Increasing Trustworthiness of Power BI Content in Microsoft Fabric

Melissa Coates

Data Architect | Consultant | Trainer

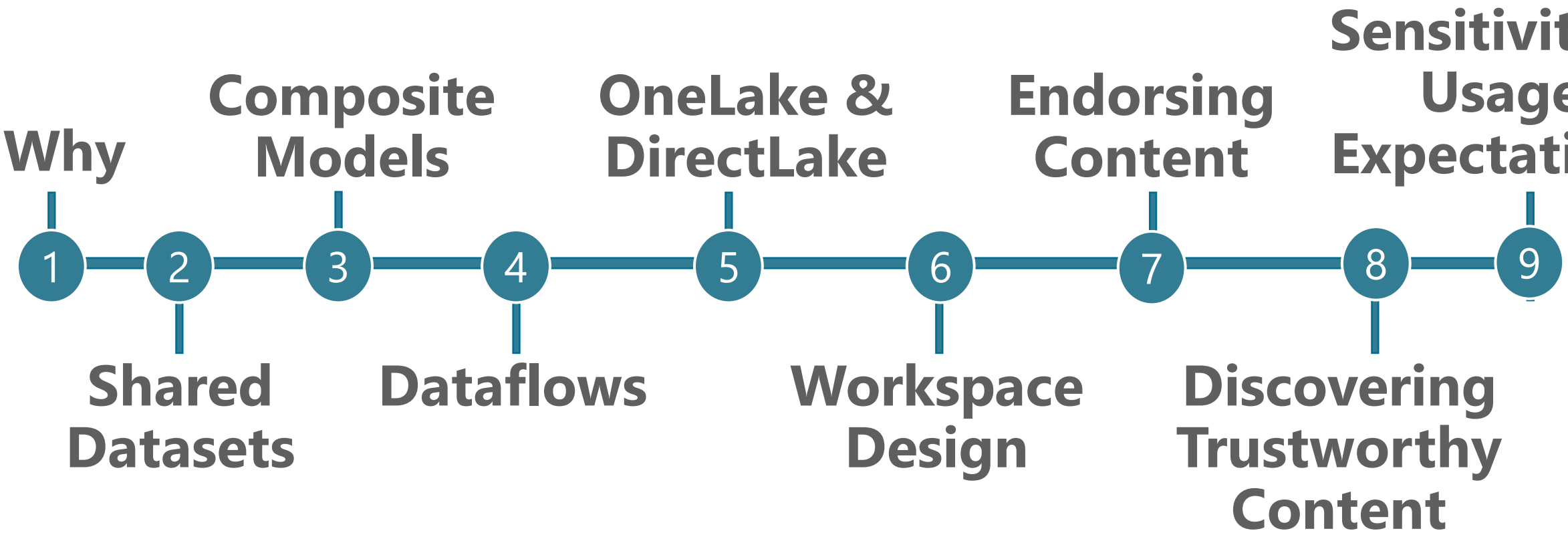
Coates Data Strategies

Slides & recordings: CoatesDS.com/Presentations

Content last updated: June 6, 2023

Increasing Trustworthiness of Power BI Content

Agenda:



Melissa Coates

Data architect specializing in Power BI governance & administration

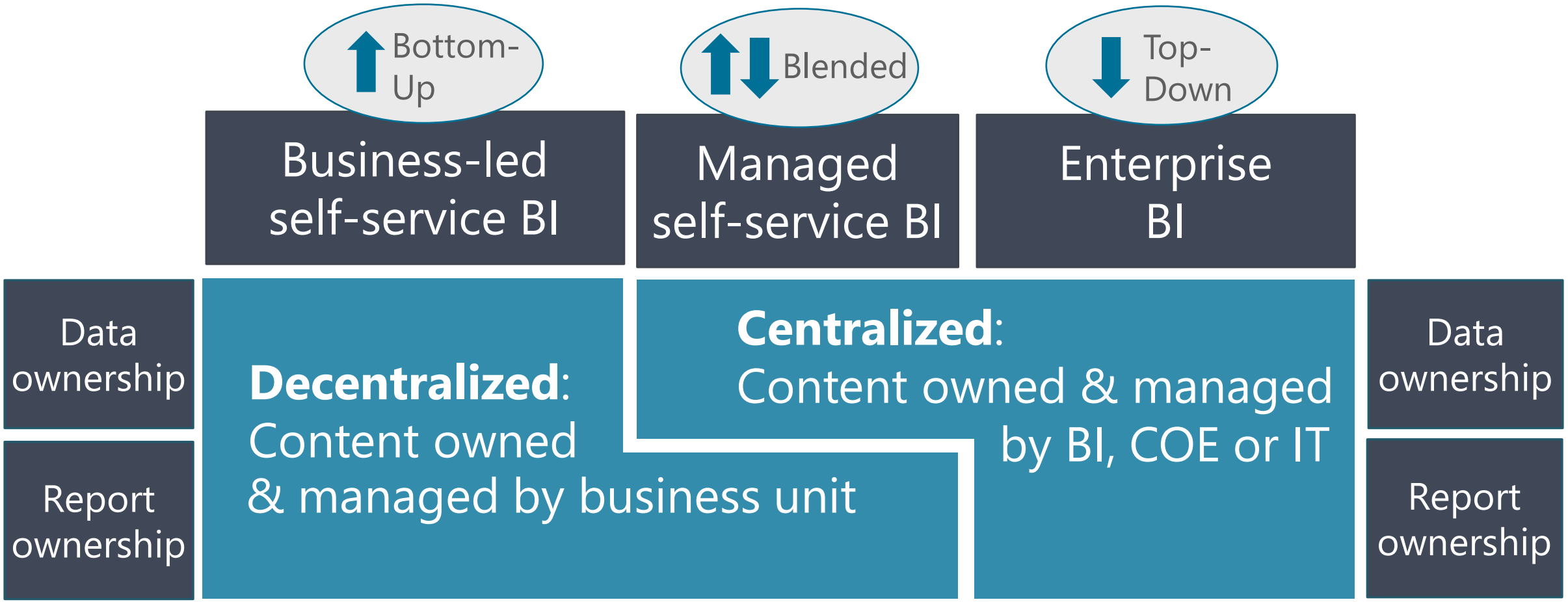
Author of Microsoft guidance:
Power BI Adoption Roadmap

Power BI Implementation Planning

Original creator of the Power BI Deployment & Governance training course (now run by Mike Carlo)

Why Are We Concerned with Trustworthiness?

Business Intelligence Approaches

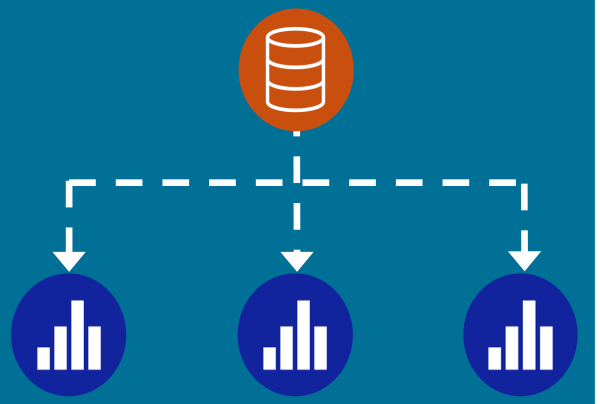


For more info, see Power BI Adoption Roadmap:
[Content ownership and management]

What Do We Mean by Trustworthy Content?

Considerations such as:

- Content quality is assured
- Approved, common definitions and KPIs are used
- Owner, SME, and steward responsibilities are clear
- Data provenance & lineage are clear
- Data duplication is minimized
- Data & load processes are managed & secured
- Awareness of content which is discoverable & findable
- Clear expectations about responsible data use

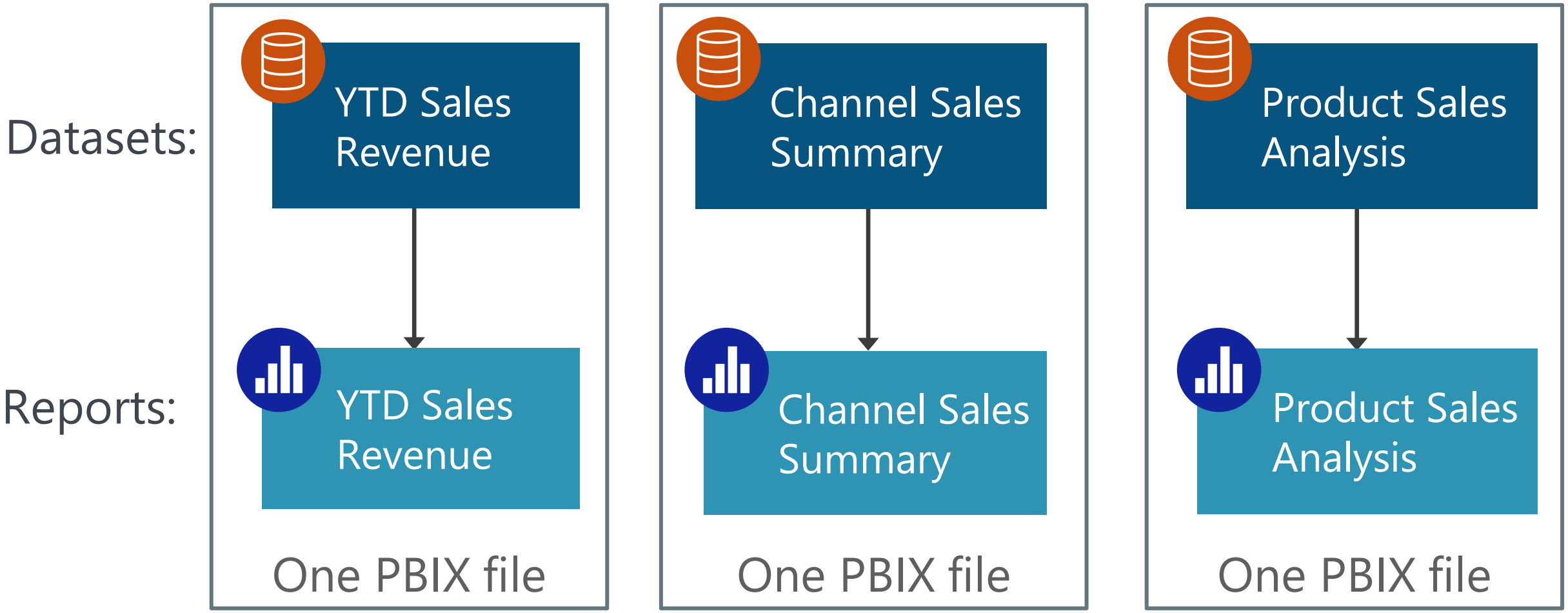


Shared Datasets

Achieving Data Model Reuse in Power BI

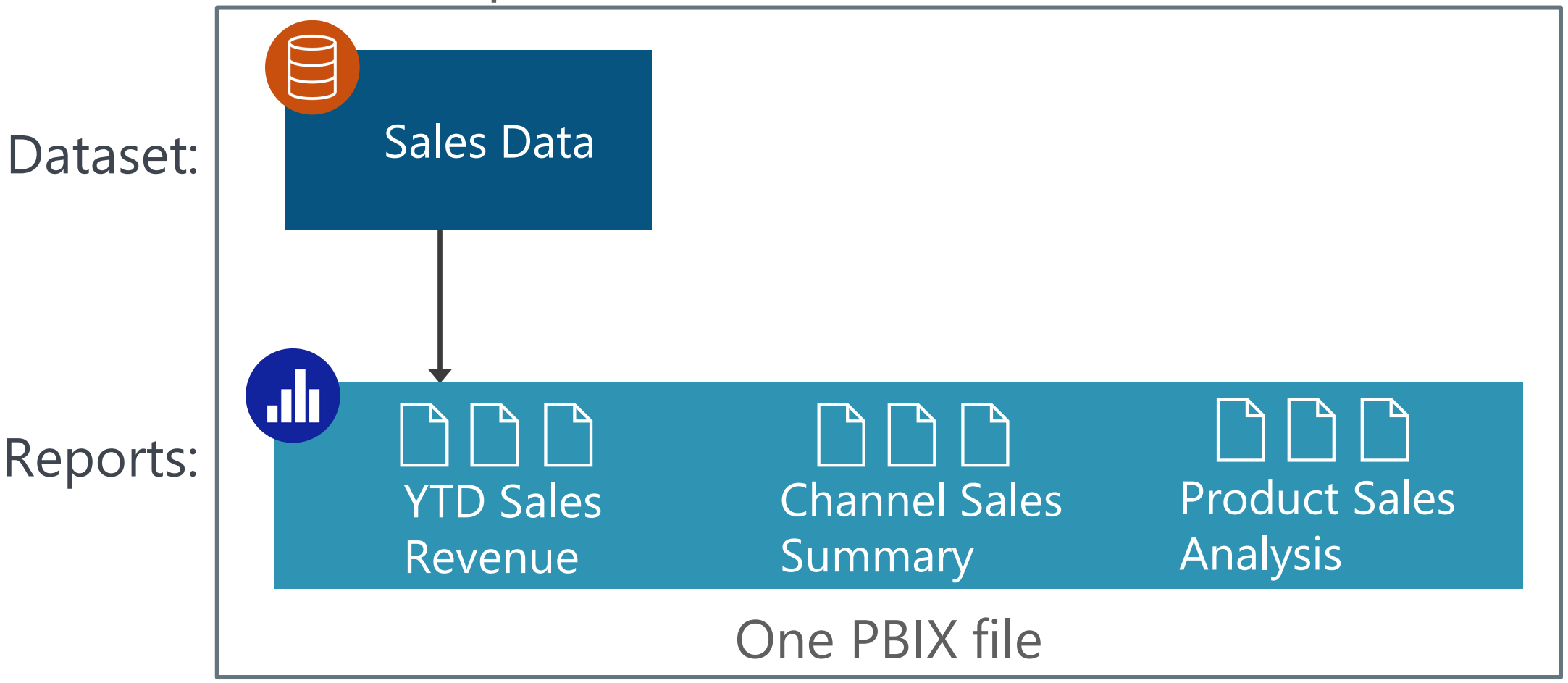
Where A Lot of Authors Start

In Power BI Desktop:

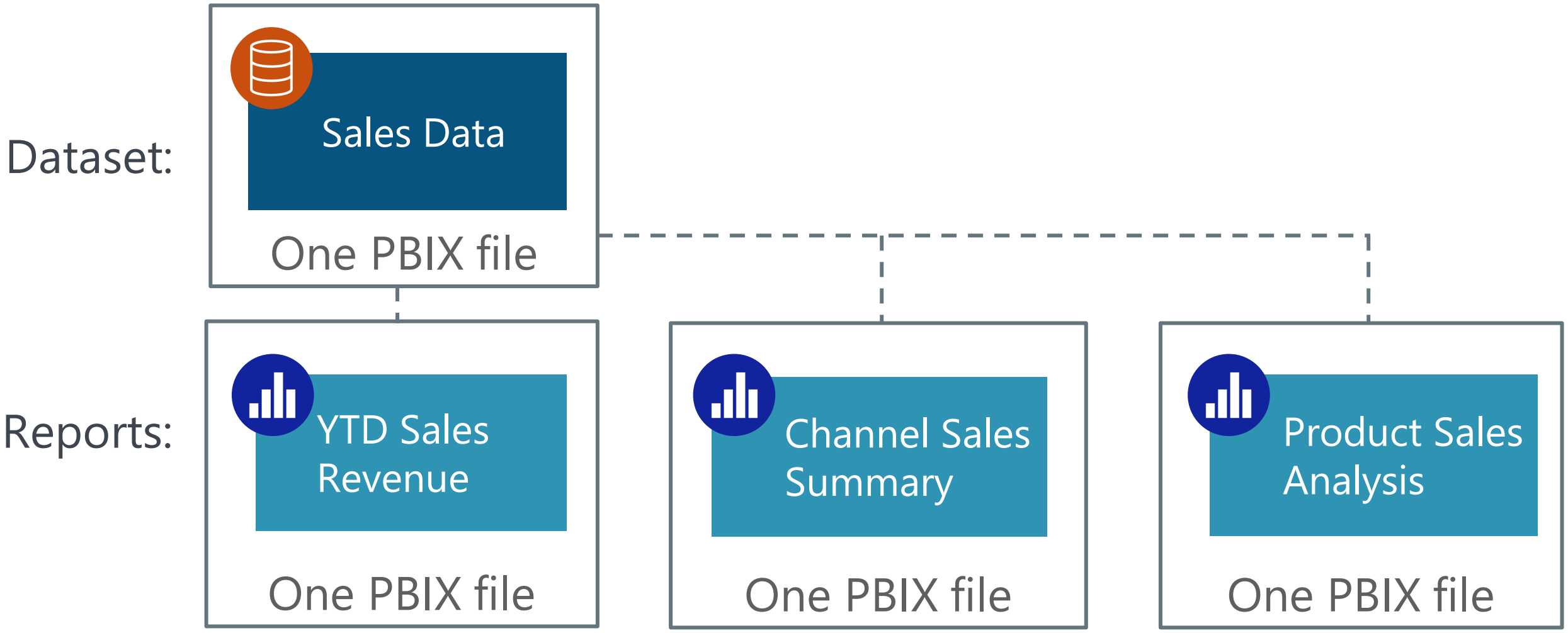


Where A Lot of Authors Start

In Power BI Desktop:

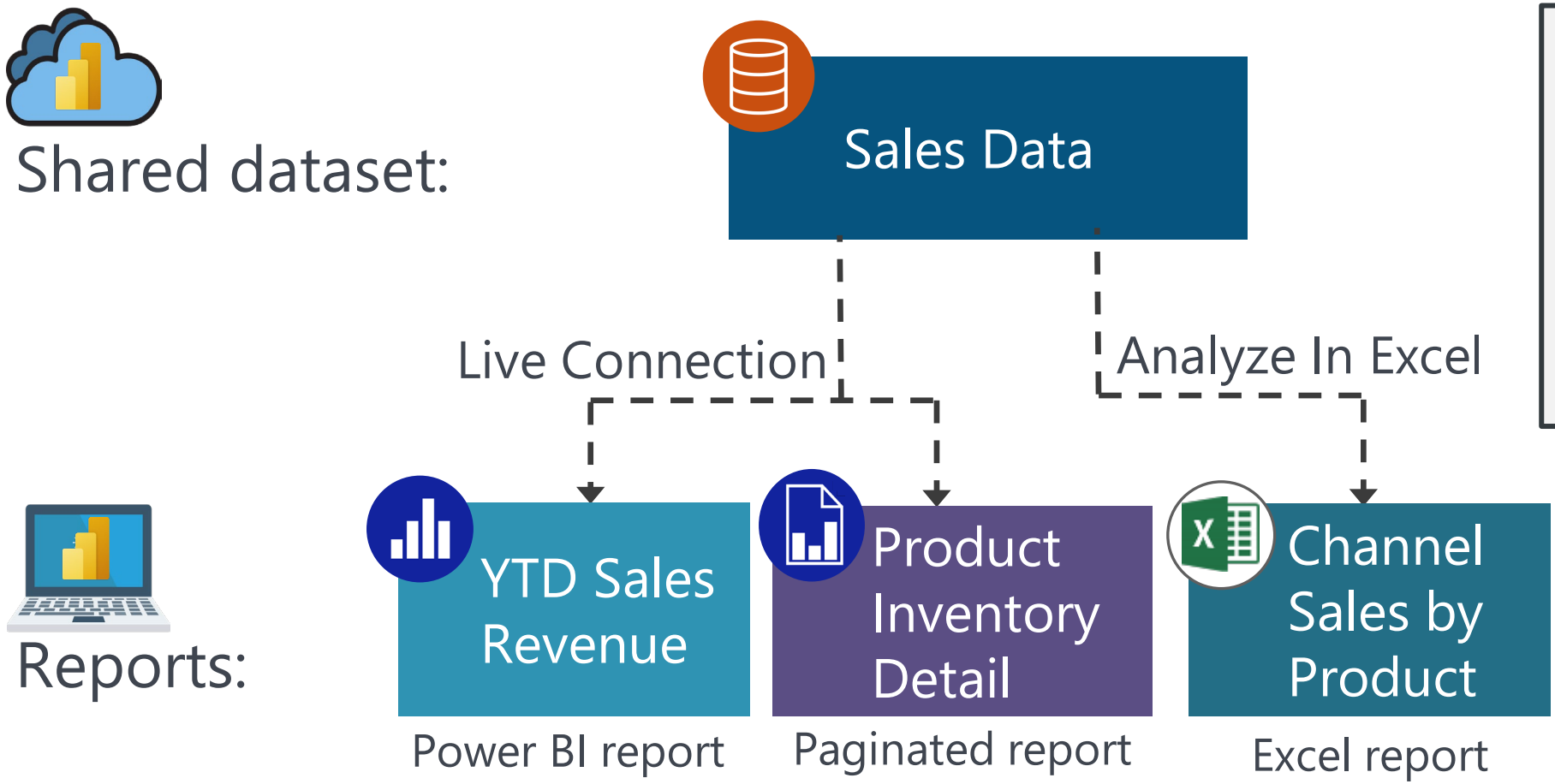


The Goal: Decoupling Data & Reports



Decoupling Data & Reports

A **shared dataset** is intended for reuse across multiple reports



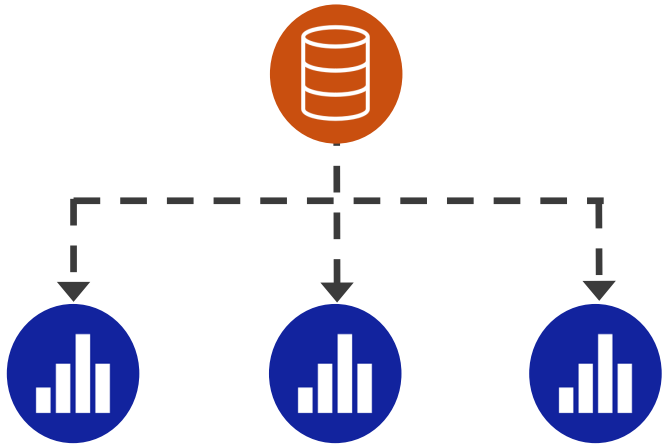
Also known as:

Golden dataset & thin reports

Hub & spoke method

More info: CoatesDS.com/blog/5-tips-for-separating-power-bi-datasets-and-reports

Key Advantages of Shared Datasets



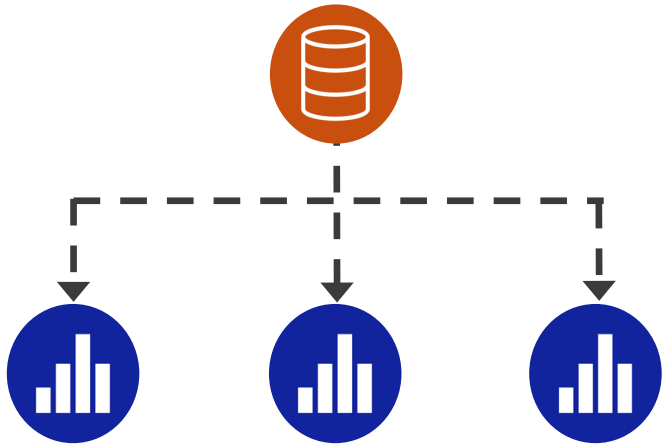
☑ Self-Service BI Enablement

- Different people can handle data modeling and report creation
- Promotes discipline at the core & flexibility at the edge

☑ Follows the Fabric model

- Decoupled datasets follows with how Fabric artifacts get created

Key Advantages of Shared Datasets



☑ Reduced Risk

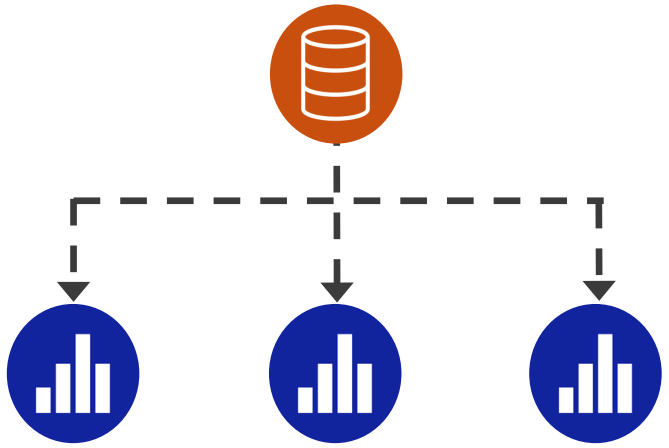
Fewer datasets results in:

- Fewer inconsistencies
- Reduced maintenance & change effort
- Reduced level of data validations

Less duplication of data results in:

- Lessened governance compliance concerns

Key Advantages of Shared Datasets



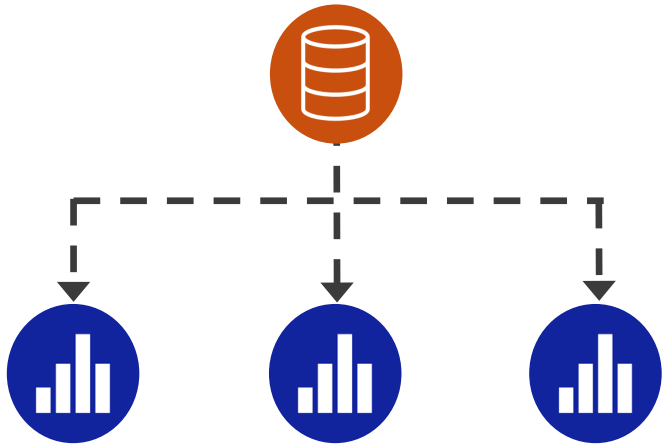
☑ Improved Usability for Report Consumers

- Reports can each address individual use cases
- Reports contain only as many report pages as are needed

☑ Performance

- Fewer dataset refreshes need to run

Key Advantages of Shared Datasets



☑ Additional Security Options

- Separate PBIX files may be deployed to different workspaces (which may be secured differently)
- Separate reports may be shared individually

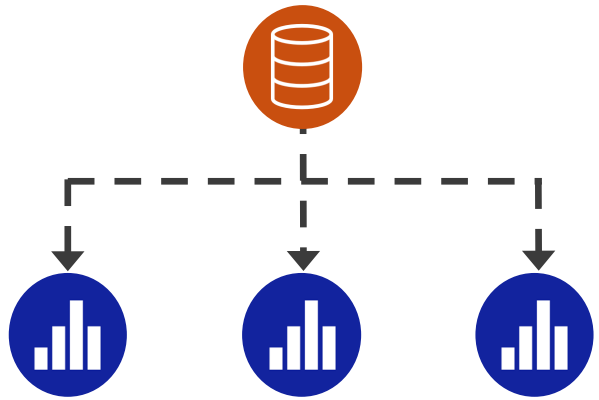
Key Disadvantages of Shared Datasets

☒ Additional oversight

- Favors process over speed
- Should have additional oversight & change control, which isn't desirable for every single dataset

☒ Ownership & responsibility

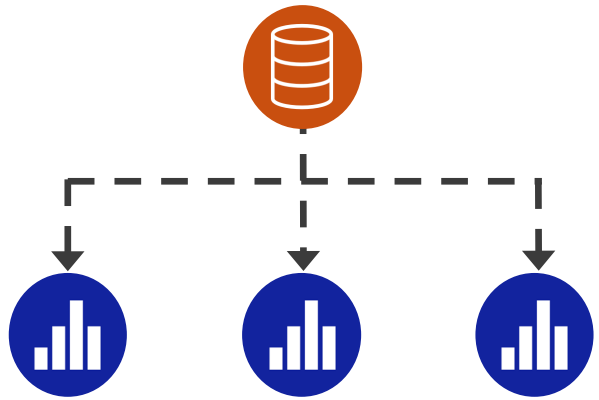
- Changes may be decided by another team who have different priorities



Watch Out For With Shared Datasets

☒ A habit change

- Involves a habit change for dataset authors
- Not an intuitive way to work for brand new self-service authors – need to teach them this concept fairly early in their journey

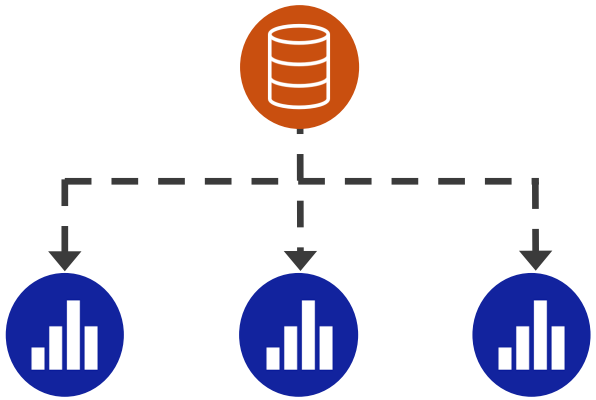


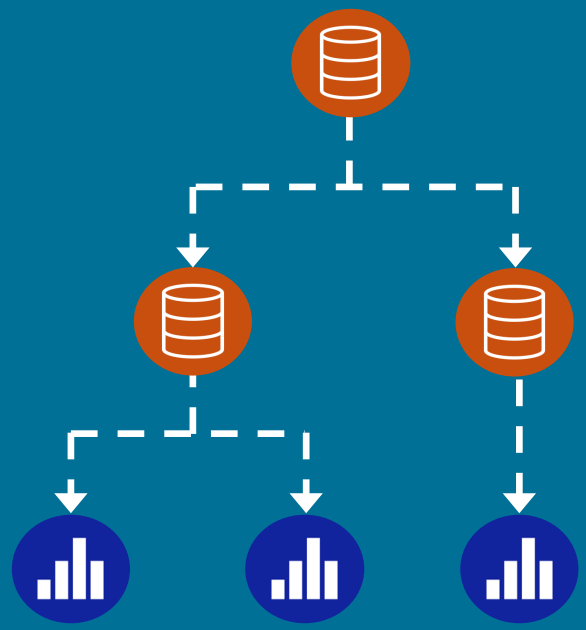
Watch Out For With Shared Datasets

☒ **Complacency, leading to inadequate report testing**

No guarantees that there aren't report errors:

- Report visuals can still be misleading
- Report-level measures can still be incorrect
- Composite models allow extending & customizing the original model, which can introduce confusion or issues

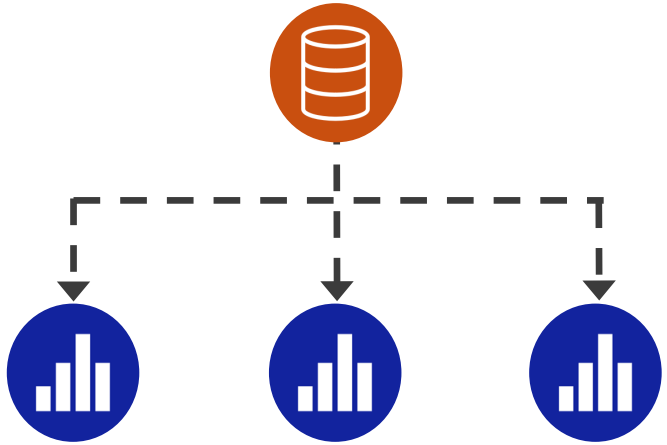




Composite Models

Bridging Enterprise BI and Self-Service BI

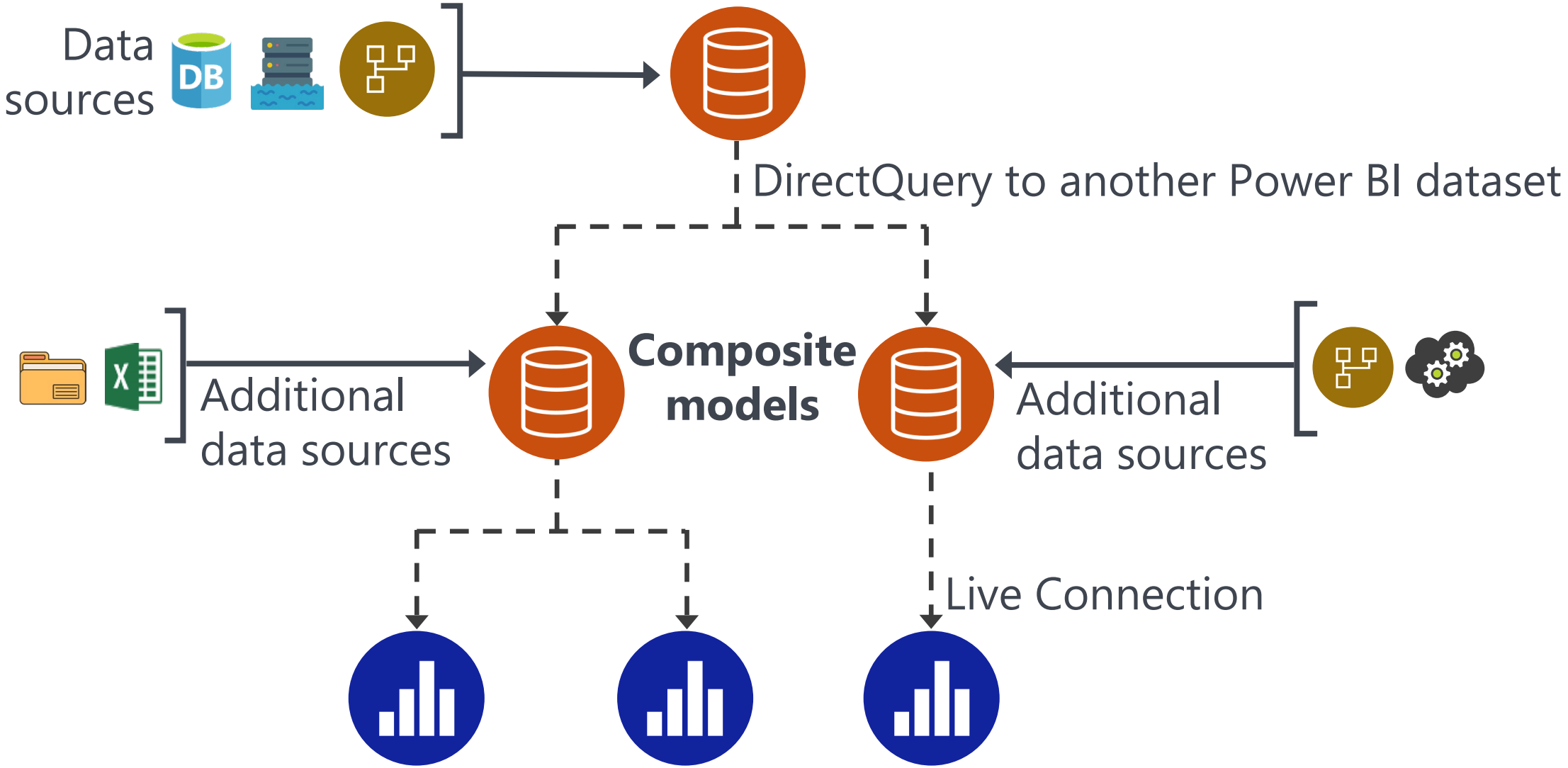
What If a Shared Dataset Doesn't Have All the Data That's Needed?



Live connection:

Requires all of the data to be in the data source

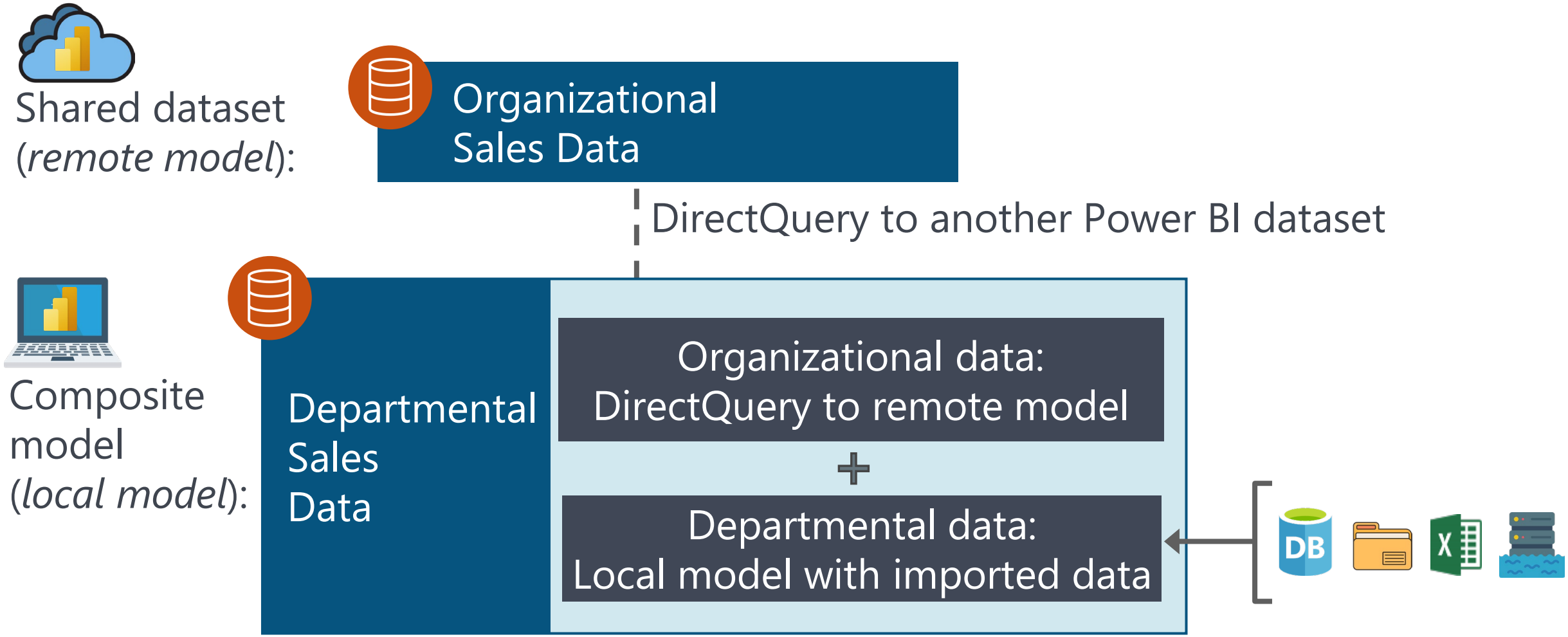
Taking Things a Step Further



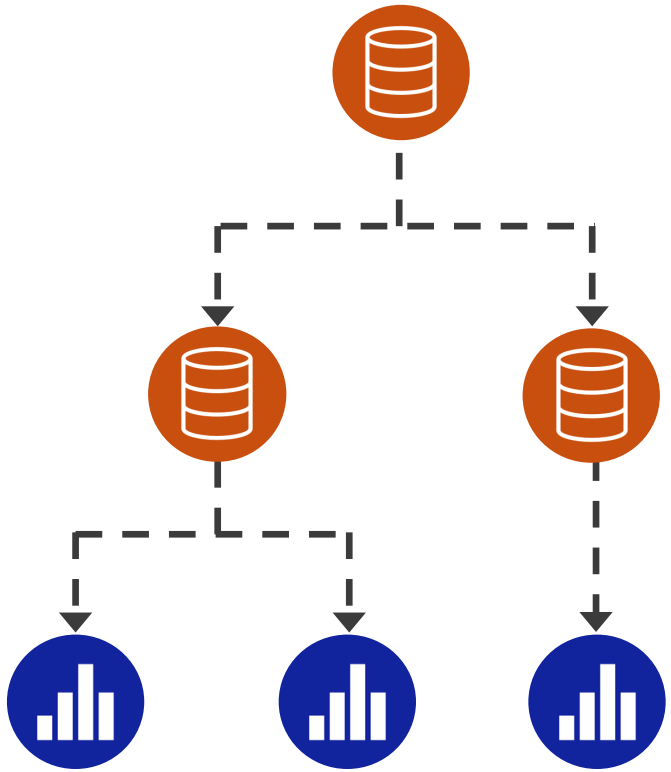
For more info, see the General Availability announcement

Extending & Customizing a Shared Dataset

A **composite model** combines multiple connection types



Key Advantages of Composite Models



☑ Self-Service BI Enablement

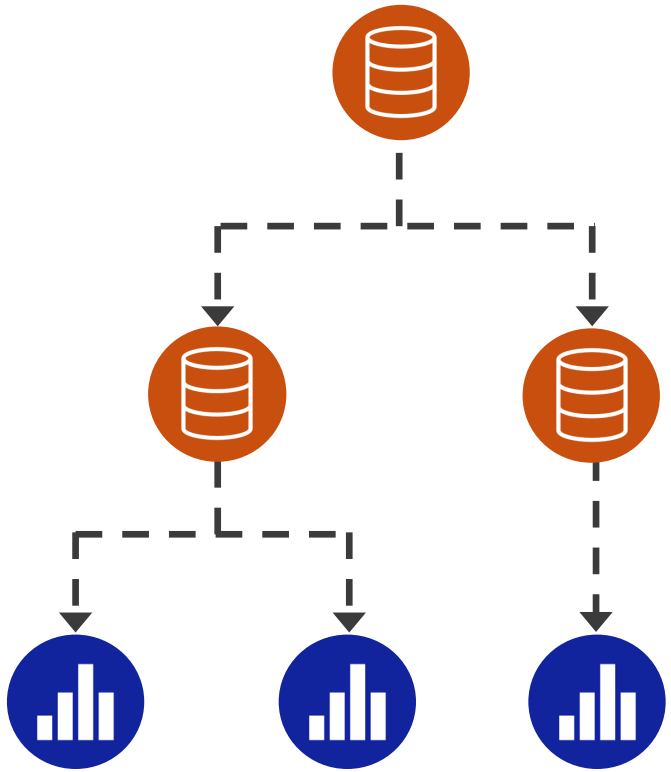
Extend a centralized data model:

- New data sources, tables, relationships
- New measures & calculated columns
- New hierarchies

Personalize a centralized data model:

- Reformatting & renaming

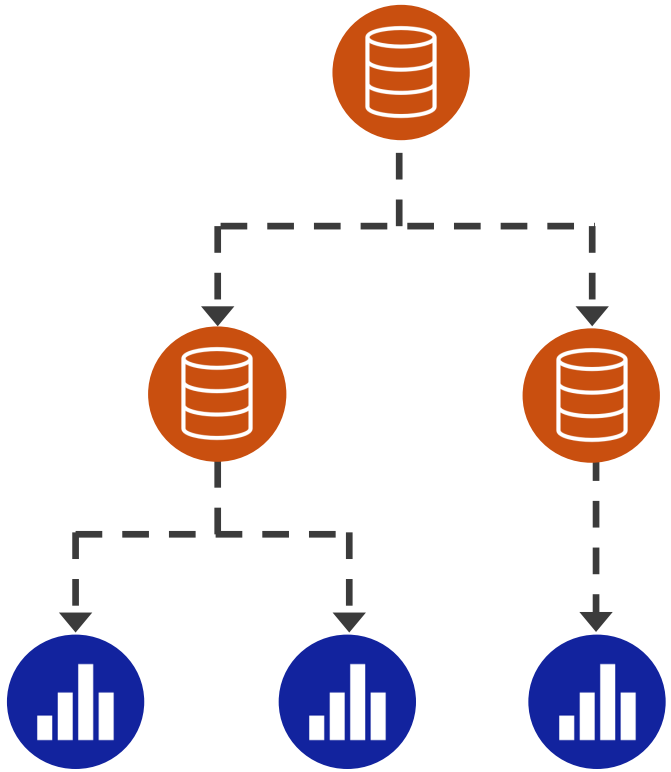
Key Advantages of Composite Models



Supports Enterprise BI + Self-Service BI

- Different people managing different datasets
- Promotes discipline at the core & flexibility at the edge
- Less duplication of effort for dataset authors

Key Advantages of Composite Models



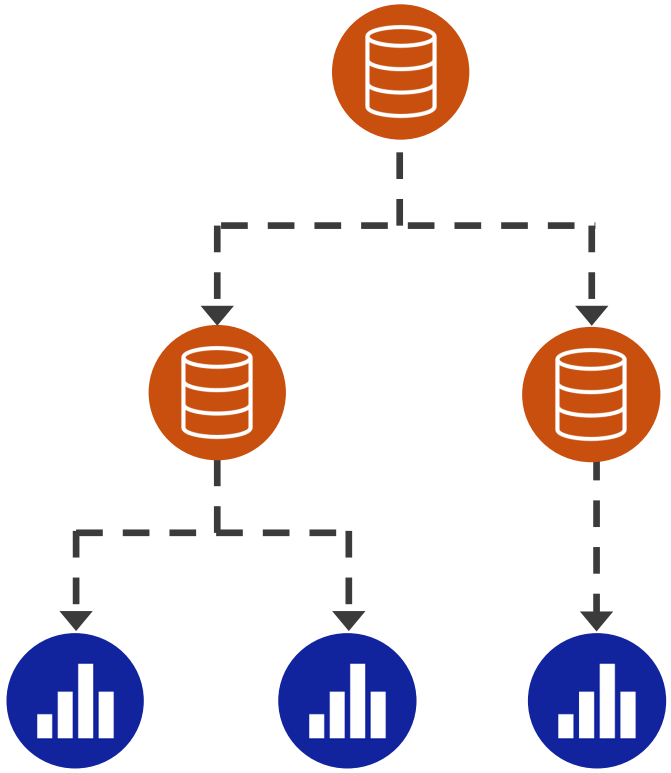
☑ Reduced Risk

- Reuse of existing data without duplicating it again
- Fewer inconsistencies
- Lessened governance compliance concerns

Key Disadvantages of Composite Models

☒ Performance

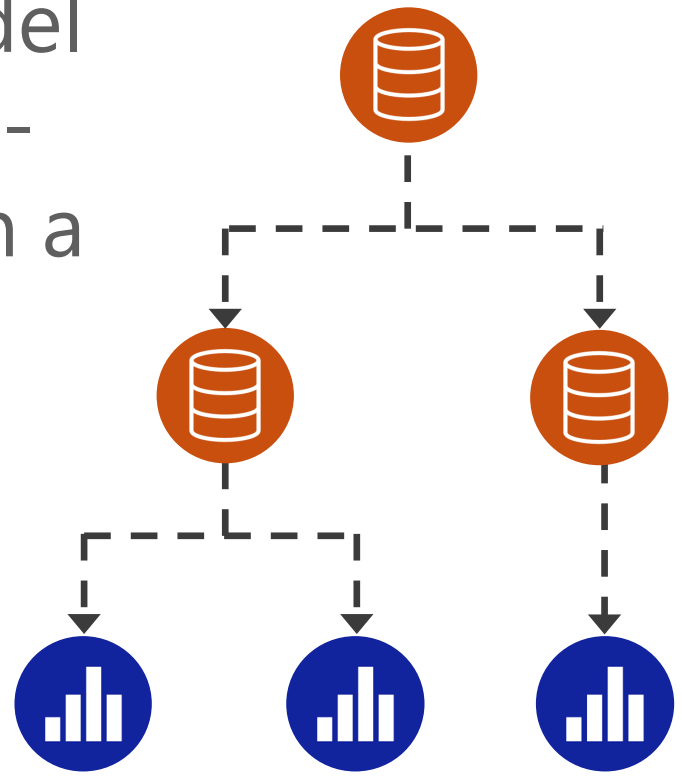
- Querying a remote model will never perform as well



Key Disadvantages of Composite Models

☒ Unexpected Results

- You might not know what the original author had intended for calculations
- DAX formulas created in the remote model might rely on context or other columns -- yielding unexpected results when used in a different way



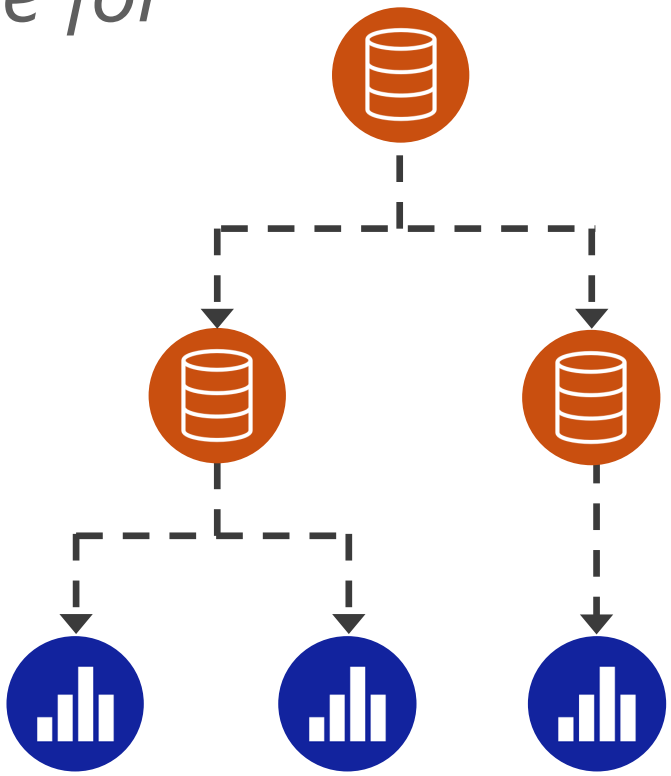
Key Disadvantages of Composite Models

☒ Permissions Requirement *(temporary issue)*

- Requires 'build' permission on all datasets in the chain for all ***consumers*** of the downstream reports.
  (Resolved for Premium/PPU -- still an issue for some Pro workspaces)

☒ Row-Level Security

- RLS needs to be defined on the remote model first (cannot propagate RLS from the local model)



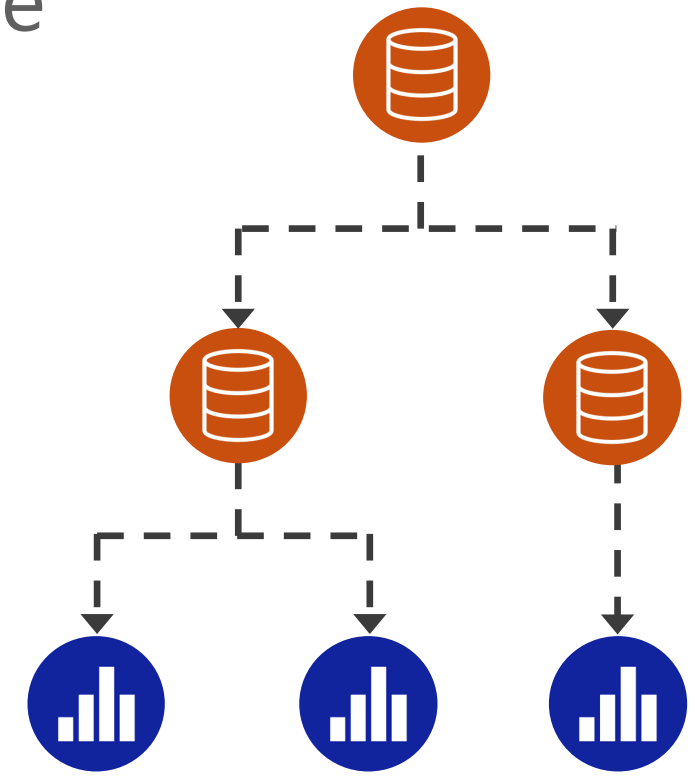
Key Disadvantages of Composite Models

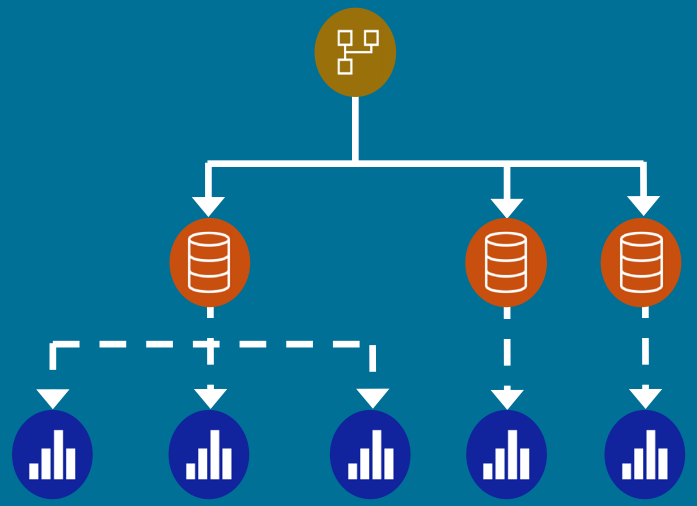
☒ Expectations for Accuracy

- Unreasonable expectations that downstream content is accurate
- Perceived quality of original model can be reduced if a lot of downstream issues

☒ Additional Layers of Technical Debt

- Sub-optimal downstream models

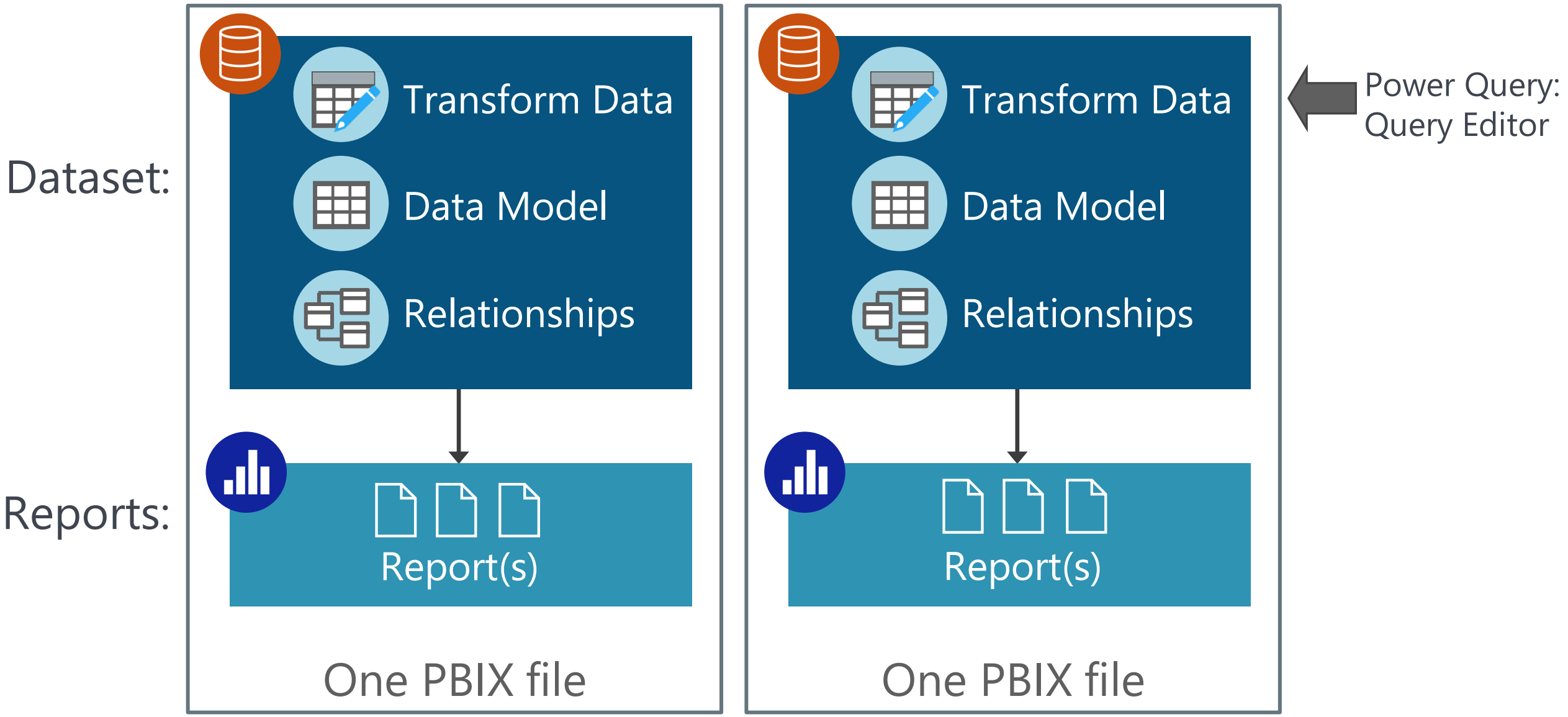




Dataflows

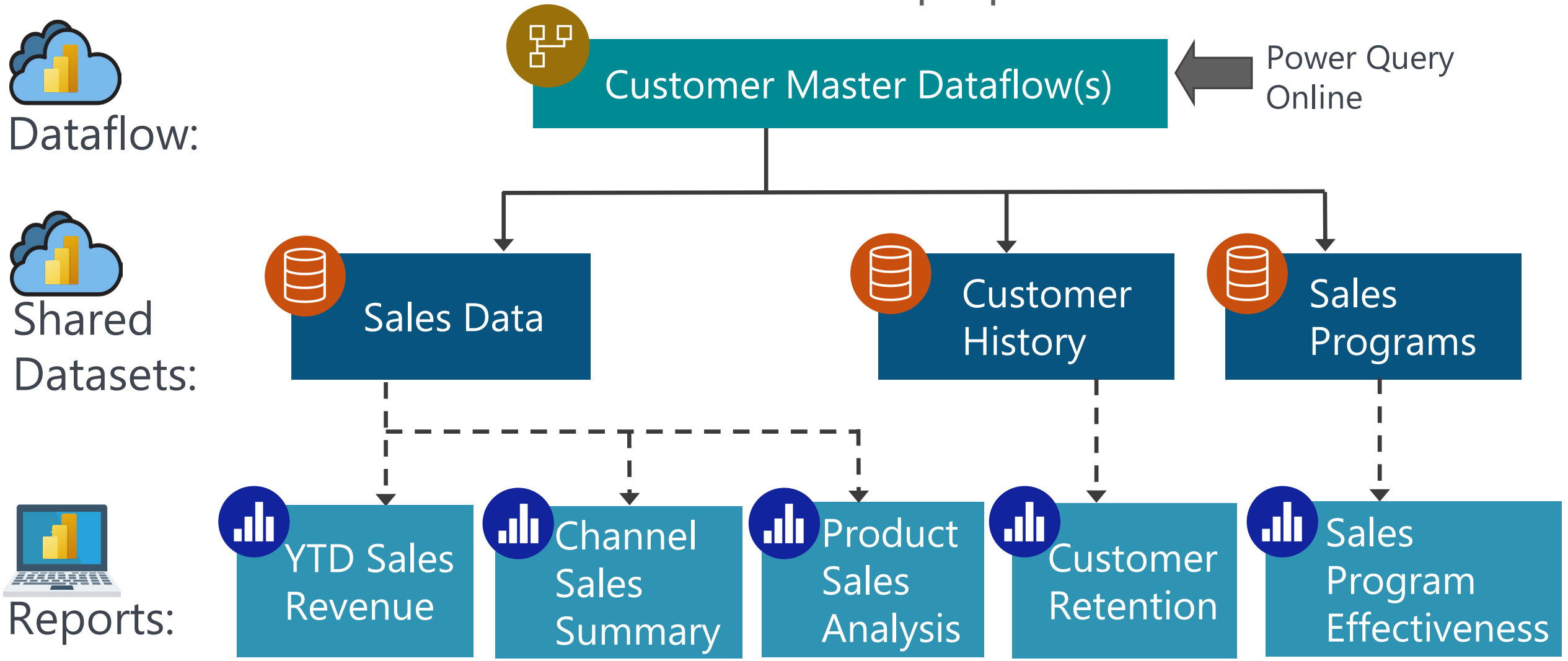
Achieving Data Preparation Reuse in Power BI

What if There's a Lot of Data Preparation?

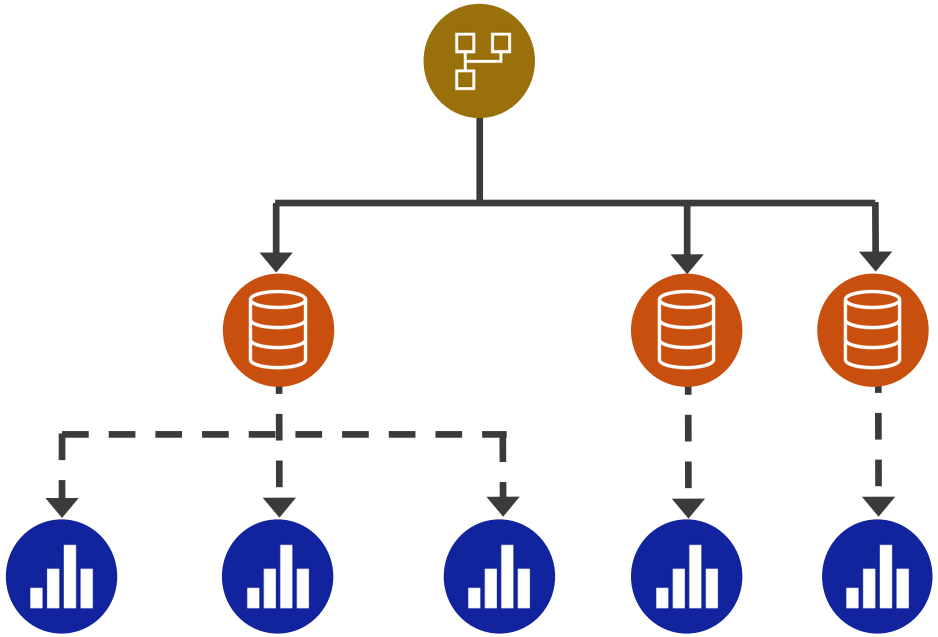


Decoupling Data Preparation from the Model

Dataflows are intended for reuse of data preparation across datasets



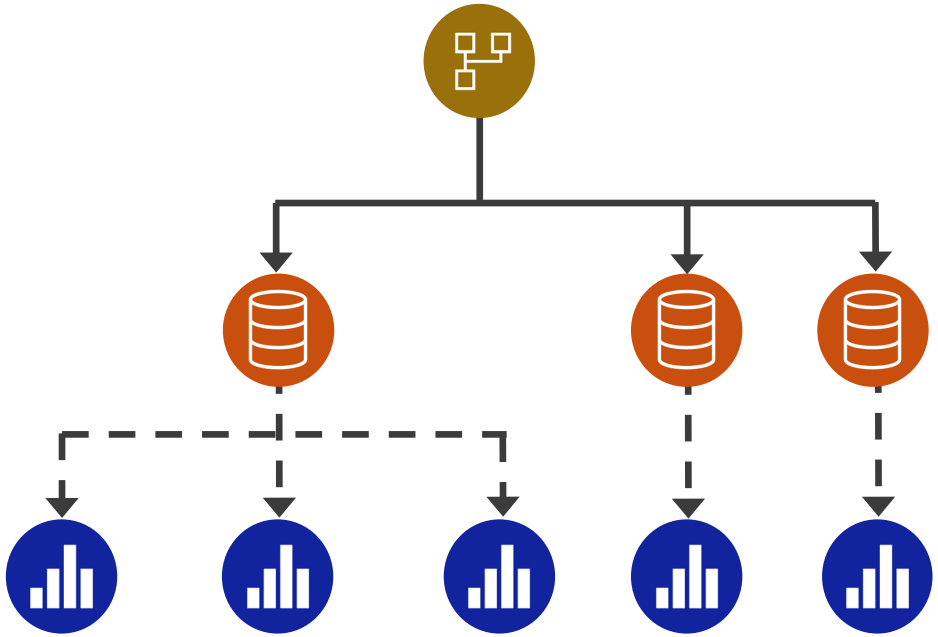
Key Advantages of Dataflows



☑ Self-Service BI Enablement

- Different people can handle data preparation vs. data modeling vs. report creation
- Shields dataset authors from complex source systems
- Reduced effort for dataset authors because data is staged/pre-processed

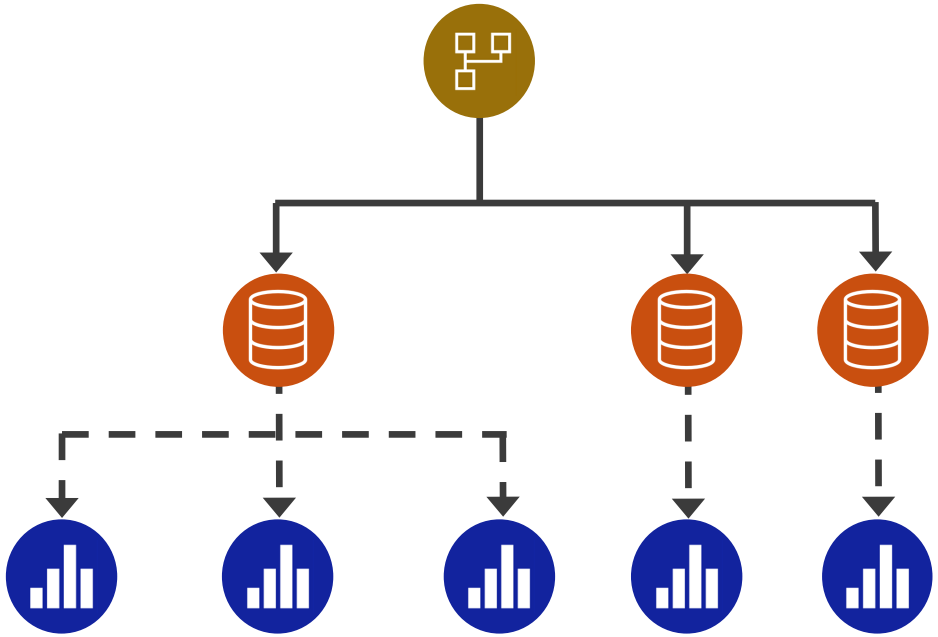
Key Advantages of Dataflows



☑ Follows the Fabric model

- Dataflows Gen 2 are one option for accessing, preparing, and loading data into OneLake *(not shown in the image)*

Key Advantages of Dataflows



☑ Reusability of Common Data

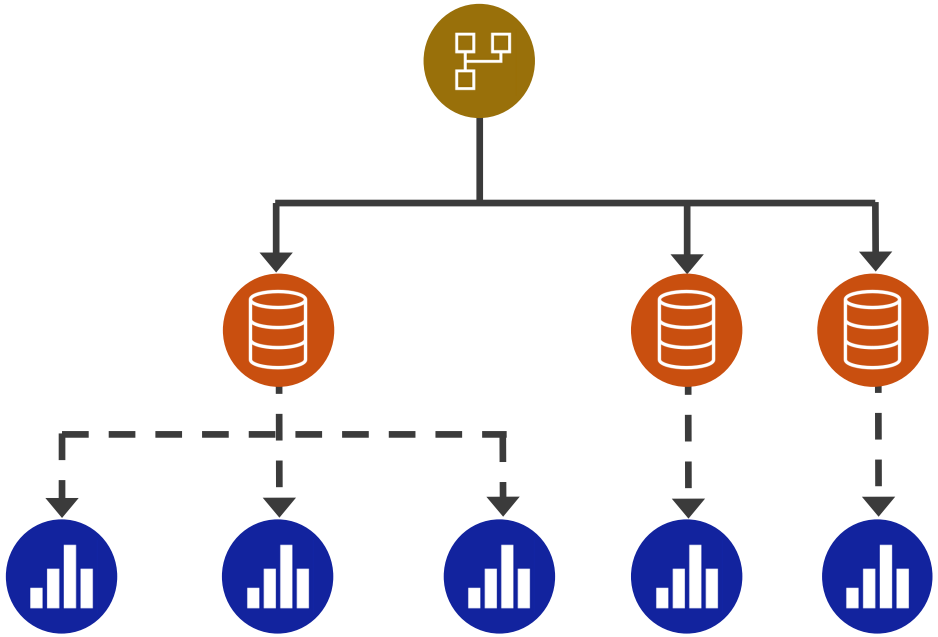
Dimensional data

- Date
- Customer
- Region
- Product

Subject areas useful to many teams
(ex: annual budgets)

Standard security mappings
(ex: for row-level security)

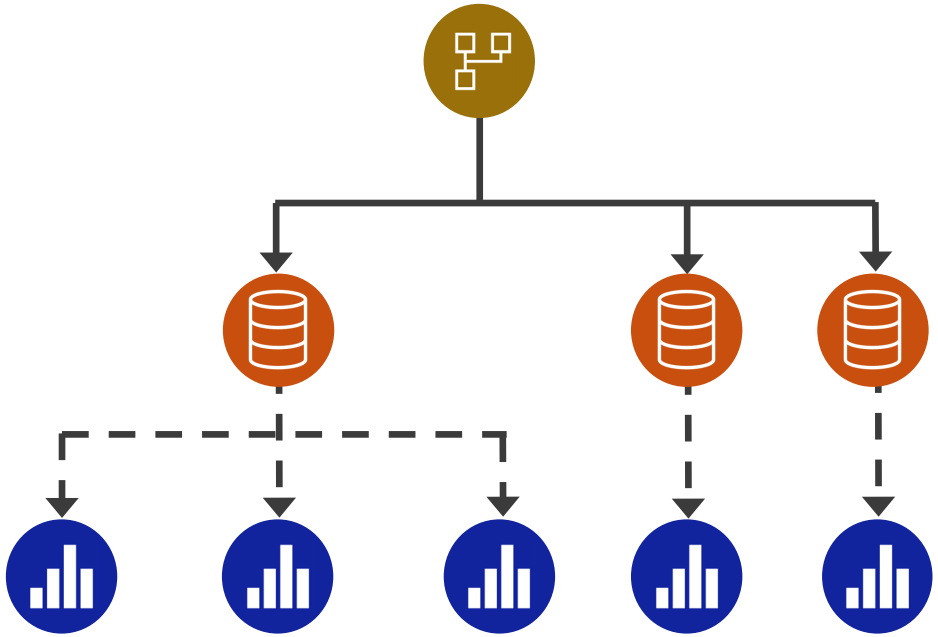
Key Advantages of Dataflows



☑ Fewer Queries to Source Systems

- Reduced load on source systems during data refresh
- Fewer dataset authors need direct access or permissions to access source systems

Key Advantages of Dataflows



☑ Improved Consistency

Usability is improved when consistency is a priority for:

- Friendly table & column names
- Data types
- Calculated columns
- Removal of unneeded columns

OneLake

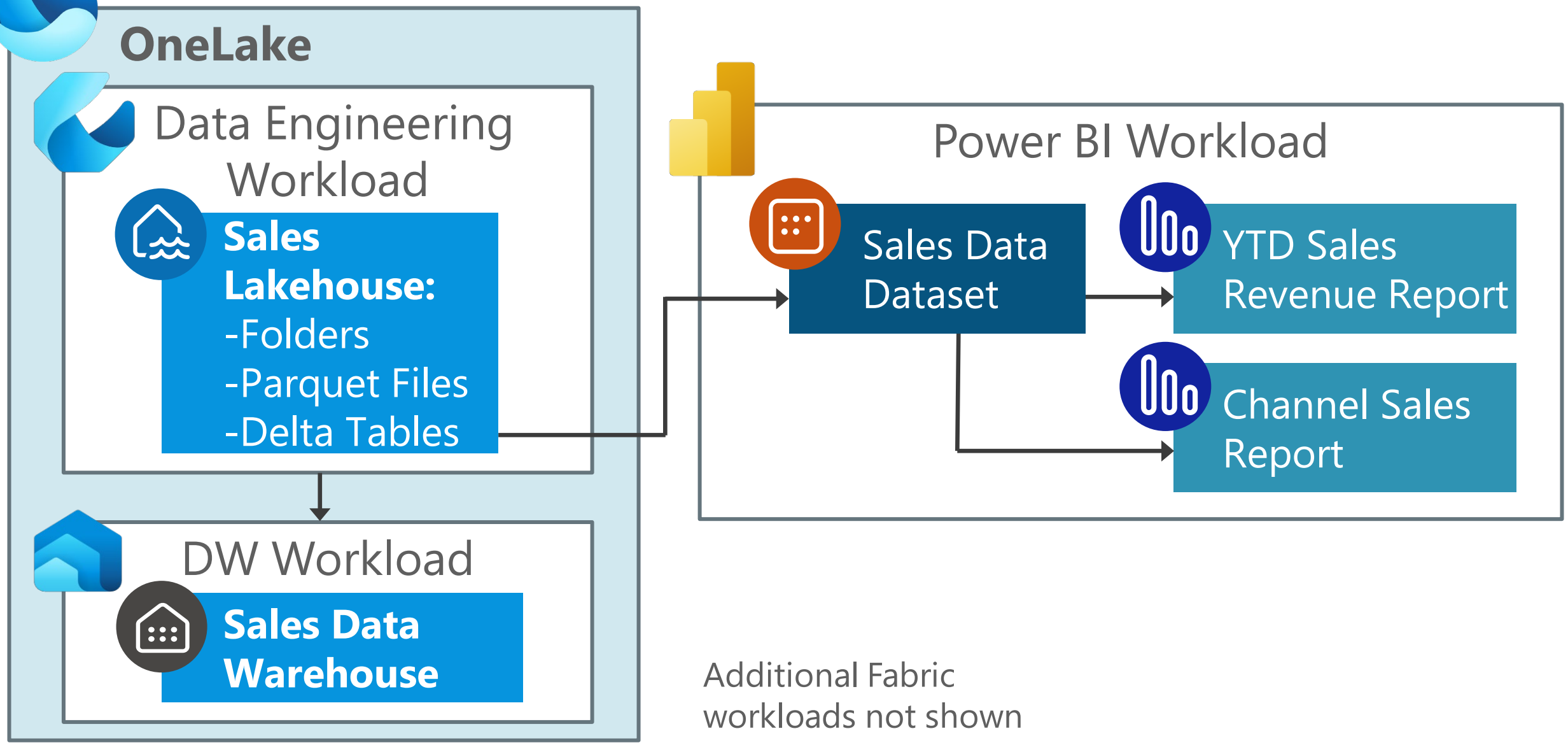
A centralized data storage location

DirectLake

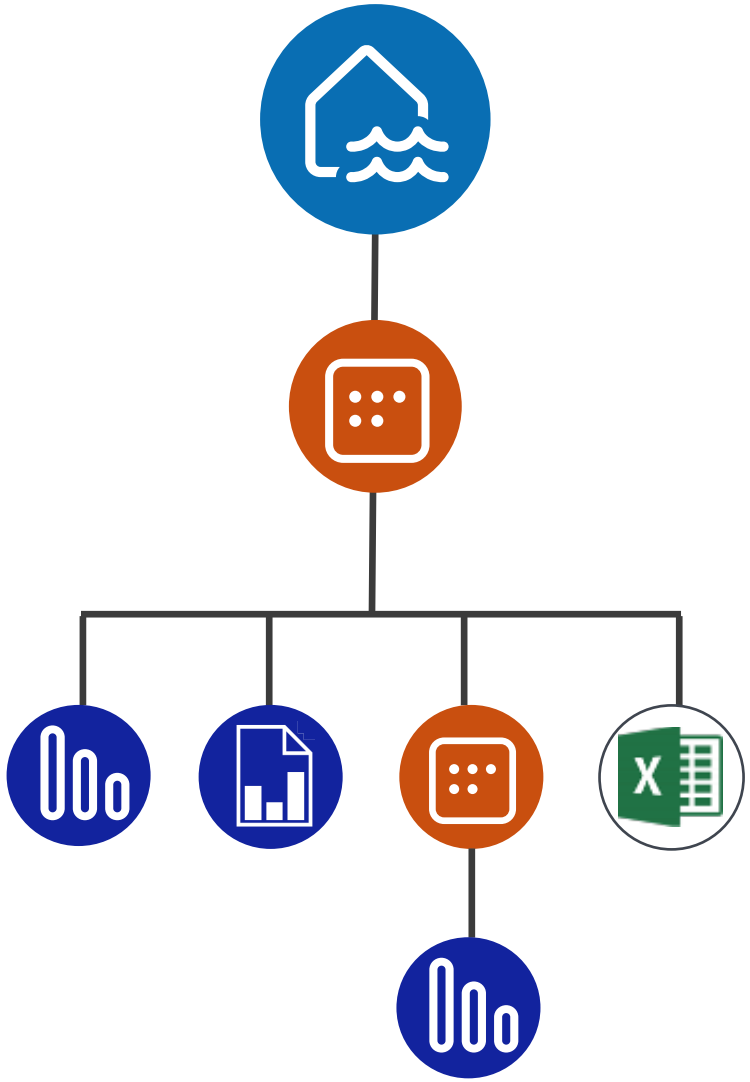
Accessing data from Power BI directly in OneLake without data refreshes or duplicating the data again

Reducing Copies of Data with OneLake

Ultimate goal: a single storage location for all Fabric workloads



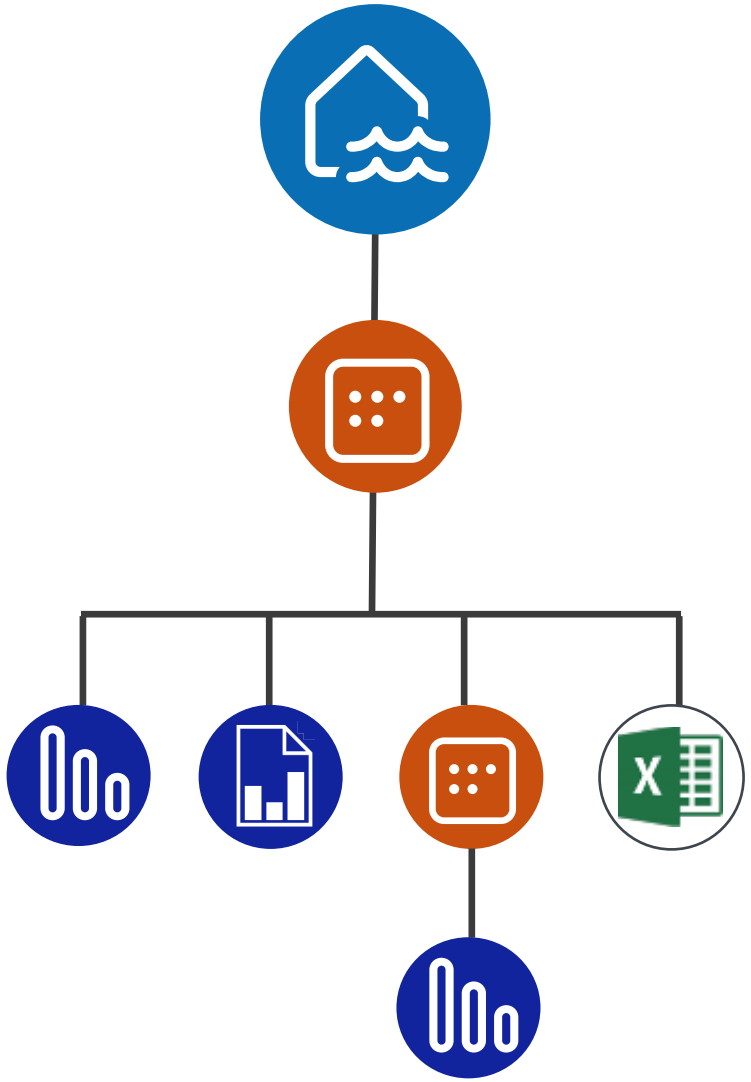
Key Advantages of Using OneLake



☑ Governance

- All Fabric workloads will end up using OneLake as its default data location (“the OneDrive for data”)
- Data storage can be centrally managed, audited, monitored, secured
- Less data proliferation and fewer data silos leads to reduced governance concerns

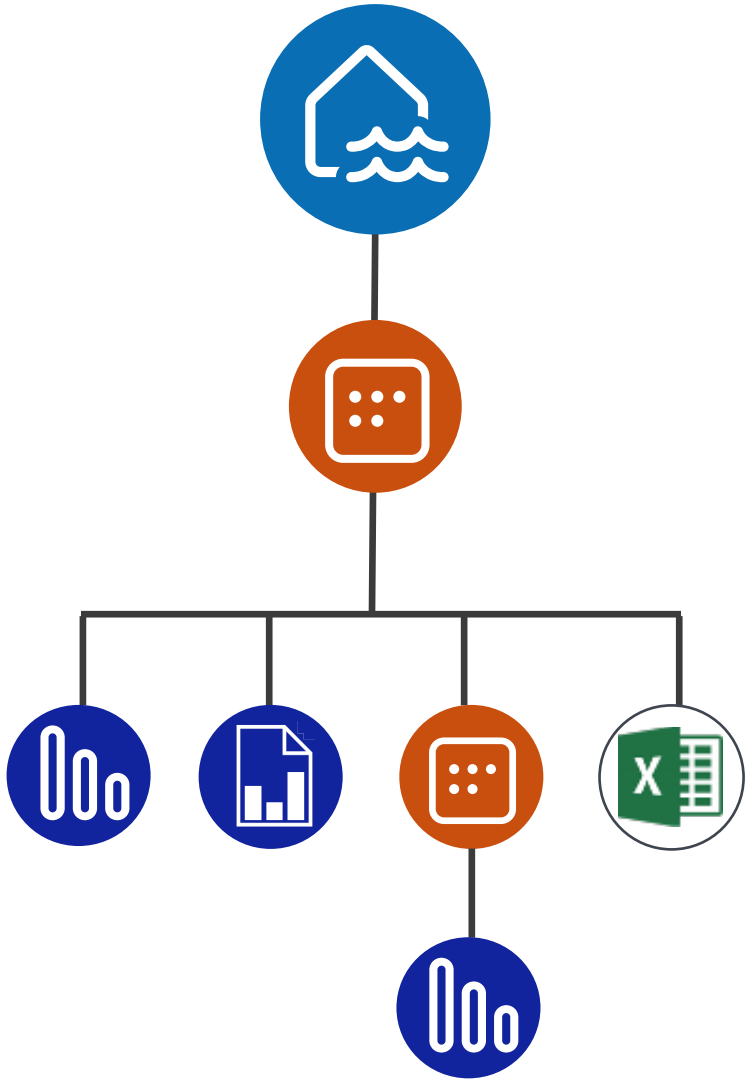
Key Advantages of Using OneLake



☑ Fewer ‘copies’ of the data

- OneLake has the data and we don’t have to copy it again into a Power BI dataset
- Far less data integration work
- Many types of artifacts can reuse the same data in OneLake – not just Power BI (ex: Data Warehouse uses a SQL endpoint)

Key Advantages of Using OneLake



☑ User enablement

- Makes it easier for users to work with data in the right ways (within your governance guardrails and with administrator visibility)
- Data can (optionally) be ‘discoverable’ through the OneLake Data Hub

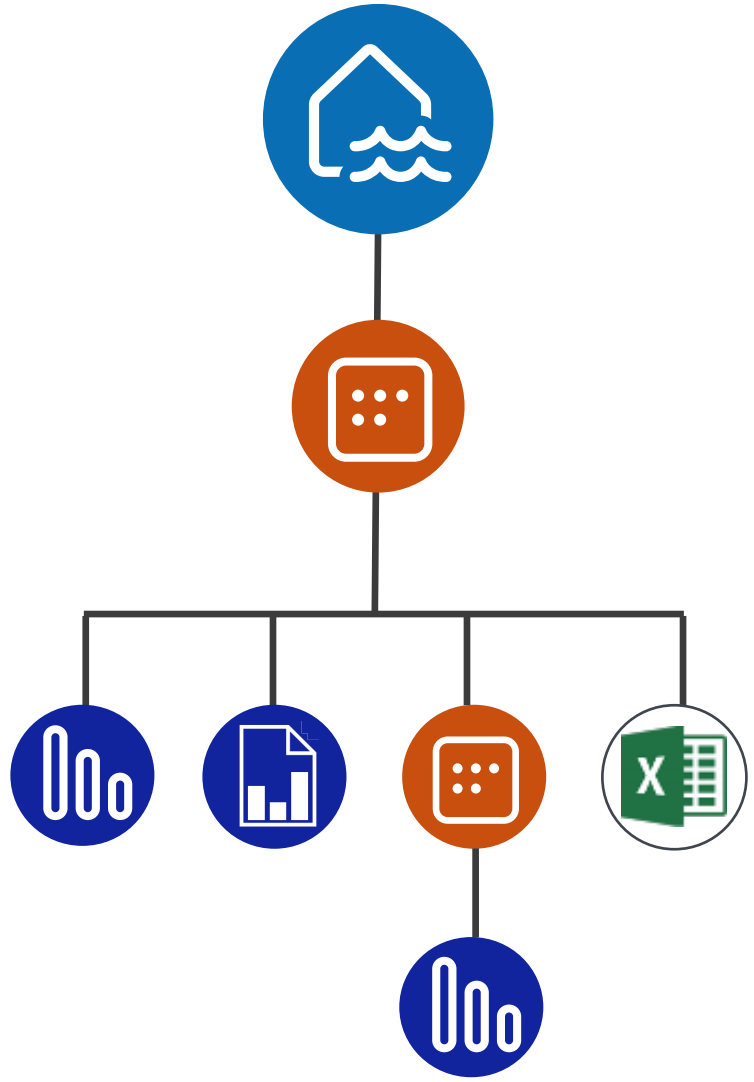
Possible Disadvantages of Using OneLake

☒ Greater dependencies

- Additional risk if the data isn't well managed with proper change control (organizational readiness concerns)

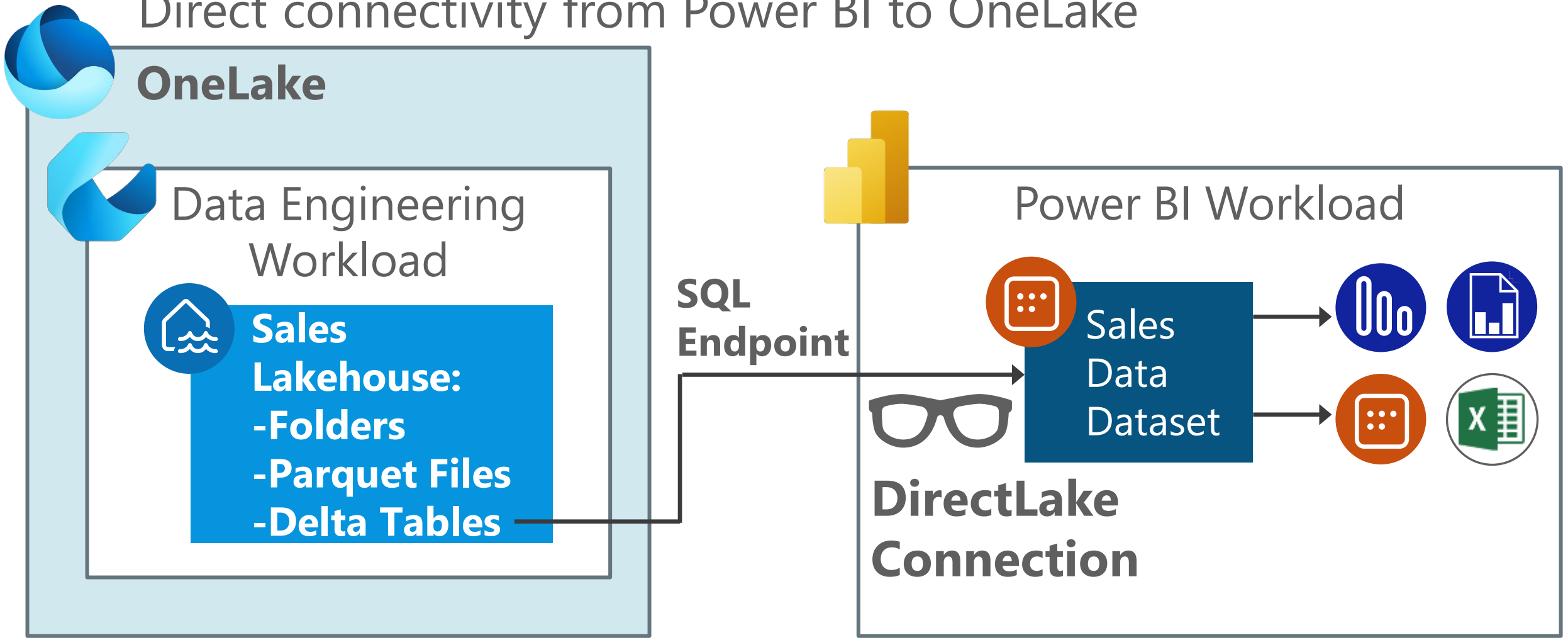
☒ Ownership & responsibility

- Maintaining the lakehouse might be managed by another team (separate from Power BI creators) who have different skills (ex: Spark) and different priorities (people & process considerations)

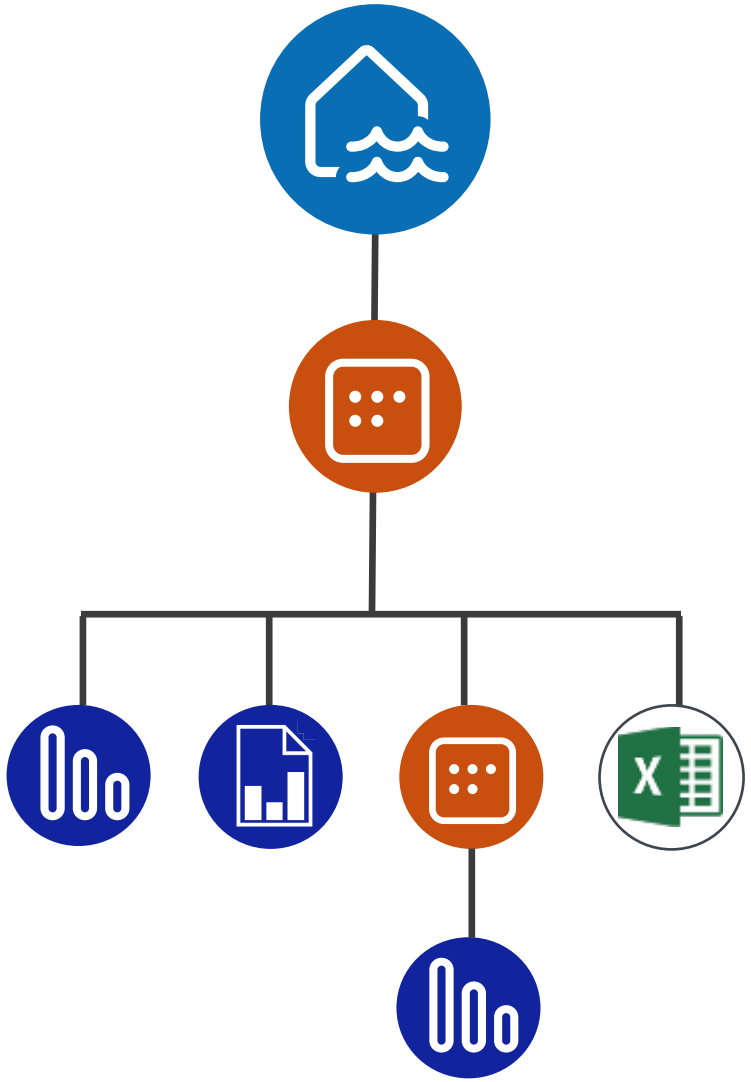


Accessing OneLake with DirectLake

Direct connectivity from Power BI to OneLake



Key Advantages of DirectLake in Power BI



☑ Best of both worlds

Goal is for reports to be able to access Power BI datasets:

- With the speed of *import mode* (because the data is in memory)
- With less data latency like *DirectQuery mode* (without tons of queries to the source)
- So we don't need to duplicate data
- We don't need to schedule refresh

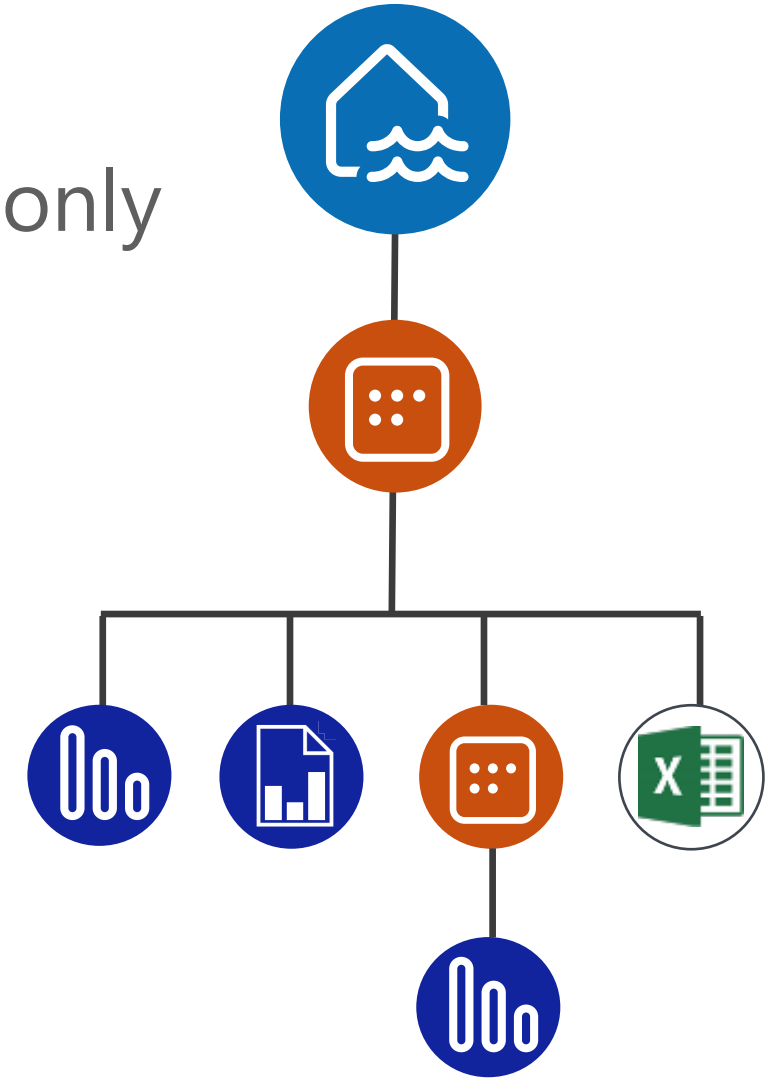
Things to Know About Using DirectLake

☑ Two types of datasets

- Decide if you'll use the 'default' read-only dataset or a 'custom' dataset

☑ Technical limitations

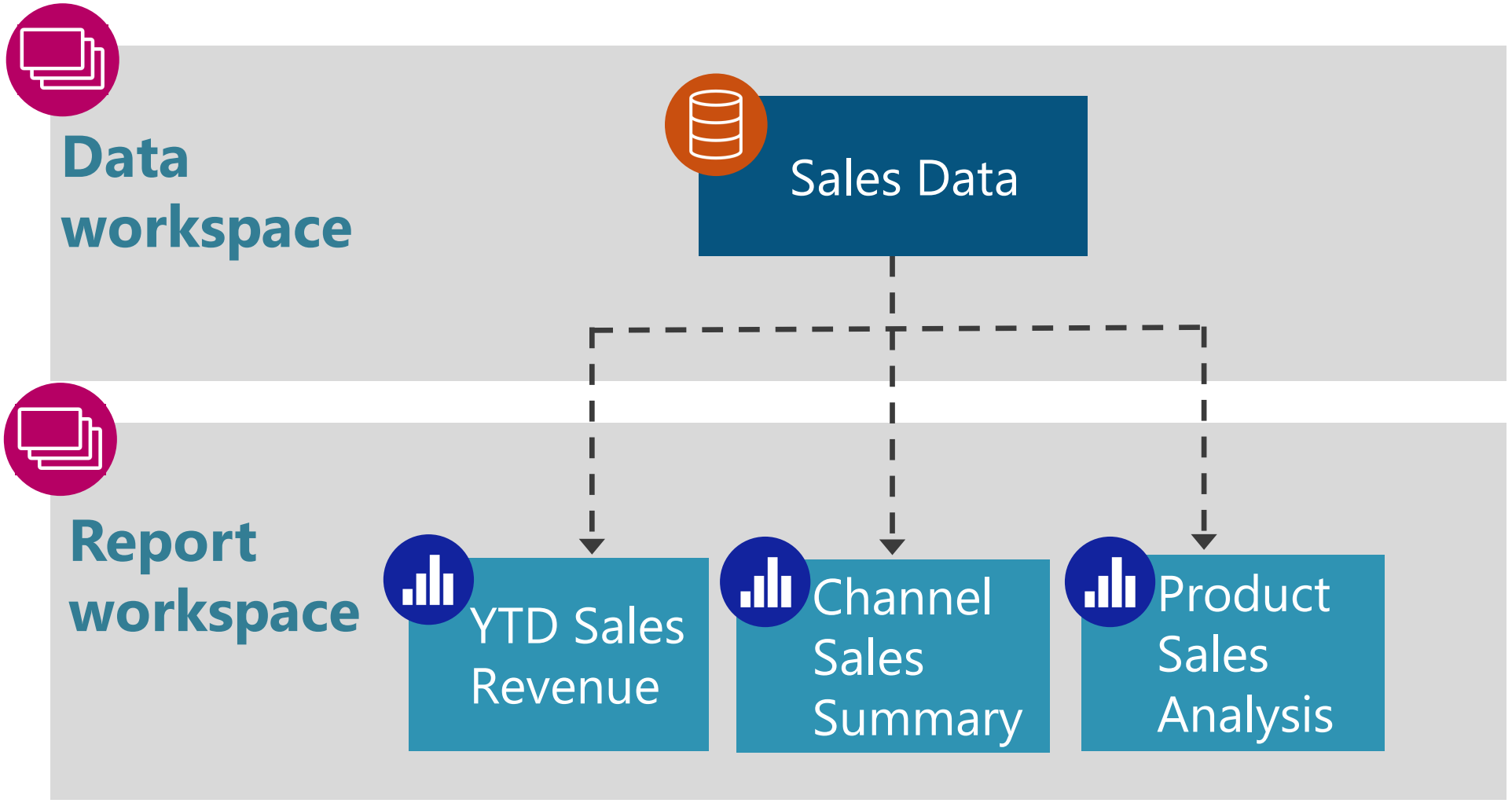
- Requires Delta tables
- Not everything is supported; sometimes it might revert to DirectQuery mode



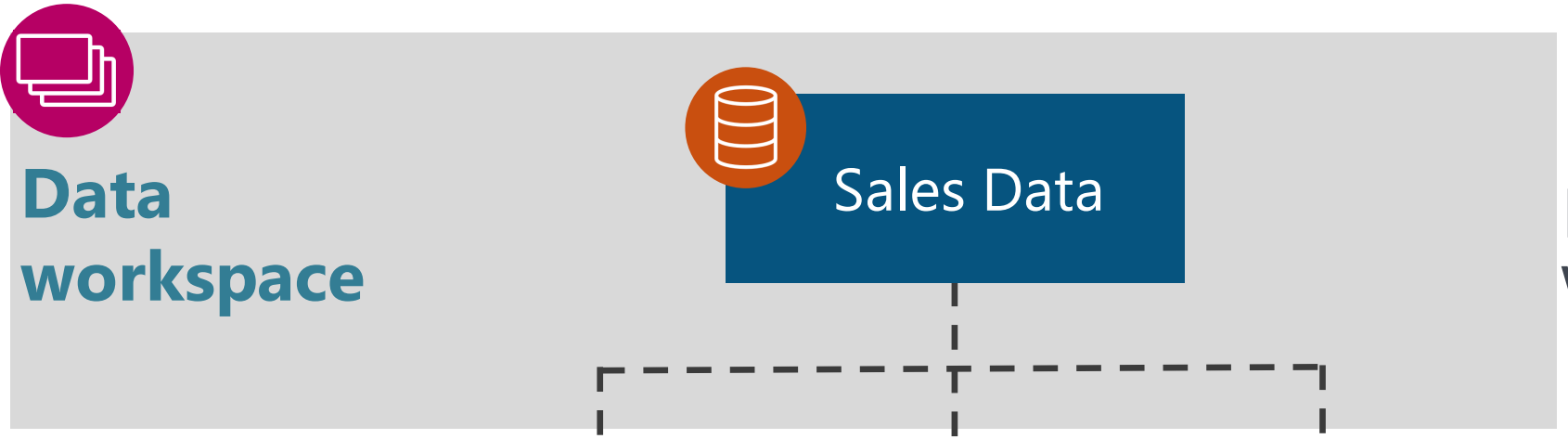
Workspace Design

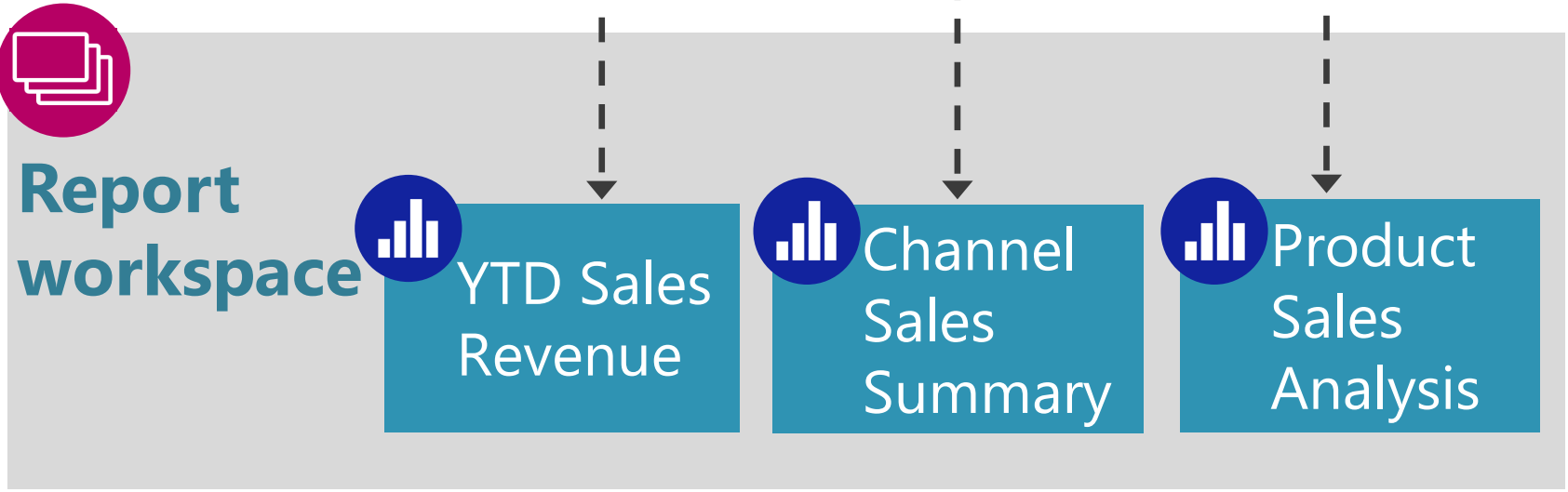
Advantages of separating ‘data’ and ‘reporting’ workspaces

How to Organize Content in Workspaces?



Assigning Permissions



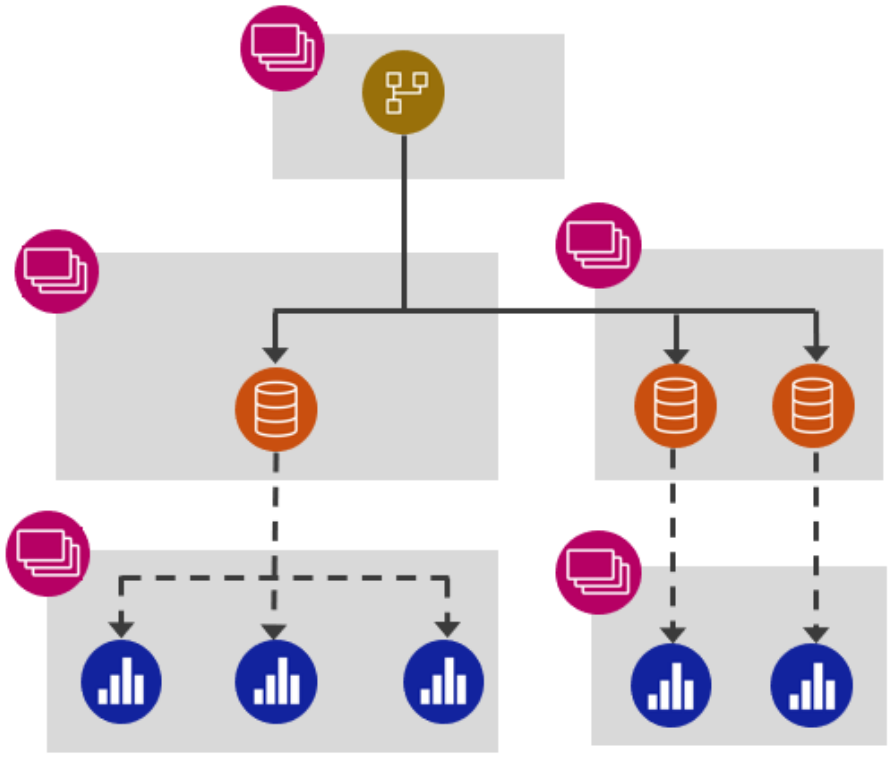


BI, IT, COE, data authors
Workspace role

Report authors
Workspace role
AND
Build on the dataset

More info: https://docs.microsoft.com/en-us/power-bi/collaborate-share/service-new-workspaces
https://docs.microsoft.com/en-us/power-bi/connect-data/service-datasets-build-permissions

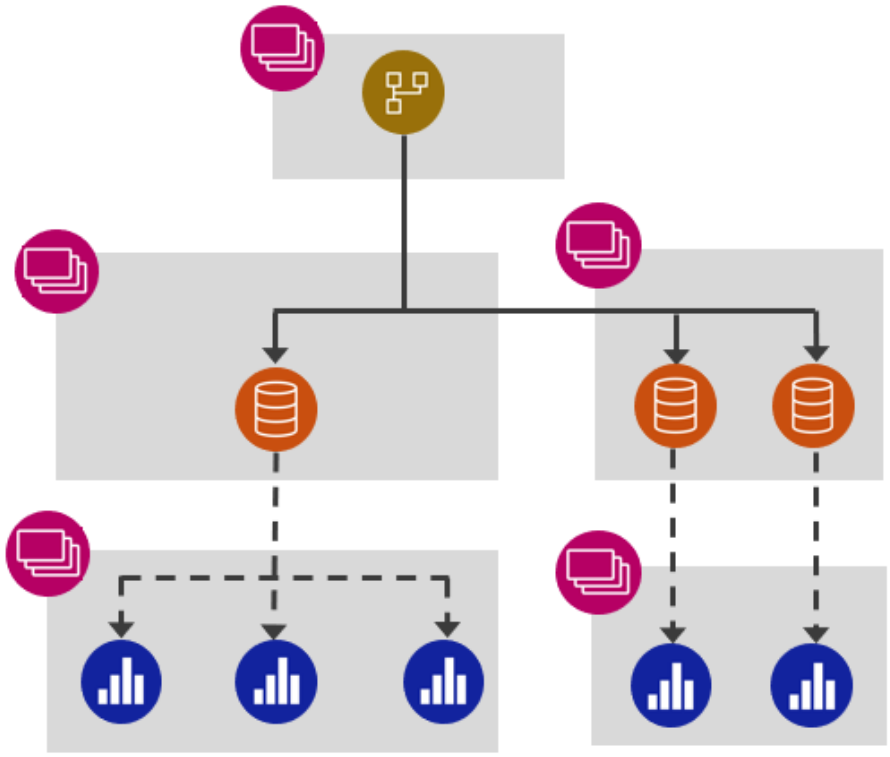
Key Advantages of Data vs. Report Workspaces



☑ Clear Ownership & Responsibilities

- Clarity on who owns & manages different content
- Separate access control (edit vs. view workspace roles)
- Enables distributed ownership and cross-team collaboration

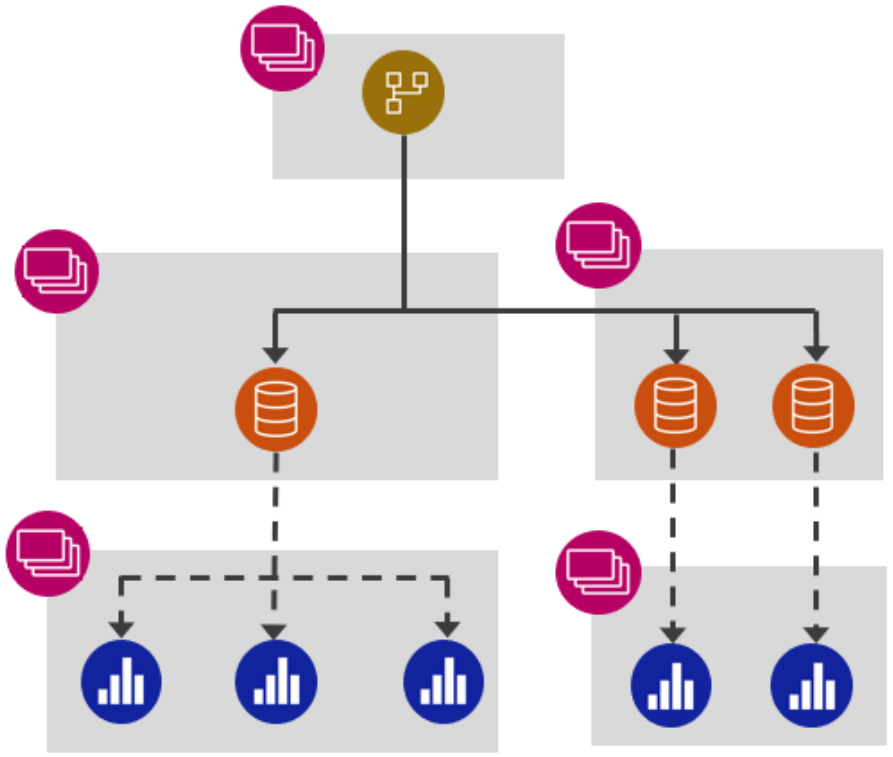
Key Advantages of Data vs. Report Workspaces



☑ Simplify for Report Creators

- Fabric creates a lot of data artifacts, so it's even more important to shield report creators from seeing a ton of items that they don't need to see in a workspace *(not shown in image)*

Key Advantages of Data vs. Report Workspaces



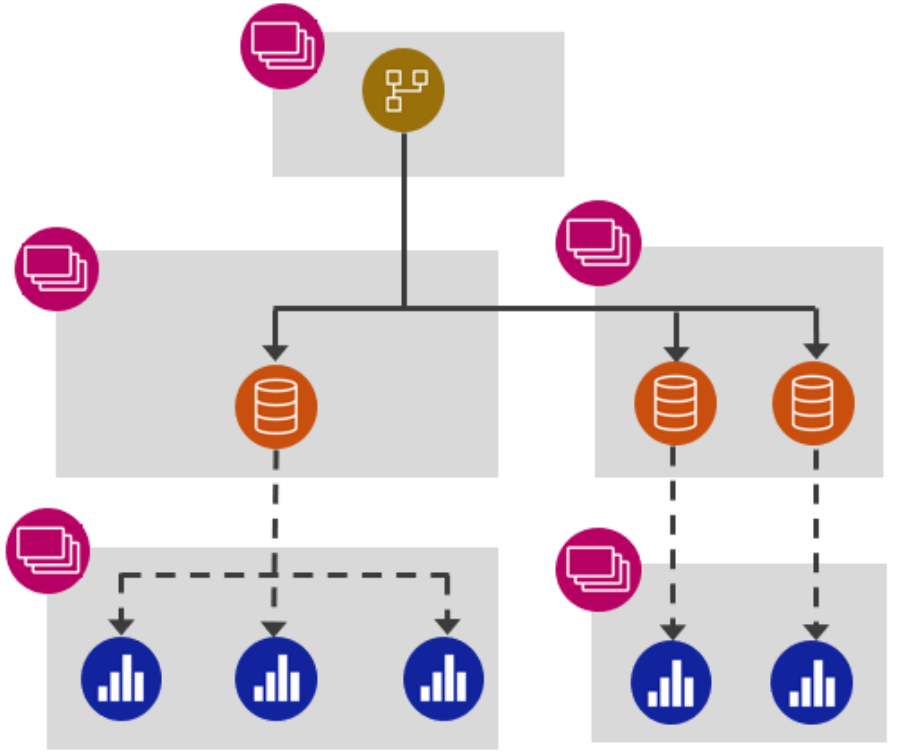
☑ Security

- Assign the ‘build’ permission on the dataset to allow certain people to build their own reports
- Row-level security works (!) for report authors who only have ‘read’ and/or ‘build’ permissions on the dataset

Key Disadvantages of Data vs. Report Workspaces

☒ Additional Workspaces

- More workspaces for admins to manage & monitor
- More workspaces for users to navigate

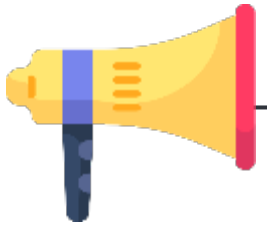


Endorsing Content

Two Types of Endorsements

Certified

The author wants to signify the content is trustworthy

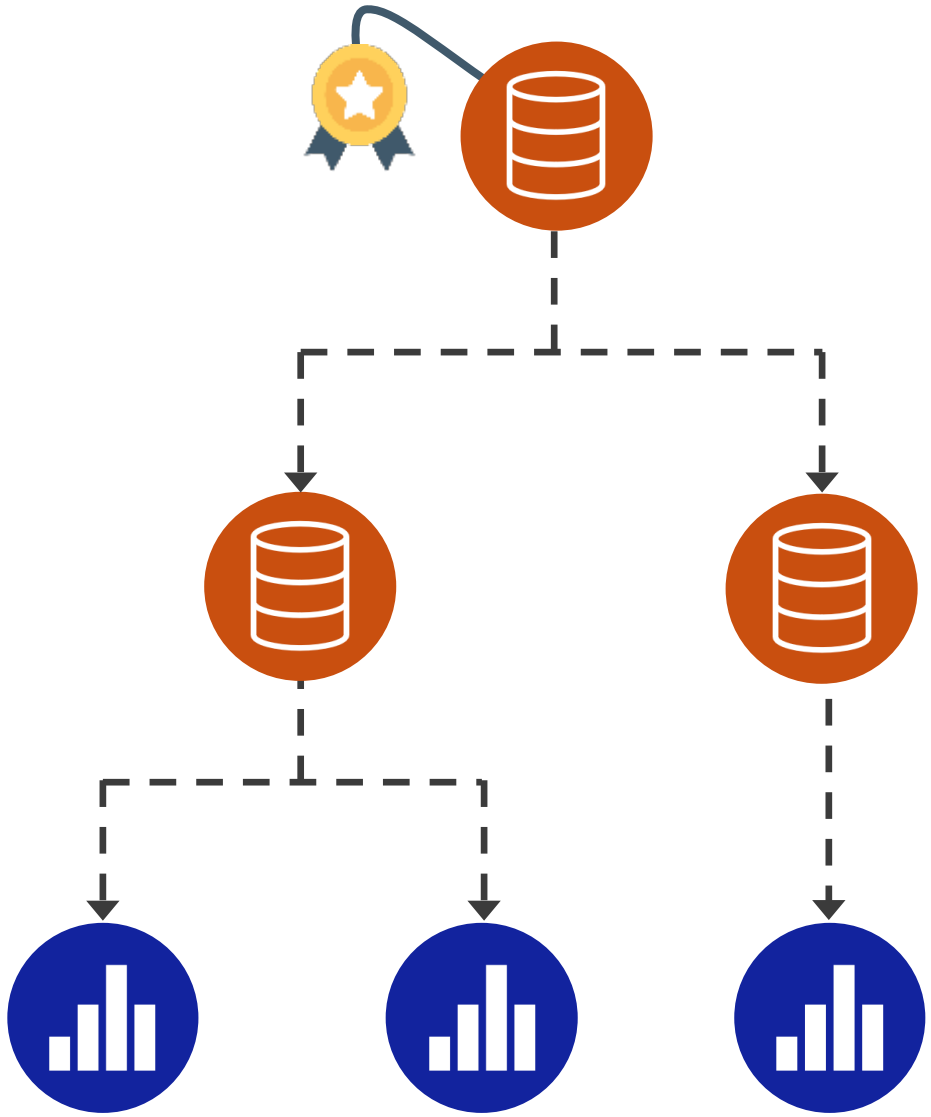


Promoted

The author wants to amplify that the content is available

For more info: CoatesDS.com/blog/power-bi-certification

What Should Be Certified?



As we encourage data reuse, we want to let other users know how trustworthy the content is.

What Can Be Endorsed?

Data:

- Dataset
- Dataflow
- Datamart
- Lakehouse

Reports:

- Power BI Report
- Paginated Report
- Scorecard

Set of Reports & Dashboards:

- App

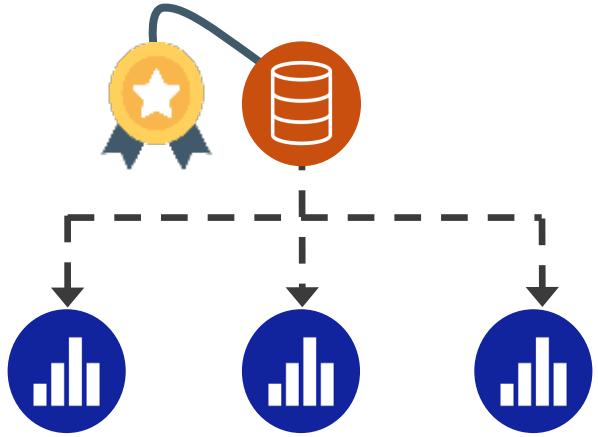
Not supported (yet?):

Warehouse, SQL Endpoint, Dashboards, Workbooks, Workspaces

The Meaning of ‘Data’ Certification

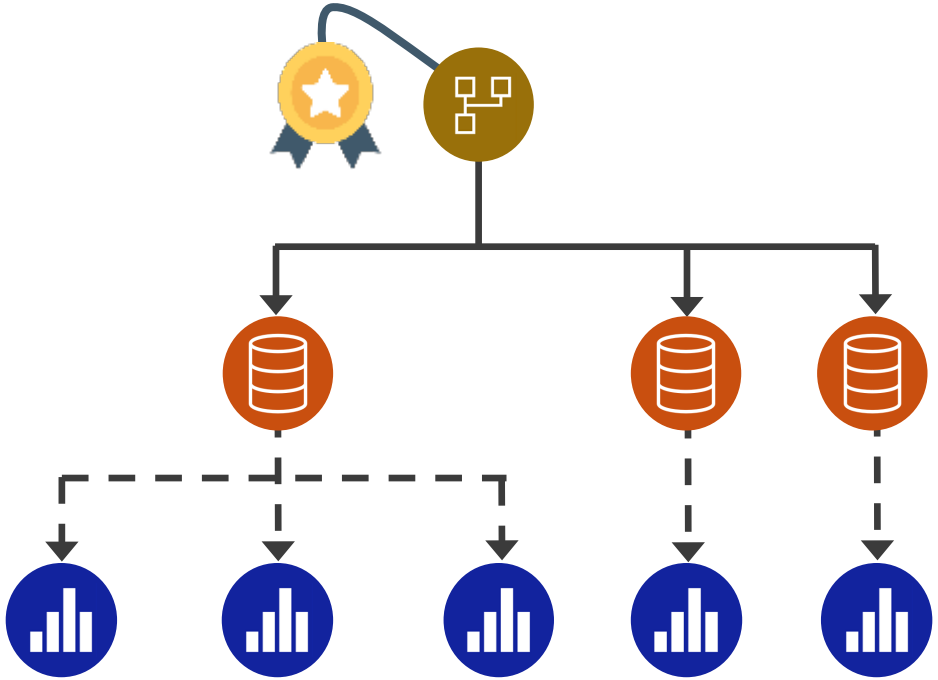
Certified dataset:

Report authors accept a dependency and trust the dataset as a source for building their own reports.



Certified dataflow:

Dataset authors accept a dependency and trust the dataflow as a source for creating their own datasets.



The Meaning of ‘Report’ Certification

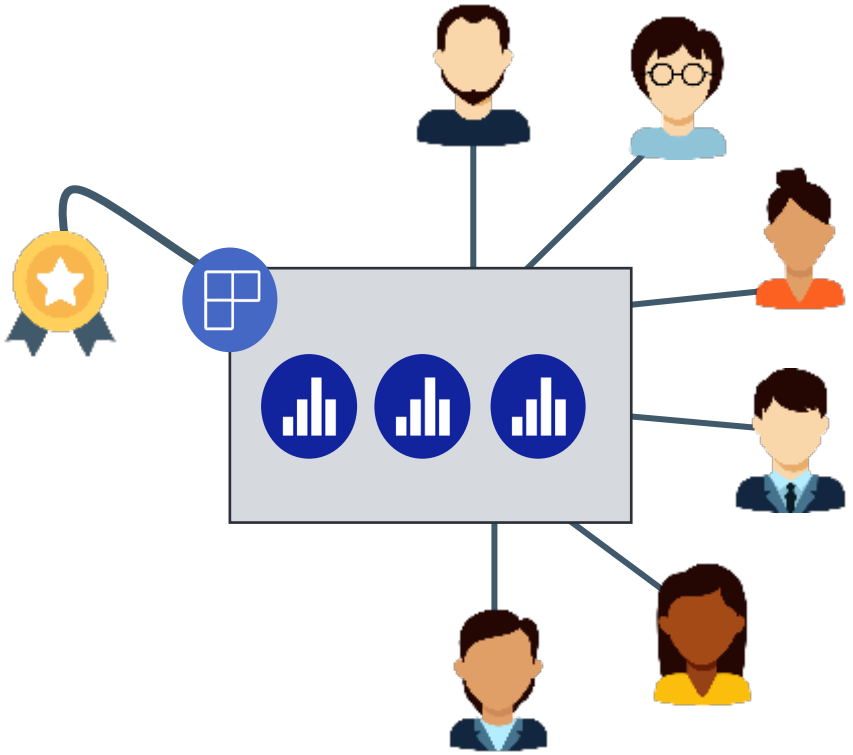
Certified report:

Consumers of the report trust the data to make business decisions based on it.

The Meaning of ‘App’ Certification

Certified app:

Consumers of the app trust **all** of the reports within the app to make business decisions.



Process to Certify Content

Ensure the process to endorse content is meaningful and consistently applied. At a minimum, include:

- Data source & lineage review
- Data model tech review
- Report/visualization review
- Security review
- Data accuracy validation
- Documentation review

More info:
https://docs.microsoft.com/en-us/power-bi/collaborate-share/service-endorsement-overview

Discovering Trustworthy Content

Finding Trustworthy Content

Why make your endorsed data discoverable?

- Help people understand the data exists
- Reduce potential for duplicate effort
- Increase investment in existing data assets
- Allow people to request access to the data

More info: https://docs.microsoft.com/en-us/power-bi/collaborate-share/service-discovery
https://docs.microsoft.com/en-us/power-bi/connect-data/service-datasets-hub

OneLake Data Hub

Targeted primarily to content creators (rather than consumers) because it displays ‘data’ items only:

- Datasets
- Dataflows
- Datamarts
- Lakehouses
- SQL Endpoints
- Warehouses
- KQL Databases

Standard search in Power BI relies on having existing permissions to the content, so search isn’t the same as discovery

Sensitivity Labels & Usage Expectations

Sensitivity Labels

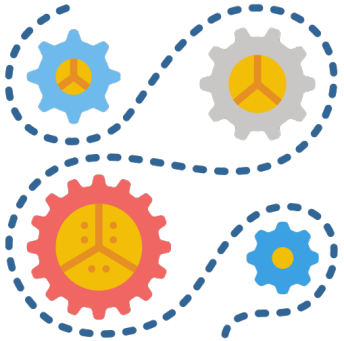
Why assign sensitivity labels?

- Help people understand how sensitive or confidential it is so they understand ***what they can or can't do***
- Help people take responsible actions
- Assist with right-sizing user permissions
- Create automated data loss prevention policies

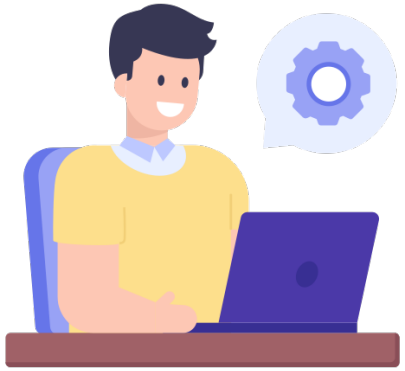
More info: https://learn.microsoft.com/en-us/power-bi/guidance/powerbi-implementation-planning-info-protection-data-loss-prevention-overview

Ways to Apply Sensitivity Labels

Manually applied by a user



Auto-applied programmatically



Applied by a Power BI administrator with REST API

Sensitivity Label Stays With the Content

Power BI Desktop file (.pbix)

Power BI Service artifacts:

- -Reports
- -Datasets
- -Dataflows
- -Dashboards

Power BI mobile apps

Live connection, or

Analyze in Excel, or

Download PBIX file, or

Report/data exports to:

- -Excel -PowerPoint
- -PDF -Word

Q&A

More Information from Melissa Coates

Slides:
CoatesDS.com/Presentations

Diagrams:
CoatesDS.com/Diagrams

Power BI Governance Training:
CoatesDS.com/Training

Blog:
CoatesDS.com/Blog-Posts

YouTube:
YouTube.com/CoatesDataStrategies

APPENDIX:
Documenting, Testing, and Helping User Recognize Trustworthy Content

Documentation

Include report documentation to help both **consumers and creators**, such as:

- Purpose
- Target audience
- Business rules & calculation logic
- Definition of key performance indicators
- Data sources (lineage)
- Who to contact with questions

Change Log

Consider publishing a change log so consumers can easily see:

- What changes occurred
- When
- By whom

Artifact Properties

Set helpful workspace & artifact properties:

- Contact
- Description
- Sensitivity label
- Endorsement
- Discoverability
- Custom instructions for requesting access

Data Quality

Consider publishing a data quality dashboard so report consumers can verify:

- Data validation results
- Repeatable reconciliation process
- Authoritative data sources (lineage)

Consider including an easy way for report consumers to submit a data issue they see or ask questions.

How will you catch data quality issues before consumers do?

Testing Processes

Create a list of ***repeatable data quality unit tests*** such as:

- Row counts
- Totals
- Data meets certain expectations
- Good, maintainable, testable coding standards are used

Separate Environments

Safeguard the production content with:

- Development
- Test
- Production