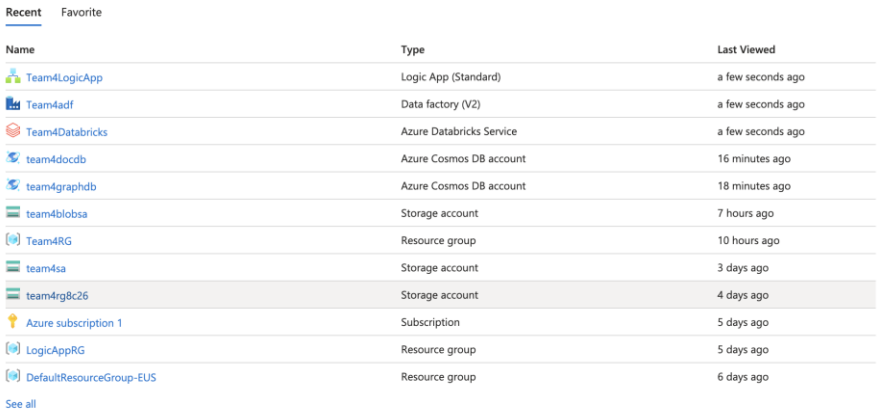
P-3: Implementation

Global Inbound and Outbound Travel

Introduction:

To implement the project, we have followed our architecture diagram. We have used many services available on Azure cloud platform which are listed below.

1. Azure Logic Apps
2. Azure Databaricks
3. Azure Blob Storage
4. Azure Data Factory
5. Azure Cosmos DB



Recent Favorite

Name	Type	Last Viewed
Team4LogicApp	Logic App (Standard)	a few seconds ago
Team4adf	Data factory (V2)	a few seconds ago
Team4Databricks	Azure Databricks Service	a few seconds ago
team4docdb	Azure Cosmos DB account	16 minutes ago
team4graphdb	Azure Cosmos DB account	18 minutes ago
team4blobsa	Storage account	7 hours ago
Team4RG	Resource group	10 hours ago
team4sa	Storage account	3 days ago
team4rg8c26	Storage account	4 days ago
Azure subscription 1	Subscription	5 days ago
LogicAppRG	Resource group	5 days ago
DefaultResourceGroup-EUS	Resource group	6 days ago

See all

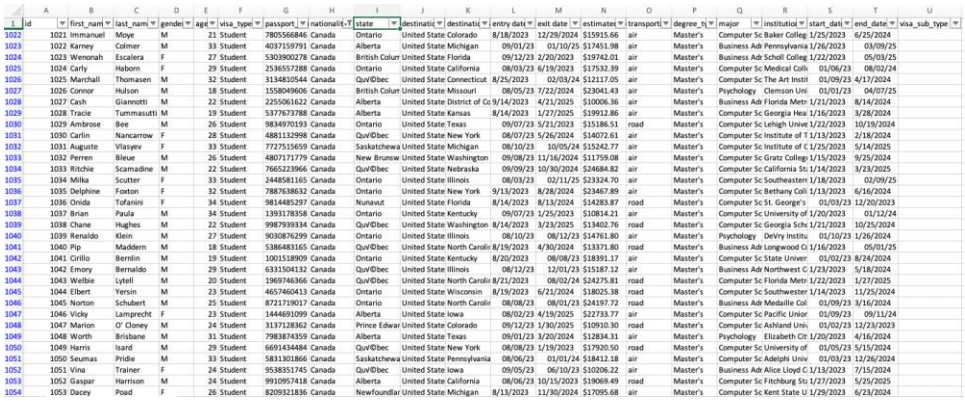
Dataset Details:

Our primary dataset has been loaded from CSV files. It contains the details if inbound travel and travel information to and across different countries. The data dictionary is listed below.

Column Name	Details
id	Numeric id
First_name	Tourist's first name
Last_name	Tourist's last name
gender	Tourist's gender
age	Tourist's age
Visa_type	Type of visa that tourist holds (business, pleasure, student)
Passport_number	Tourist's passport number

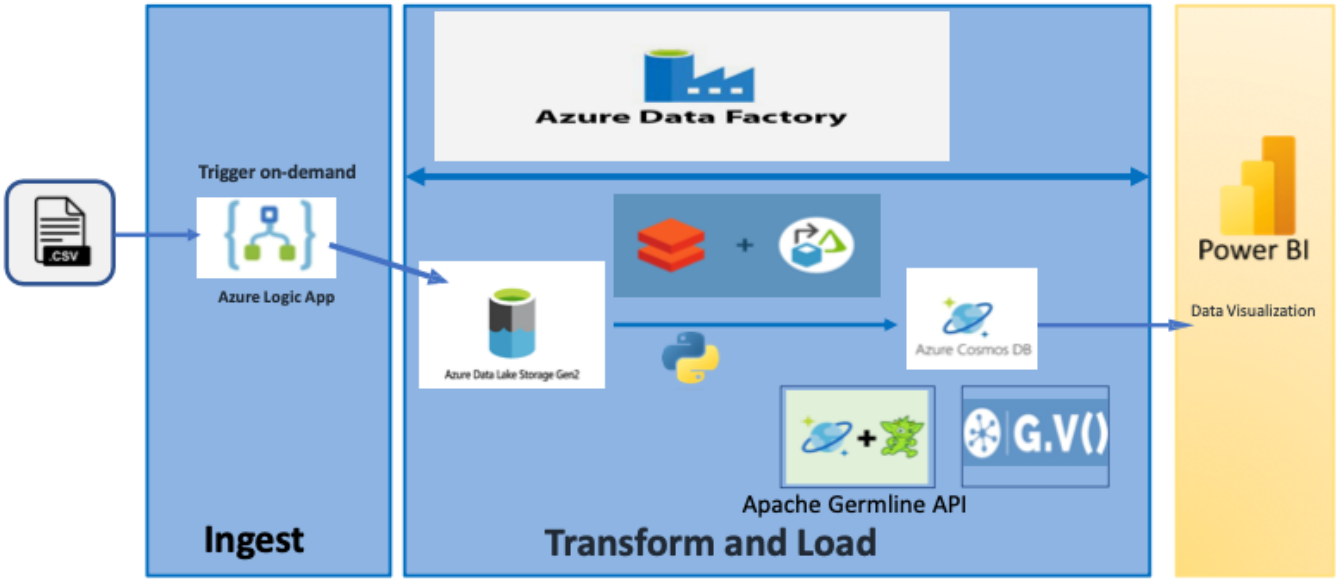
nationality	Tourist's country of origin
state	Tourist's state of origin
Destination_country	Destination country travelled
Destination_state	Destination state travelled
Entry date	Entry date in the destination country
Exit_date	Exit date from the destination country
Estimated expenses	Estimated expenditure in the destination country
transportation	Transport mode used to reach destination country
Degree_type	Degree type that student is pursuing (only available for student visa type)
major	Major that student is pursuing (only available for student visa type)
institution	Institution where the student is enrolled(only available for student visa type)
Start date	Program start date for students (only available for student visa type)
End date	Program end date for students (only available for student visa type)
Visa_Sub_type	visa sub category (only available for business visa type)

Data sample



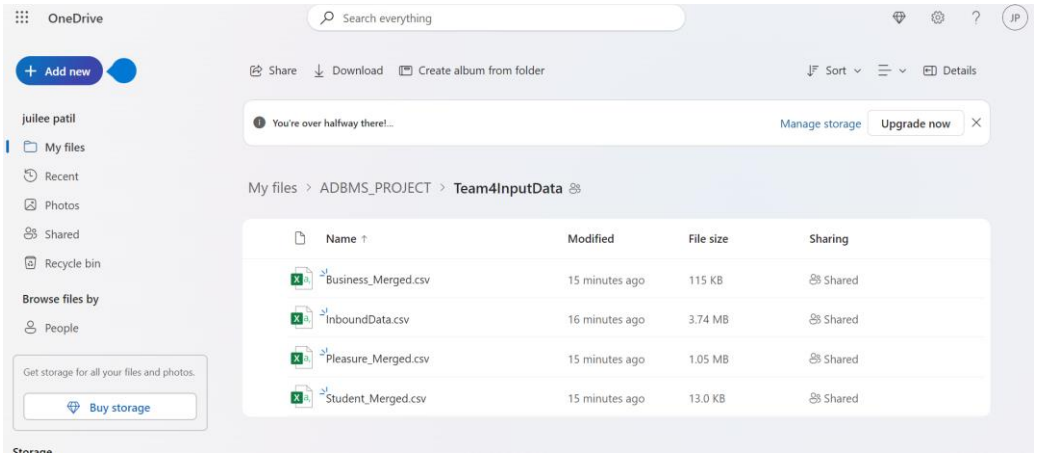
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	first_nam	last_nam	gende	age	visa_type	passport_	nationalit	state	destinatic	destinatic	entry dat	exit date	estimate	transport	degree_t	major	institutio	start_dat	end_date	visa_sub_type
1022	1021	Immanuel	Moye	M	21	Student	7805566846	Canada	Ontario	United State	Colorado	8/18/2023	12/29/2024	$15915.66	air	Master's	Computer Sc	Baker Colleg	1/25/2023	6/25/2024	
1023	1022	Karney	Colmer	M	33	Student	4037159791	Canada	Alberta	United State	Michigan	09/01/23	01/10/25	$17451.98	air	Master's	Business Adr	Pennsylvania	1/26/2023	03/09/25	
1024	1023	Wenonah	Escalera	F	27	Student	5303900278	Canada	British Colun	United State	Florida	09/12/23	2/20/2023	$19742.01	air	Master's	Business Adr	Scholl Colleg	1/22/2023	05/03/25	
1025	1024	Carly	Haborn	F	29	Student	2536557288	Canada	Ontario	United State	California	08/03/23	6/19/2023	$17532.39	air	Master's	Computer Sc	Medical Colle	01/06/23	08/02/24	
1026	1025	Marchall	Thomasen	M	32	Student	3134810544	Canada	Qué bec	United State	Connecticut	8/25/2023	02/03/24	$12117.05	air	Master's	Computer Sc	The Art Instit	01/09/23	4/17/2024	
1027	1026	Connor	Hulson	M	18	Student	1558049606	Canada	British Colun	United State	Missouri	08/05/23	7/22/2024	$23041.43	air	Master's	Psychology	Clemson Uni	01/01/23	04/07/25	
1028	1027	Cash	Giannotti	M	22	Student	2255061622	Canada	Alberta	United State	District of Co	9/14/2023	4/21/2025	$10006.36	air	Master's	Business Adr	Florida Metr	1/21/2023	8/14/2024	
1029	1028	Tracie	Tummasutti	M	19	Student	5377673788	Canada	Alberta	United State	Kansas	8/14/2023	1/27/2025	$19912.86	air	Master's	Computer Sc	Georgia Hea	1/16/2023	3/28/2024	
1030	1029	Ambrose	Bee	M	26	Student	9834970193	Canada	Ontario	United State	Texas	09/07/23	5/21/2023	$15186.51	road	Master's	Computer Sc	Lehigh Unive	1/22/2023	10/19/2024	
1031	1030	Carlin	Nancarrow	F	28	Student	4881132998	Canada	Qué bec	United State	New York	08/07/23	5/26/2024	$14072.61	air	Master's	Computer Sc	Institute of T	1/13/2023	2/18/2024	
1032	1031	Auguste	Vlasyev	F	33	Student	7727515659	Canada	Saskatchewa	United State	Michigan	08/10/23	10/05/24	$15242.77	air	Master's	Computer Sc	Institute of C	1/25/2023	5/14/2025	
1033	1032	Perren	Bleue	M	26	Student	4807171779	Canada	New Brunsw	United State	Washington	09/08/23	11/16/2024	$11759.08	air	Master's	Computer Sc	Gratz Colleg	1/15/2023	9/25/2024	
1034	1033	Ritchie	Scamadine	M	22	Student	7665223966	Canada	Qué bec	United State	Nebraska	09/09/23	10/30/2024	$24684.82	air	Master's	Computer Sc	California St	1/14/2023	3/23/2025	
1035	1034	Milka	Scutter	F	33	Student	2448581165	Canada	Ontario	United State	Illinois	08/03/23	02/11/25	$23324.70	air	Master's	Computer Sc	Southeastern	1/18/2023	02/09/25	
1036	1035	Delphine	Foxton	F	32	Student	7887638632	Canada	Ontario	United State	New York	9/13/2023	8/28/2024	$23467.89	air	Master's	Computer Sc	Bethany Coll	1/13/2023	6/16/2024	
1037	1036	Onida	Tofanini	F	34	Student	9814485297	Canada	Nunavut	United State	Florida	8/14/2023	8/13/2024	$14283.87	road	Master's	Computer Sc	St. George's	01/03/23	12/20/2023	
1038	1037	Brian	Paula	M	34	Student	1393178358	Canada	Ontario	United State	Kentucky	09/07/23	1/25/2023	$10814.21	air	Master's	Computer Sc	University of	1/20/2023	01/12/24	
1039	1038	Chane	Hughes	M	22	Student	9987939334	Canada	Qué bec	United State	Washington	8/14/2023	3/23/2025	$13402.76	road	Master's	Computer Sc	Georgia Sch	1/21/2023	10/25/2024	
1040	1039	Renaldo	Klein	M	27	Student	9030876299	Canada	Ontario	United State	Illinois	08/10/23	08/12/23	$14761.80	air	Master's	Psychology	DeVry Institu	01/10/23	1/26/2024	
1041	1040	Pip	Maddern	M	18	Student	5386483165	Canada	Qué bec	United State	North Carolir	8/19/2023	4/30/2024	$13371.80	road	Master's	Business Adr	Longwood Cc	1/16/2023	05/01/25	
1042	1041	Cirillo	Bernlin	M	19	Student	1001518909	Canada	Ontario	United State	Kentucky	8/20/2023	08/08/23	$18391.17	air	Master's	Computer Sc	State Univer	01/02/23	8/24/2024	
1043	1042	Emory	Bernaldo	M	29	Student	6331504132	Canada	Qué bec	United State	Illinois	08/12/23	12/01/23	$15187.12	air	Master's	Business Adr	Northwest C	1/23/2023	5/18/2024	
1044	1043	Welbie	Lytell	M	20	Student	1969746366	Canada	Qué bec	United State	North Carolir	8/21/2023	08/02/24	$24275.81	road	Master's	Computer Sc	Florida Met	1/22/2023	1/27/2025	
1045	1044	Elbert	Yersin	M	23	Student	4657460413	Canada	Ontario	United State	Wisconsin	8/19/2023	6/21/2024	$18025.38	road	Master's	Computer Sc	Southwester	1/14/2023	11/25/2024	
1046	1045	Norton	Schubert	M	25	Student	8721719017	Canada	Ontario	United State	North Carolir	08/08/23	08/01/23	$24197.72	road	Master's	Business Adr	Medaille Col	01/09/23	3/16/2024	
1047	1046	Vicky	Lamprecht	F	23	Student	1444691099	Canada	Alberta	United State	Iowa	08/02/23	4/19/2025	$22733.77	air	Master's	Computer Sc	Pacific Unior	01/09/23	09/11/24	
1048	1047	Marion	O' Cloney	M	24	Student	3137128362	Canada	Prince Edwar	United State	Colorado	09/12/23	1/30/2025	$10910.30	road	Master's	Computer Sc	Ashland Univ	01/02/23	12/23/2023	
1049	1048	Worth	Brisbane	M	31	Student	7983874359	Canada	Alberta	United State	Texas	09/01/23	3/20/2024	$12834.31	air	Master's	Psychology	Elizabeth Cit	1/20/2023	4/16/2024	
1050	1049	Harris	Isard	M	29	Student	6691434484	Canada	Qué bec	United State	New York	08/08/23	1/19/2023	$17920.50	road	Master's	Computer Sc	University of	01/05/23	5/15/2024	
1051	1050	Seumas	Pridie	M	33	Student	5831301866	Canada	Saskatchewa	United State	Pennsylvania	08/06/23	01/01/24	$18412.18	air	Master's	Computer Sc	Adelphi Univ	01/03/23	12/26/2024	
1052	1051	Vina	Trainer	F	24	Student	9538351745	Canada	Qué bec	United State	Iowa	09/05/23	06/10/23	$10206.22	air	Master's	Business Adr	Alice Lloyd C	1/13/2023	7/15/2024	
1053	1052	Gaspar	Harrison	M	24	Student	9910957418	Canada	Alberta	United State	California	08/06/23	10/15/2023	$19069.49	road	Master's	Computer Sc	Fitchburg St	1/27/2023	5/25/2025	
1054	1053	Dacey	Poad	F	26	Student	8209321836	Canada	Newfoundlar	United State	Michigan	8/13/2023	11/30/2024	$17095.68	air	Master's	Computer Sc	Kent State U	1/29/2023	6/23/2024	

Implementation process:

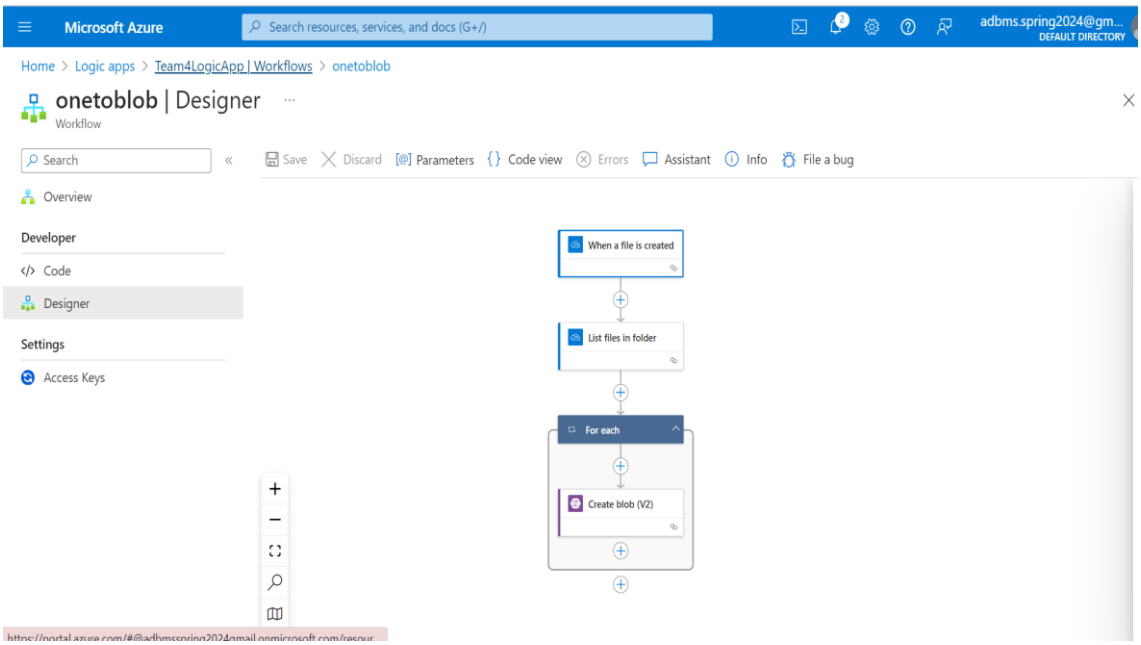


The implementation follows the architecture diagram submitted as part of P2. The process starts with ingesting the data present in the csv file located in the shared onedrive folder to Azure Blob Storage using Azure Logic app. To streamline data from OneDrive to a graph database, a Logic App triggers when a file is created in OneDrive and copies the file to Azure Blob Storage. The process is shown below.

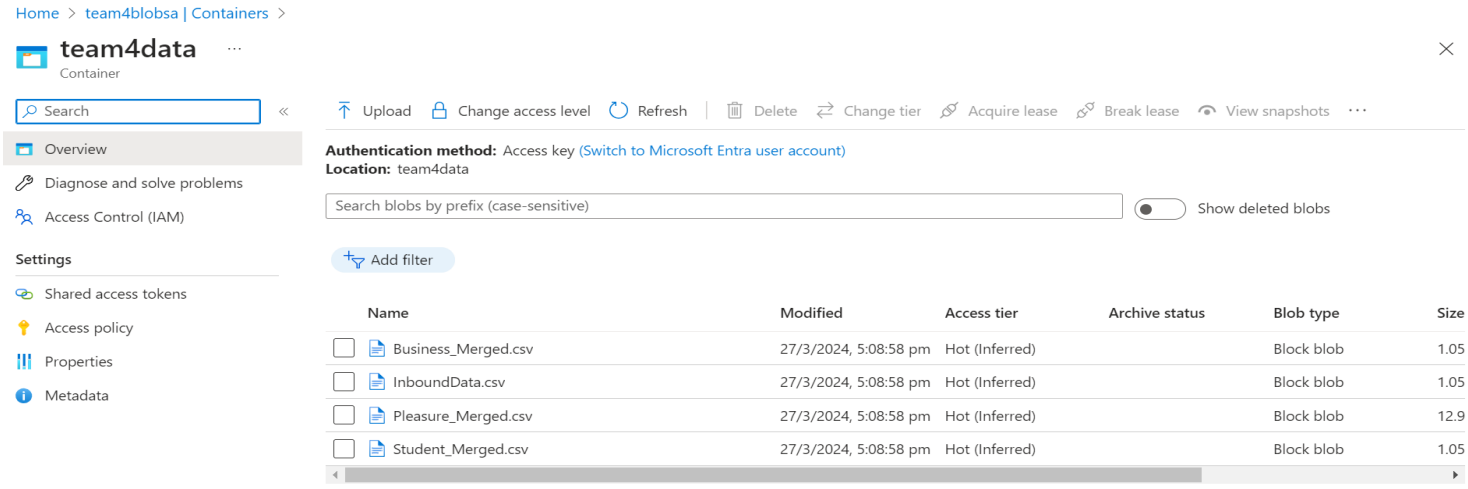
Files present in OneDrive folder locaion



Created Logic App workflow to copy files from one drive to blob storage when a new file is added

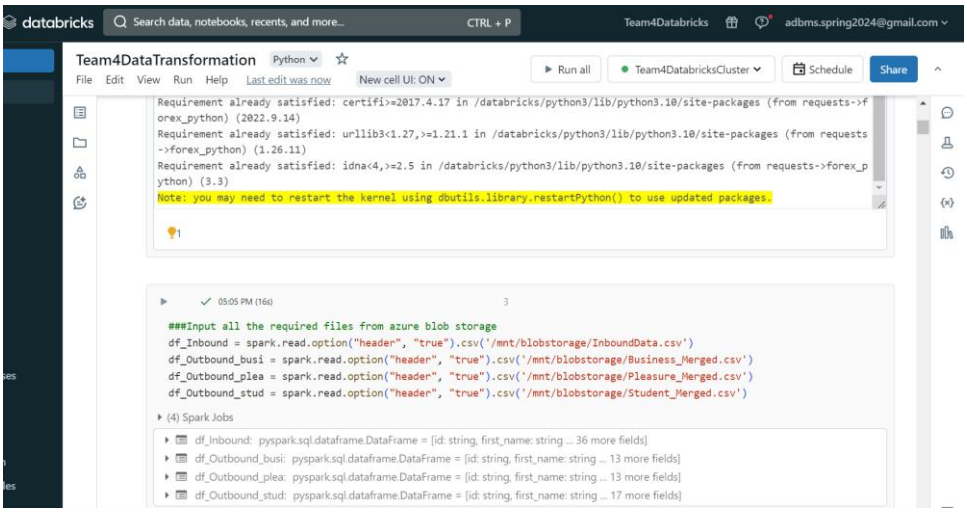


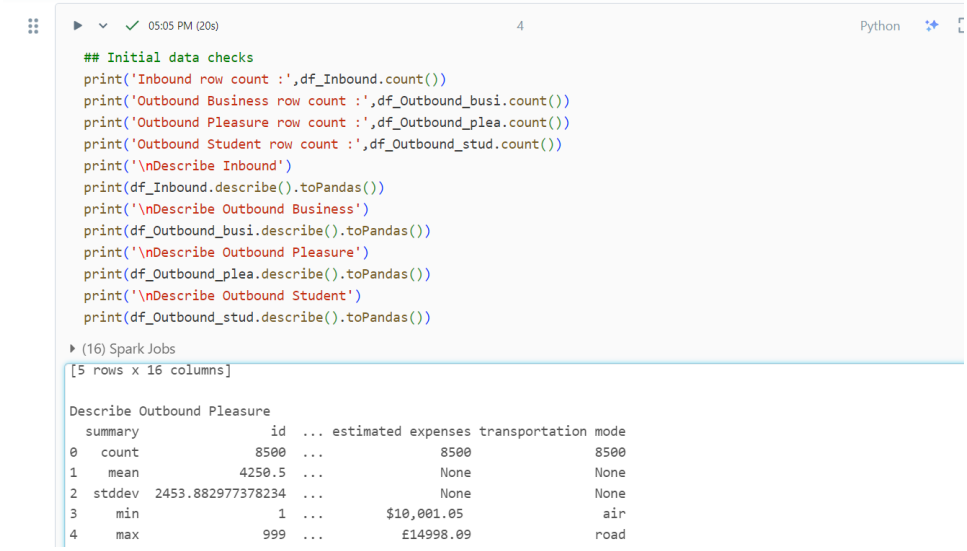
Files is Azure Blob Storage :



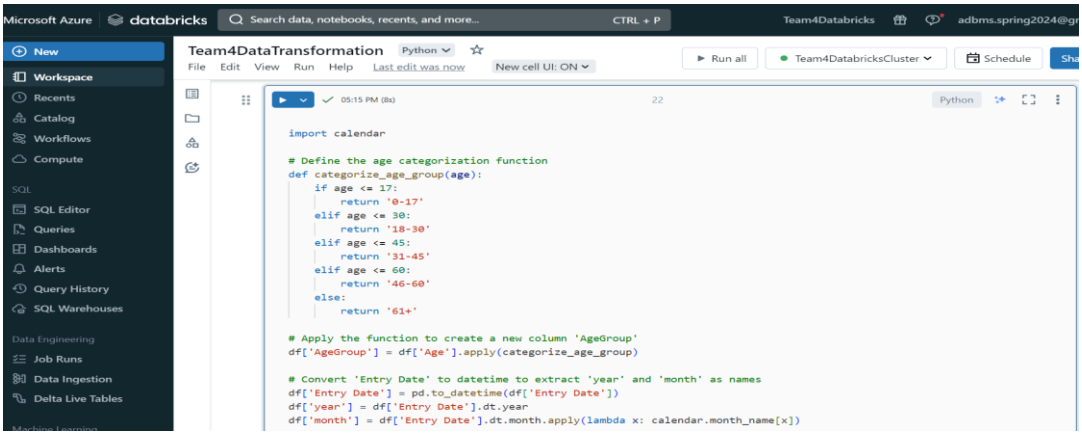
Once the data is loaded in the Blob Storage, Azure Databricks then processes and transforms the data, merging it into a single file suitable for graph database ingestion. Finally, the prepared data is loaded into the graph database and document database for analysis and querying. This automated pipeline requires careful setup and testing to ensure data integrity throughout the flow.

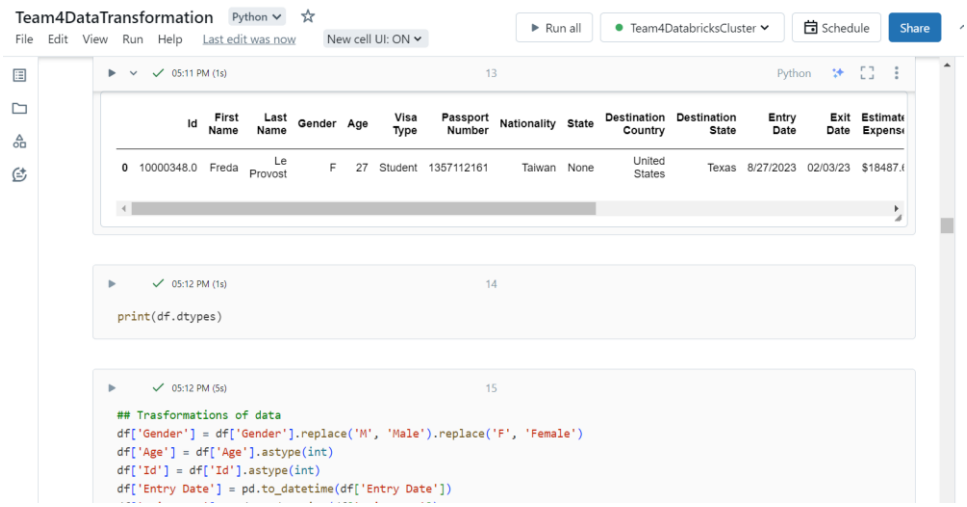
Processing data through databricks . We created a cluster for compute and used databricks notebook for the cleaning , transformation and loading of data into databases. We created a mount to access the data files in blod storage.



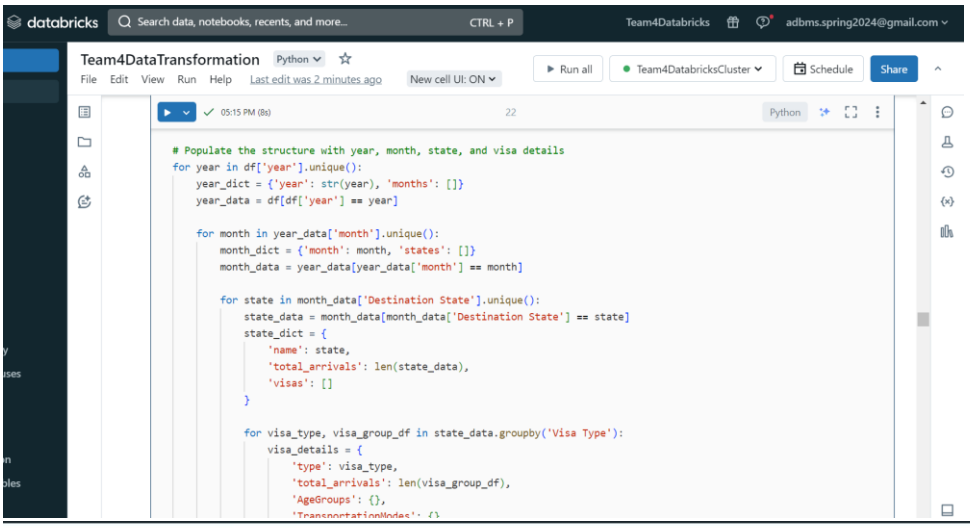


Cleaning and transformation

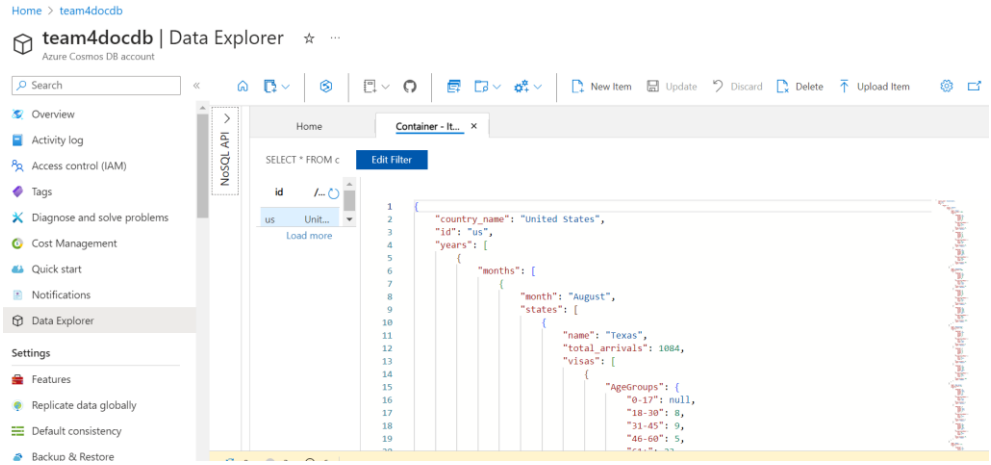




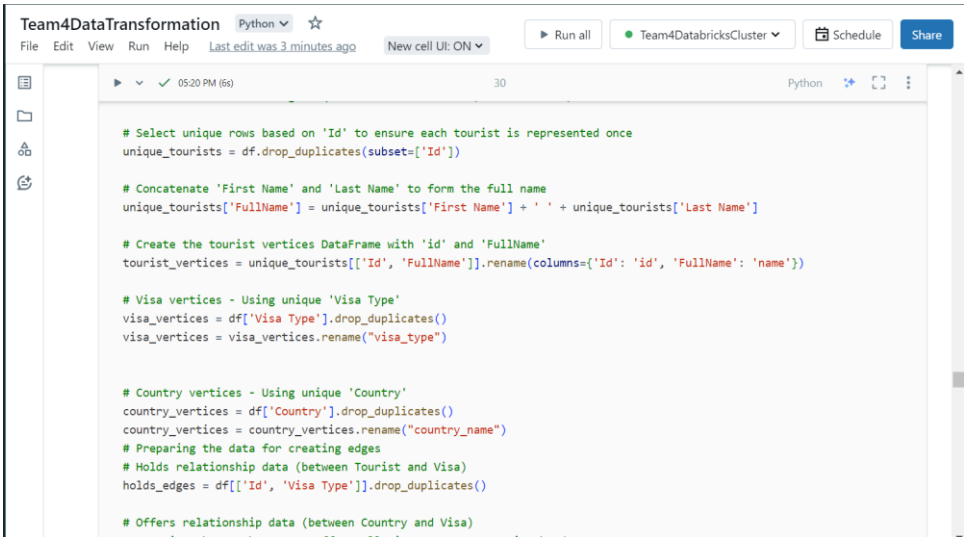
Create a spark notebook to do data transformations and load data in cosmos db for document database

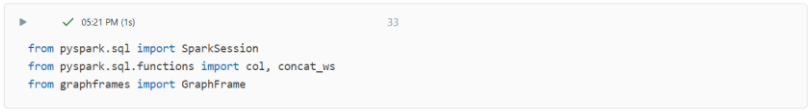


Screenshot of data loaded in Document db



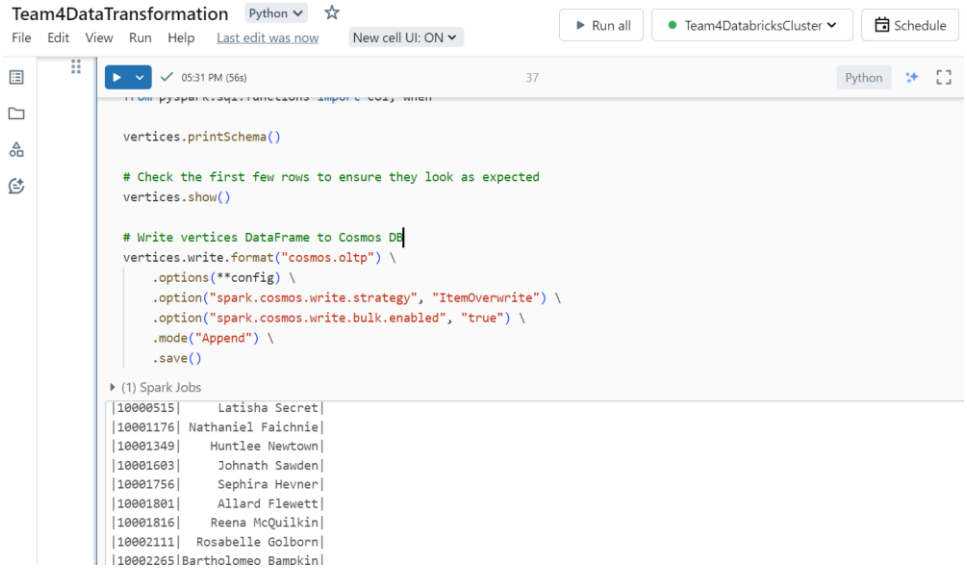
Using same notebook create vertices and edges and then Loaded into Cosmos Graph DB

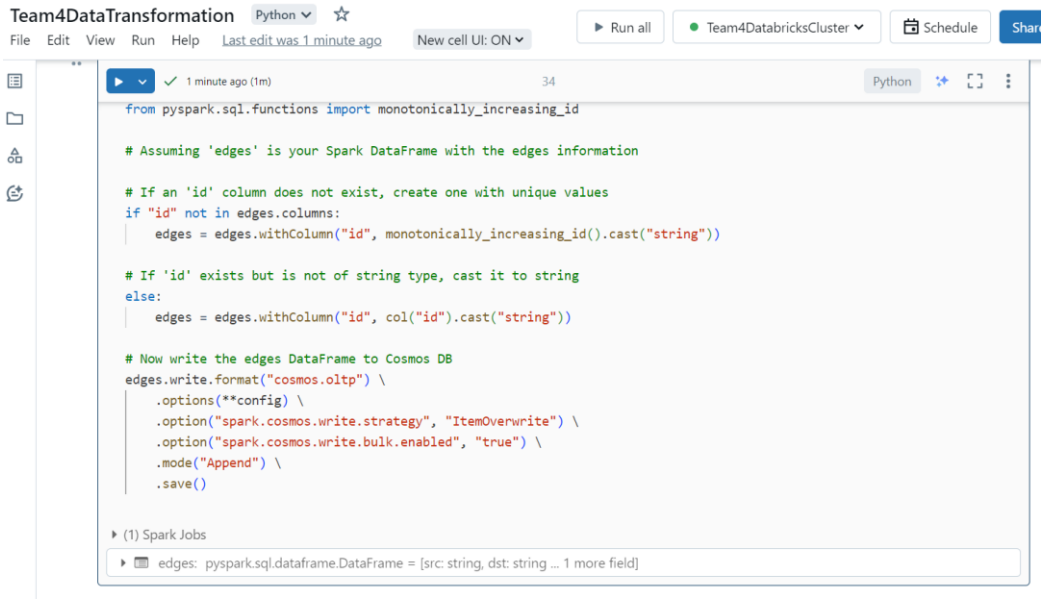




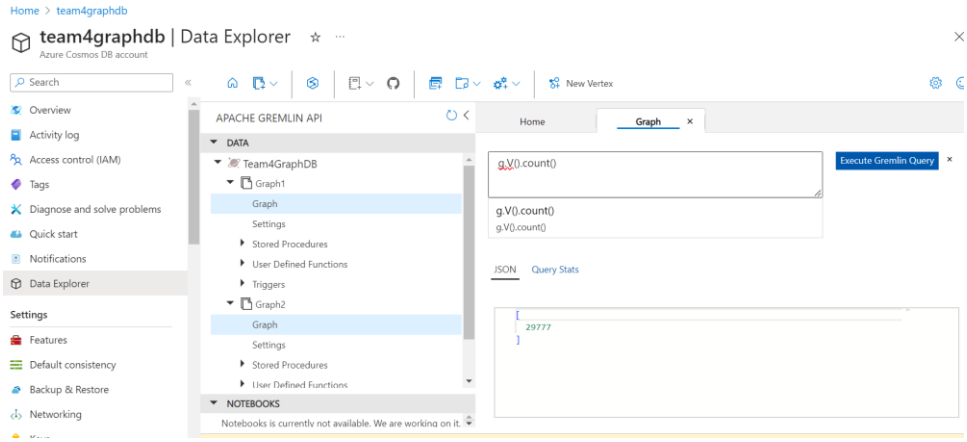
```python
# Convert the vertices and edges to Spark DataFrames
vertices = spark.createDataFrame(vertices_pd)
visa_vertices_df = spark.createDataFrame(visa_vertices.to_frame(name='VisaType'))
country_vertices_df = spark.createDataFrame(country_vertices.to_frame(name='Country'))
holds_edges_df = spark.createDataFrame(holds_edges)
offers_edges_df = spark.createDataFrame(offers_edges)
is_from_visits_edges_df = spark.createDataFrame(is_from_visits_edges)

# Create an edges DataFrame with src and dst columns
edges = (holds_edges_df
        .withColumnRenamed("Id", "src")
        .withColumnRenamed("Visa Type", "dst")
        .unionByName(offers_edges_df.withColumnRenamed("Country", "src")
                                    .withColumnRenamed("Visa Type", "dst"))
        .unionByName(is_from_visits_edges_df.withColumnRenamed("Id", "src")
                                            .withColumnRenamed("Country", "dst")))

edges =edges.withColumnRenamed("src", "src")

# Create a GraphFrame
graph = GraphFrame(vertices, edges)
```

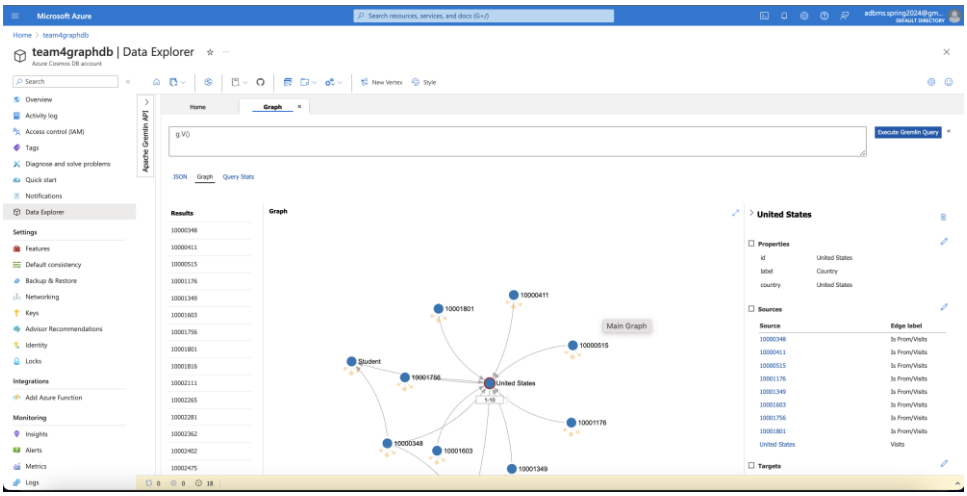
Import data into graph db

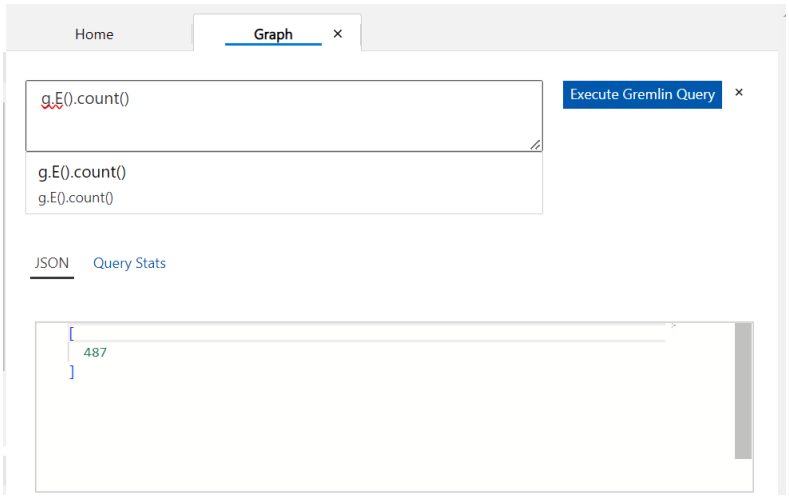




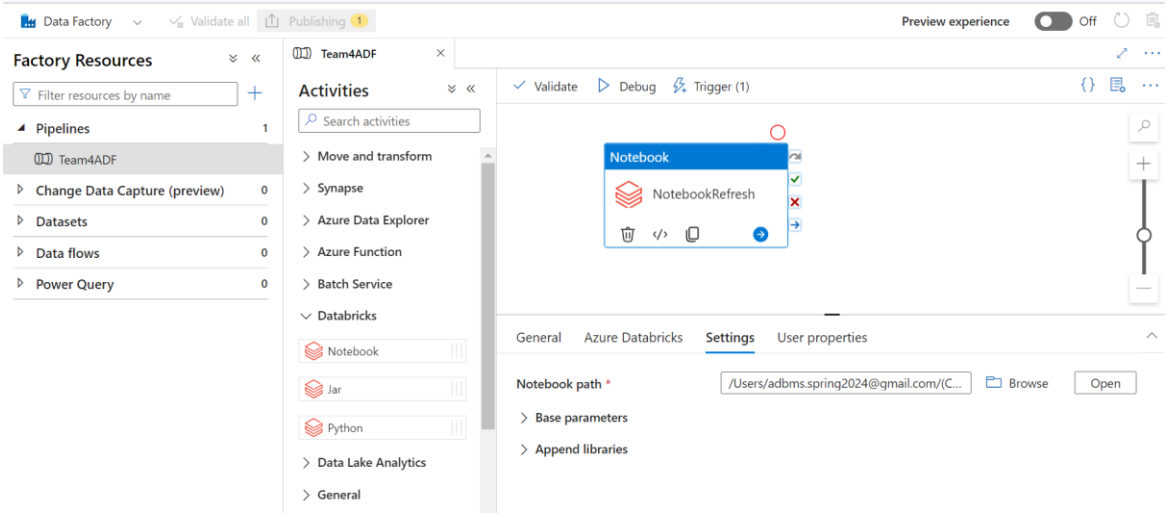
Loaded in graph db as vertices and edges


Home > team4graphdb
team4graphdb | Data Explorer
Azure Cosmos DB account
Search
New Vertex
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Notifications
Data Explorer
Settings
Features
Default consistency
Backup & Restore
Networking
APACHE GREMLIN API
DATA
Team4GraphDB
Graph1
Graph
Settings
Stored Procedures
User Defined Functions
Triggers
Graph2
Graph
Settings
Stored Procedures
NOTEBOOKS
Notebooks is currently not available. We are working on it
Home
Graph
g.V().count()
Execute Gremlin Query
g.V().count()
g.V().count()
JSON
Query Stats
[
29777
]


Microsoft Azure
team4graphdb | Data Explorer
Azure Cosmos DB account
Apache Gremlin API
Home
Graph
g.V()
Execute Gremlin Query
JSON
Graph
Query Stats
Results
Graph
Main Graph
Student
United States
United States
Properties
Sources
Targets


Home
Graph
g.E().count()
Execute Gremlin Query
g.E().count()
g.E().count()
JSON
Query Stats
[
487
]

Scheduling notebook for run using ADF :



Once the data is loaded in there respective databases then the notebook is scheduled to run every day to have latest data loaded in the databases . For this we have used Azure data factory on schedule the trigger. Refresh can be done using databricks but we choose ADF because it provides robust monitoring and logging capabilities. It tracks the status of each pipeline run and can log detailed execution information, which can be viewed directly in the Azure Portal or consumed via Azure Monitor potentially leading to better cost management by shutting down clusters when not in use or scaling them according to the workload.