

Modernizing Enterprise Analytics at IDEM

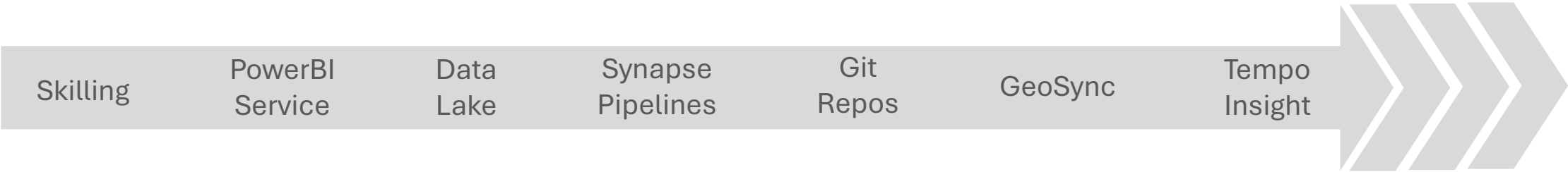
IDEM Data Operations

- Vanessa Williams, Cloud Solutions Manager
- Muhammad Azimi, Enterprise Data Engineer
- Kayla Sturgeon, Enterprise Data Analyst
- Chris Waldron, Enterprise Data Scientist



Tooling at the Enterprise Level

2022



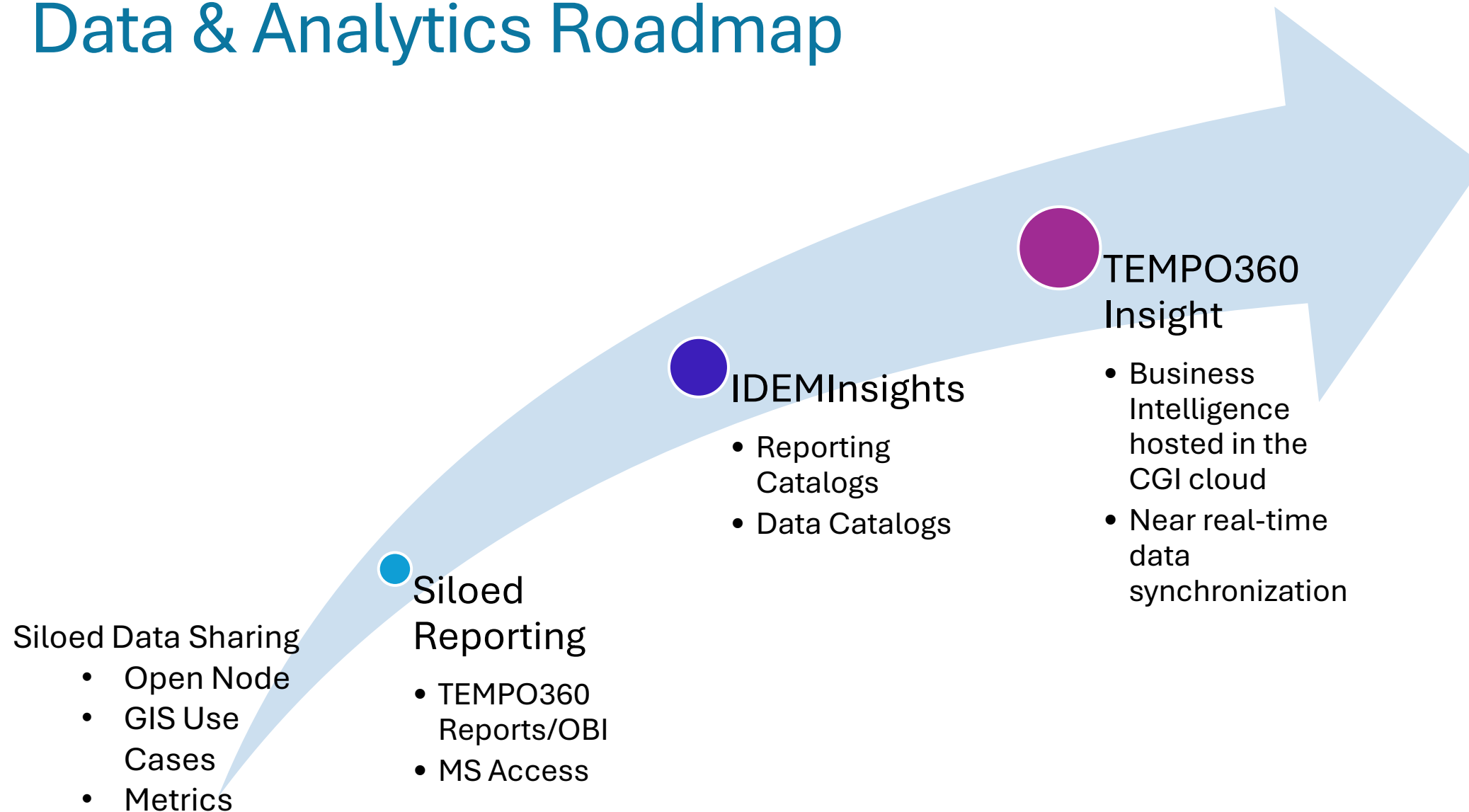
Operational Value

- Reusable and Repeatable Data Solutions (Automation)
- Shared Management of Data Processes
- Consolidated Data Source
- Process and Code Tracking and Storage

Business Value

- Data Accessibility and Democratization (Improved Reporting)
- Consistent Data Across all Data Solutions (Improved Data Accuracy/Single Source of Truth)
- Enables Data Governance Implementation
- Improved Reporting, Insights, and Decisions

Data & Analytics Roadmap





- Permits
- Registrations
- Notifications
- Determinations
- Memberships

Authorizations ✓

- Field Inspections
- Audits
- Record Reviews

Compliance ✓

- Warning Letters
- Violation Letters
- NOV/AO
- CO

Enforcement ✓

- IDEM Lead
- Oversight

Remediation ✓

- Complaints
- Spill Reports
- Media Requests
- Assistance

Incidents ✓

- Fee Assessments
- Invoices
- Payments

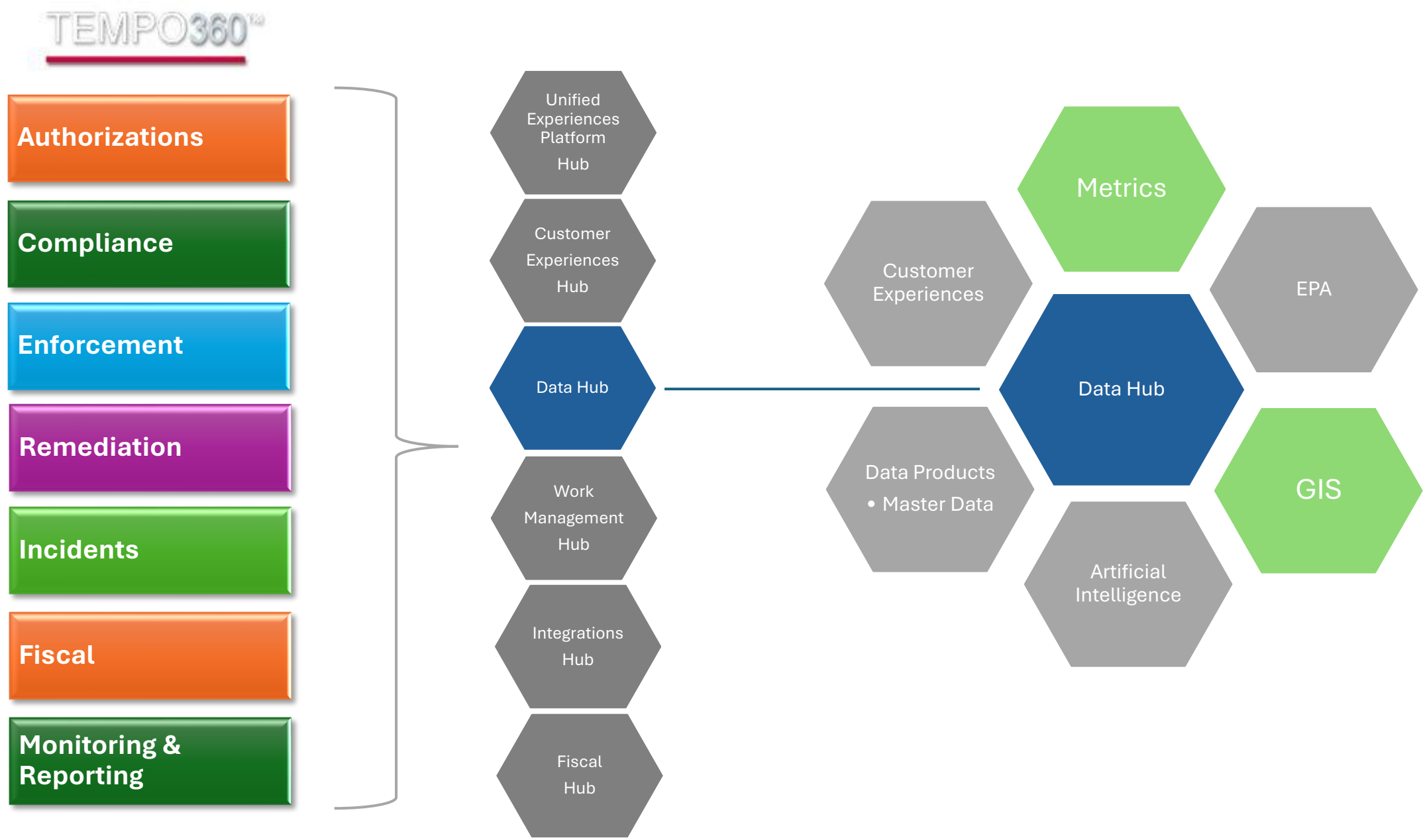
Fiscal ✓

- Reports Submitted to IDEM by Regulated Entity
- Monitoring or Sampling Conducted by IDEM

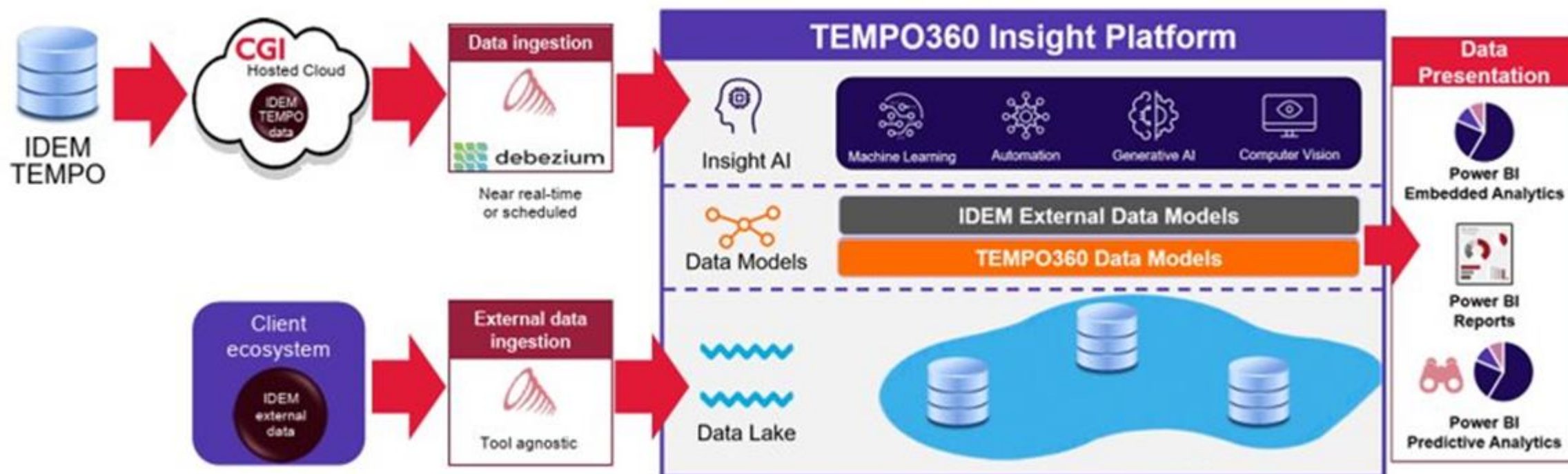
Monitoring & Reporting ✓

- Investigations
- Institutional Controls
- Support Services
- Grants & Loans
- Exams & Licensing

And More... ✓



Tempo 360 Insight Platform





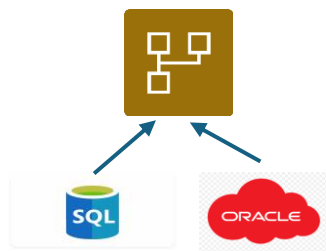
AUTOMATION

Automating Insights in Power BI

Data Lifecycle – Premium Capacity

Dataflows

- Oracle Connector
- Executed stored procedures and packages



Semantic Models

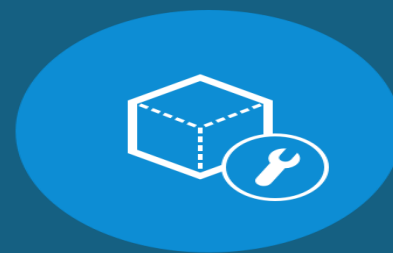
- Set up parameters
- DAX expressions
- Power Query transformations
- Visuals
- Publish to DEV

Testing

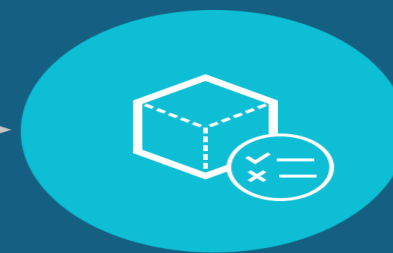
- Deploy to UAT for user testing
- Deploy to PROD when testing is complete

Deployments

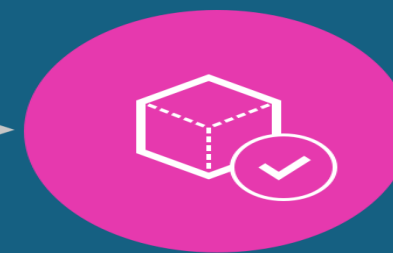
- Set up parameter rules to point to the correct workspace storing the correct data source. (DEV, TEST, PROD)



Development



Testing/QA



Production

IDEM Data Catalog

OLQ Data Catalog

OWQ Data Catalog

OLQ
Branches/Sections
Data Catalogs

OLQ
Branches/Sections
Reporting

OWQ
Branches/Sections
Data Catalogs

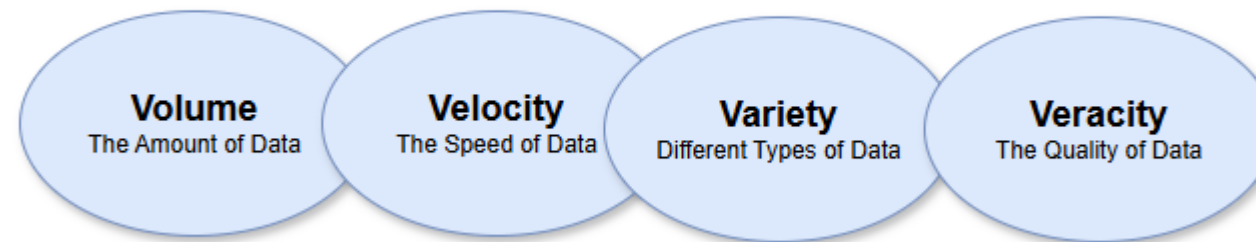
OWQ
Branches/Sections
Reporting

Lost your work? No problem!

- Track changes to reports and semantic models over time.
- Facilitate collaboration among developers through version control.
- Easily revert to previous versions if needed.

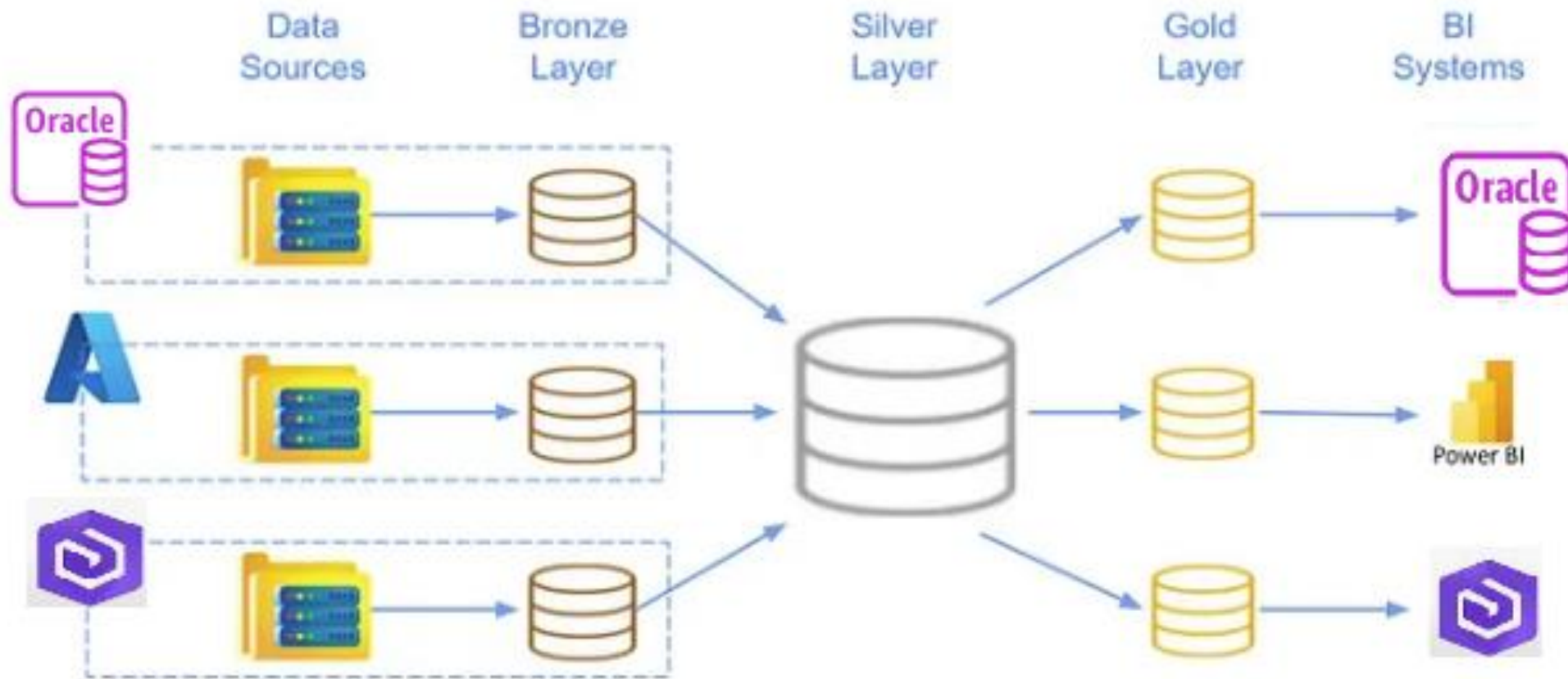


The Need for Efficient Data Integration and Analytics



- **Volume:** Amount of data generated from various sources
- **Velocity:** the speed at which data is being generated, collected, and processed.
- **Variety:** the diverse types and formats of data that are generated.
- **Veracity:** the uncertainty or trustworthiness of the data

Medallion Architecture



Azure Synapse Pipelines

Notebooks

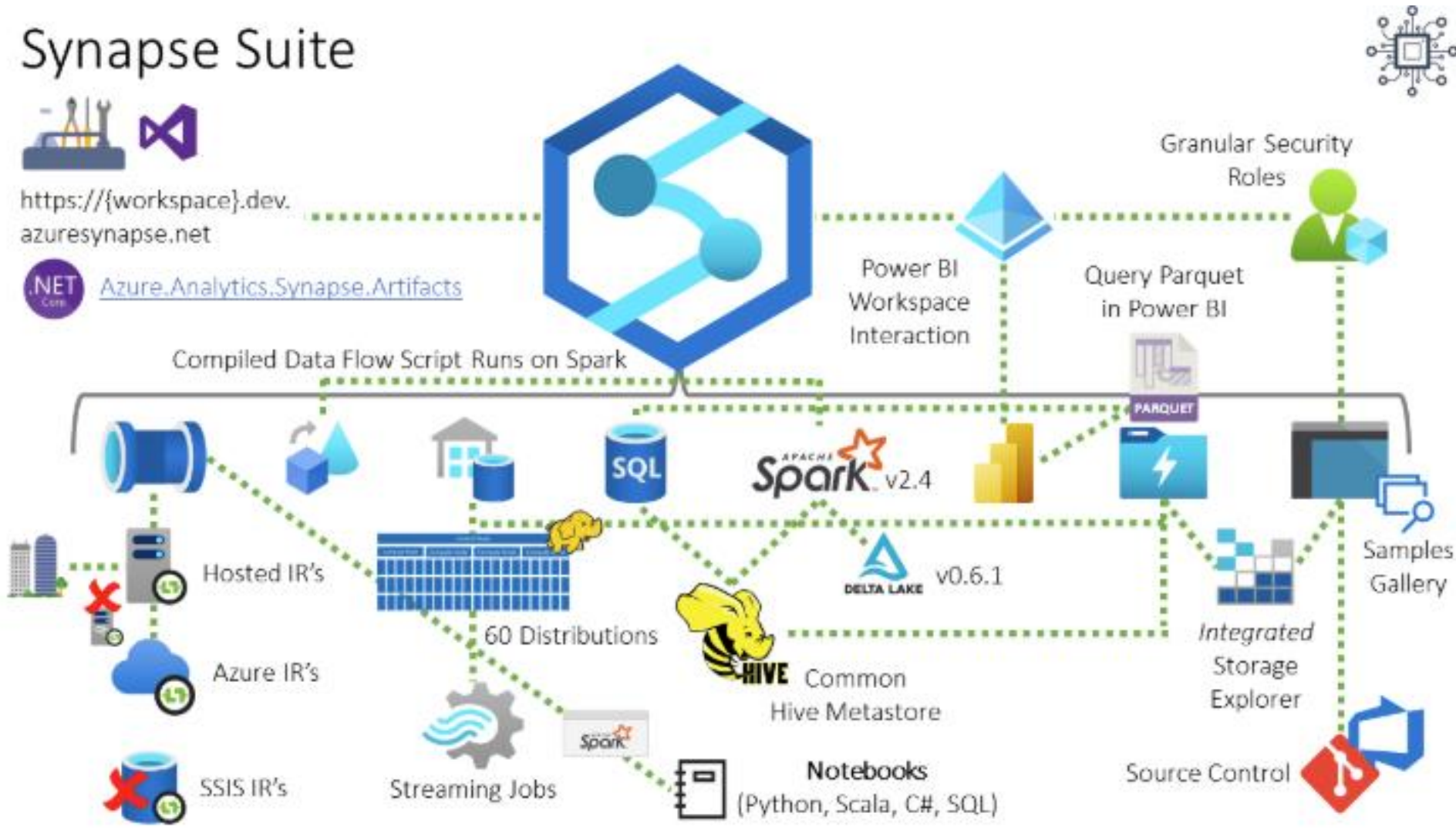
Apache Spark:

Dedicated SQL Pool/Data Warehousing:

Serverless SQL

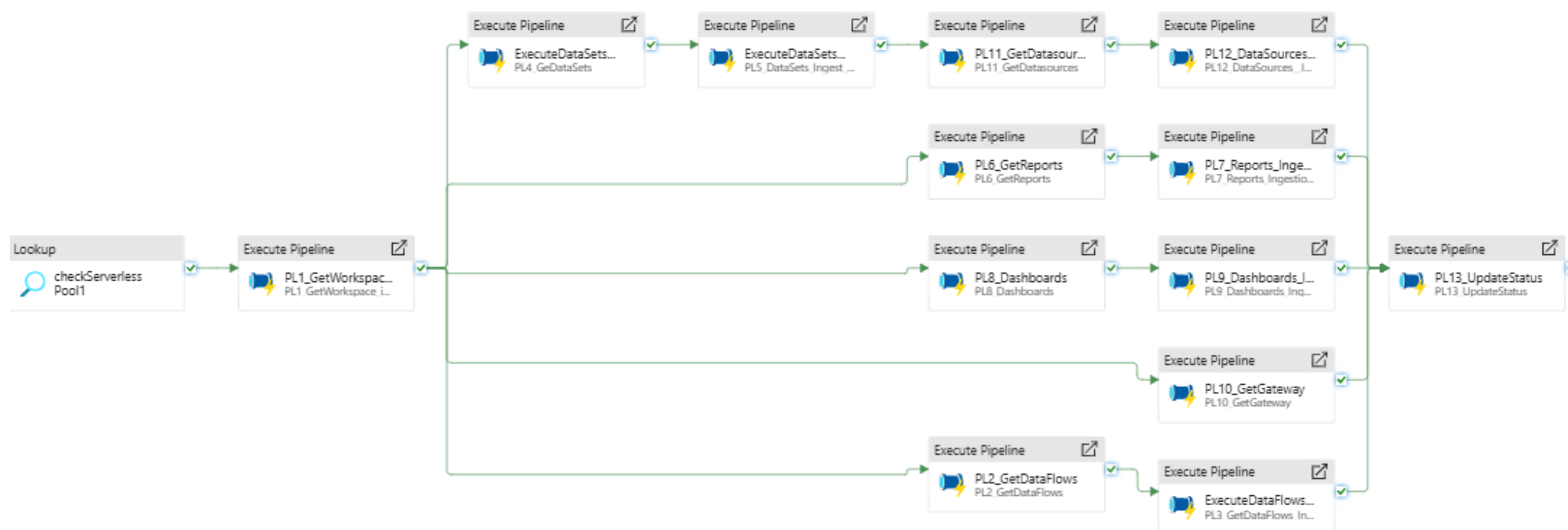
Data Exploration Tools

Other Tools: IR (Azure/Default IR, SHIR, SSIS IR), Azure Key Vault to store password/secrets



Azure Synapse Analytics Pipelines

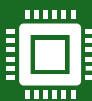
- Ideal for complex ETL processes and data orchestration, almost same UI and functionality as of Azure Data Factory
 - Facilitates nightly updates from on-premises Oracle databases, including the ERP system TEMPO.
 - Enables integration of TEMPO data with various data sources
 - Supports comprehensive analysis and linking of data across systems.



Challenges



In our ERP system, TEMPO, we could not find a watermark column in many of the tables needed for performing delta refreshes. Consequently, we have to rely on full refreshes, which may present future challenges due to the growing data volume.



An alternative solution may include the implementation of database replication technology. IDEM is presently engaged in discussions with the TEMPO vendor, CGI, to facilitate the ingestion of data into the vendor's cloud utilizing replication technology. Subsequently, there are plans to implement Medallion Architecture surrounding the replicated data.

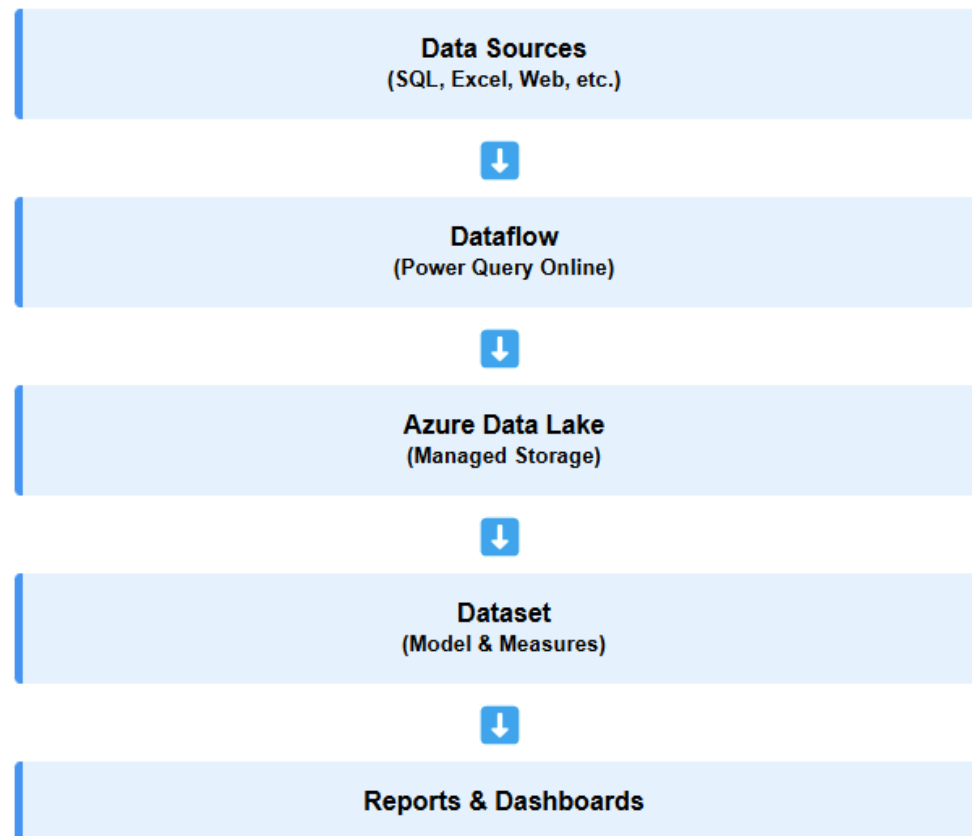


Additionally, IDEM has previously invested considerable effort in developing Oracle Business Intelligence (OBI) reports that are based on PL/SQL packages; however, we are unable to utilize these Oracle programs to transfer data into ADLS through Synapse pipelines technical limitations.



On a positive note, we have successfully used these oracle PL/SQL packages in Power BI data flows.

Simplifying ETL and Data Preparation with Power BI



- Dataflows in Power BI are cloud-based ETL (Extract, Transform, Load) solutions used to prepare data. They allow you to define reusable data transformation logic using Power Query (M language) and store the cleaned data in Azure Data Lake (managed by Power BI). Empowers users to prepare data and build reports directly in Power BI.
- Example Use Case: OLQ leveraging Power BI Dataflows for analytics.

Comparison: Azure Synapse pipelines vs. Power BI Dataflows

Feature	Azure Synapse Pipelines	Power BI Dataflows
Complexity of ETL	Ideal for complex ETL	Suitable for simple ETL
User Audience	Data engineers, IT professionals	Business users and IT Prof/Report-Developers(Data Analyst)
Integration	Multi-source (Oracle, GIS, etc.)	Multi source





SOI Ecosystem for GIS: Desktop Software, SaaS, PaaS

The collage displays various ArcGIS Pro components:

- Top Left:** ArcGIS Pro - PortlandDO - Scene. A 3D city model with a 'Trees' layer selected. A 'My Identify' window shows a table of tree data.
- Bottom Left:** ArcGIS Pro - Geosoft Grids. A 2D map view showing a 'Slope' layer with a color-coded symbology.
- Center:** ArcGIS Pro - PortlandDO - Scene. A 3D city model with a 'Trees' layer selected. A 'My Identify' window shows a table of tree data.
- Bottom Center:** ArcGIS Pro - PortlandDO - Scene. A 3D city model with a 'Trees' layer selected. A 'My Identify' window shows a table of tree data.
- Right:** ArcGIS Pro - Add_data_to_a_project - V. A 2D map view showing a 'Slope' layer with a color-coded symbology.

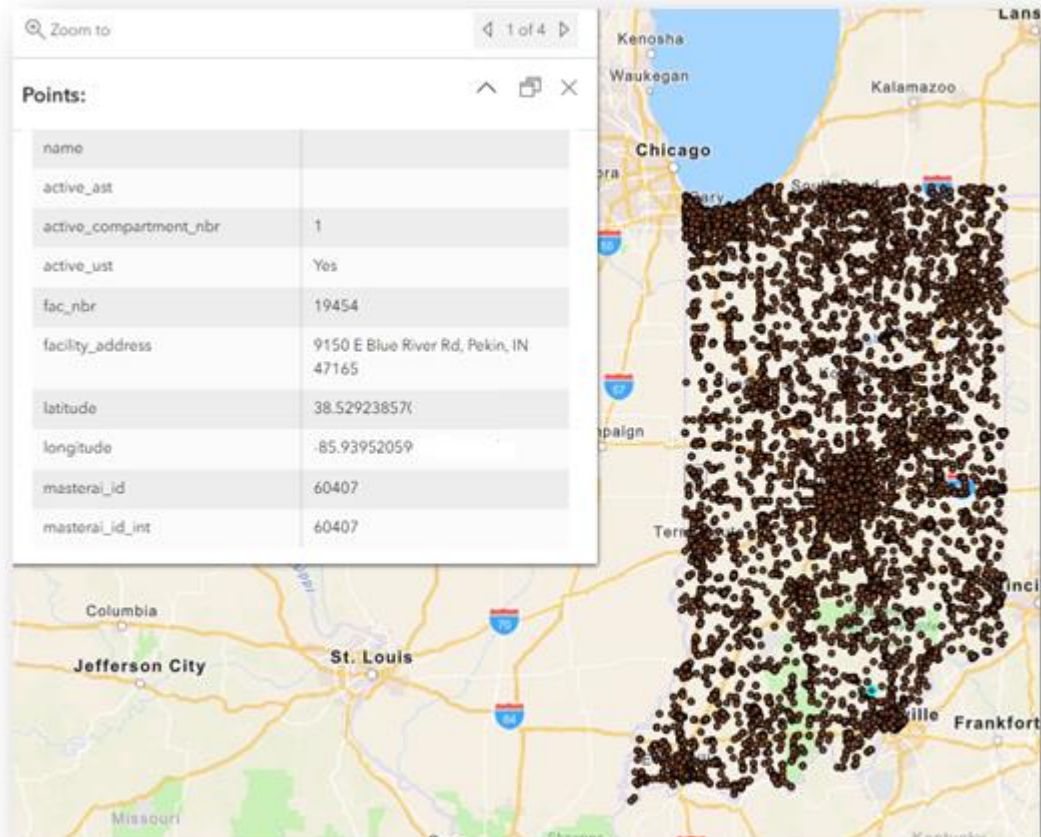
Analysis Tools Panel (Right):

- 3D Analyst Tools
- Analysis Tools
 - Extract
 - Clip
 - Select
 - Split
 - Split By Attributes
 - Table Select
 - Overlay
 - Pairwise Overlay
 - Proximity
 - Statistics
- Aviation Tools
- Business Analyst Tools
- Cartography Tools
- Conversion Tools
- Crime Analysis and Safety Tools
- Data Interoperability Tools

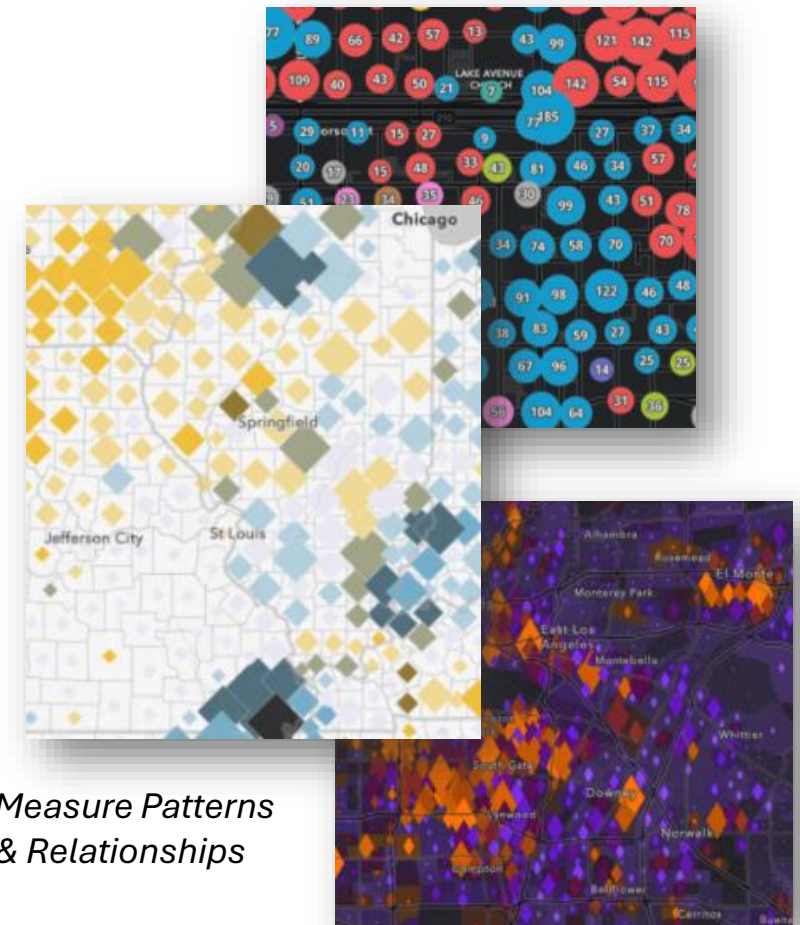
Python Window (Bottom Center):

- Python Notebook: Open a new notebook.
- Python Window: Show python window.

GIS for Reference: “Points on a Map”

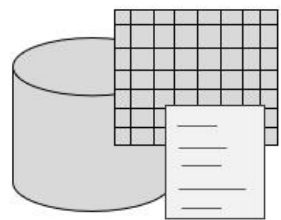


GIS for Business Value: Spatial Analysis, Viz, Insights, Apps



GIS Data Sources

(at IDEM)

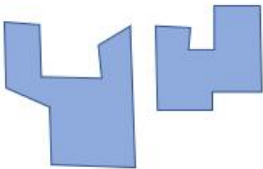


Address/
Lat & Long

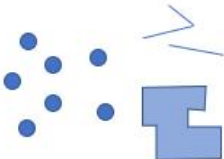


Beginning at the North
Range 1 West; thence S
the Ohio River to Log
Creek to the West line
thence North with said
of said Section; thence
to the Northeast corner
of Section 36, Township

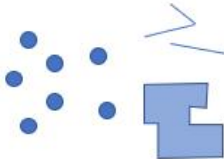
Property
Description



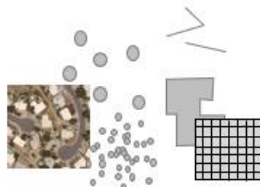
CAD/
Images



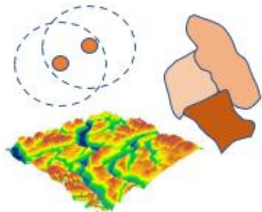
High Accuracy
GPS



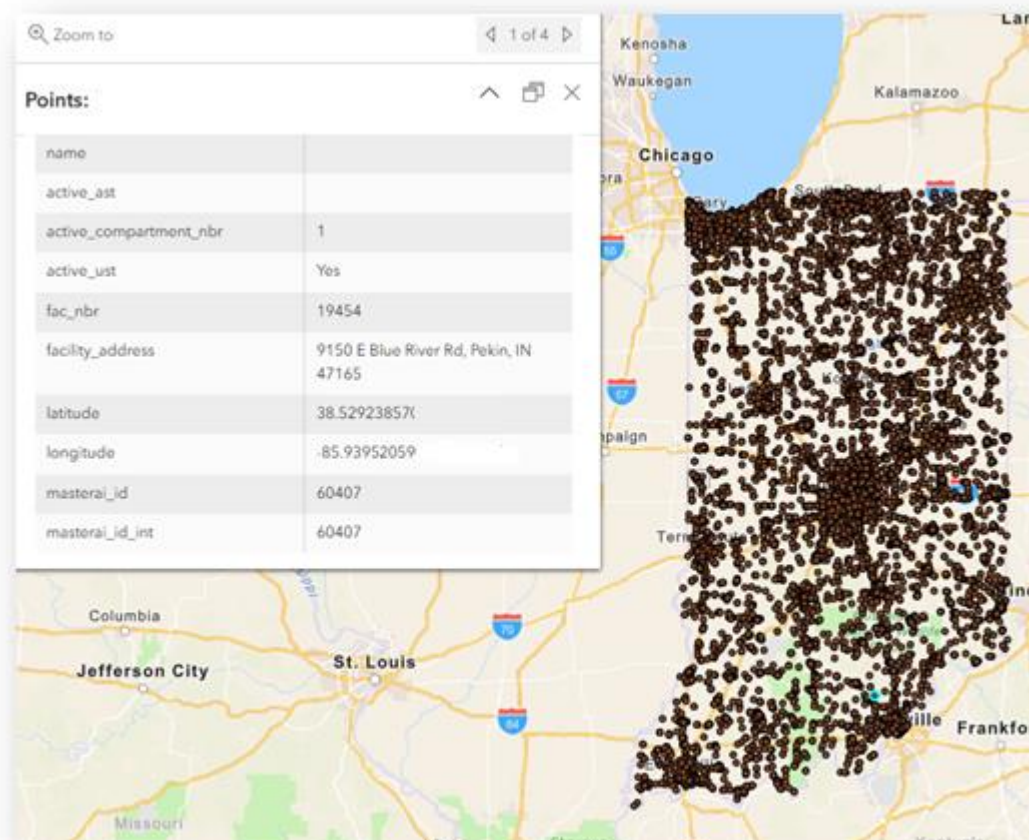
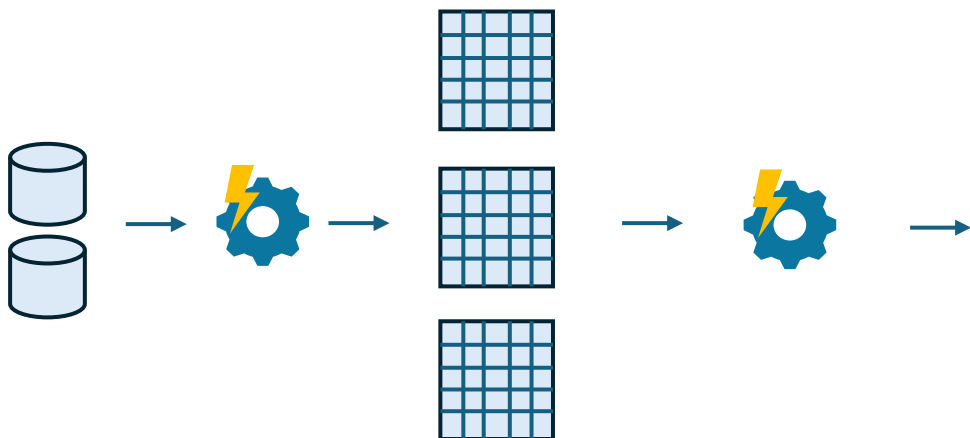
Ortho/LiDAR/
Drone

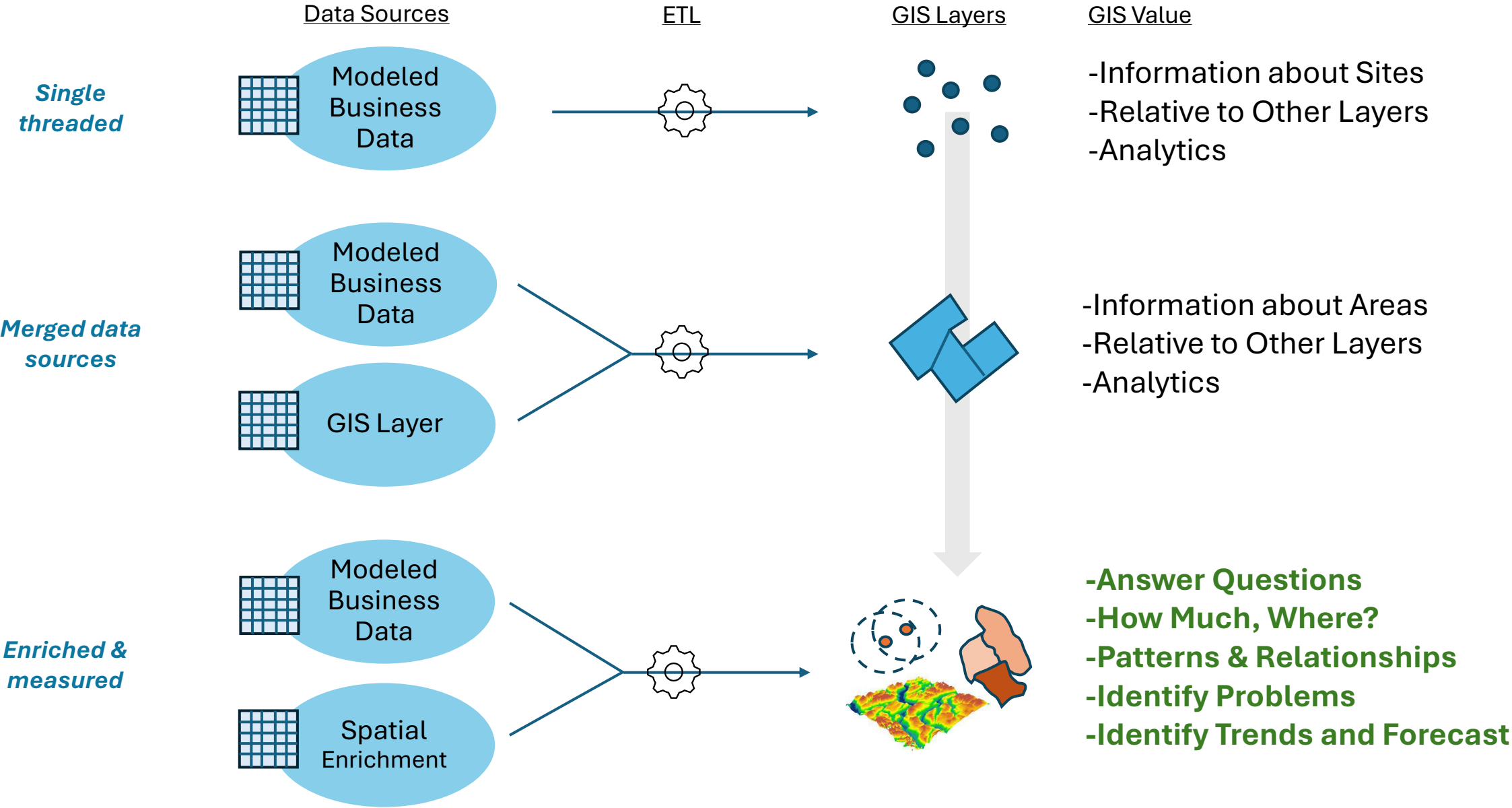


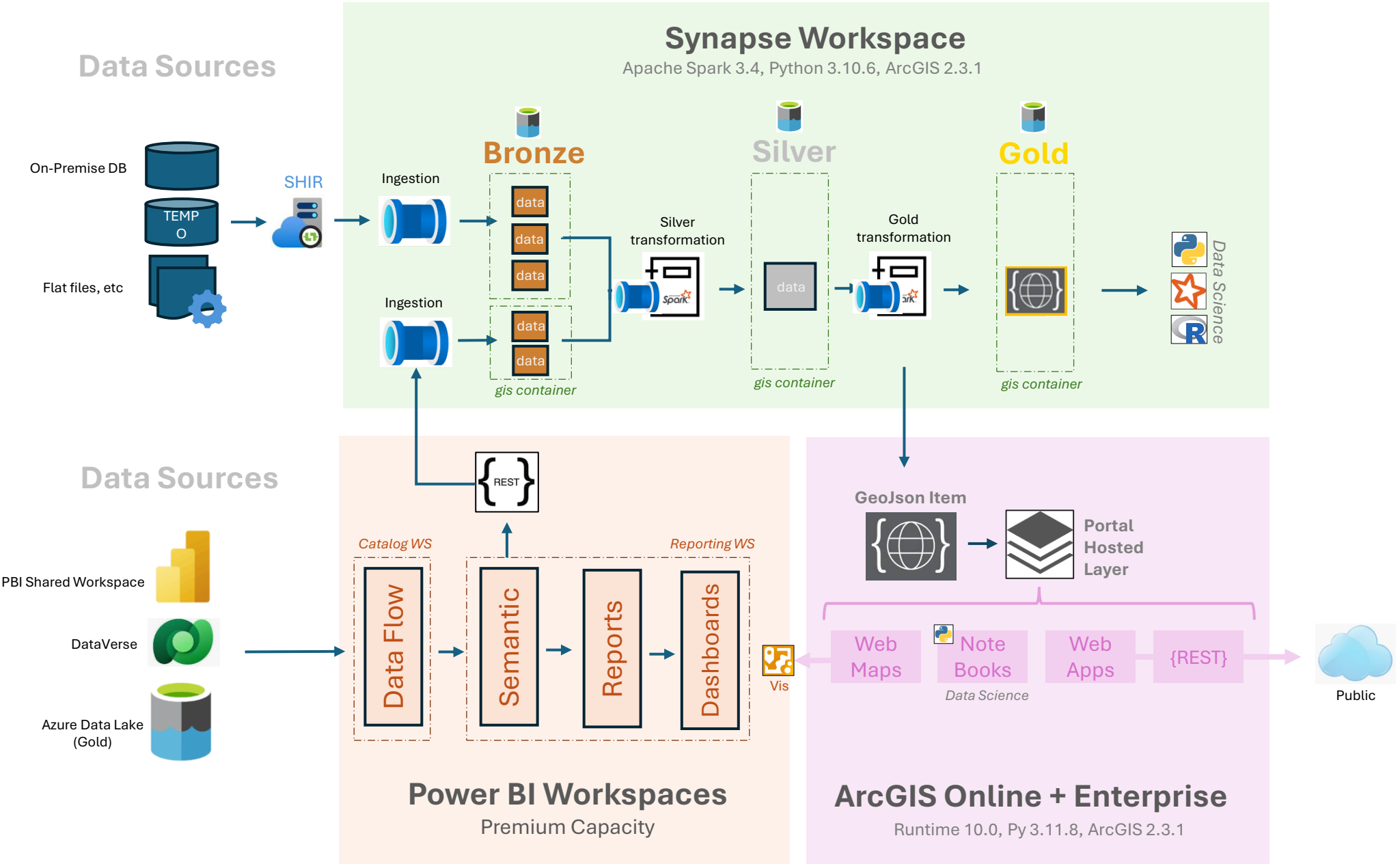
Derived/
Analytic



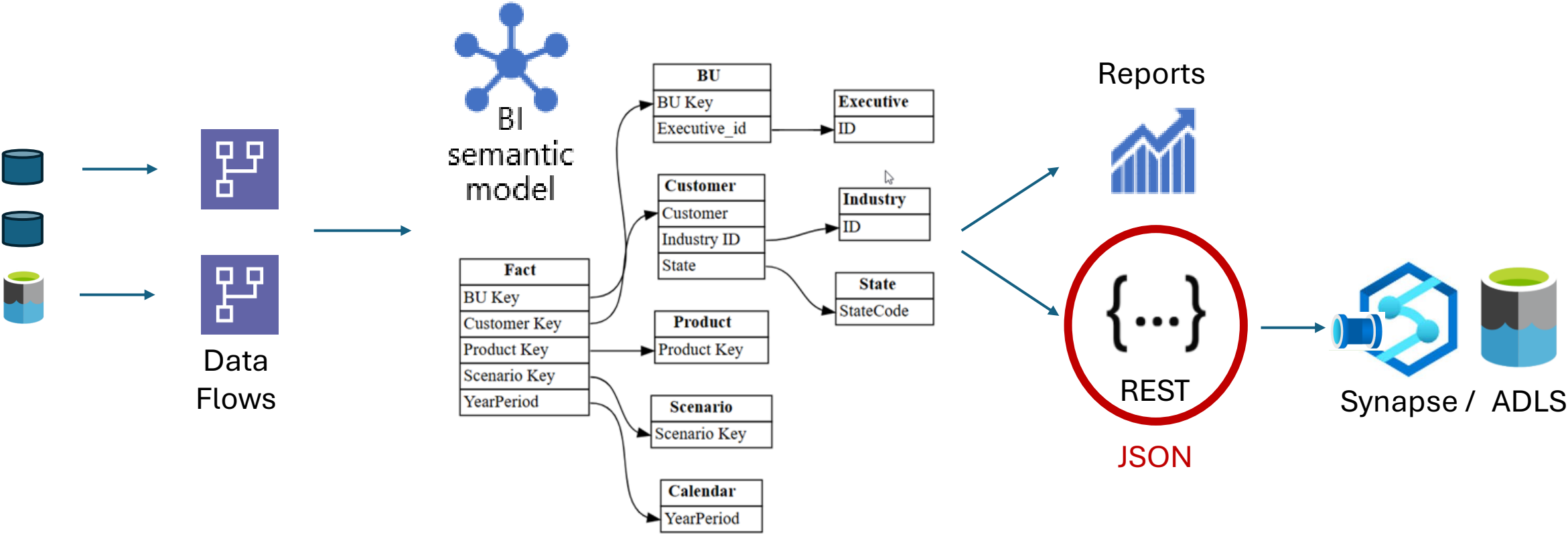
A Better GIS Layer Refresh (from business systems)



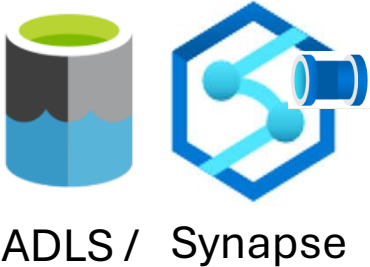




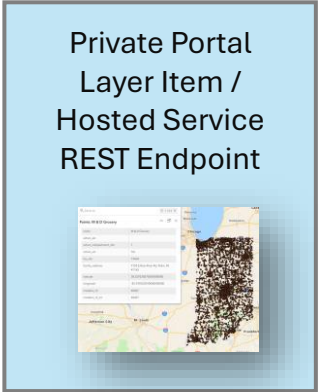
 Power BI Service Semantic Model and API
(for ingestion)



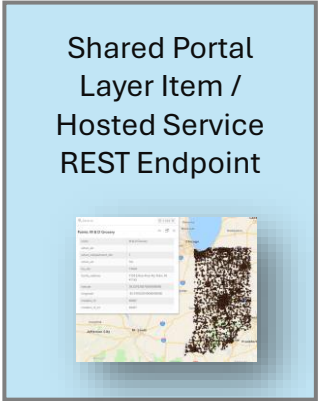
 ArcGIS Layer and API
(for ingestion and publishing)



User-Managed GIS Layer Item



Authoritative GIS Layer Item

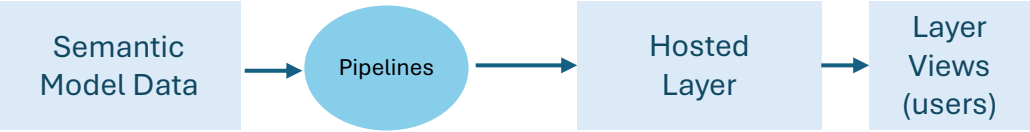


**Authoritative Layer Views
For GIS User and Maps**

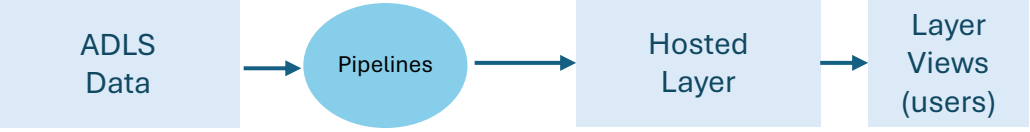


- Where records...
- Fields[]
- Area
- Groups or Public

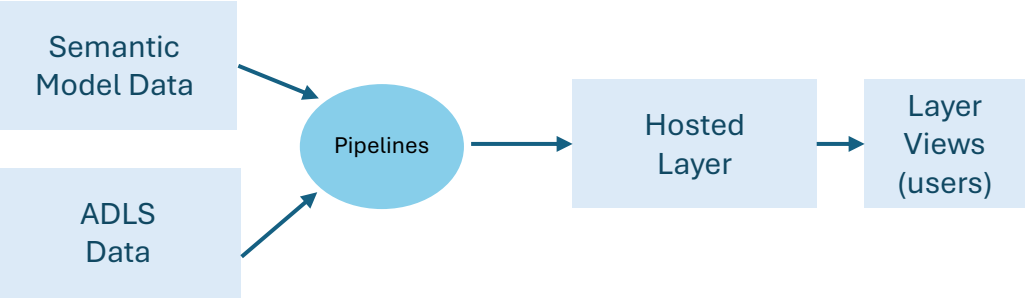
PowerBI-Modeled Data with Coordinates



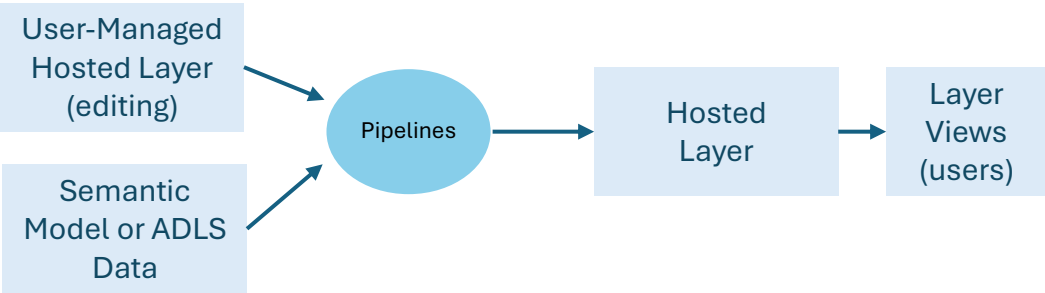
ADLS Data with Coordinates or Geometry



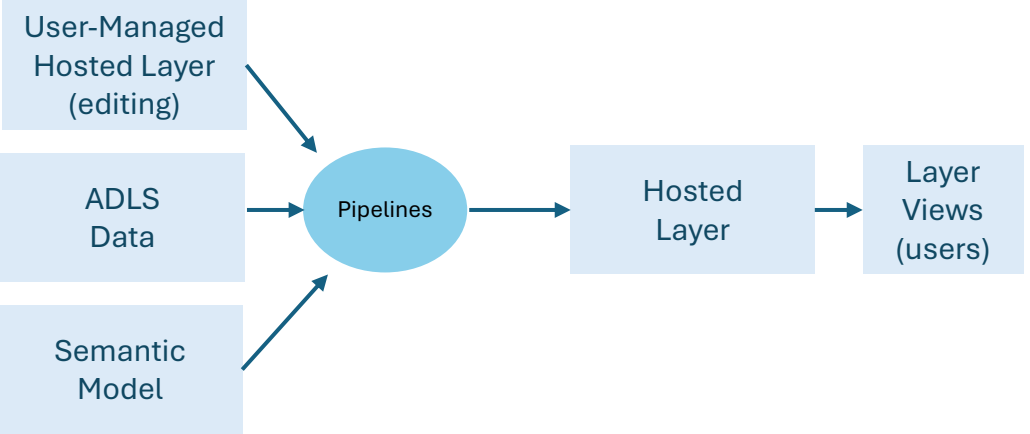
PowerBI-Modeled Data and/or ADLS Data



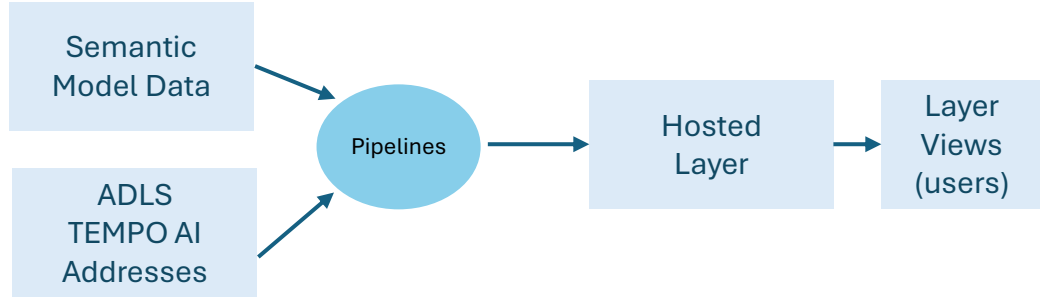
Merging and Enrichment of GIS-Managed Data



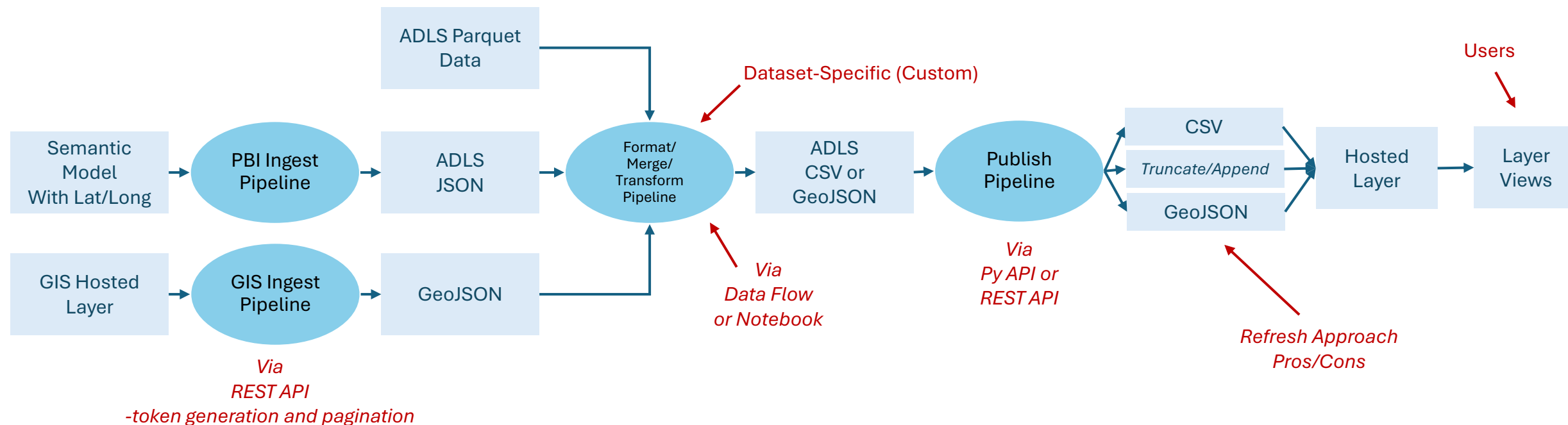
Multiple Data Sources



PowerBI-Modeled Data without Coordinates



Common Composite GeoSync Pattern



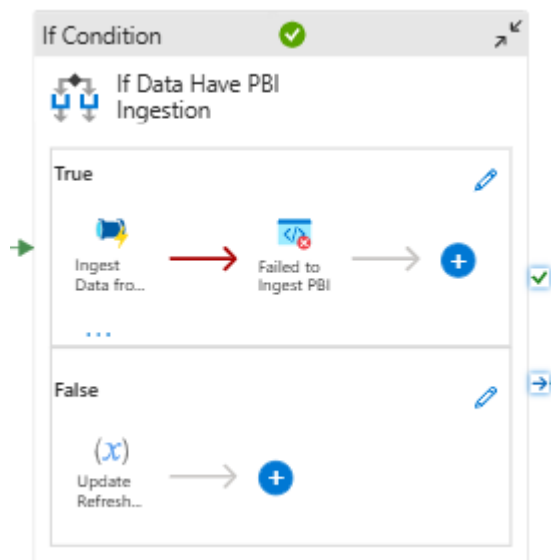
Current Limitations (without Customization)

- 1 semantic model table source
- 1 hosted layer source
- 1 hosted layer with 1 sub layer published

GeoSync: Orchestration of Pipeline Modularization

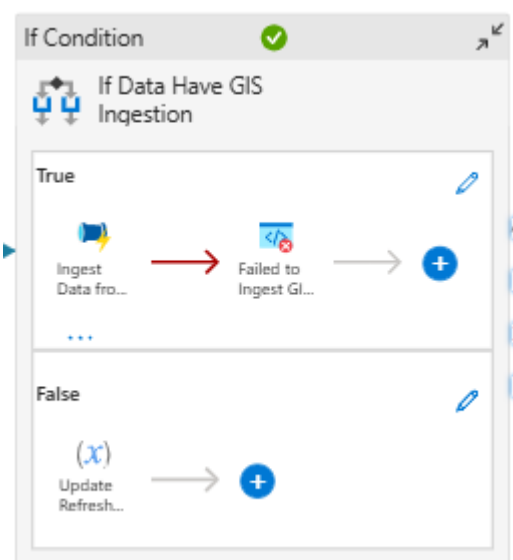
(for each data solution)

Ingest PowerBI Semantic Model

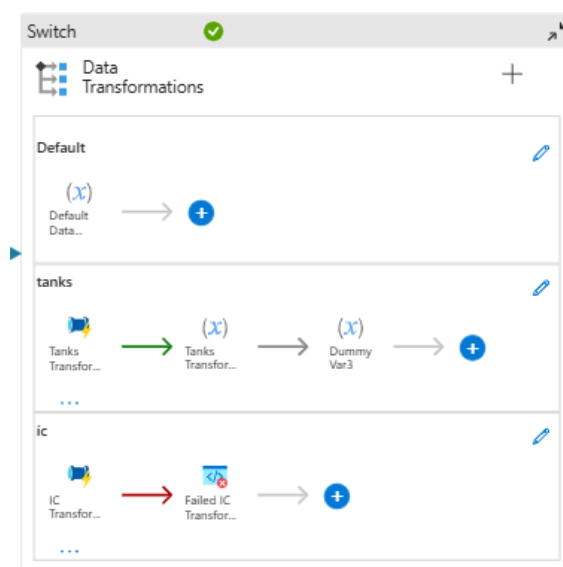


- Pagination
- Nested Json
- Bronze Storage

Ingest GIS Hosted Layer

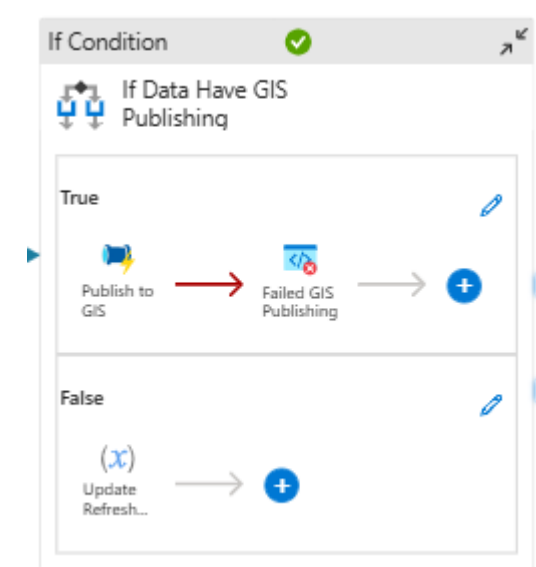


Transformations (Dataset-Specific)



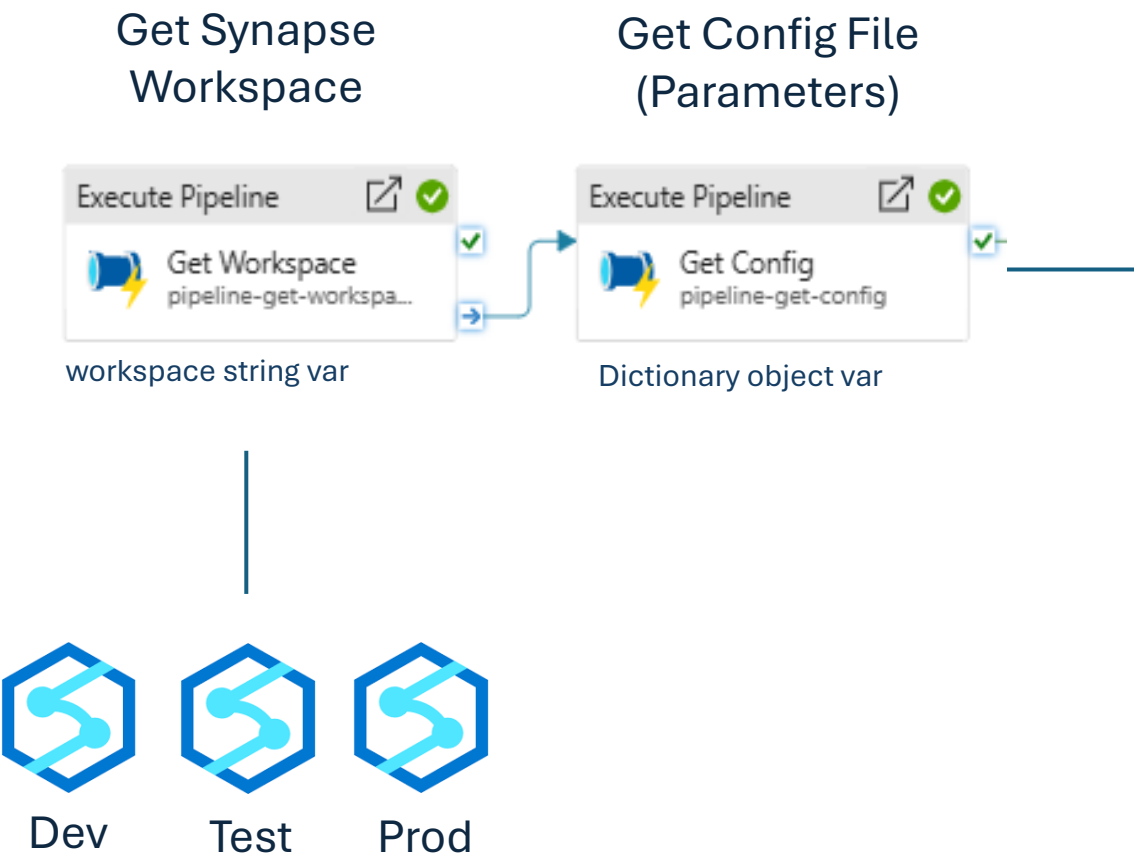
- Field Mappings
- Merging/Enrichment
- Custom Analysis
- Silver Storage

Publish to GIS Hosted Layer



- Publication Methods
- Gold Storage

GeoSync: Parameterization of Pipeline Modules



[dataset1.json, dataset2.json,...]

```
"has_pbi_ingestion":true,
"has_gis_ingestion":true,
"has_gis_anonymous":false,
"has_transformation":true,
"has_gis_publishing":true,
"has_null_geometry":false,
"pbi_source_group":"9670cc09-e457-43a4-a301-6d2f236faf52",
"pbi_source_dataset":"20d60760-832a-457f-b0b5-de5ab3910d90",
"pbi_source_query":"accounts",
"pbi_sink_folder":"elocator",
"pbi_sink_file":"staffing.json",
"gis_source_base":"https://gisdata.in.gov/server/rest/services/...",
"gis_source_guid":"bb1475eaf38b46fd8bf606156ab121ad",
"gis_source_layerid":"9902",
"gis_source_query_where":"1=1",
"gis_source_query_fields":"*",
"gis_source_query_format":"geojson",
"gis_source_query_returngeometry":true,
"gis_source_query_returncentroid":false,
"gis_source_query_string":"","
.
.
.
```

Observability of Pipelines

GIS Server Testing/Coordination
Needed for Parallelism

2- Each orchestration is run sequentially

Name	Value
pipelineRunId	63a840ce-1119-4f94-9052-...
pipelineName	pipeline-orchestration

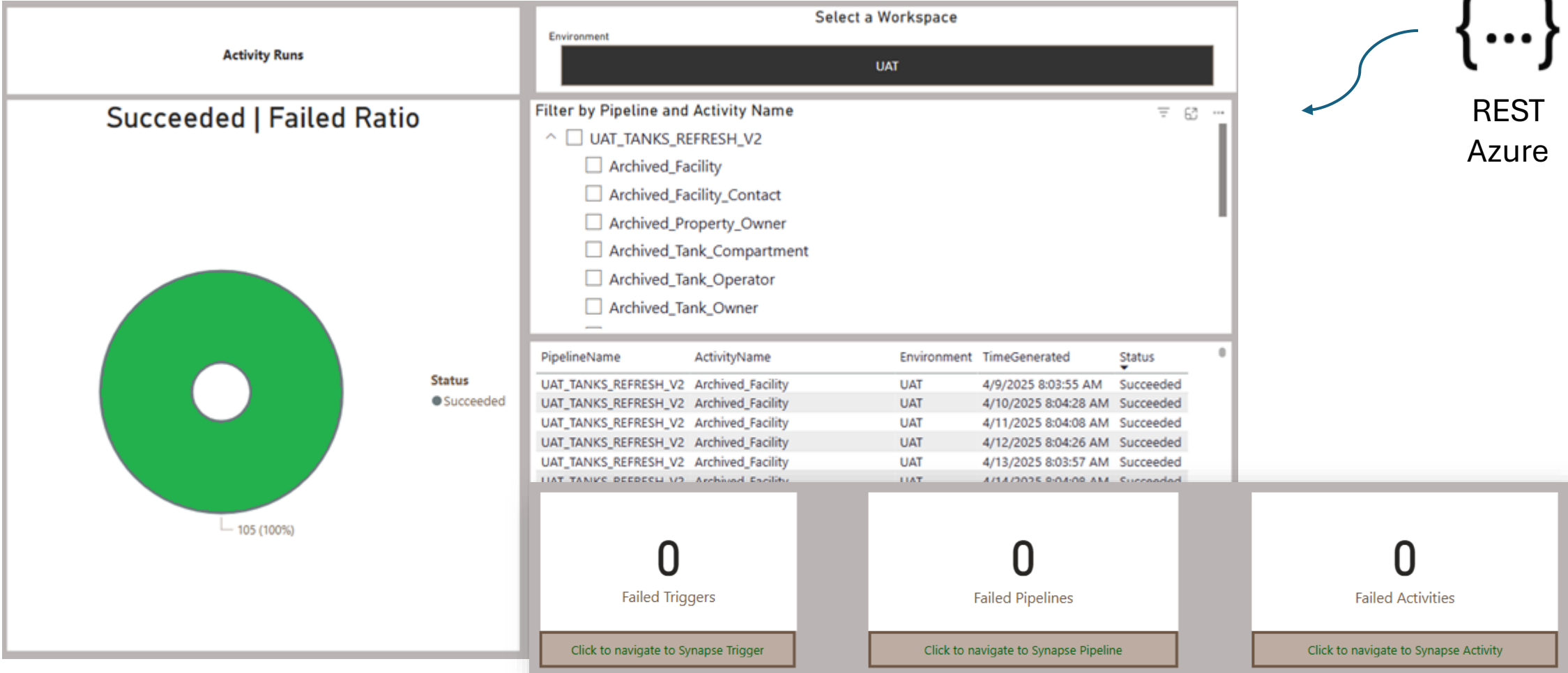
1 - Hub/Iterator

Activity name	↑↓	Activity sta...	↑↓	Activity type	↑↓
Set variable2		✔ Succeeded		Set variable	
GIS Refresh Pipeline		✔ Succeeded		Execute Pipeline	
Set variable2		✔ Succeeded		Set variable	
GIS Refresh Pipeline		✔ Succeeded		Execute Pipeline	
For Each GIS Refresh Item		✔ Succeeded		ForEach	

3 - Individual Data/Layer Solution

Activity name	↑↓	Activity sta...	↑↓	Activity type	↑↓
Pipeline Return Status for Com...		✔ Succeeded		Set variable	
Publish to GIS		✔ Succeeded		Execute Pipeline	
If Data Have GIS Publishing		✔ Succeeded		If Condition	
IC Tranformation Successful		✔ Succeeded		Set variable	
IC Transformation		✔ Succeeded		Execute Pipeline	
Data Transformations		✔ Succeeded		Switch	
Update Refresh Status for GIS		✔ Succeeded		Set variable	
Ingest Data from GIS		✔ Succeeded		Execute Pipeline	
If Data Have GIS Ingestion		✔ Succeeded		If Condition	
Update Refresh Status for PBI I...		✔ Succeeded		Set variable	
If Data Have PBI Ingestion		✔ Succeeded		If Condition	
Status is In Progress per Valid ...		✔ Succeeded		Set variable	
Config is Valid		✔ Succeeded		Set variable	
Test If Invalid Configs		✔ Succeeded		If Condition	
Get Config		✔ Succeeded		Execute Pipeline	
Get Workspace		✔ Succeeded		Execute Pipeline	

Observability Reports and Alerts in PowerBI



Enhancing GeoSync (future use cases)

- Pipeline Optimization and Parallelism for Scalability
- Geocoding for datasets without spatial data (e.g., addresses only)
- Socio-economic attribute enrichment
- Spatial Aggregation/Summarization/Analysis
- Archiving
- Streaming
- Delta Updates (for larger datasets)

IDEM Data Operations

- Vanessa Williams, vwilliams@idem.in.gov
- Muhammad Azimi, meazimi@idem.in.gov
- Kayla Sturgeon, klsturgeon@idem.in.gov
- Chris Waldron, cjwaldro@idem.in.gov

