TRANSFORMING TRANSPORT DATA BY INTEGRATING SPATIAL AND ASPATIAL DATA

Shenuka Abeysena, Danny Wong
Jun 2024

WHO ARE WE?

Danny Wong

Senior Solutions Architect, Databricks

Shenuka Abeysena

Director, Data & Digital Architecture

Data is the key to unlocking our future and Databricks is one of the foundations from which we can achieve our goals.

TRANSFORMING TRANSPORT DATA

Transport domain in the public sector

The transport domain within the public sector refers to the area of government responsibility that deals with the planning, regulation, management, and provision of transportation services and infrastructure for the public. This domain can encompass various modes of transportation, including road, rail, air, and waterways, as well as associated services such as public transport, road transport, freight, traffic management, and infrastructure maintenance.

Overall, the transport domain in the public sector plays a vital role in shaping transportation policies, regulations, and investments to ensure the safe, efficient, and sustainable movement of people and goods within communities and across regions.

Investments to improve, update and add capability to a transport network can lengthy and expensive which makes accurate, timely and informative data critical to the outcome.

SPATIAL VS ASPATIAL DATA

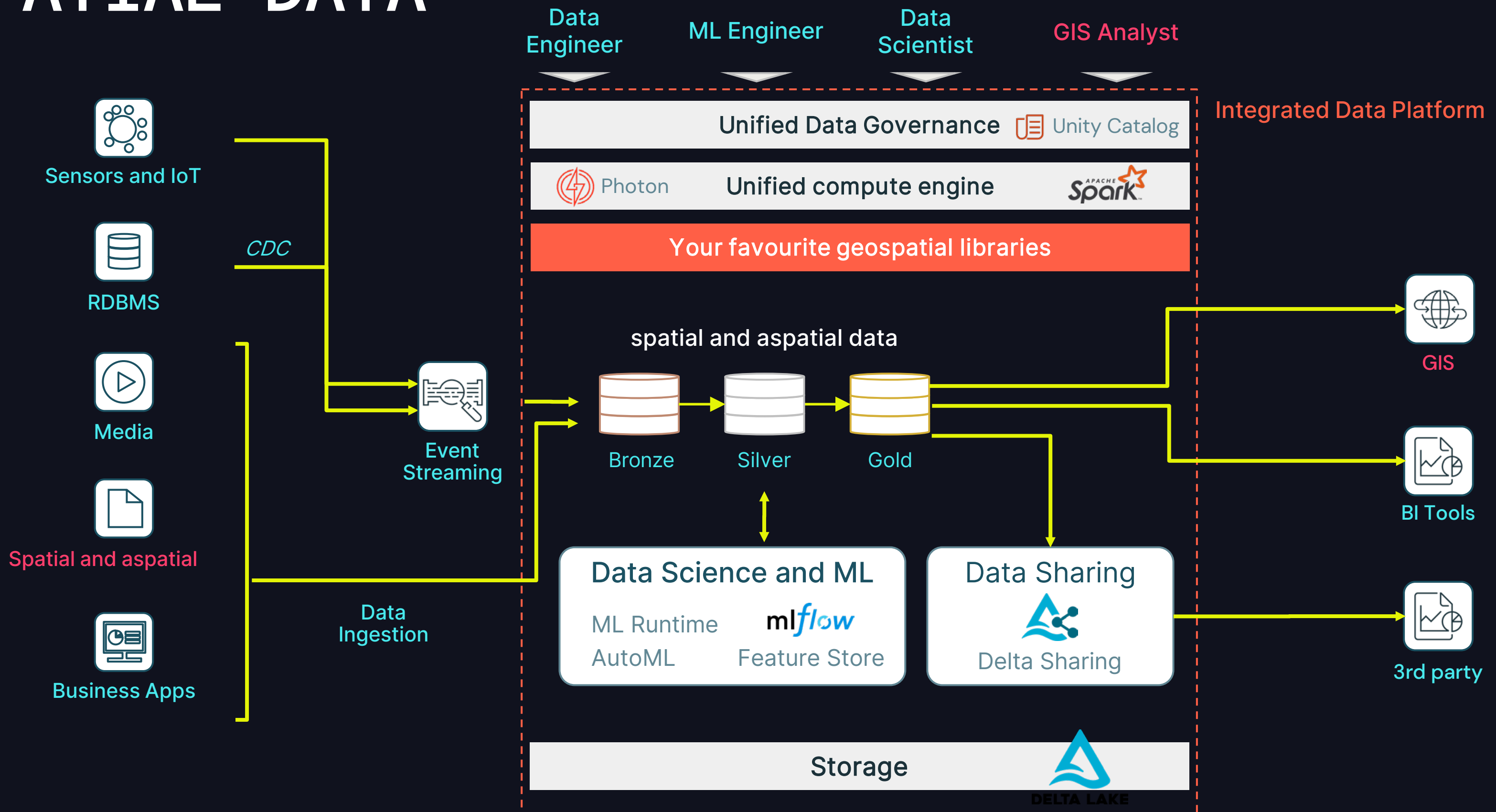
Utilising data to transform how we use and operate transport

Data can be understood through many different perspectives. In the context of the transport domain, we find that the spatial vs aspatial aspect of the data matters quite significantly.

Aspatial data, or non-spatial data, lacks inherent geographic components, representing attributes or characteristics such as demographics, finances, text, or numerical datasets without specific spatial references.

Spatial data, or geospatial data, refers to information tied directly or indirectly to specific geographic locations, represented by geographic features, used for analysing and visualising spatial features, relationships and patterns.

UNIFIED ARCHITECTURE FOR SPATIAL AND ASPATIAL DATA



DATABRICKS GEOSPATIAL LAKEHOUSE

Flexibility to choose your own adventure for geospatial processing

WHY DATABRICKS FOR GEOSPATIAL

Scalable, Flexible and Simplified

Scalability, cost effective and optimized

The H3-centric approach is significantly more cost-effective than geometry-centric or hybrid methods for spatial analytics at full data scale

Flexible Geospatial Data Processing

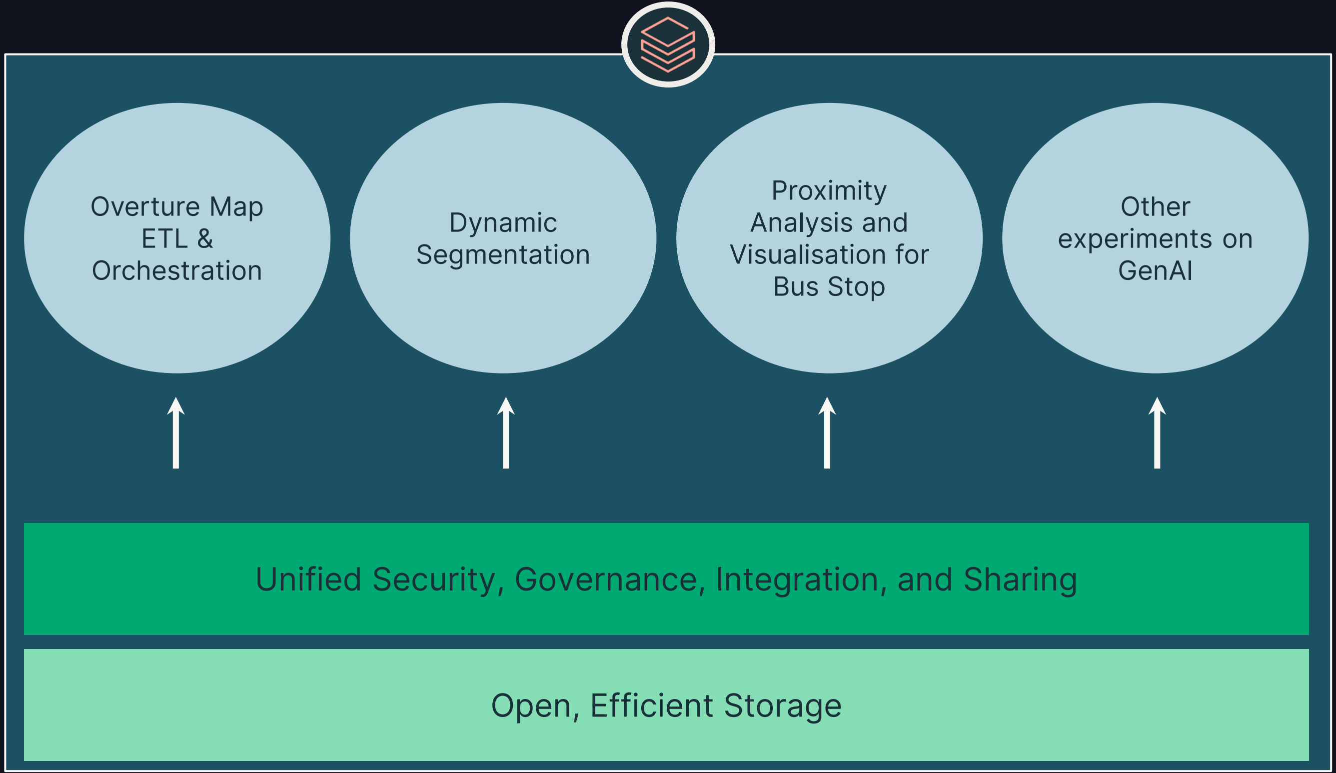
Databricks supports a wide range of geospatial data formats and integrates with various libraries, frameworks, and external GIS tools, providing flexibility in data processing

Govern Data Access and Ensure Compliance

Unity Catalog's governance capabilities help manage access to ingested data, safeguarding sensitive information and ensuring compliance with data privacy regulations.

WHAT ARE WE COVERING TODAY

Transport use cases



Laying the foundation

OVERTURE MAPS SPATIAL DATA INGESTION

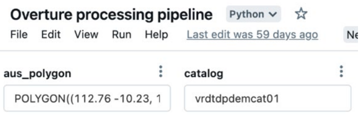
World-wide data

~ 356 GB
raw data

Volumes

Filter on Geometry

DELTA LAKE



Widgets

Multitask Jobs for Maps

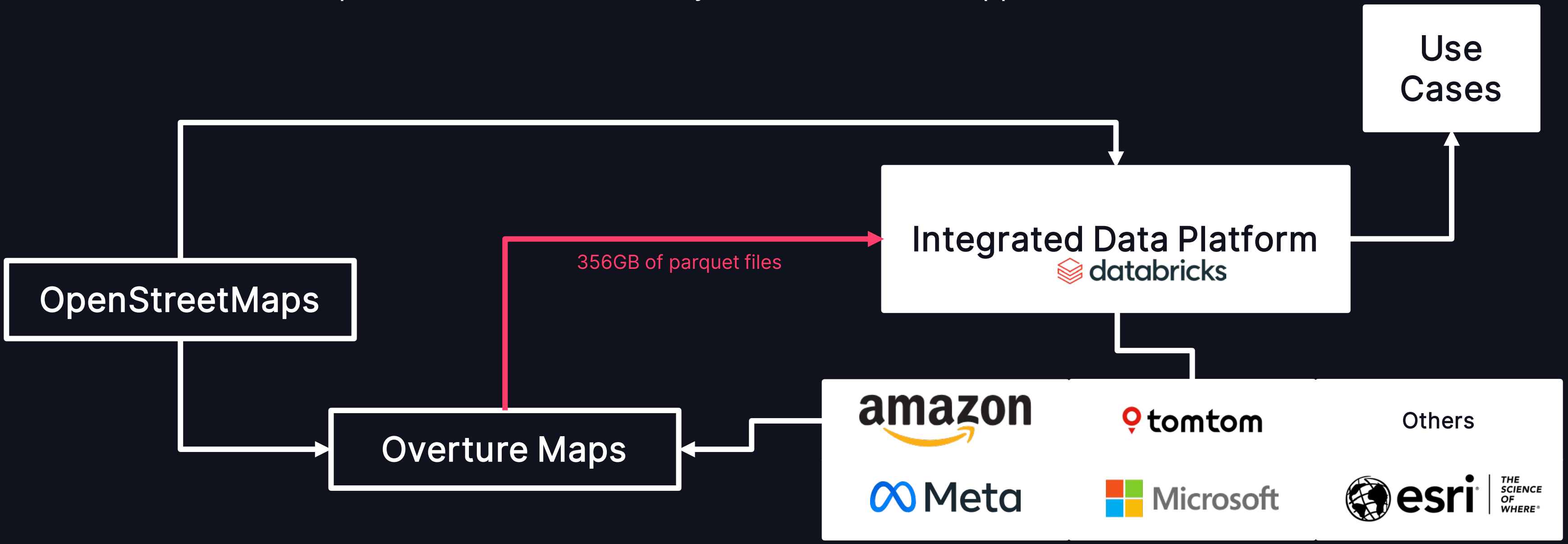
Workflows

OVERTURE MAPS SPATIAL DATA INGESTION

Laying the foundation

Creating and maintaining accurate and up -to date network maps of the world.

Use and reference multiple network simultaneously or on demand to support various use cases.



COPYING THE WORLD-WIDE OPEN MAP DATASET

Overture maps spatial data ingestion

2/17/2024 (37m) 4 Python

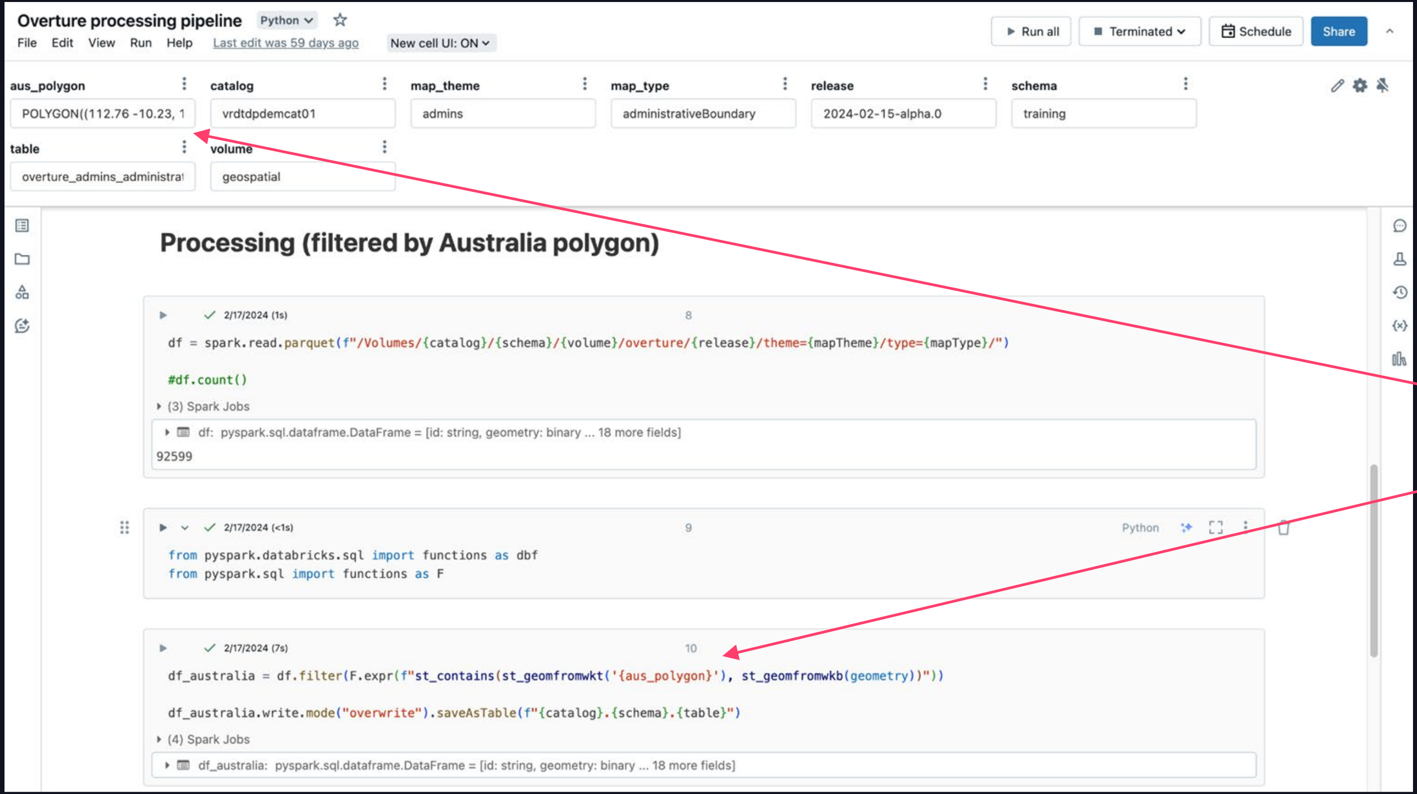
```
%sh
mkdir -p /Volumes/vrdtdpdemcat01/training/geospatial/overture
azcopy copy "https://overturemapswestus2.dfs.core.windows.net/release/2024-02-15-alpha.0/" "/Volumes/vrdtdpdemcat01/training/geospatial/overture"
--recursive
```

```
*** WARNING: max output size exceeded, skipping output. ***

100.0 %, 378 Done, 0 Failed, 0 Pending, 0 Skipped, 378 Total,


Job 46e5df92-3072-844e-685f-9236fa6db88b summary
Elapsed Time (Minutes): 37.3227
Number of File Transfers: 361
Number of Folder Property Transfers: 17
Number of Symlink Transfers: 0
Total Number of Transfers: 378
Number of File Transfers Completed: 361
Number of Folder Transfers Completed: 17
Number of File Transfers Failed: 0
Number of Folder Transfers Failed: 0
Number of File Transfers Skipped: 0
Number of Folder Transfers Skipped: 0
TotalBytesTransferred: 381301554836
Final Job Status: Completed
```

~356GB of geoparquet files

PARAMETERISE THE PIPELINE NOTEBOOK

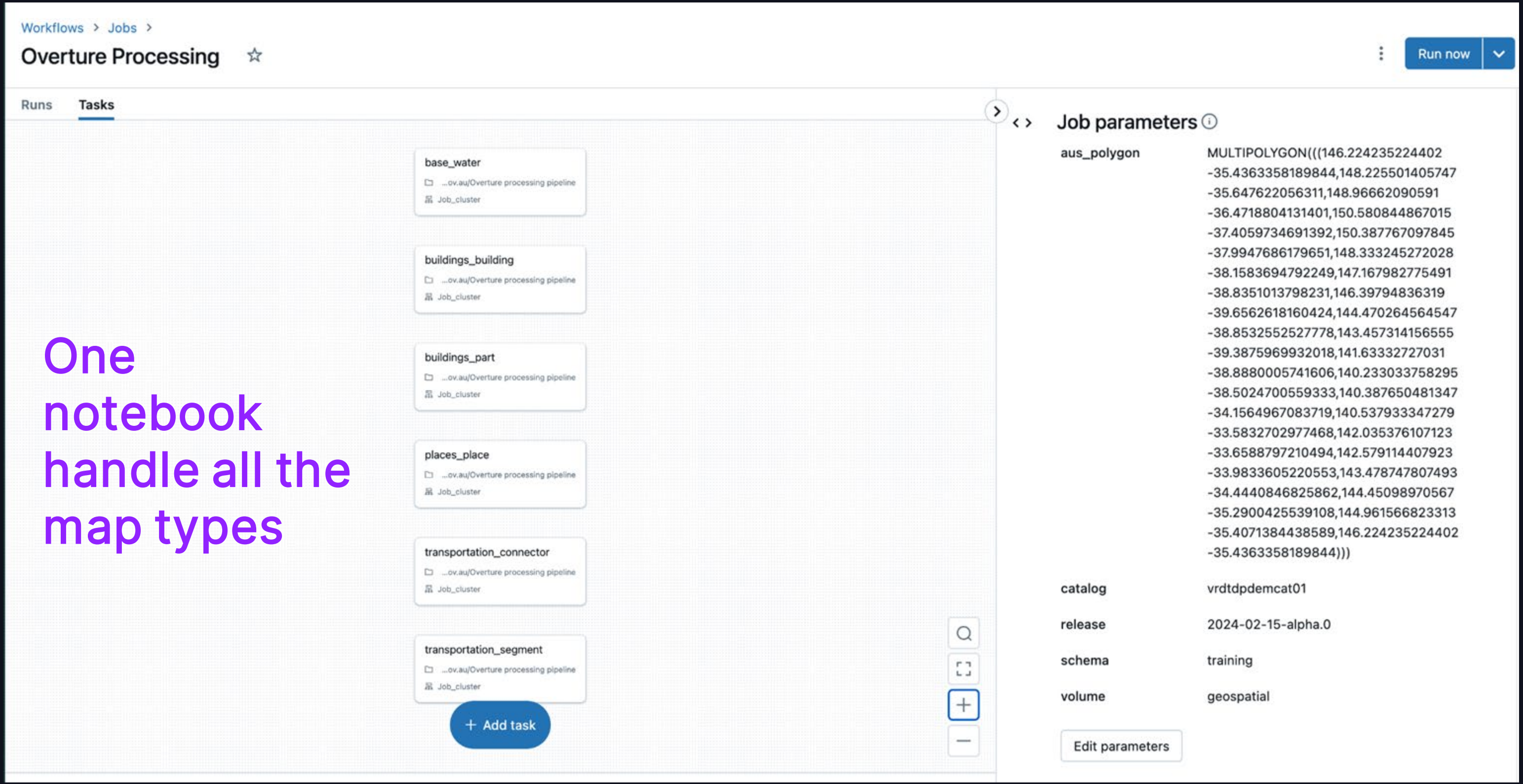
Overture maps spatial data ingestion



Filtered based on Victoria boundary geometry

PRODUCTIONIZED WITH DATABRICKS WORKFLOW

Overture maps spatial data ingestion



One notebook handle all the map types

PROCESS THE WORLD'S GEO DATA

For less than the price of an ice cream scoop 🍧

Select the appropriate cluster size to align with the specific time constraints and budgetary considerations

Cluster size	Time	Databricks Cost
Single node	1 hour 33 minutes	$0.7
2 worker nodes	45 minutes	$1.01
4 worker nodes	27 minutes	$1.01
8 worker nodes	17 minutes	$1.15

🚀🚀 With Photon enabled!

WHAT IT BRINGS US?

A cost efficient framework to bring in new geo data sources

Enrich data with geospatial context

Integrate Overture map data to add geospatial context to existing datasets, enabling more comprehensive insights

Leverage Regularly Updated Geospatial Data

Overture Maps provides a comprehensive, regularly updated dataset to support robust geospatial analysis in Databricks

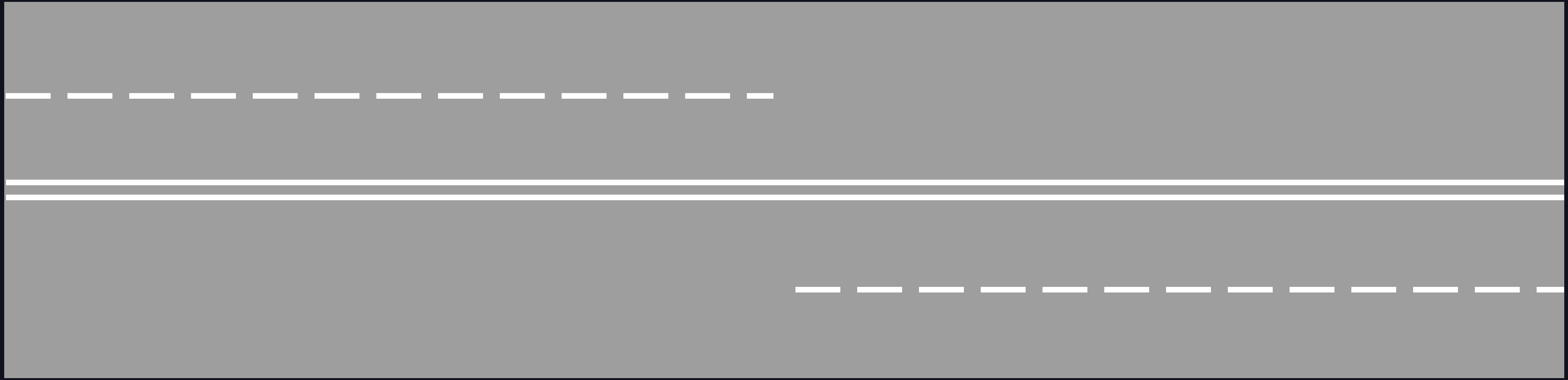
Safeguarded sensitive information

Maintain public trust and prevent potential potential data breaches by restricting access to authorized personnel by UC

Analyzing the road

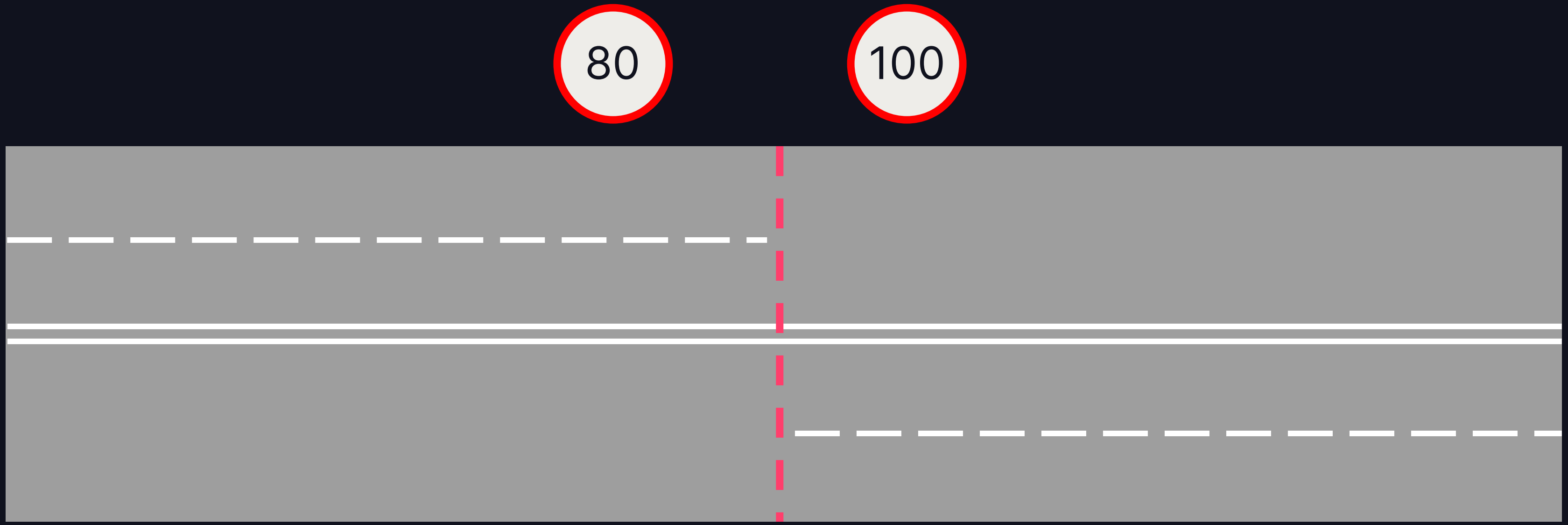
LET'S START WITH A ROUTE

It looks simple, isn't it?



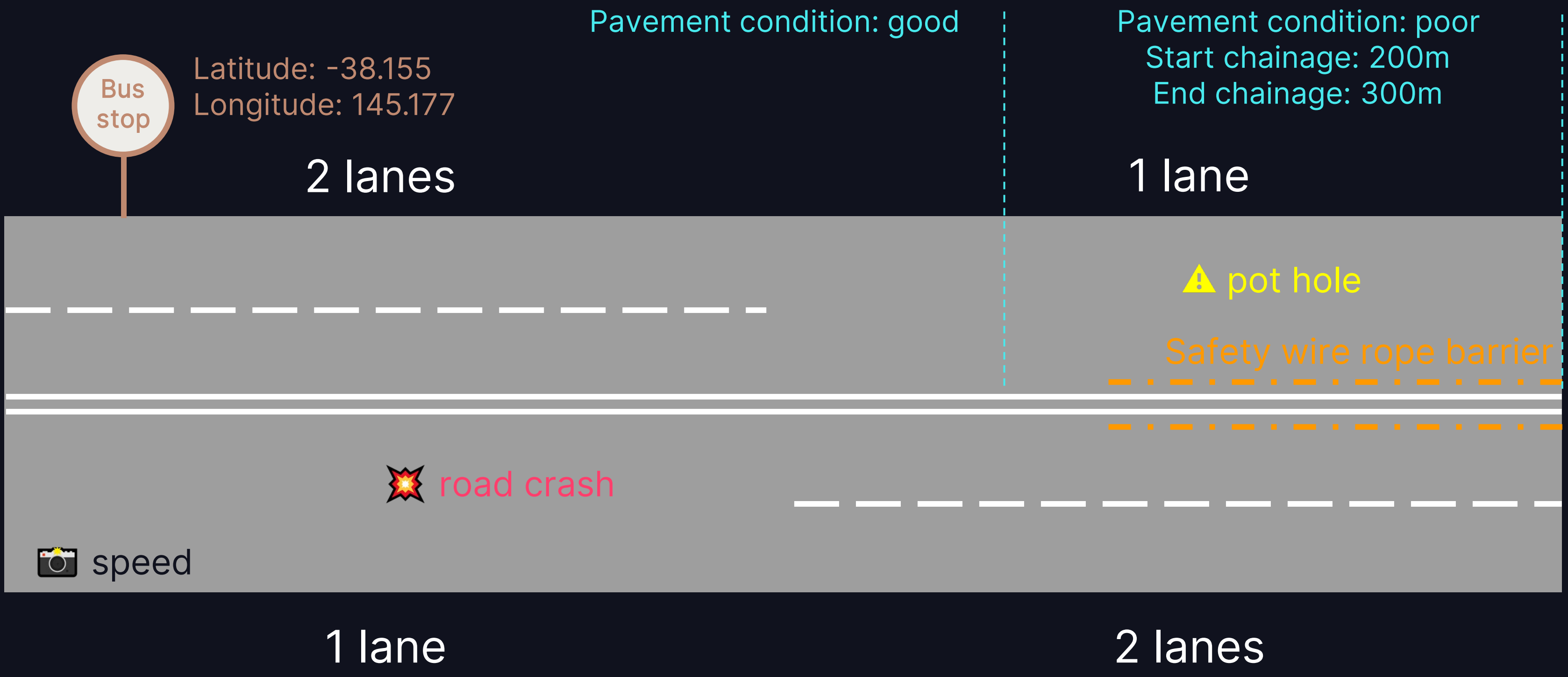
SEGMENT BASED ON SPEED ZONES

Can change over time (e.g. school zones, road maintenance)



BUT IT'S NOT THAT SIMPLE

How to segment them into logical sections based on attributes or events?



DYNAMIC SEGMENTATION FOR STRATEGIC ASSET MANAGEMENT

To understand patterns and identify trends

Dynamic segmentation allows for the integration of various type of transportation related data along the same route for different purposes, such as speed limits, pavement conditions, traffic volumes, number of lanes and asset management.

Route	ID	Start Measure	End Measure	Features Description	Geometry
2000F	A	50		Vehicle breakdown	Point
2000F	B	70	250	Wire Rope Barrier	Line
2000F	C	70	200	60 kph zone	Line

BEFORE

Challenges in scalability, data integration, data freshness

- Traditional GIS tools struggle to handle the large volumes of data generated by modern transportation networks
- Managing and integrating diverse datasets that include various spatial and non-spatial attributes is a significant challenge
- The ability to process and analyze data in real-time is essential for dynamic segmentation to be effective in transportation planning and management

NOW

Geospatial Lakehouse

- Functions like ST_LineSubstring are designed to handle geospatial computations efficiently, scaling to meet the demands of extensive datasets. These functions can be executed in parallel across multiple workers to optimize performance and resource utilization
- A centralized repository serves as the authoritative reference for all spatial and non-spatial (aspatial) data, ensuring consistency and reliability across the entire dataset
- Utilizing structured streaming technology, real-time data feeds, such as those from speed sensors, are processed continuously, enabling immediate data analysis and decision-making

BULK SEGMENTATION AT SCALE

ST_LINESUBSTRING (Sedona)

	route	Start_Chainage_m	End_Chainage_m	geometry
1	57650	11400	11500	> LINESTRING (145.17733810504564 -38.1557111635
2	57651	12300	12400	> LINESTRING (145.24851468539188 -38.1246144723
3	57650	1800	1900	> LINESTRING (145.2633897238568 -38.11067906330
4	57651	7800	7900	> LINESTRING (145.20417415645278 -38.1450013675

SQL

```
SELECT PC.Classified_Road_Number,PC.Direction,
PC.route, PC.Surface_Type, PC.Roughness_Category,
PC.Start_Chainage_m, PC.End_Chainage_m,
ST_LineSubstring(Rte.geometry,
PC.Start_Chainage_m/Rte.ARCLENGTH,
PC.End_Chainage_m/Rte.ARCLENGTH) as geometry
FROM pavement_condition PC, routes Rte
WHERE PC.route = Rte.ROUTE_ID and
Classified_Road_Number = 5765
```

Explanation

In transportation planning or traffic analysis, ST_LineSubstring can be used to extract a particular segment of a road or path for detailed study, such as a stretch of road where frequent accidents occur.

For large LINESTRING geometries (e.g.routes), ST_LineSubstring can be used to create smaller, more manageable segments (e.g. Pavement conditions) for analysis, which can be particularly useful when working with large datasets or when only a specific section of the data is relevant.

FINDING MEASURE VALUE ALONG A ROUTE

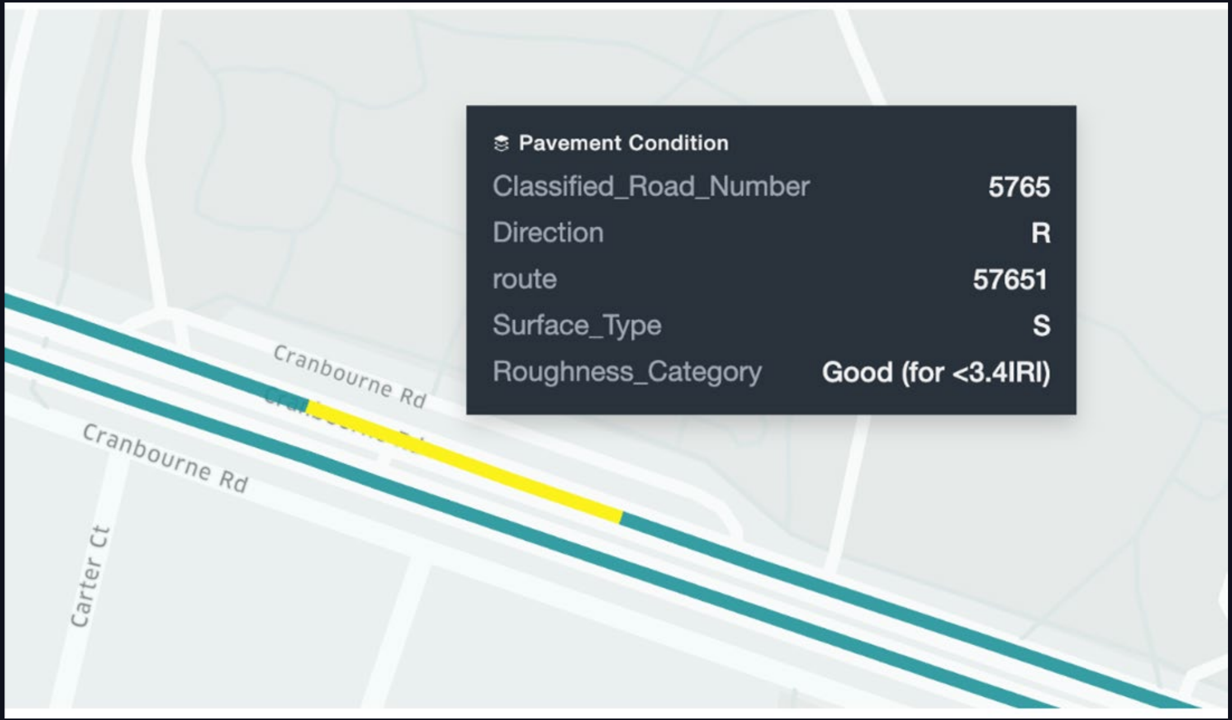
ST_LINELOCATEPOINT (Sedona)

	ROUTE_ID	Location_Description	Metlink_ID	measure
1	57650	Os 140-154 S/S Sladen St W/O South Gippsland Hwy	16344	112.86717559342574
2	57650	DUFF ST W/O RAISELL RD N/O	16344	112.86717559342574
3	57650	Sladen St E/O Lamb St N/S	16520	152.31258653072476
4	57650	Sladen St E/O Cherryhills Drive S/S	21404	1192.629476467152

SQL	Explanation
`SELECT RTE.ROUTE_ID, BS.Location_Description, BS.Metlink_ID, ST_LineLocatePoint(RTE.geometry, BS.geometry) * ARCLENGTH as measure FROM route_bus_stops BS, routes RTE WHERE RTE.ROUTE_ID = 57650 ORDER BY measure`	This SQL query is designed to calculate the distance along a specific bus route to each bus stop on that route. For managing transportation assets such as bus stops, signage, and maintenance points, ST_LineLocatePoint can help in pinpointing their exact locations on the road network. This aids in asset inventory management, maintenance scheduling, and optimizing the placement of new assets.

VISUALISING SEGMENTS BY ATTRIBUTE

Infrastructure analysis and asset maintenance planning



WHAT IT BRINGS US?

Perform segmentation and analysis at scale

Improved Operational Efficiency

The scalability of geospatial functions allows for faster processing of large datasets, enabling the transport agencies to update and analyze their data more frequently

Reduced Data Management Costs

By eliminating data silos and centralizing data management, a unified data repository can help reduce the costs associated with managing and maintaining multiple, disparate datasets

Improved Situational Awareness

The ability to process and analyze real-time geo data streams enables the transport agencies to respond more quickly to changing conditions on the ground

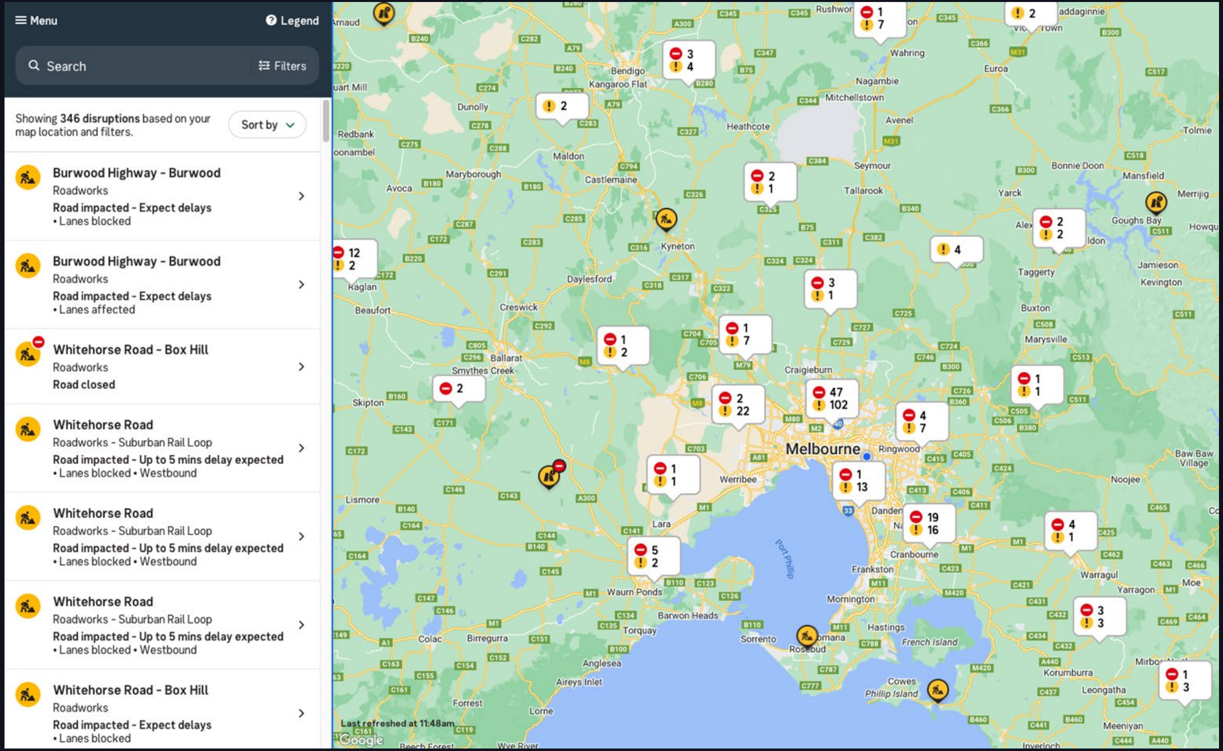
Bringing data to life

FINDING A NEEDLE IN A HAYSTACK

Visualization simplifies the understanding of complex data

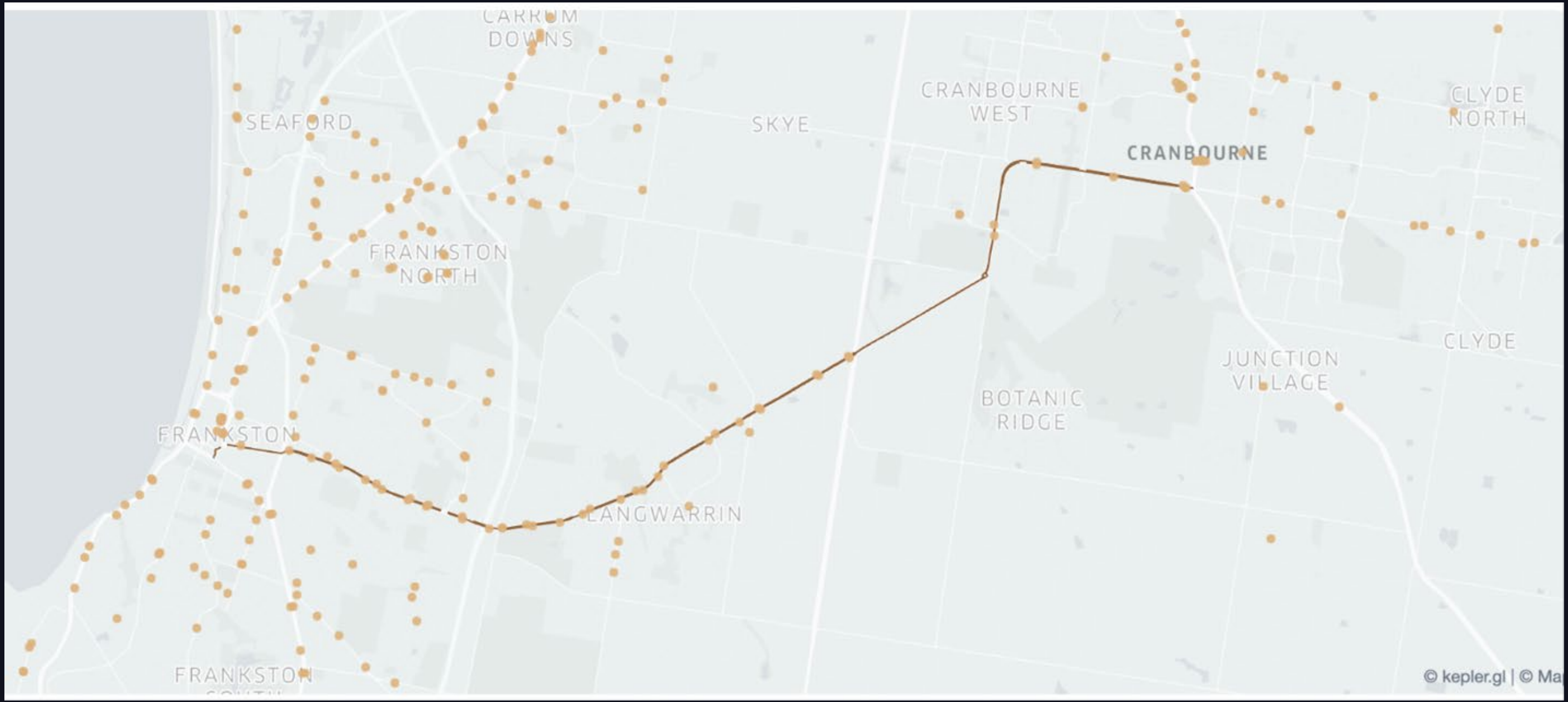
Numerous dimensions

- Bus stop locations
- Passenger flow
- Traffic volume
- Public transport time table
- Emergency response
- Planned and unplanned disruption, etc



VISUALIZATION OF BUS STOP ANALYTICS

Challenge: How can we easily find all the bus stops along a specific route?



THE ST_BUFFER APPROACH

Creates a buffer zone around a route with a distance (angular units)

THE ST_BUFFER APPROACH

Lessons learned

😍

- **Familiarity** - Commonly used in GIS software and supported by many spatial databases, ST_Buffer is well-documented and widely understood by GIS professionals
- **Flexibility** - The buffer's size can be easily adjusted to reflect different definitions of "nearness"

🥲

- **Performance** - Buffering can be computationally expensive, especially when dealing with complex linestrings or large datasets
- **Complex** - using angular units (degrees) for buffer distances can lead to inaccuracies, especially over larger distances or near the poles, because degrees do not represent consistent physical distances across the globe → Lots of experimentation 🧪

THE H3 K-RING APPROACH

Generates a set of hexagons around hexagon within a specified grid distance. Approximates a circle.

THE H3 K-RING APPROACH

K = 5 (5 steps away from the center as if you were walking through the grid)

- Generating K-ring indexes for the route
    - Helps in understanding the spatial distribution of bus stops and their proximity to the route
- When generating a k-ring with k = 4, no bus stops are included within the ring
- When k = 5, only the bus stops on one side of the road are being captured

THE H3 K-RING APPROACH

K = 7

- When k = 7, the bus stops on both sides of the road are being captured
- When visualizing bus stop locations from various sources, it's crucial to recognize that slight coordinate discrepancies can occur, highlighting the importance of visualization for identifying and reconciling these differences

THE H3 K-RING APPROACH

Lessons learned

😍

- **Scalability** - H3 is designed to optimize and scale geospatial analysis and it can efficiently handle large datasets, which is beneficial for statewide transportation analysis
- **Multi-resolution** - H3 supports multiple levels of resolution, allowing for flexible granularity in analysis. This can be useful for zooming in on high-interest areas or zooming out for a broader overview

🥲

- **Approximation** - While H3's hexagonal grids offer many benefits, they are an approximation of space and may not perfectly align with the actual shapes and paths of transportation routes
- **Integration** - H3 might require additional integration effort if the existing geospatial stack is not designed to work with H3 indices

GENERATING KRING INDEX FOR THE ROUTE

H3_KRING

SQL	Explanation
`SELECT b.*, r.* FROM (SELECT *, explode(h3_kring(cellid, 7)) as kring FROM routes_wkt_h3) as r JOIN busstops_geom_wkt_h3 as b ON r.kring == b.cellid`	The query is about spatially joining bus stop points with the nearest route linestrings, while also considering the use of H3's k-ring function with a k value of 7 for geospatial analysis. The k value of 7 in the H3 k-ring function indicates the grid distance from an origin cell, identifying all hexagons within that proximity, which could be applied to determine the bus stops' vicinity to routes

DISPLAY ALL BUS STOPS OF INTEREST

Separating the signal from the noise

WHAT IT BRINGS US?

Rendering data on a map for solving complex questions

Better Data Quality Assurance

Visualize bus stop data from multiple sources, identify the discrepancies in the geometry or other attributes among multiple data sources

Smarter Infrastructure Maintenance

Leverage proximity analysis to inform cost-effective maintenance schedules

Accessibility Analysis

Allows planners to analyze the accessibility of public transit for the community, ensure that stops are strategically located to serve high-demand areas

Other experiments

STREAMLINING DISRUPTION INSIGHTS

DBRX + AI_QUERY() to summarize unplanned disruptions

Raw results +

	wkt	closedRoadName	closedRoadSESRegion
1	POINT(143.890262102896 -35.8587189745485)	WOOD LANE	LODDON MALLEE (NORTH WEST)
2	> LINESTRING(143.88916136		
3	POINT(146.779975338703 -3		
4	> LINESTRING(146.77944470		
5	POINT(144.346174832714 -3		
6	> LINESTRING(144.34821814		
7	POINT(142.670228015771 -3		
8	> LINESTRING(142.67600025		
9	POINT(144.95415960551 -36		
10	POINT(144.522684744554 -3		
11	> LINESTRING(144.47518615		
12	POINT(145.117981791237 -3		
13	> LINESTRING(145.11940858		

Raw results Table 1 +

analysis

Unplanned Disruption Analysis

- There are a significant number of unplanned disruptions across various regions in Victoria, with a total of 20 incidents recorded in the provided data.
- The most common cause of disruptions is flooding, accounting for 9 out of 20 incidents (45%). This is likely due to the data being collected during a period of heavy rain and flooding in the state.
- Other causes of disruptions include roadworks (3 incidents), hazards (3 incidents), and bridge damage (2 incidents).
- The majority of disruptions (12 out of 20) involve closures of the entire road, which can significantly impact traffic and travel times.
- The data also shows that some disruptions have been ongoing for several months, indicating that they may be complex and require significant time and resources to resolve.

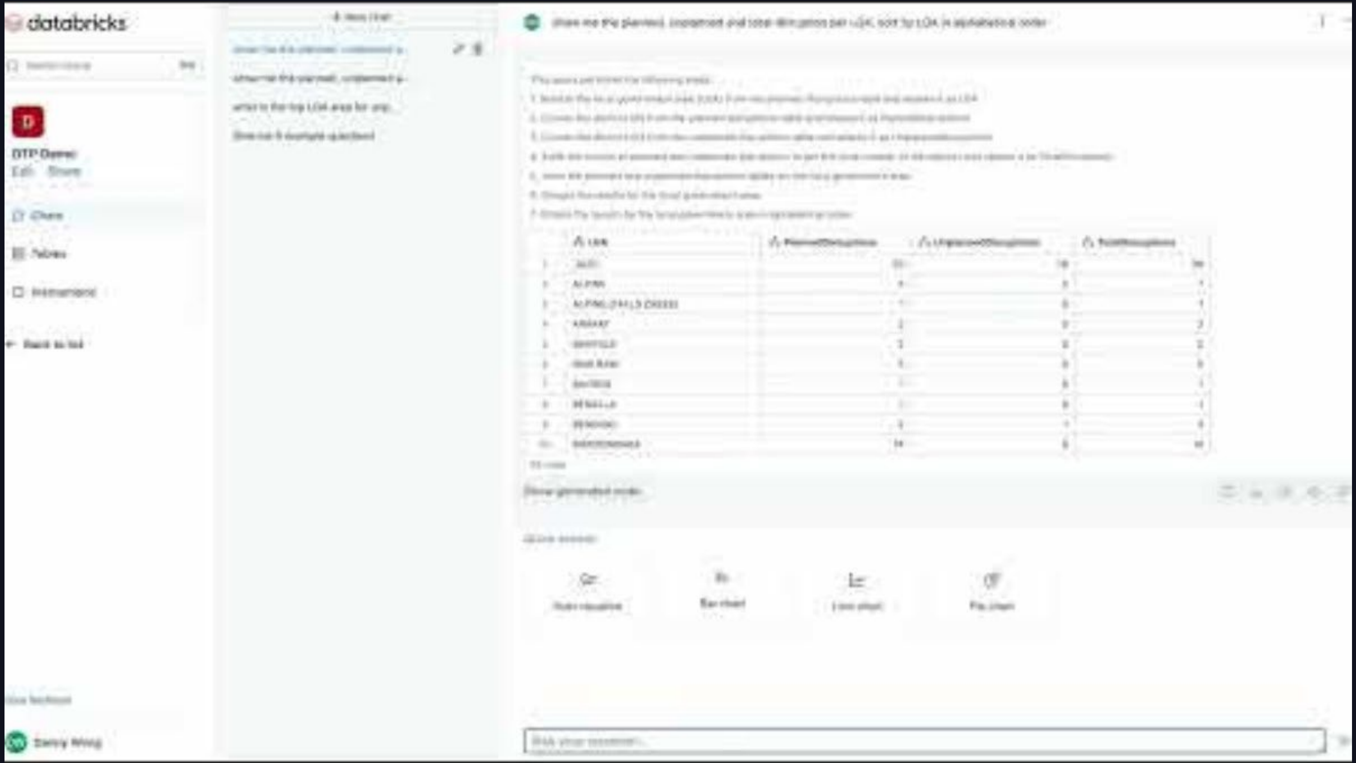
Cause of Disruption	Number of Incidents
Flooding	9
Roadworks	3
Hazard	3
Bridge Damage	2

Severity of Disruption	Number of Incidents
Entire Road Closed	12
Lane Closure	5
Speed Limit Reduction	3
No Blockage	0

Duration of Disruption	Number of Incidents
Less than 1 month	5
1-3 months	7
More than 3 months	8

STREAMLINING DISRUPTION INSIGHTS

Conversational Analytics with Genie Data Room



THE FUTURE

What next for spatial data

Key Takeaways:

- Data interoperability drives innovation in transport - combining spatial and aspatial data aspects can yield benefits in multiple ways
- Democratisation of data - collaboration is essential for realising the full potential of data-driven innovation
- Flexible geospatial data processing at scale in a way that is cost effective, optimised and adds intelligence

Call to Action:

- Embrace data interoperability as the foundation for future innovation and improvement
- Foster collaboration across teams, functions and capabilities to drive transformative change in transport

Data is the key to unlocking our future

DATA+AI SUMMIT

databricks

THANK YOU!