

P-3: Implementation

Global Inbound and Outbound Travel

Juilee Patil – NUID 002724809

Raksha Israni – NUID: 002925990













Dristi Dani – NUID: 002756885

Ashwin Kumar Kuchibhotla – NUID: 002655594

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 Databricks
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 of inbound travel and travel information to and across different countries. The data dictionary is listed below.

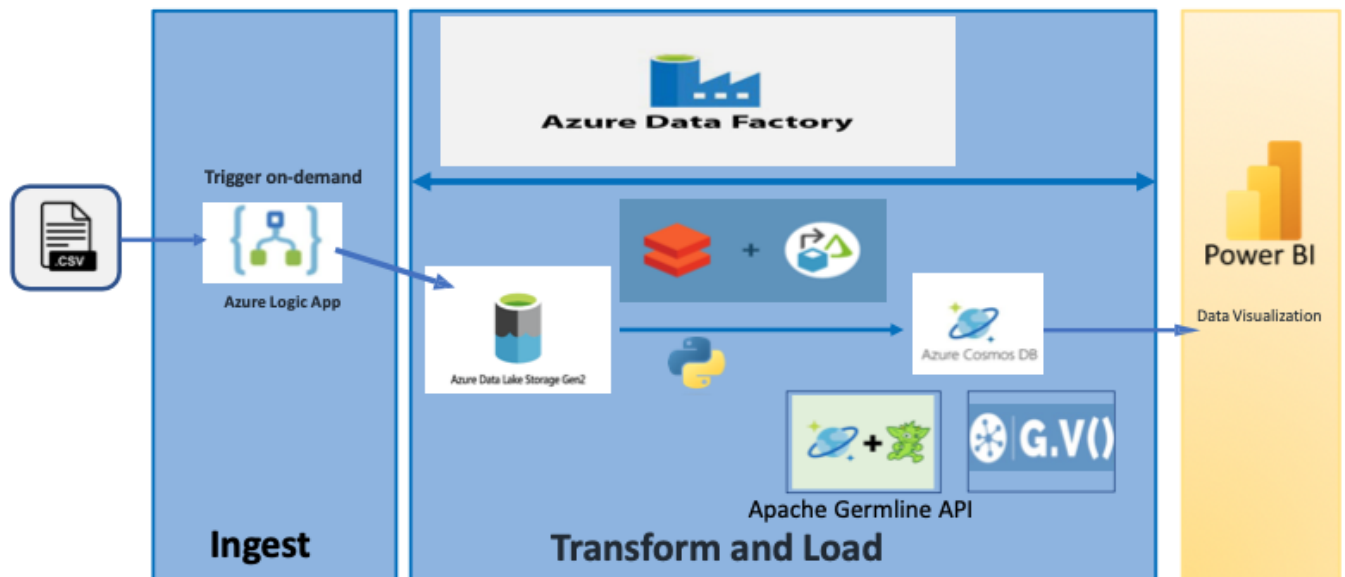
Column Name	Details
id	Numeric id

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 subcategory (only available for business visa type)

Data sample

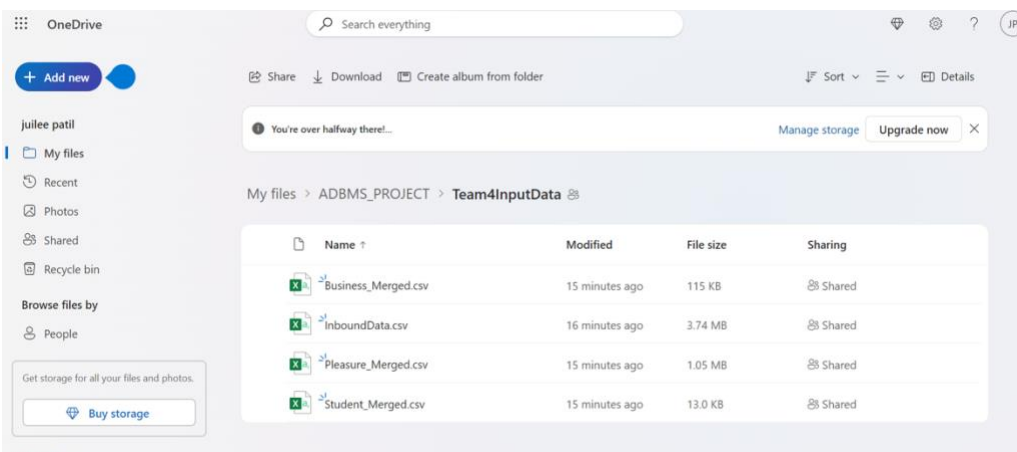
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	first_name	last_name	gender	age	visa_type	passport	nationality	state	destination	destination	entry_date	exit_date	estimated	transport	degree_type	major	institution	start_date	end_date	visa_sub_type
1022	1021	Immanuel	Moye	M	21	Student	780556886	Canada	Ontario	United State Colorado	8/18/2023	12/29/2024	515915.66	air	Master's	Computer Sc	Baker Collg	1/25/2023	6/25/2024		
1023	1022	Karney	Colmer	M	33	Student	4037159791	Canada	Alberta	United State Michigan	09/01/23	07/10/25	517451.98	air	Master's	Business Adm	Pennsylvania	1/26/2023	03/09/25		
1024	1023	Weronah	Escobara	F	27	Student	5303903279	Canada	British Colum	United State Florida	09/12/23	2/28/2024	519742.01	air	Master's	Business Adm	Schell Collg	1/22/2023	05/03/25		
1025	1024	Carly	Habern	F	29	Student	2536557288	Canada	Ontario	United State California	08/03/23	4/18/2023	517532.39	air	Master's	Computer Sc	Medical Coll	01/06/23	08/02/24		
1026	1025	Marshall	Thomassen	M	32	Student	3114810444	Canada	QueBec	United State Connecticut	8/25/2023	02/09/24	512117.05	air	Master's	Computer Sc	The Art Instit	01/09/23	4/17/2024		
1027	1026	Conner	Hudson	M	18	Student	1518049006	Canada	British Colum	United State Missouri	08/05/23	7/22/2024	520241.43	air	Master's	Psychology	Clemson Univ	01/01/23	04/07/25		
1028	1027	Cash	Giametti	M	22	Student	2255061622	Canada	Alberta	United State District of Co	9/14/2023	4/21/2025	510006.36	air	Master's	Business Adm	Florida Metr	1/21/2023	8/14/2024		
1029	1028	Tracie	Thummasutti	M	19	Student	5377673788	Canada	Alberta	United State Kansas	8/14/2023	1/27/2025	519912.86	air	Master's	Computer Sc	Georgia Hes	1/16/2023	3/16/2024		
1030	1029	Amelrose	Ree	M	26	Student	9834970193	Canada	Ontario	United State Texas	09/07/23	5/21/2023	515386.51	road	Master's	Computer Sc	Lehigh Univ	1/22/2023	10/19/2024		
1031	1030	Carlin	Nancarrow	F	28	Student	4881132998	Canada	QueBec	United State New York	08/07/23	1/26/2024	514072.61	air	Master's	Computer Sc	Institute of I	1/23/2023	1/18/2024		
1032	1031	Auguste	Vlashev	F	33	Student	7727516659	Canada	Saskatchewan	United State Michigan	08/10/23	10/05/24	515242.77	air	Master's	Computer Sc	Institute of C	1/25/2023	5/14/2025		
1033	1032	Perren	Blues	M	26	Student	4807171779	Canada	New Brunsw	United State Washington	09/08/23	11/16/2024	511759.08	air	Master's	Computer Sc	Grazia Collg	1/15/2023	9/25/2024		
1034	1033	Rioche	Scamradine	M	22	Student	7665227666	Canada	QueBec	United State Nebraska	09/09/23	10/30/2024	524684.82	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	523324.70	air	Master's	Computer Sc	Southeastern	1/18/2023	02/09/25		
1036	1035	Delphine	Foxton	F	32	Student	788768632	Canada	Ontario	United State New York	9/13/2023	8/28/2024	523467.89	air	Master's	Computer Sc	Bethany Coll	1/13/2023	4/16/2024		
1037	1036	Onida	Tafanini	F	34	Student	9814482297	Canada	Nunavut	United State Florida	8/14/2023	8/23/2024	514283.87	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	510814.21	air	Master's	Computer Sc	University of	1/20/2023	03/12/24		
1039	1038	Chane	Hughes	M	22	Student	9987939334	Canada	QueBec	United State Washington	8/14/2023	3/23/2025	513402.76	road	Master's	Computer Sc	Georgia Schs	1/21/2023	10/25/2024		
1040	1039	Renald	Klein	M	27	Student	9030976299	Canada	Ontario	United State Illinois	08/10/23	08/12/23	514761.80	air	Master's	Psychology	Derry Instit	01/10/23	1/16/2024		
1041	1040	Pip	Maddren	M	18	Student	5386483165	Canada	QueBec	United State North Caroli	8/19/2023	4/30/2024	513373.80	road	Master's	Business Adm	Longwood C	1/16/2023	05/01/25		
1042	1041	Girillo	Berslin	M	19	Student	1001518909	Canada	Ontario	United State Kentucky	8/20/2023	08/08/23	518391.17	air	Master's	Computer Sc	State Univ	01/02/23	8/24/2024		
1043	1042	Emery	Bernardo	M	29	Student	6311504132	Canada	QueBec	United State Illinois	08/12/23	12/01/23	515187.12	air	Master's	Business Adm	Northwest C	1/23/2023	1/18/2024		
1044	1043	Wetase	Lylell	M	20	Student	1969746366	Canada	QueBec	United State North Caroli	8/11/2023	08/02/24	524275.81	road	Master's	Computer Sc	Florida Metr	1/22/2023	1/21/2025		
1045	1044	Elbert	Yersin	M	23	Student	4657460413	Canada	Ontario	United State Wisconsin	8/19/2023	6/21/2024	518025.38	road	Master's	Computer Sc	Southwestern	1/14/2023	11/25/2024		
1046	1045	Norton	Schubert	M	25	Student	8721719017	Canada	Ontario	United State North Caroli	08/08/23	08/01/23	524197.72	road	Master's	Business Adm	Medaille Coll	01/09/23	3/16/2024		
1047	1046	Vicky	Lamprecht	F	23	Student	1444691099	Canada	Alberta	United State Iowa	08/02/23	4/18/2025	522733.77	air	Master's	Computer Sc	Pacific Univ	01/01/23	09/11/24		
1048	1047	Marion	O'Cloney	M	24	Student	3137128362	Canada	Prince Edward	United State Colorado	09/12/23	1/30/2025	510910.30	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	512834.31	air	Master's	Psychology	Elizabeth Cr	1/20/2023	4/16/2024		
1050	1049	Harris	Isard	M	29	Student	6681943484	Canada	QueBec	United State New York	08/06/23	1/26/2023	517300.50	road	Master's	Computer Sc	University of	01/05/23	5/15/2024		
1051	1050	Seumas	Pride	M	33	Student	5831305865	Canada	Saskatchewan	United State Pennsylvania	08/06/23	01/01/24	518412.18	air	Master's	Computer Sc	Adelphi Univ	01/03/23	12/26/2024		
1052	1051	Vina	Trainer	F	24	Student	9538351745	Canada	QueBec	United State Iowa	09/05/23	06/10/23	510206.22	air	Master's	Business Adm	Alice Lloyd C	1/13/2023	7/15/2024		
1053	1052	Gasper	Harrison	M	24	Student	9910957418	Canada	Alberta	United State California	08/06/23	10/15/2023	519069.49	road	Master's	Computer Sc	Fitchburg St	1/27/2023	5/25/2025		
1054	1053	Deacy	Powd	F	26	Student	8209321836	Canada	NewFoundl	United State Michigan	8/13/2023	11/10/2024	517095.48	air	Master's	Computer Sc	Kent State U	1/29/2023	6/28/2024		

Implementation process:

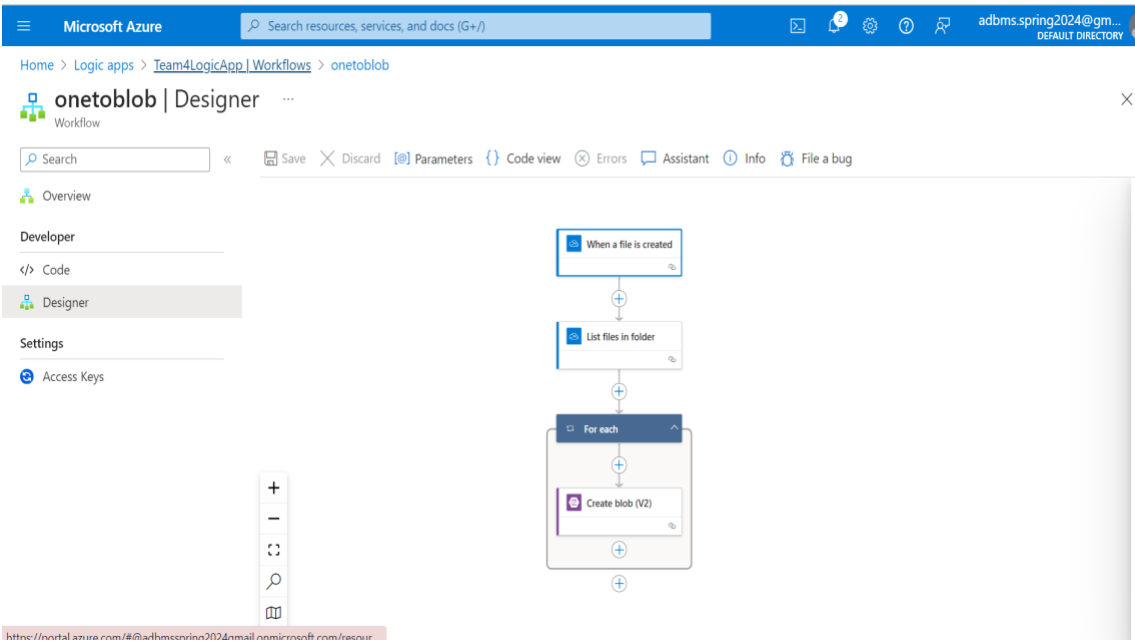


The implementation follows the architecture diagram submitted as part of P2. The process starts with ingesting the data present in the csv file located in the shared OneDrive folder to Azure Blob Storage using Azure Logic app. To streamline data from OneDrive to a graph database, a Logic App triggers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



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



Files in Azure Blob Storage:

The screenshot shows the Azure Blob Storage container "team4data". The container is located at "team4blobsa | Containers". The authentication method is "Access key (Switch to Microsoft Entra user account)". The location is "team4data". The search filter is "Search blobs by prefix (case-sensitive)". The "Show deleted blobs" toggle is turned off. The table below lists the files in the container:

Name	Modified	Access tier	Archive status	Blob type	Size
<input type="checkbox"/> Business_Merged.csv	27/3/2024, 5:08:58 pm	Hot (Inferred)		Block blob	1 KB
<input type="checkbox"/> InboundData.csv	27/3/2024, 5:08:58 pm	Hot (Inferred)		Block blob	1 KB
<input type="checkbox"/> Pleasure_Merged.csv	27/3/2024, 5:08:58 pm	Hot (Inferred)		Block blob	1 KB
<input type="checkbox"/> Student_Merged.csv	27/3/2024, 5:08:58 pm	Hot (Inferred)		Block blob	1 KB

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 blob storage.

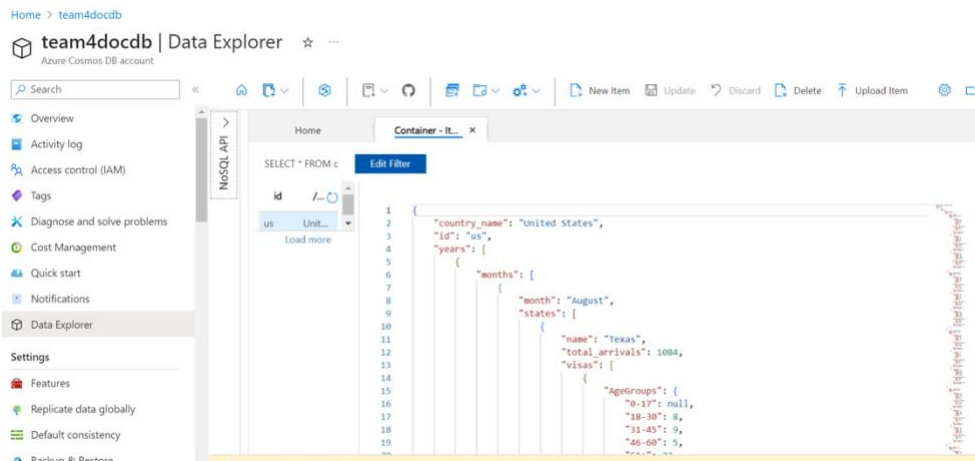
The screenshot shows a Databricks notebook titled 'Team4DataTransformation'. The top section displays package requirements for 'requests' and 'forex_python'. A note indicates that the kernel may need to be restarted. The main code block, executed at 05:05 PM (16s), reads four CSV files from Azure Blob Storage into Spark DataFrames: 'InboundData.csv', 'Business_Merged.csv', 'Pleasure_Merged.csv', and 'Student_Merged.csv'. Below the code, the Spark Jobs output shows the successful creation of these DataFrames. The bottom section, executed at 05:05 PM (20s), performs initial data checks by printing row counts and descriptive statistics for each DataFrame. The output for 'Describe Outbound Pleasure' is shown as a table below.

summary	id	... estimated expenses	transportation mode
0 count	8500	...	8500
1 mean	4250.5	...	None
2 stddev	2453.882977378234	...	None
3 min	1	...	\$10,001.05
4 max	999	...	£14998.09

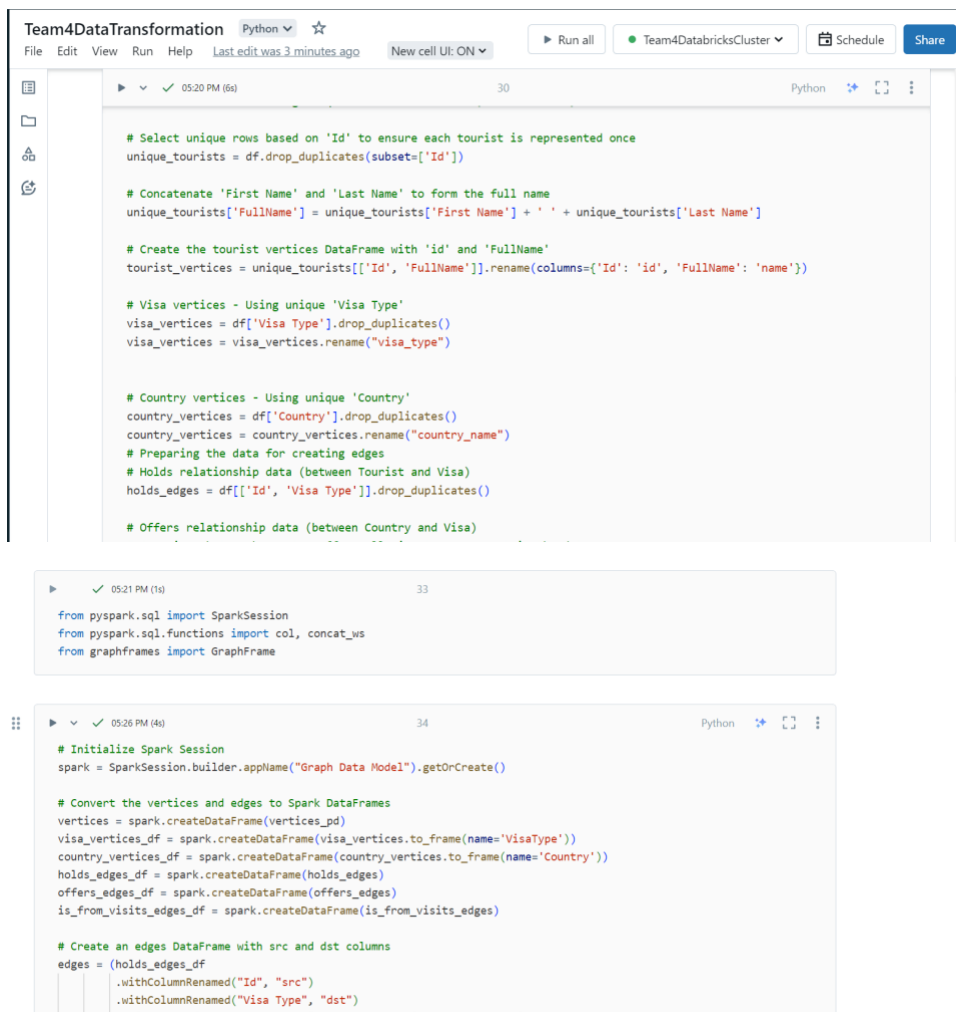
Cleaning and transformation

The screenshot shows the same Databricks notebook at a later stage. The code block, executed at 05:15 PM (8s), defines an age categorization function using a 'calendar' module. The function categorizes ages into groups: '0-17', '18-30', '31-45', '46-60', and '61+'. It then applies this function to create a new column 'AgeGroup' in the DataFrame. Additionally, the code converts the 'Entry Date' to a datetime format and extracts the 'year' and 'month' as new columns. The output shows the successful execution of these transformations.

Screenshot of data loaded in Document db



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




```

# 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

Team4DataTransformation Python   Run all   Schedule

File Edit View Run Help Last edit was now New cell UI: ON 

 05:31 PM (56s) 37 Python  

```

vertices.printSchema()

# Check the first few rows to ensure they look as expected
vertices.show()

# Write vertices DataFrame to Cosmos DB
vertices.write.format("cosmos.oltp") \
    .options(**config) \
    .option("spark.cosmos.write.strategy", "ItemOverwrite") \
    .option("spark.cosmos.write.bulk.enabled", "true") \
    .mode("Append") \
    .save()


```

▶ (1) Spark Jobs

10000515	Latisha Secret
10001176	Nathaniel Faichnie
10001349	Huntlee Newtown
10001603	Johnath Sawden
10001756	Sephira Hevner
10001801	Allard Flewett
10001816	Reena McQuilkin
10002111	Rosabelle Golborn
10002265	Bartholomeo Bamokin

Team4DataTransformation Python   Run all   Schedule 

File Edit View Run Help Last edit was 1 minute ago New cell UI: ON 

 1 minute ago (1m) 34 Python   

```

from pyspark.sql.functions import monotonically_increasing_id

# Assuming 'edges' is your Spark DataFrame with the edges information

# If an 'id' column does not exist, create one with unique values
if "id" not in edges.columns:
    edges = edges.withColumn("id", monotonically_increasing_id().cast("string"))

# If 'id' exists but is not of string type, cast it to string
else:
    edges = edges.withColumn("id", col("id").cast("string"))

# Now write the edges DataFrame to Cosmos DB
edges.write.format("cosmos.oltp") \
    .options(**config) \
    .option("spark.cosmos.write.strategy", "ItemOverwrite") \
    .option("spark.cosmos.write.bulk.enabled", "true") \
    .mode("Append") \
    .save()

```

▶ (1) Spark Jobs

edges: pyspark.sql.dataframe.DataFrame = [src: string, dst: string ... 1 more field]

Loaded in graph db as vertices and edges

Home > team4graphdb

team4graphdb | Data Explorer

Search

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
 - User Defined Functions
- NOTEBOOKS

g.V().count()

g.V().count()
g.V().count()

Execute Gremlin Query

JSON Query Stats

```
{
  "count": 29777
}
```

NOTEBOOKS

Notebooks is currently not available. We are working on it.

Microsoft Azure

Search resources, services, and docs (Ctrl+K)

Home > team4graphdb

team4graphdb | Data Explorer

Search

Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems
Quick start
Notifications
Data Explorer

Settings
Features
Default consistency
Backup & Restore
Networking
Keys
Advisor Recommendations
Identity
Locks
Integrations
Add Azure Function
Monitoring
Insights
Alerts
Metrics
Logs

APACHE GREMLIN API

Home Graph

g.V()

Execute Gremlin Query

Results

```
100003948
10000411
10000515
10001176
10001349
10001403
10001756
10001801
10001816
10002111
10002265
10002281
10002362
10002402
10002475
```

Graph

Main Graph

United States

Properties

- id: United States
- label: Country
- country: United States

Sources

Source	Edge label
100003948	Is From/Visits
10000411	Is From/Visits
10000515	Is From/Visits
10001176	Is From/Visits
10001349	Is From/Visits
10001801	Is From/Visits
10001756	Is From/Visits
10001816	Is From/Visits
United States	Visits

Targets

Home Graph

g.E().count()

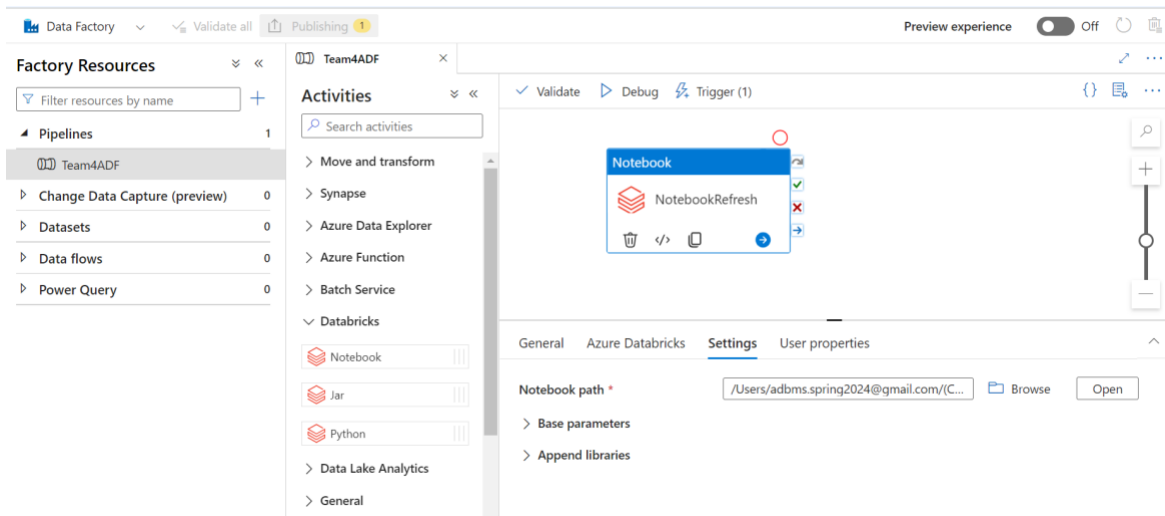
g.E().count()
g.E().count()

Execute Gremlin Query

JSON Query Stats

```
{
  "count": 487
}
```

Scheduling notebook for run using ADF:



Once the data is loaded in their respective databases then the notebook is scheduled to run every day to have the latest data loaded in the databases. For this we have used Azure data factory to schedule the trigger. Refresh can be done using databricks but we chose 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.