BlueGranite Office Hours

Overview of Power BI Dataflows

Melissa Coates
March 28, 2019

1

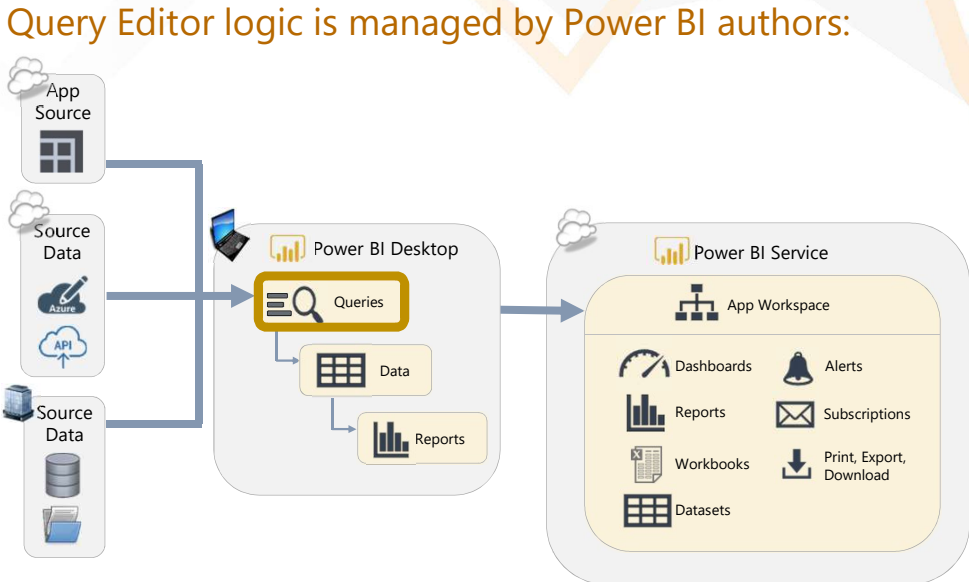## Current State: Self-Service BI Without Dataflows

Pros:
- Self-service BI freedom
- SMEs handle data cleansing

Limitations:
- Query Editor is confined to a single PBIX file
- Additional tools cannot access the Query Editor output

Query Editor logic is managed by Power BI authors:

2

Current State: Corporate BI Without Dataflows

Pros:
- Centralized data delivery for corporate BI

Limitations:
- DW & semantic layer don't meet all needs
- Large DWs can't react fast enough
- Requires data engineering pros

ETL and data centralization handled by IT:

3

Future State: Self-Service Data Preparation

- Data prep in a familiar, business user-friendly tool
- Prepared data for reuse among many datasets

4

What are Power BI Dataflows?

Capability	Targeted To	Purpose
Reusable, self-service data preparation Reuse of a dataflow across numerous datasets	Analysts responsible for data acquisition & data cleansing who want to maintain this work once Analysts who want to consume complex data prep work done by colleagues	Promote data consistency Reduce cost, time, & expertise required

5

Common Use Cases for Dataflows

Standardized / Reusable Data
Data is valuable enough that it has use cases for many datasets, different types of analysis, many types of apps

Pre-Processing
Process larger sets of data which exceed resources available on local laptop or Power BI Desktop

Staged Data
Deliver staged data for Power BI data modelers to finish preparing

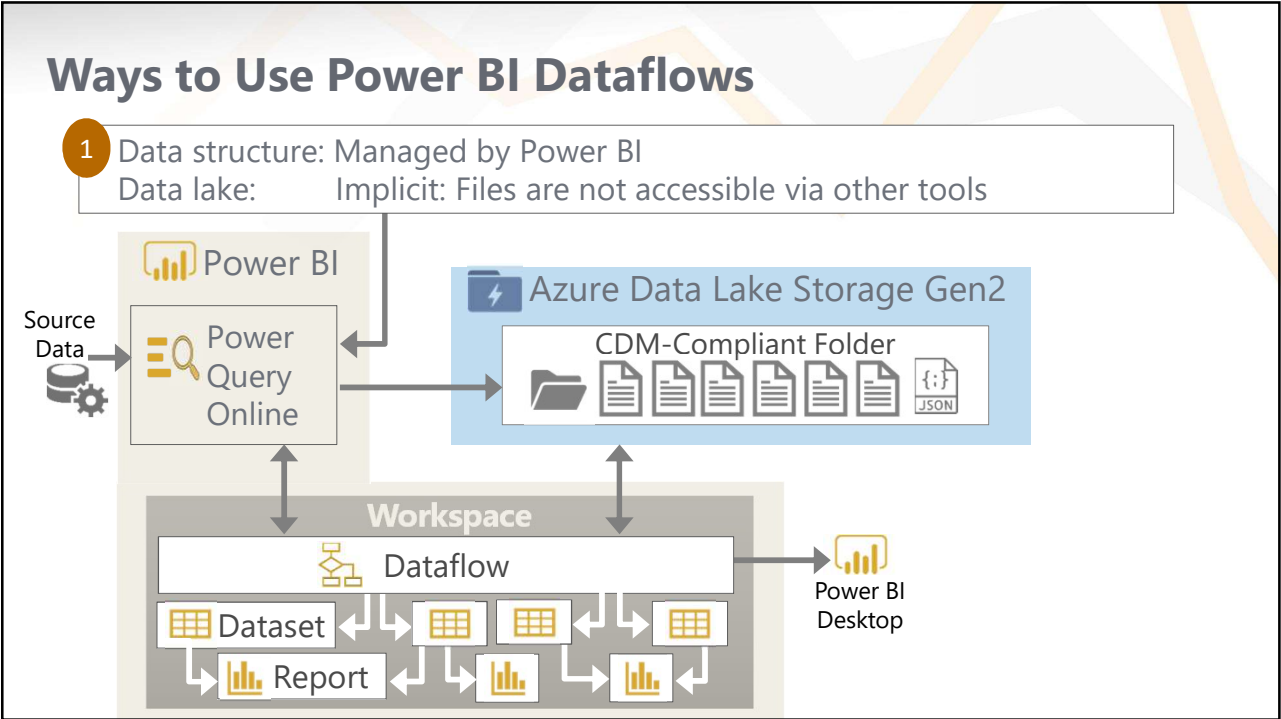### Reduce Load on Source System
Minimize number of queries sent to originating source system

6

Three Ways to Use Power BI Dataflows

7

Ways to Use Power BI Dataflows

1 Data structure: Managed by Power BI
Data lake: Implicit: Files are not accessible via other tools

8

Ways to Use Power BI Dataflows

2 Data structure: Managed by Power BI
Data lake: Explicit: Data lake has been associated to Power BI tenant &
                           workspace storage has been enabled for dataflows

9

Ways to Use Power BI Dataflows

3 Data structure: Managed by the tool of choice
Data lake: Explicit: Data lake has been associated to Power BI tenant &
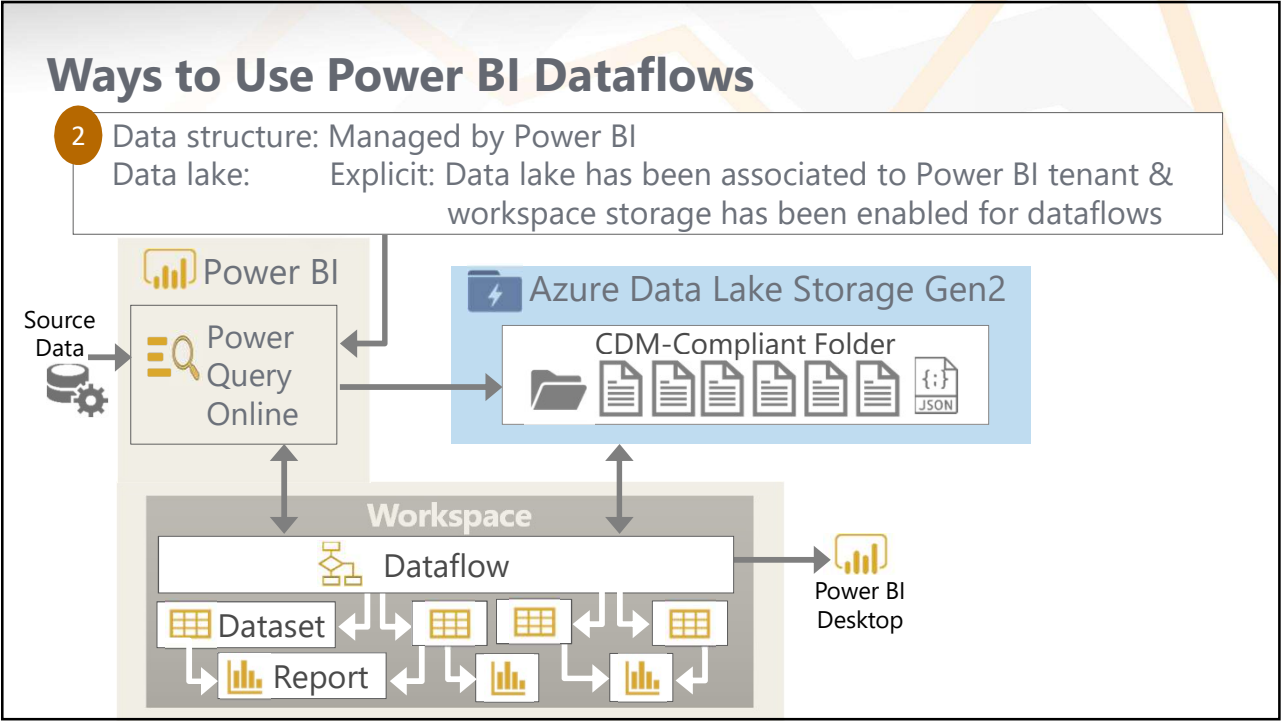   workspace storage has been enabled for dataflows
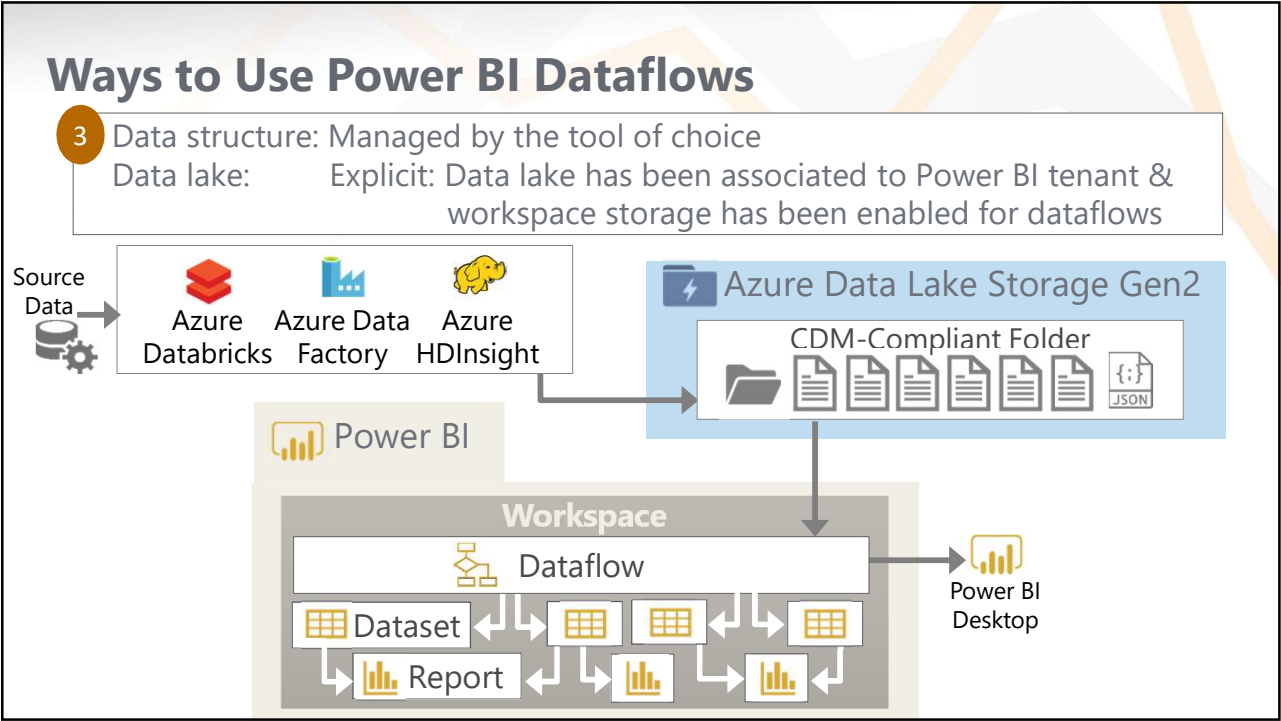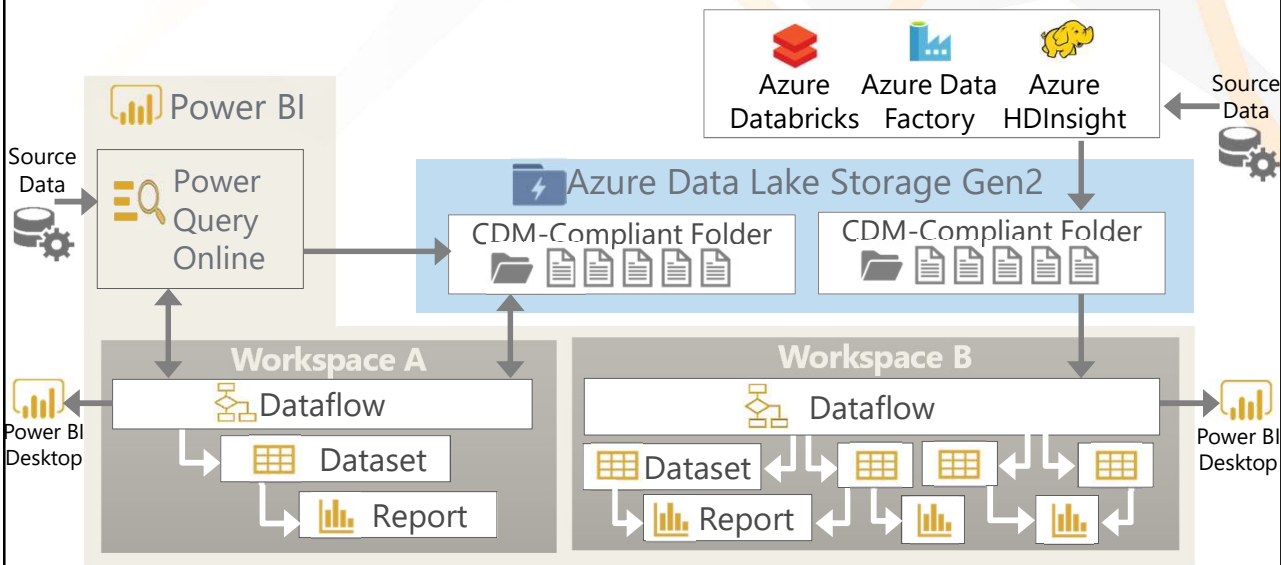

10

Ways to Use Power BI Dataflows

Power BI's Role	Data Managed by Power Query Online	Data Managed by Tool of Choice	Data Access in Azure Data Lake Storage Gen2	
1. Fully manages the dataflow	✔	Unavailable	Implicit (not accessible)	Self-Service BI
2. Fully manages the dataflow	✔	No	✔ Explicit	Managed Self-Service BI
3. Consumer of dataflow	No	✔	✔ Explicit	Managed Self-Service & Corporate BI

All 3 can be used simultaneously in the same Power BI tenant.

11

Ways to Use Power BI Dataflows

12

6

Dataflows Demo

13

Q&A
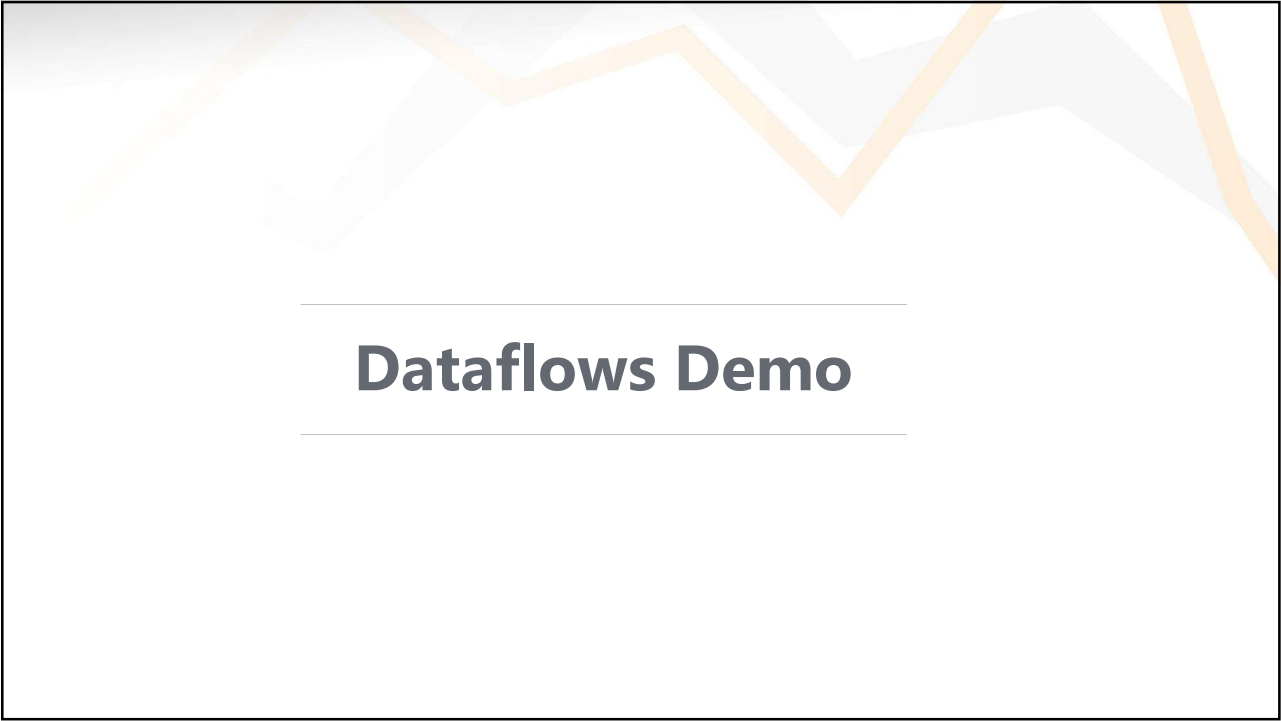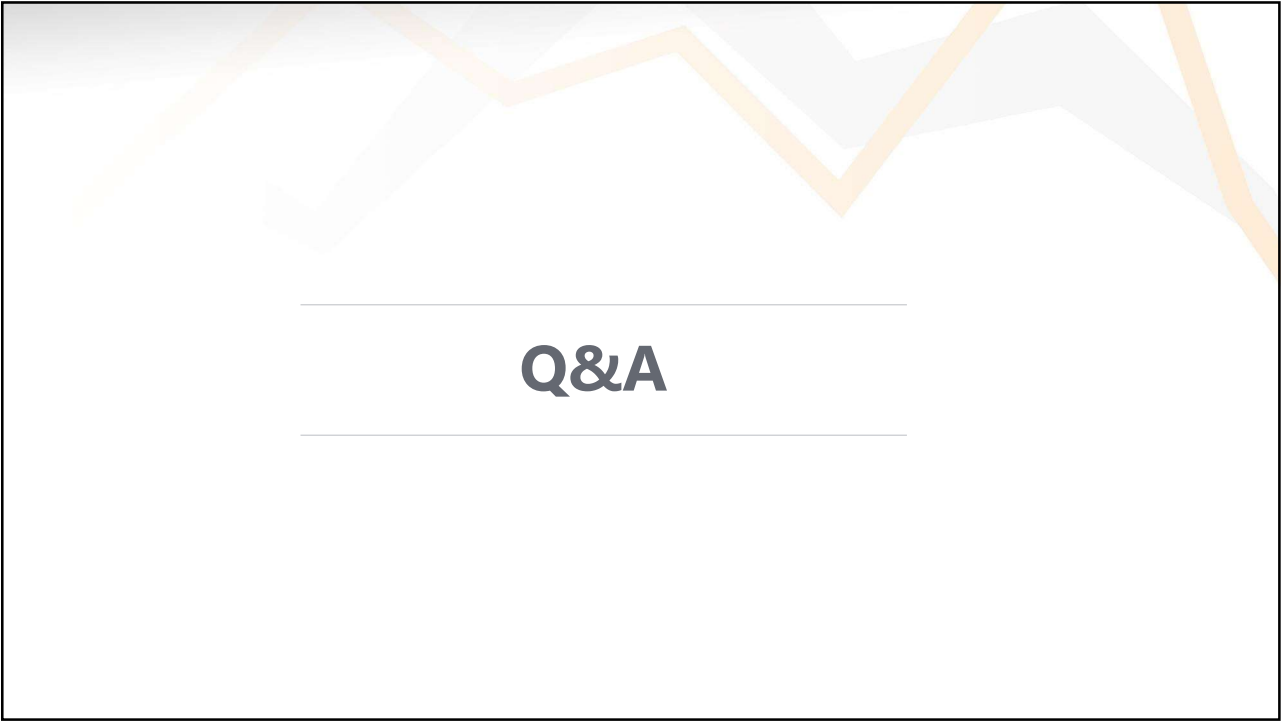
14

Reusability Scenarios in Power BI

15

Relationship of Dataflow to Dataset

A single dataflow can feed data to many datasets

A dataset is still defined which contains calculations, relationships, RLS, etc.

There are two separate refresh schedules
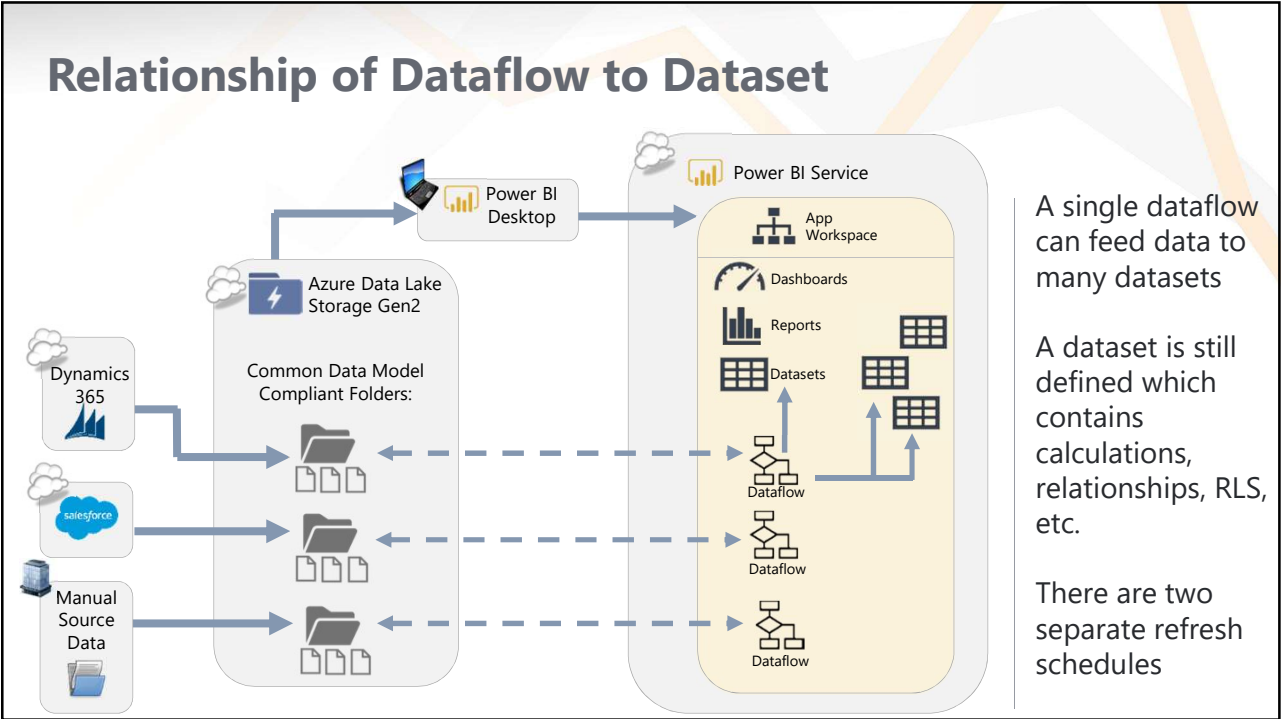
16

Evolving to Reusability

Focus: One Report
o Imports simple, flat datasets and creates reports.
o No concern for reusable data warehousing or ETL processes.

Focus: One Dataset
o Comfortable working with tables, relationships & calculations.
o A single dataset may serve multiple reports.

Focus: Reusability
o Embraces approaches to reduce data redundancy and duplication of business rules.
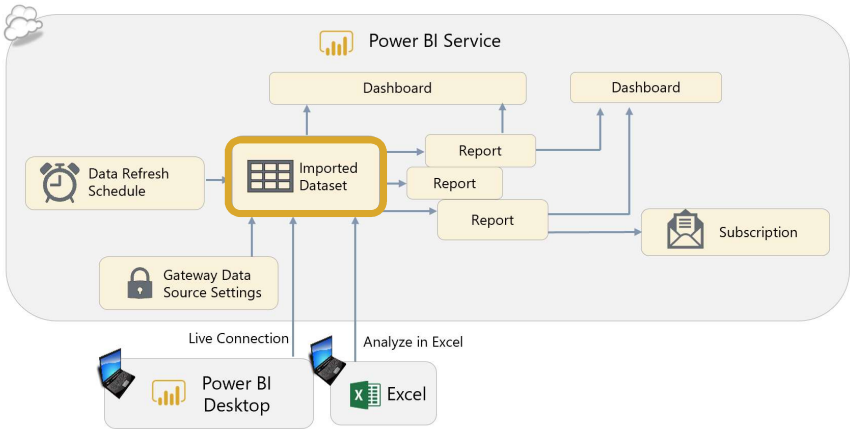
17

Reusability in Power BI Desktop: Datasets

18

19

20

Data is Stored in ADLS Gen2

Fully managed by Power BI: 10GB/user and 100TB per P1 node limits apply
Bring your own data lake: No dataflow size limits (same dataset limits apply)

Data + accompanying metadata are stored in CDM-compliant folders in the data lake

21

CDM Folders

A folder in the data lake conforming to specific, well-defined and standardized metadata structures and self-describing data.

CDM = Common Data Model

Image source: Power BI Dataflows whitepaper

22

11

Using CDM Folders

The existence of the model.json file indicates compliance with the CDM metadata format

- CDM folder
- Metadata file
- Data files

Image source: https://docs.microsoft.com/en-us/power-bi/service-dataflows-azure-data-lake-integration

23

Using CDM Folders

"File system" in ADLS Gen 2
=
"Container" in Azure Storage

- File systems
- Optional sub folders, for better organization of CDM folders and disambiguation
- CDM folders

CDM folder in ADLS Gen 2
=
Dataflow in Power BI

Image source: https://docs.microsoft.com/en-us/power-bi/service-dataflows-azure-data-lake-integration

24
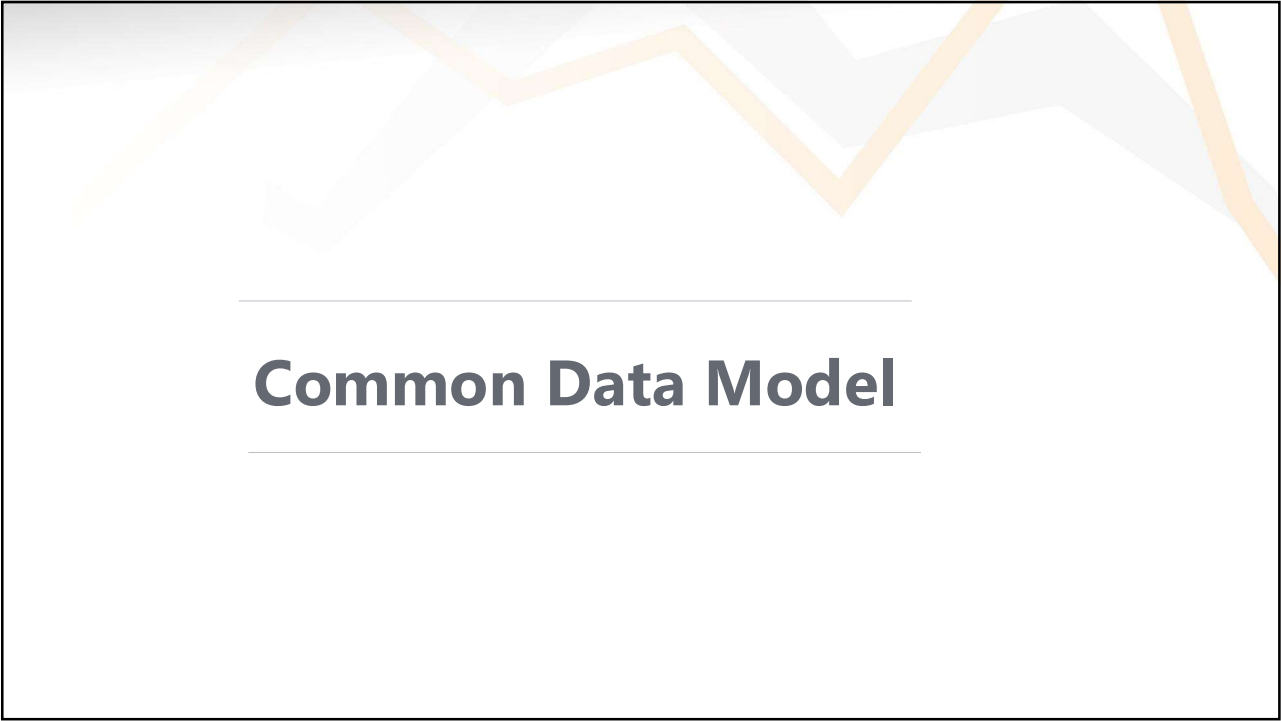
Common Data Model

25

What is the Common Data Model?

Open source

Declarative specification

Part of the Open Data Initiative

26

What is the Common Data Model?

Modular & extensible entities across common business processes

Goal is consistent semantic meaning & interoperability across toolsets

Started as part of Dynamics 365 & is evolving

Entities: Can be standard (from CDM) or custom
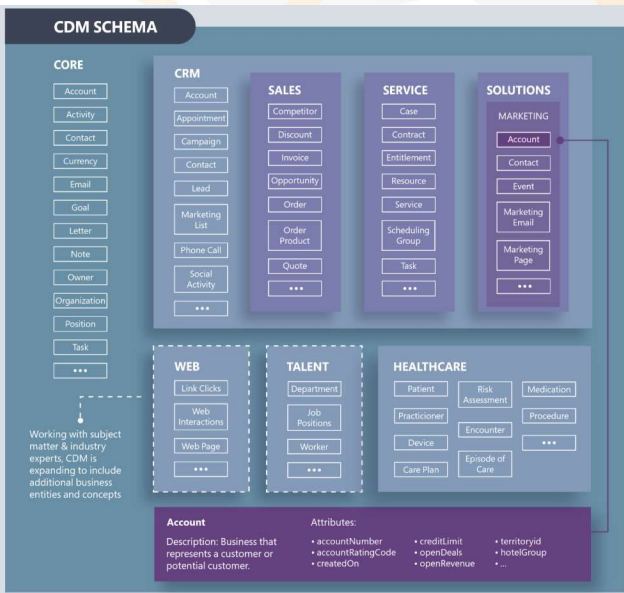
27

What is the Common Data Model?

More and more services will support the CDM over time

Designed for openness and adoption by a variety of services and applications

28

More Details
on
Dataflows
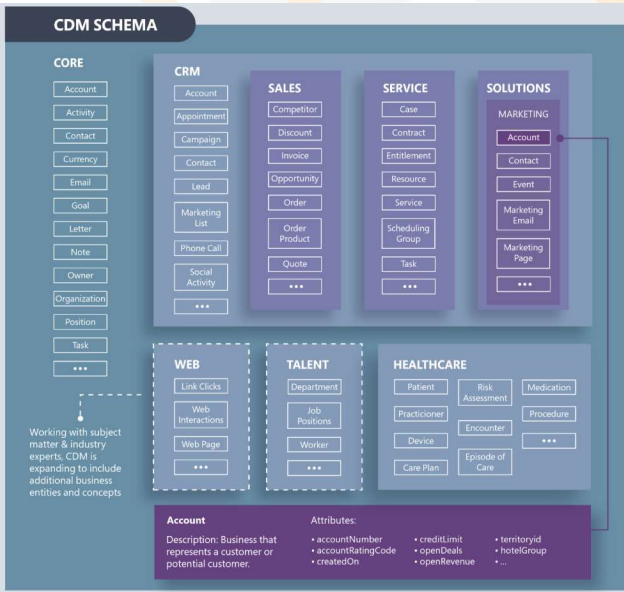
29

Dataflows "Lingo"

Dataflow	Analogous to a database; a collection of entities
Entity	Analogous to a table which holds data: represents one or more files in ADLS Gen 2 (1 file unless incremental refresh is configured, in which case multiple files); strongly typed for downstream consumption by datasets
Dataflow Formula	M expressions, defined in Power Query Online, which define how an entity gets computed
Auto Orchestration with Dataflows Calculation Engine	Builds a dependency graph between entities in the same workspace. Automatically triggers a refresh (parallel or sequential) for all dependent entities for transactional consistency.

30

Linked Entities and Computed Entities

Linked entity (aka referenced entity):
Read-only reference to another entity, already created from another dataflow (like a select stmt)
No data movement or duplication

Computed entity:
Also considered a linked entity...but also relies on "in-storage calculations" to reload data to a new entity
(like a select>insert operation)
The results of the computed entity are persisted in another CDM folder in ADLS Gen 2

31

Linked Entities Data Refresh

Links in the same workspace:
Automatically triggers a dataflow refresh operation for any dependent entities – this is the "auto orchestration" core principle

Links in another workspace:
No automatic refreshes; is an external data source

Start creating your dataflow

Define new entities
Choose a data source to define the entities for your dataflow. You can map your data to standard Common Data Model entities, or define custom entities instead.
Learn more
[Add new entities]

Link entities from other dataflows
Linking to entities from other dataflows reduces duplication and helps maintain consistency across your organization.
Learn more
[Add linked entities]

Attach a Common Data Model folder
Attach a Common Data Model folder from your Azure Data Lake Storage Gen2 account to a new dataflow, so you can use it in Power BI.
Learn more
[Create and attach]

Option 1 or 2 discussed earlier

Option 3 discussed earlier

32

Additional Dataflows Features Available with Premium

Power BI Premium opens up the full featureset for Power BI Dataflows:

Feature	Pro	Premium
Storage allocation	10GB per user	100TB per Premium node
Data ingestion	Serial ingestion	Parallel ingestion
Refresh frequency	Up to 8x/day	Up to 48x/day
Incremental updates	--	Yes
Linked entities	--	Yes
Computed entities	--	Yes
Cognitive Services AI	--	Yes

33

Bring Your Own Data Lake

Requirements

Azure and Power BI Tenants
- Power BI is in the same AAD tenant as ADLS Gen2
- Power BI is in the same region as ADLS Gen2

Power BI Admin Center
- Dataflows are enabled
- Data lake account associated to the Power BI tenant
- Allow workspace admins to assign workspaces to the storage account

Workspace
- Storage settings enabled

Azure Data Lake Storage Gen2
- Account with hierarchical namespace enabled
- File system name = powerbi

Permissions
- RBAC for Power BI Service in the data lake
- ACLs for Power BI Service & Power Query Online in the data lake
- ACLs for dataset creator to read data in the lake

If Premium

Allow dataflows workload for the capacity

34

Resources to
Learn More

35

Resources to Learn More

Whitepaper:
https://go.microsoft.com/fwlink/?linkid=2034388&clcid=0x409

Official blogs:
https://powerbi.microsoft.com/en-us/blog/power-bi-delivers-dataflows-enterprise-reporting-and-major-updates-to-power-bi-desktop/
https://powerbi.microsoft.com/en-us/blog/introducing-power-bi-data-prep-wtih-dataflows/

Matthew Roche's Blog:
https://ssbipolar.com/2018/10/23/dataflows-in-power-bi/
https://ssbipolar.com/2018/11/27/power-bi-dataflows-faq/

Documentation site:
https://docs.microsoft.com/en-us/power-bi/service-dataflows-overview

36

Resources to Learn More

Release notes & roadmap:
https://docs.microsoft.com/en-us/business-applications-release-notes/April19/business-intelligence/power-bi-service/power-bi-dataflows/self-service-data-prep-with-dataflows

Community ideas:
https://ideas.powerbi.com/forums/265200-power-bi-ideas?category_id=341638

Community forum:
https://community.powerbi.com/t5/forums/searchpage/tab/message?filter=location&q=dataflow

Connecting ADLS Gen2 to Dataflow Storage in the Power BI Service:
https://docs.microsoft.com/en-us/power-bi/service-dataflows-connect-azure-data-lake-storage-gen2

Power BI Dataflows and Azure Data Lake Integration:
https://docs.microsoft.com/en-us/power-bi/service-dataflows-azure-data-lake-integration

37

Resources to Learn More

Common Data Model documentation:
https://docs.microsoft.com/en-us/common-data-model/

Common Data Model standard:
https://github.com/Microsoft/CDM

Common Data Model tutorials:
https://aka.ms/cdmadsblog
https://github.com/Azure-Samples/cdm-azure-data-services-integration/blob/master/Tutorial/CDM-Azure-Data-Services-Integration-Tutorial.md

Model.json schema:
https://docs.microsoft.com/en-us/common-data-model/model-json

38