 4Cyl	10.20	7.40	8.90	
TLX SH-AWD - 3.5L - 6Cyl	11.27	7.57	9.67	
TLX SH-AWD A-SPEC - 3.5L - 6Cyl	12.00	8.20	10.30	
TLX SH-AWD A-SPEC/Limited Edition - 3.5L - 6Cyl	12.00	8.20	10.30	
TSX - 2.4L - 4Cyl	10.90	7.80	9.50	
TSX - 3.5L - 6Cyl	12.10	8.30	10.40	
MID-SIZE	11.25	8.22	9.88	
Total	12.74	9.18	11.14	



Q. Will there be any difference in the CO2 emissions when Fuel Consumption for City and Highway are considered separately and when their weighted variable interaction is considered?

Q. Comparison b/w -- CarTypes, Company Wise, Different Models





# GitHub link

[https://github.com/RheasCode/Grp5\\_CS2/tree/ADF\\_collab\\_branch](https://github.com/RheasCode/Grp5_CS2/tree/ADF_collab_branch)

The screenshot shows a GitHub repository page for the URL [https://github.com/RheasCode/Grp5\\_CS2/tree/ADF\\_collab\\_branch](https://github.com/RheasCode/Grp5_CS2/tree/ADF_collab_branch). The repository has 0 forks and no releases published. It features a dark theme. A list of commits is displayed, all made by the user 'RheasCode' 29 minutes ago. The commits include:

- dataset: Add files via upload, 32 minutes ago
- factory: Adding pipeline: PL\_Grp5\_CS2, 46 minutes ago
- linkedService: Adding pipeline: PL\_Grp5\_CS2, 46 minutes ago
- pipeline: Adding pipeline: PL\_Grp5\_CS2, 46 minutes ago
- ADB\_Notebook (1).ipynb: Add files via upload, 29 minutes ago
- Aggregation\_Notebook.ipynb: Add files via upload, 29 minutes ago
- CS2\_Grp5\_Report.pbix: Add files via upload, 30 minutes ago
- publish\_config.json: Update publish\_config.json, 46 minutes ago
- readme.md: Initial commit, 46 minutes ago

Below the commits is a 'readme.md' file containing the text: "Initialized by Azure Data Factory!"



# Challenges faced & Feedback

1. The data isn't as precise as needed, as vehicles with the same specifications have different fuel consumption levels (on both city and highway) as well as different carbon emission levels. Further going through the data, we can also notice that the values in "fuel consumption combined" column are not the same as calculated using the "fuel consumption city" and "fuel consumption hwy".
2. The different fuel types (N, D, E, Z, X) were not labeled as to what fuel type they represented in the data dictionary, so educated guesses need to be made as to what they could mean.
3. The conversion of Fuel Consumption Combined (L/100 km) to Miles per Gallon doesn't equal the values given in the Fuel Consumption Combined (mpg).



# THANK YOU

