ASSESSMENT 2
DATA IN THE CLOUD

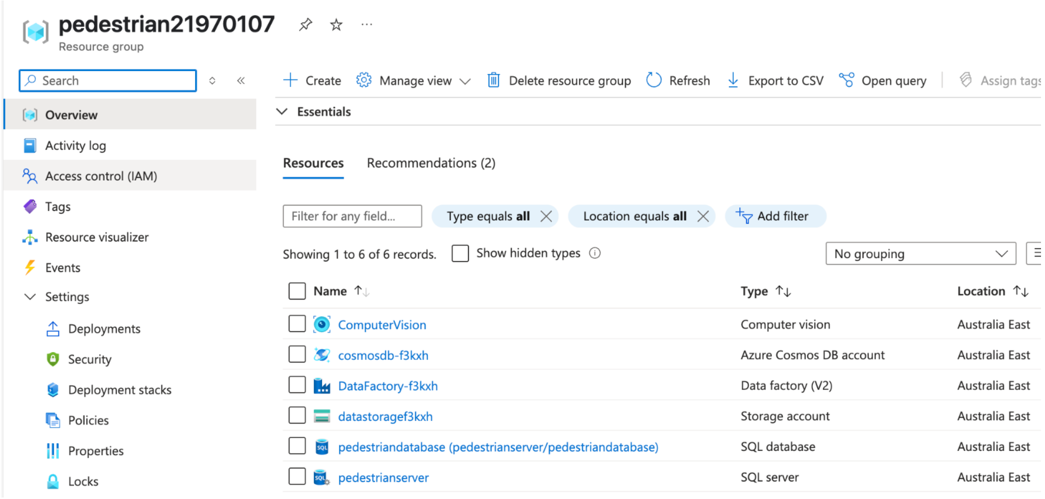
ARISH KARTHIKEYAN - 21970107

BUS5001 CLOUD PLATFORMS AND ANALYTICS, LA TORBE BUSINESS SCHOOL

Table of Contents

A Big Data Processing Pipeline

Resource Group Configuration



pedestrian21970107
Resource group

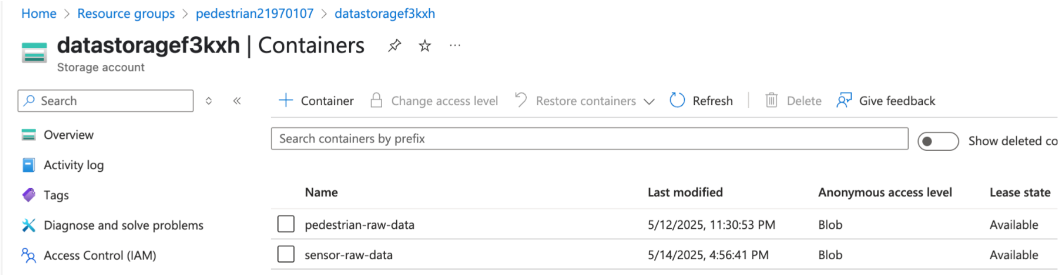
Showing 1 to 6 of 6 records.

Name	Type	Location
ComputerVision	Computer vision	Australia East
cosmosdb-f3kxh	Azure Cosmos DB account	Australia East
DataFactory-f3kxh	Data factory (V2)	Australia East
datastoragef3kxh	Storage account	Australia East
pedestriandatabase (pedestrianserver/pedestriandatabase)	SQL database	Australia East
pedestrianserver	SQL server	Australia East

All Azure resources used in this pipeline are grouped under a centralized resource container titled`pedestrian21970107`. This resource group facilitates logical organization, streamlined management, and cost tracking of related services. It includes the following key services:

- Azure Data Factory (DataFactory-f3kxh): Orchestrates ETL activities.
- Azure Blob Storage (datastoragef3kxh): Serves as the landing zone for raw JSON data.
- Azure SQL Server (pedestrianserver) and SQL Database (pedestriandatabase): Hosts the final structured dataset for downstream analysis.
- Azure Cosmos DB and Computer Vision (provisioned but not used in this specific pipeline).

Data Ingestion and Storage



datastoragef3kxh | Containers
Storage account

Name	Last modified	Anonymous access level	Lease state
pedestrian-raw-data	5/12/2025, 11:30:53 PM	Blob	Available
sensor-raw-data	5/14/2025, 4:56:41 PM	Blob	Available

The ingestion layer comprises two external REST APIs:

- Pedestrian API: Provides hourly pedestrian counts from city sensors.
- Sensor API: Delivers sensor metadata such as installation location, type, and status.

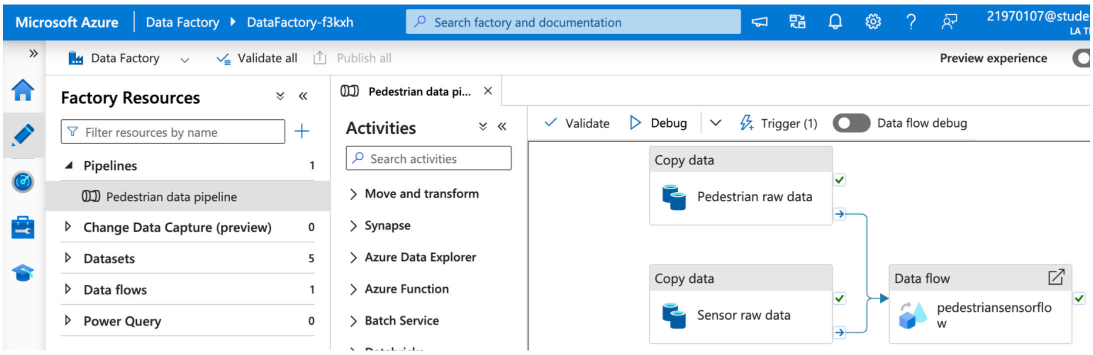
Data is ingested into Azure Blob Storage containers:

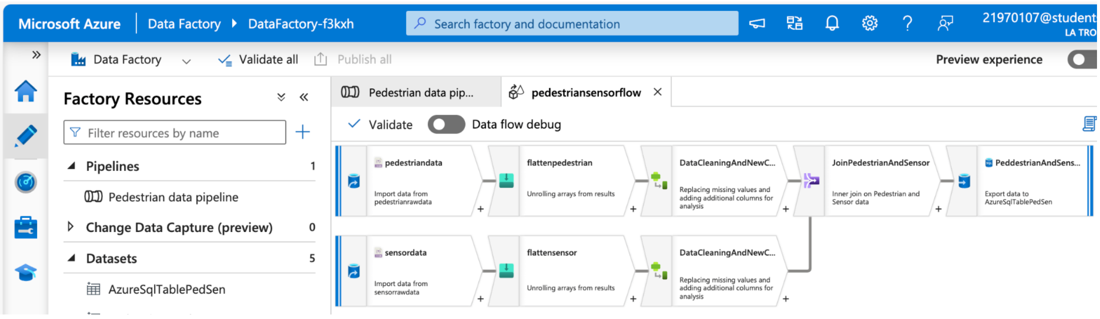
- `pedestrian-raw-data`
- `sensor-raw-data`

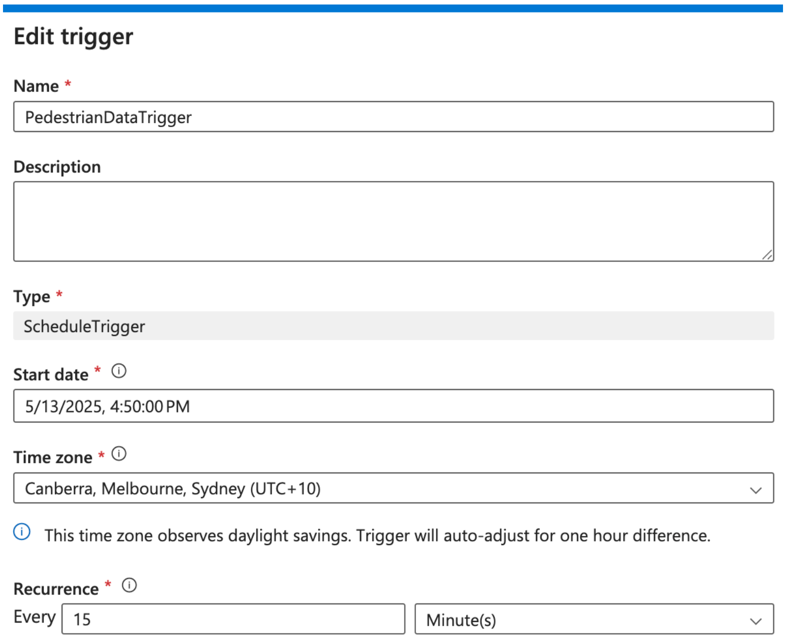
These containers reside within the storage account `datastoragef3kxh`, located in the Australia East region. Ingestion is facilitated by Azure Data Factory's Copy Data activity, configured through linked

REST services (`pedestrianrestservice`, `sensorrestservice`). The data is saved in JSON format with UTF-8 encoding and no compression, using datasets named `pedestrianrawdata` and `sensorrawdata`.

Data Transformation in Azure Data Factory







Edit trigger

Name *

PedestrianDataTrigger

Description

Type *

ScheduleTrigger

Start date *

5/13/2025, 4:50:00 PM

Time zone *

Canberra, Melbourne, Sydney (UTC+10)

This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.

Recurrence *

Every 15 Minute(s)

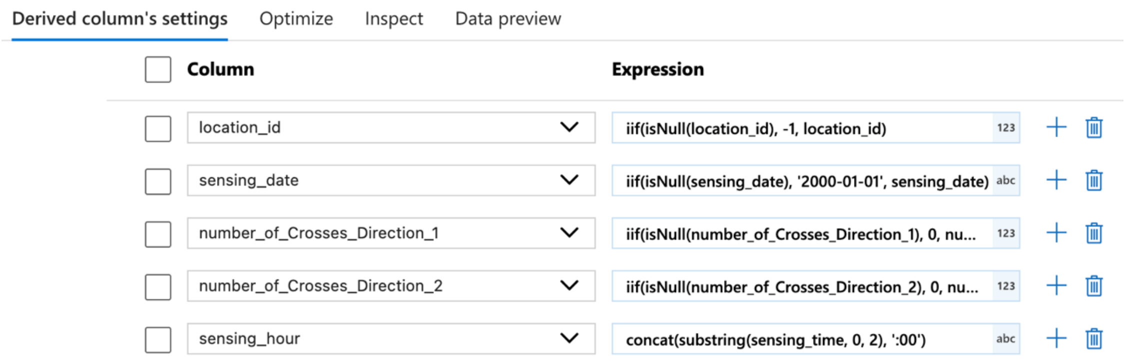
A schedule trigger named “PedestrianDataTrigger” was configured in Azure Data Factory to initiate the pipeline every 15 minutes, simulating a real-time data ingestion and transformation process.

The transformation process is orchestrated within a pipeline titled `Pedestrian data pipeline`. The pipeline invokes a Data Flow (`pedestriansensorflow`) to process and clean data through the following stages:

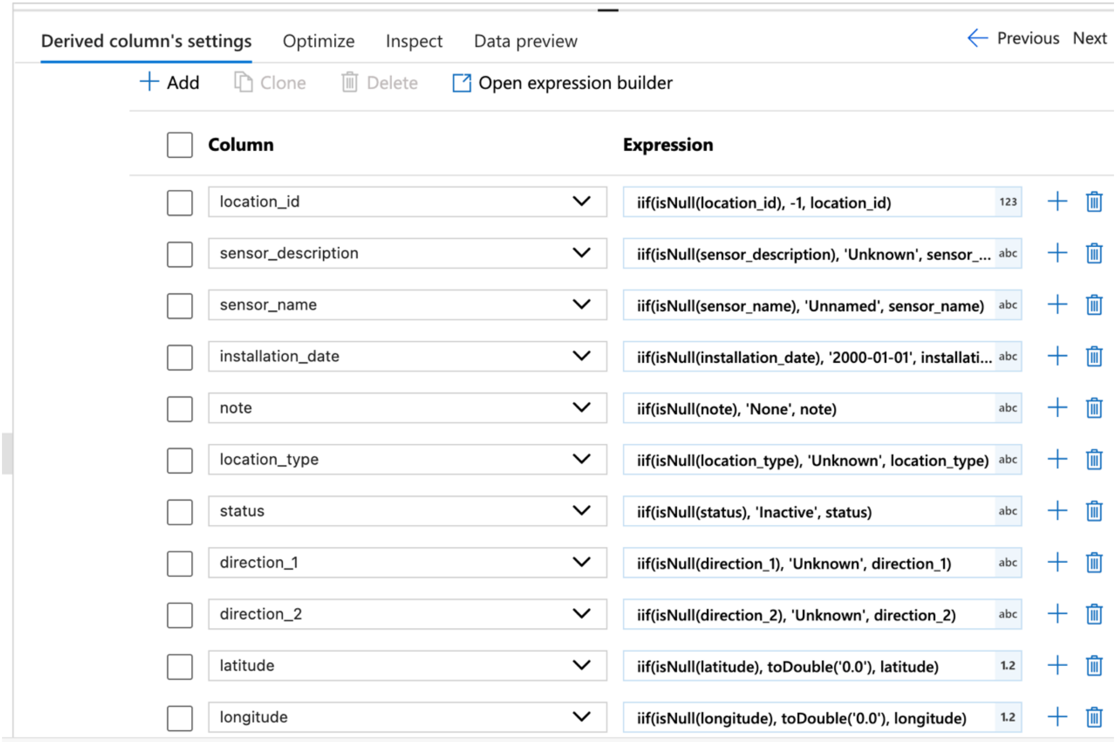
Flattening

Nested JSON arrays from the blob containers are normalized using flattening transformations (`flattenpedestrian`, `flattensensor`), allowing easier downstream operations on individual fields.

Data Cleaning and Feature Engineering



□ Column	Expression
□ location_id	iif(isNull(location_id), -1, location_id)
□ sensing_date	iif(isNull(sensing_date), '2000-01-01', sensing_date)
□ number_of_Crosses_Direction_1	iif(isNull(number_of_Crosses_Direction_1), 0, nu...
□ number_of_Crosses_Direction_2	iif(isNull(number_of_Crosses_Direction_2), 0, nu...
□ sensing_hour	concat(substring(sensing_time, 0, 2), ':00')



□ Column	Expression
□ location_id	iif(isNull(location_id), -1, location_id)
□ sensor_description	iif(isNull(sensor_description), 'Unknown', sensor_...
□ sensor_name	iif(isNull(sensor_name), 'Unnamed', sensor_name)
□ installation_date	iif(isNull(installation_date), '2000-01-01', installati...
□ note	iif(isNull(note), 'None', note)
□ location_type	iif(isNull(location_type), 'Unknown', location_type)
□ status	iif(isNull(status), 'Inactive', status)
□ direction_1	iif(isNull(direction_1), 'Unknown', direction_1)
□ direction_2	iif(isNull(direction_2), 'Unknown', direction_2)
□ latitude	iif(isNull(latitude), toDouble('0.0'), latitude)
□ longitude	iif(isNull(longitude), toDouble('0.0'), longitude)

The `DataCleaningAndNewColumns` stage handles missing values using conditional expressions such as:

- iif(isNull(location_id), -1, location_id)
- iif(isNull(sensor_name), 'Unnamed', sensor_name)

A new derived column `sensing_hour` is created to facilitate time-based analysis:

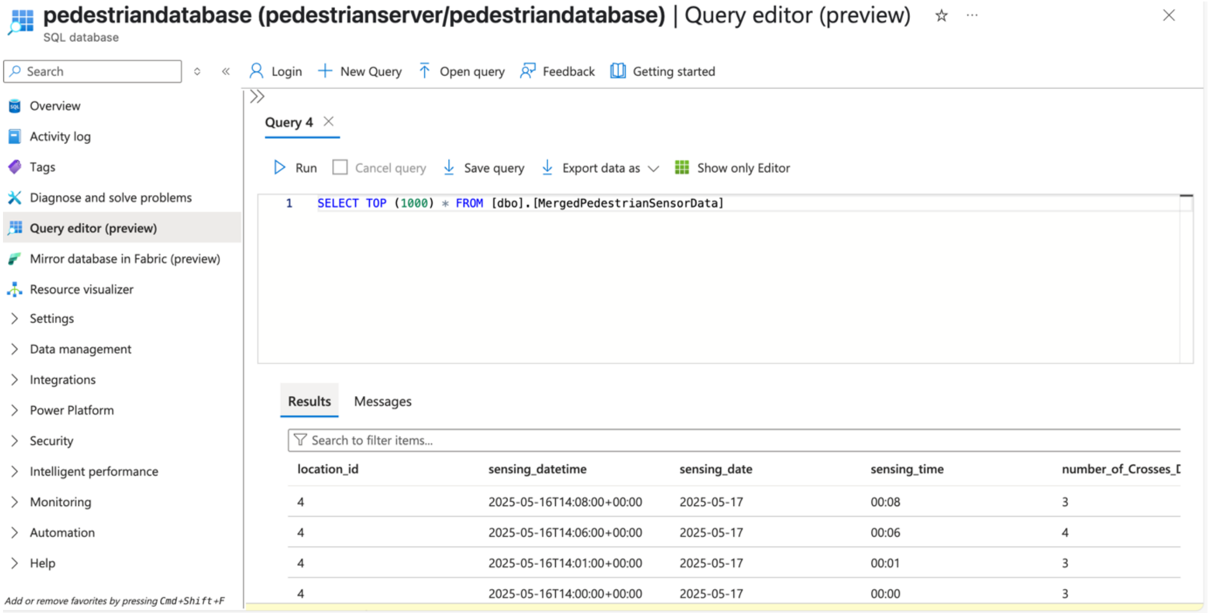
- concat(substring(sensing_time, 0, 2), ':00')

All critical fields such as `number_of_Crosses_Direction_1`, `direction_1`, and `installation_date` are assigned default fallbacks to maintain schema consistency.

Data Joining and Export

The cleaned datasets are merged using `JoinPedestrianAndSensor`, applying an inner join based on shared identifiers. The output is directed to an Azure SQL sink (`AzureSqlTablePedSen`), mapped to a table named `dbo.MergedPedestrianSensorData` in the SQL Database.

Data Storage and Access



pedestriandatabase (pedestrianserver/pedestriandatabase) | Query editor (preview)
SQL database

```
SELECT TOP (1000) * FROM [dbo].[MergedPedestrianSensorData]
```

location_id	sensing_datetime	sensing_date	sensing_time	number_of_Crosses_D
4	2025-05-16T14:08:00+00:00	2025-05-17	00:08	3
4	2025-05-16T14:06:00+00:00	2025-05-17	00:06	4
4	2025-05-16T14:01:00+00:00	2025-05-17	00:01	3
4	2025-05-16T14:00:00+00:00	2025-05-17	00:00	3

The structured dataset is persisted in the Azure SQL Database `pedestriandatabase`, hosted on the SQL server `pedestrianserver.database.windows.net`**. The table schema, accessible via the Query Editor, includes:

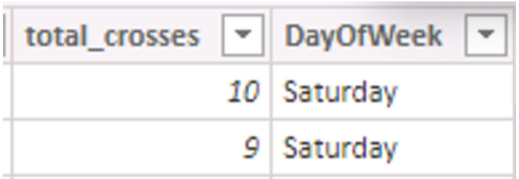
- Temporal fields: `sensing_datetime`, `sensing_hour`.
- Directional metrics: `number_of_Crosses_Direction_1`, `number_of_Crosses_Direction_2`.
- Metadata: `sensor_name`, `location_type`, `status`, `latitude`, `longitude`.

These attributes support fine-grained analysis of pedestrian flow patterns across time and space.

Power BI Integration and Data Enrichment

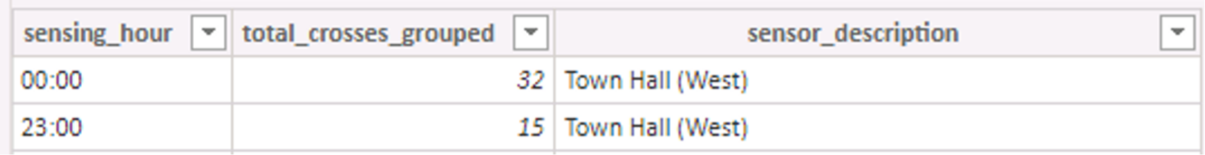
After storing the cleaned data in the Azure SQL Database, it was connected to Power BI for visualization and further transformation. The imported table "MergedPedestrianSensorData" was enriched with the following key steps:

- Day of Week Derivation: Created a "DayOfWeek" column using the DAX WEEKDAY function to enable trend analysis by weekday/weekend.



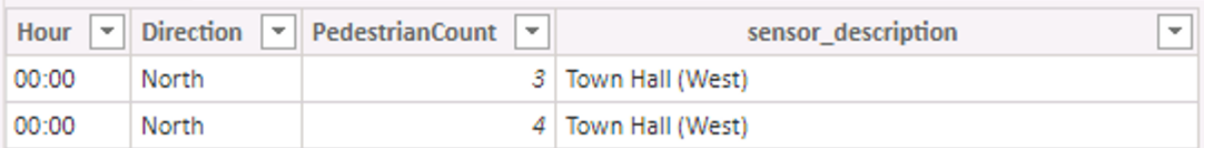
total_crosses	DayOfWeek
10	Saturday
9	Saturday

- Total Crosses: Computed total_crosses by summing values from both "number_of_Crosses_Direction_1" and "number_of_Crosses_Direction_2" fields.



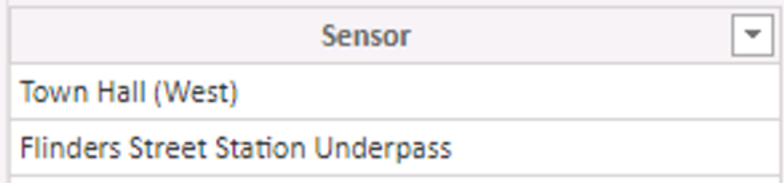
sensing_hour	total_crosses_grouped	sensor_description
00:00	32	Town Hall (West)
23:00	15	Town Hall (West)

- Hourly Aggregation: Built a summary table "SensorHourlyTotals" using SUMMARIZE to analyse pedestrian peaks by hour and location.



Hour	Direction	PedestrianCount	sensor_description
00:00	North	3	Town Hall (West)
00:00	North	4	Town Hall (West)

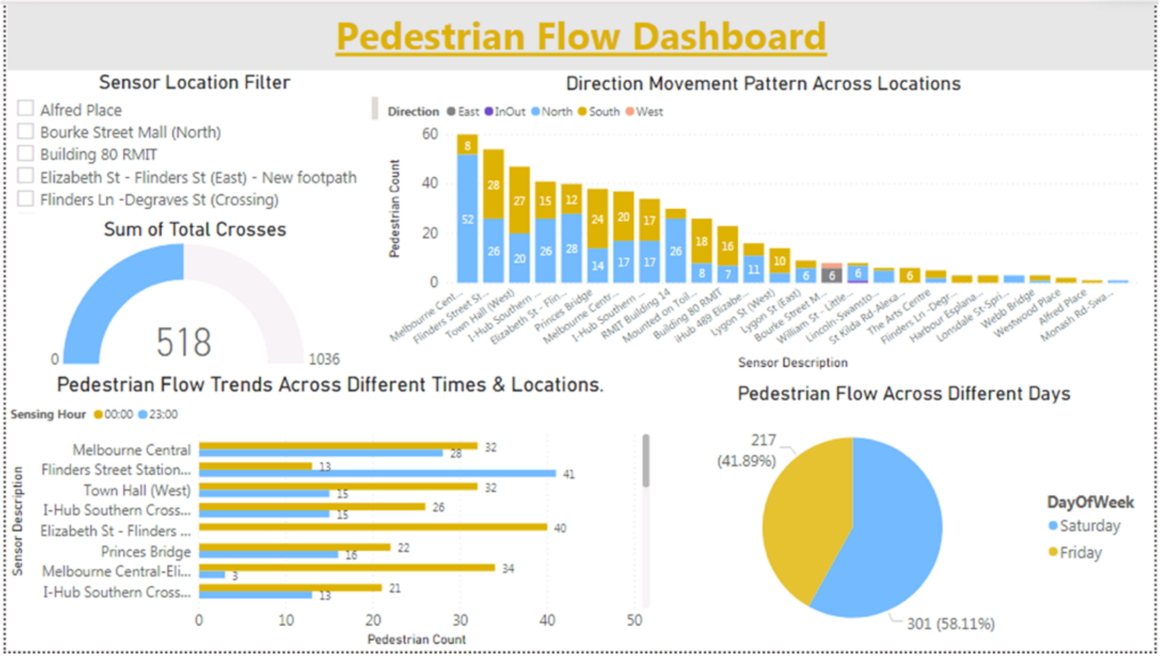
- Directional Flow Analysis: Created a unified "DirectionalFlow" table using UNION to study movement patterns by direction and time.



Sensor
Town Hall (West)
Flinders Street Station Underpass

- Sensor Lookup Table: Extracted unique sensor names into "SensorTable" to support interactive filtering in dashboards.

Power BI dashboard



Key Findings

Analysis of pedestrian data across Melbourne's key locations reveals several critical insights. First, Melbourne Central, Flinders Street Station, and Town Hall (West) reported the highest pedestrian counts, particularly in eastbound and northbound movements. Directional patterns highlight significant flow discrepancies, suggesting concentrated traffic at specific urban nodes.

Temporal analysis shows peak footfall during the late evening (23:00) at locations like Flinders Street and Princes Bridge, whereas some areas like Melbourne Central maintain relatively balanced traffic throughout the day. This highlights the impact of late-night activities and transport hubs on pedestrian volumes.

The day-wise comparison indicates higher pedestrian activity on Saturdays (58.11%) than Fridays (41.89%), emphasizing weekend foot traffic due to leisure activities and shopping trends. With a total of 518 pedestrian crossings recorded, the data suggests substantial mobility across the network, necessitating crowd control and infrastructure adjustments.

Urban Planning Insights and Recommendations for Pedestrian Infrastructure and Crowd Management

City planners must prioritize wider sidewalks at high-traffic nodes like Melbourne Central and Town Hall to accommodate volume and improve walkability. These can incorporate street furniture, trees, and lighting to enhance safety and comfort.

Implementation of raised crosswalks at busy intersections can effectively calm traffic and increase pedestrian visibility. For roads with heavy footfall, pedestrian overpasses or underpasses, designed with aesthetics like greenery and lighting, can ensure safe crossing without deterring use.

In locations with high directional variance, curb extensions and median islands should be introduced to shorten crossing distances and improve driver awareness. These are especially critical for north-south corridors where abrupt directional shifts are observed.

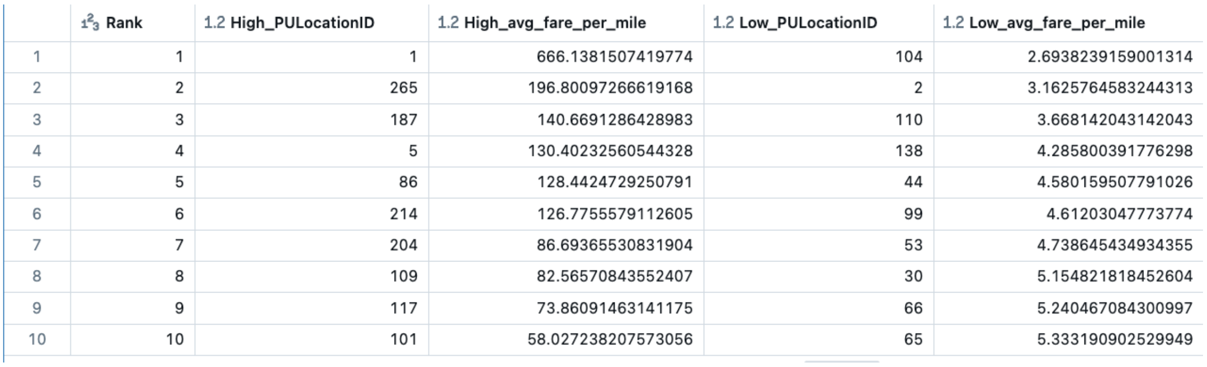
From a zoning perspective, planners should consider mixed-use developments to support walkability and reduce vehicular dependency. Ensuring grid-based street layouts with predictable intersections will further enhance pedestrian safety and accessibility.

Lastly, integrating smart crosswalks equipped with motion sensors and LED alerts can dramatically improve nighttime pedestrian safety, particularly useful in entertainment districts active during late hours.

In summary, these design and technology-led strategies will help Melbourne transition toward a safer, smarter, and more pedestrian-friendly urban future.

An Exploratory Analysis Of The NYC Taxi & Limousine Commission Dataset Using Databricks

Analysis of Top 10 Zones by Average Fare per Distance

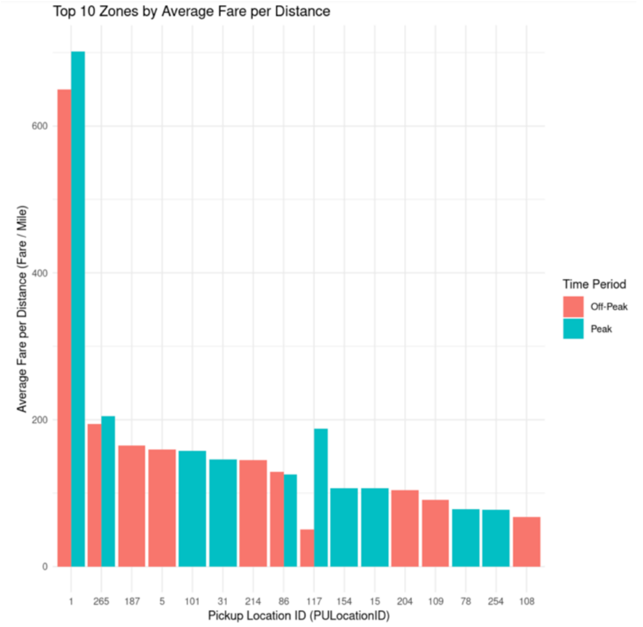


	1²3 Rank	1.2 High_PULocationID	1.2 High_avg_fare_per_mile	1.2 Low_PULocationID	1.2 Low_avg_fare_per_mile
1	1	1	666.1381507419774	104	2.6938239159001314
2	2	265	196.80097266619168	2	3.1625764583244313
3	3	187	140.6691286428983	110	3.668142043142043
4	4	5	130.40232560544328	138	4.285800391776298
5	5	86	128.4424729250791	44	4.580159507791026
6	6	214	126.7755579112605	99	4.61203047773774
7	7	204	86.69365530831904	53	4.738645434934355
8	8	109	82.56570843552407	30	5.154821818452604
9	9	117	73.86091463141175	66	5.240467084300997
10	10	101	58.027238207573056	65	5.333190902529949

The comparison between the top 10 zones with the highest and lowest average fare per mile reveals a stark contrast in fare structures. Zone 1 records the highest fare per mile at 666.14, followed by Zone 265 at 196.80, which is substantially higher than the rest of the high-fare zones ranging between 58.03 and 140.67. This indicates a sharp drop after the top two, with Zone 265 standing notably apart from others in the ranking.

In contrast, the 10 zones with the lowest average fares per mile show values ranging narrowly from 2.69 to 5.33, indicating greater consistency among low-fare zones. While high-fare zones exhibit wide variability and extreme values, the low-fare zones appear more stable and uniform. This disparity suggests that certain high-fare zones may be influenced by localized conditions, whereas low-fare zones may represent areas with longer trips or more standardized pricing.

Analysis of Top 10 Zones by Average Fare per Distance for peak vs. off peak hours



The analysis of temporal fare patterns across the top 10 pickup zones reveals distinct differences between peak and off-peak periods. Pickup Location ID 1 exhibits the highest average fare per mile in both periods, with a noticeable increase during peak hours, indicating strong demand surges. Similarly, Location ID 265 shows a significantly higher fare during peak time, suggesting sensitivity to temporal demand. In contrast, zones such as IDs 187, 5, and 86 demonstrate relatively consistent fares across both peak and off-peak periods, with only minor variations. This stability implies that these zones experience balanced demand regardless of the time, resulting in minimal temporal pricing effects.

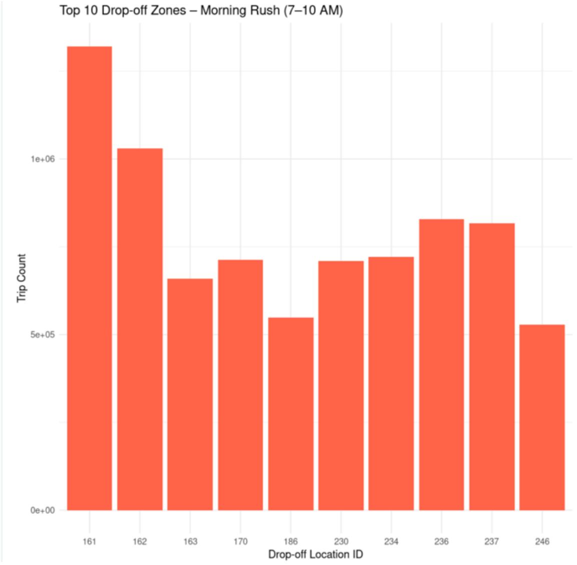
Pricing Inequities That Warrant Intervention

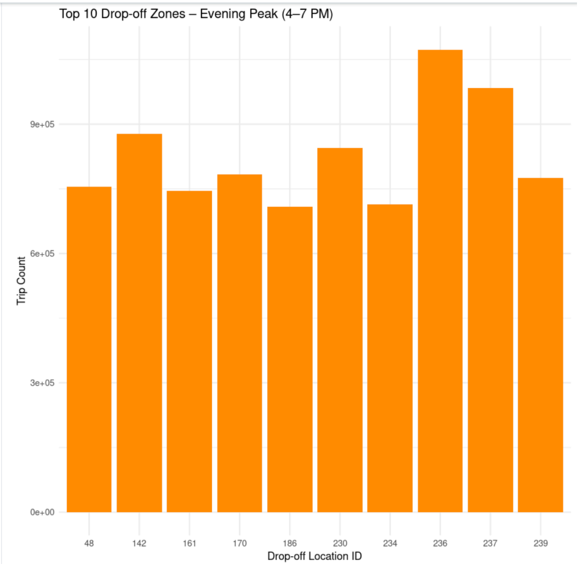
The combined analysis of average fare per mile and temporal fare variations indicates potential pricing inequities across pickup zones. The stark contrast between the highest and lowest fare zones, ranging from over 660 to under 3 units per mile—suggests geographic disparities in fare structures. Particularly, Location ID 1 and 265 exhibit unusually high fares, which may disadvantage riders relying on short but costly trips, especially during peak hours.

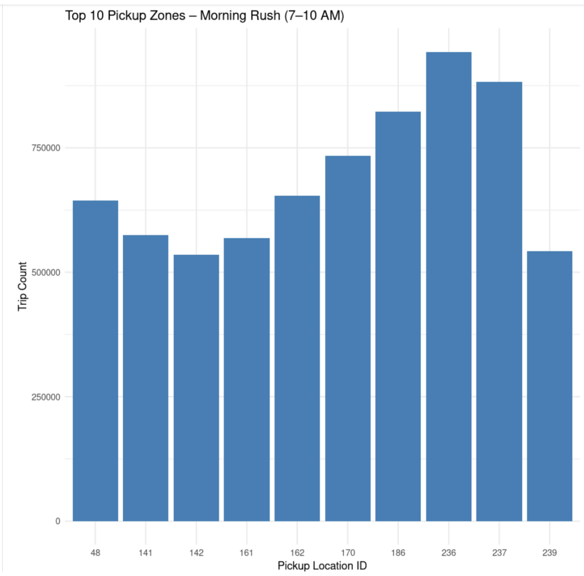
Moreover, temporal effects are not uniformly distributed. Some zones, like 265 and 117, show pronounced fare increases during peak hours, while others, such as 187, 5, and 86, maintain stable pricing. This inconsistency suggests that time-based surcharges may disproportionately affect certain areas, potentially reflecting demand-driven pricing rather than equitable access.

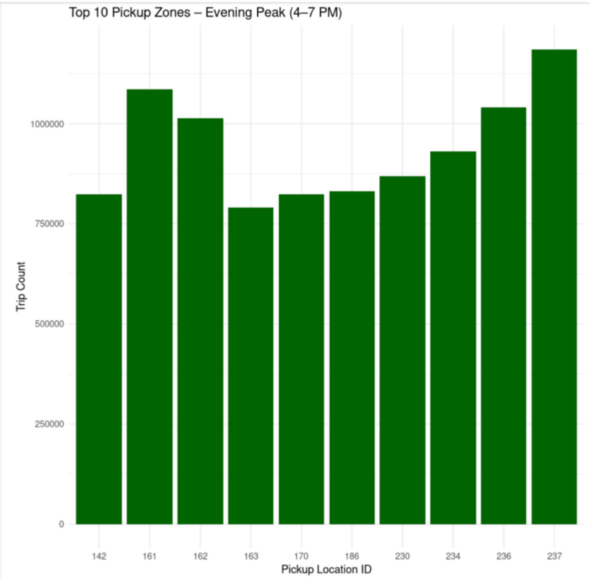
These findings support the case for regulatory or policy interventions to address fare anomalies. Standardizing peak-hour surcharges or capping fare per mile in specific zones could enhance fairness, ensure affordability, and promote equitable access to transportation services across all regions.

Top 10 drop off and pick up zones during morning rush (7AM–10AM) and Evening peak (4PM–7PM)


Top 10 Drop-off Zones – Morning Rush (7–10 AM)


Top 10 Drop-off Zones – Evening Peak (4–7 PM)





Strategic Infrastructure Recommendations for NYC Yellow Taxi Services Based on Peak-Hour Demand Patterns

New York City has successfully repurposed significant portions of its streetscape to support more equitable and efficient urban uses—introducing bike lanes, bus corridors, pedestrian plazas, and vibrant on-street dining. However, this transformation has constrained curb side space available for taxis and for-hire vehicles (FHVs), making safe and efficient passenger pick-up and drop-off increasingly challenging. Additionally, designated rest areas for drivers remain limited, impacting service reliability and driver wellbeing. In response, the Taxi and Limousine Commission (TLC), in collaboration with the Department of Transportation (DOT), aims to expand Neighbourhood Loading Zones and Taxi/FHV relief stands to address these issues strategically.

Insights from the NYC Yellow Taxi dataset offer strong empirical support for this initiative. A comparative analysis of peak-hour activity reveals a consistent pattern of concentrated demand in specific zones. During the morning rush (7–10 AM), pickup activity is highest in Zones 186, 236, and 237, indicating these are likely residential or transit-linked origins. Meanwhile, drop-offs are concentrated in Zones 161, 162, and 163, which likely correspond to central business or employment hubs. This directional flow illustrates commuter corridors that would benefit significantly from dedicated taxi loading zones to reduce congestion and improve service flow.

Conversely, the evening peak (4–7 PM) sees a reversal in this trend. Zones 237, 236, and 161 become key pickup areas, while drop-offs are more dispersed across Zones 234, 236, 237, and 239, suggesting a return toward residential neighbourhoods. Notably, Zone 237 consistently ranks among the top zones for both pickups and drop-offs across time periods, making it a clear candidate for infrastructure prioritization.

Based on these findings, it is recommended that the City expand Taxi/FHV stands and Neighbourhood Loading Zones in Zones 237, 236, 161, and 186. These locations exhibit sustained, bi-directional demand and would benefit from safe, dedicated curb space for taxi operations. Furthermore, Taxi/FHV relief stands should be increased near these zones, ensuring drivers have legal parking to rest and access amenities. Additionally, Zones 162 and 163, while not as consistent, show high traffic during peak periods and should be considered for real-time demand tracking and dynamic vehicle allocation strategies.

Aligning infrastructure investments with data-driven demand patterns will enhance commuter experience, support driver wellbeing, and ensure a more equitable, efficient urban mobility system.

Glossary Contribution To Key Concepts In Cloud Data Analytics

3 Original Concept Entries

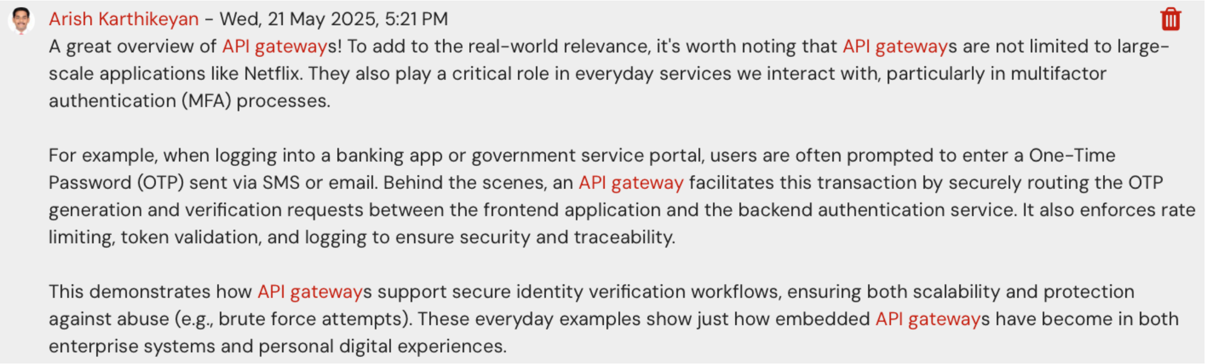
https://lms.latrobe.edu.au/mod/glossary/showentry.php?eid=72178

https://lms.latrobe.edu.au/mod/glossary/showentry.php?eid=72179

https://lms.latrobe.edu.au/mod/glossary/showentry.php?eid=72177

Comments on 3 Peer Glossary Entries

https://lms.latrobe.edu.au/mod/glossary/showentry.php?eid=71419



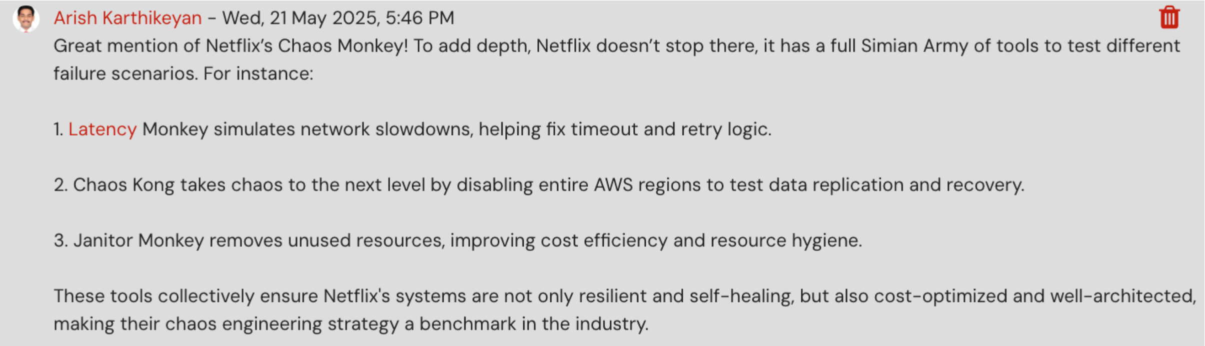
Arish Karthikeyan - Wed, 21 May 2025, 5:21 PM

A great overview of API gateways! To add to the real-world relevance, it's worth noting that API gateways are not limited to large-scale applications like Netflix. They also play a critical role in everyday services we interact with, particularly in multifactor authentication (MFA) processes.

For example, when logging into a banking app or government service portal, users are often prompted to enter a One-Time Password (OTP) sent via SMS or email. Behind the scenes, an API gateway facilitates this transaction by securely routing the OTP generation and verification requests between the frontend application and the backend authentication service. It also enforces rate limiting, token validation, and logging to ensure security and traceability.

This demonstrates how API gateways support secure identity verification workflows, ensuring both scalability and protection against abuse (e.g., brute force attempts). These everyday examples show just how embedded API gateways have become in both enterprise systems and personal digital experiences.

https://lms.latrobe.edu.au/mod/glossary/showentry.php?eid=72488



Arish Karthikeyan - Wed, 21 May 2025, 5:46 PM

Great mention of Netflix's Chaos Monkey! To add depth, Netflix doesn't stop there, it has a full Simian Army of tools to test different failure scenarios. For instance:

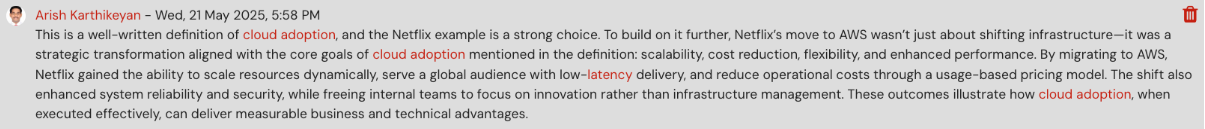
1. Latency Monkey simulates network slowdowns, helping fix timeout and retry logic.

2. Chaos Kong takes chaos to the next level by disabling entire AWS regions to test data replication and recovery.

3. Janitor Monkey removes unused resources, improving cost efficiency and resource hygiene.

These tools collectively ensure Netflix's systems are not only resilient and self-healing, but also cost-optimized and well-architected, making their chaos engineering strategy a benchmark in the industry.

https://lms.latrobe.edu.au/mod/glossary/showentry.php?eid=72154



Arish Karthikeyan - Wed, 21 May 2025, 5:58 PM

This is a well-written definition of cloud adoption, and the Netflix example is a strong choice. To build on it further, Netflix's move to AWS wasn't just about shifting infrastructure—it was a strategic transformation aligned with the core goals of cloud adoption mentioned in the definition: scalability, cost reduction, flexibility, and enhanced performance. By migrating to AWS, Netflix gained the ability to scale resources dynamically, serve a global audience with low-latency delivery, and reduce operational costs through a usage-based pricing model. The shift also enhanced system reliability and security, while freeing internal teams to focus on innovation rather than infrastructure management. These outcomes illustrate how cloud adoption, when executed effectively, can deliver measurable business and technical advantages.

PowerBI dashboard and GitHub repository links

PowerBI: https://latrobeuni-my.sharepoint.com/:u:/g/personal/21970107_students_ltu_edu_au/EUjpkxDw0LVPuO2DnjtV8e8BWrj2nIBFAgAgFfVUSCB0BA?e=Dd8vxR

GitHub: https://github.com/Ak-git-1818/21970107_BUS5001_Assignment2.git

References

Alvarez, L. (2025, February 13). *How Urban Design Affects Pedestrian Safety: The Role of Architecture in Walkable Cities*. AmazingArchitecture. https://amazingarchitecture.com/articles/how-urban-design-affects-pedestrian-safety-the-role-of-architecture-in-walkable-cities

magdy. (2025, February 27). Pedestrian Friendly Urban Planning: The Future of Urban Design. *IEREK*. https://www.ierek.com/news/pedestrian-friendly-urban-planning-the-future-of-urban-design/

NYC Taxi & Limousine Commission, & Wanttaja, R. (n.d.). Taxi Strategic Plan. In *NYC Taxi & Limousine Commission*. https://www.nyc.gov/assets/tlc/downloads/pdf/taxi_strategic_plan_2022.pdf