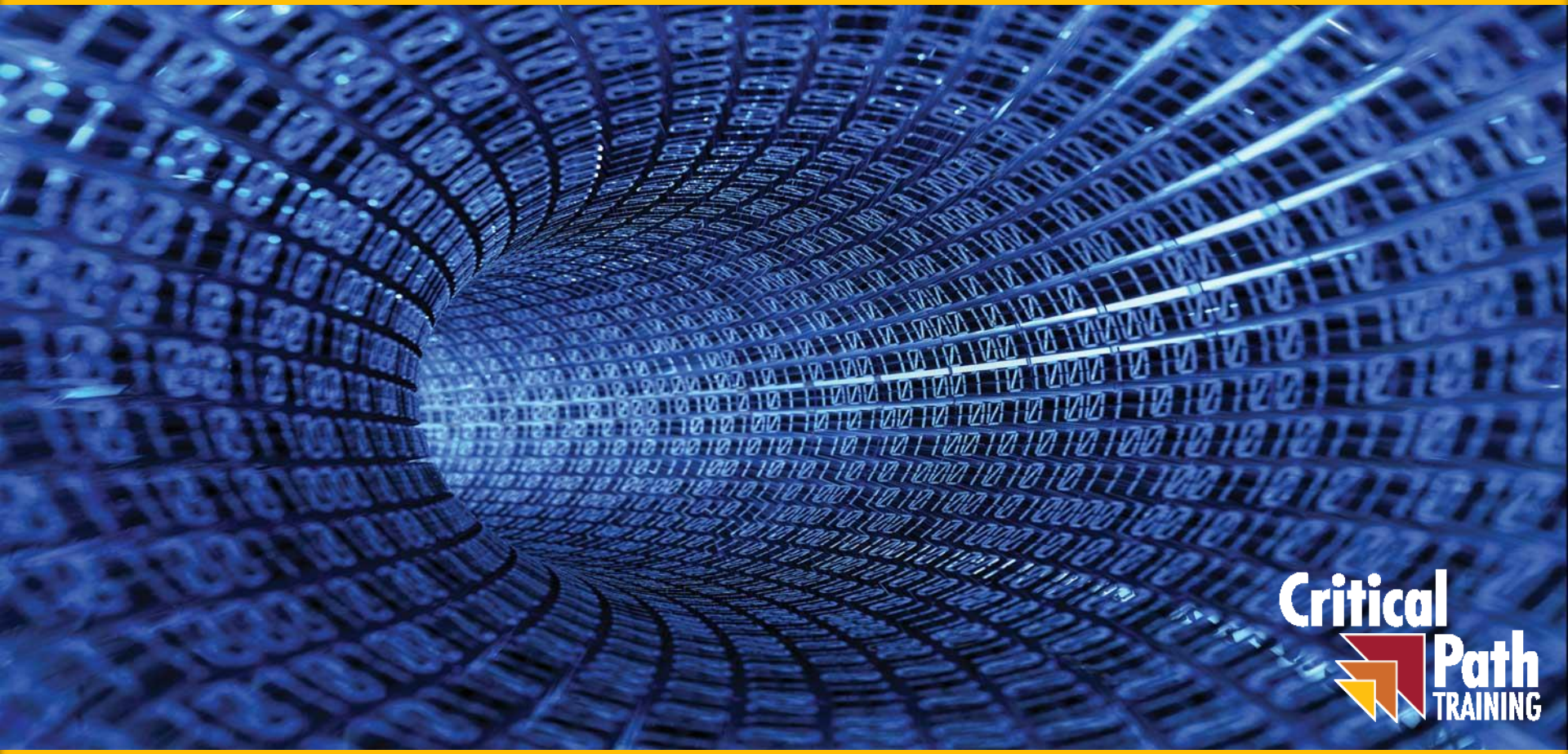
Power BI Dataflows



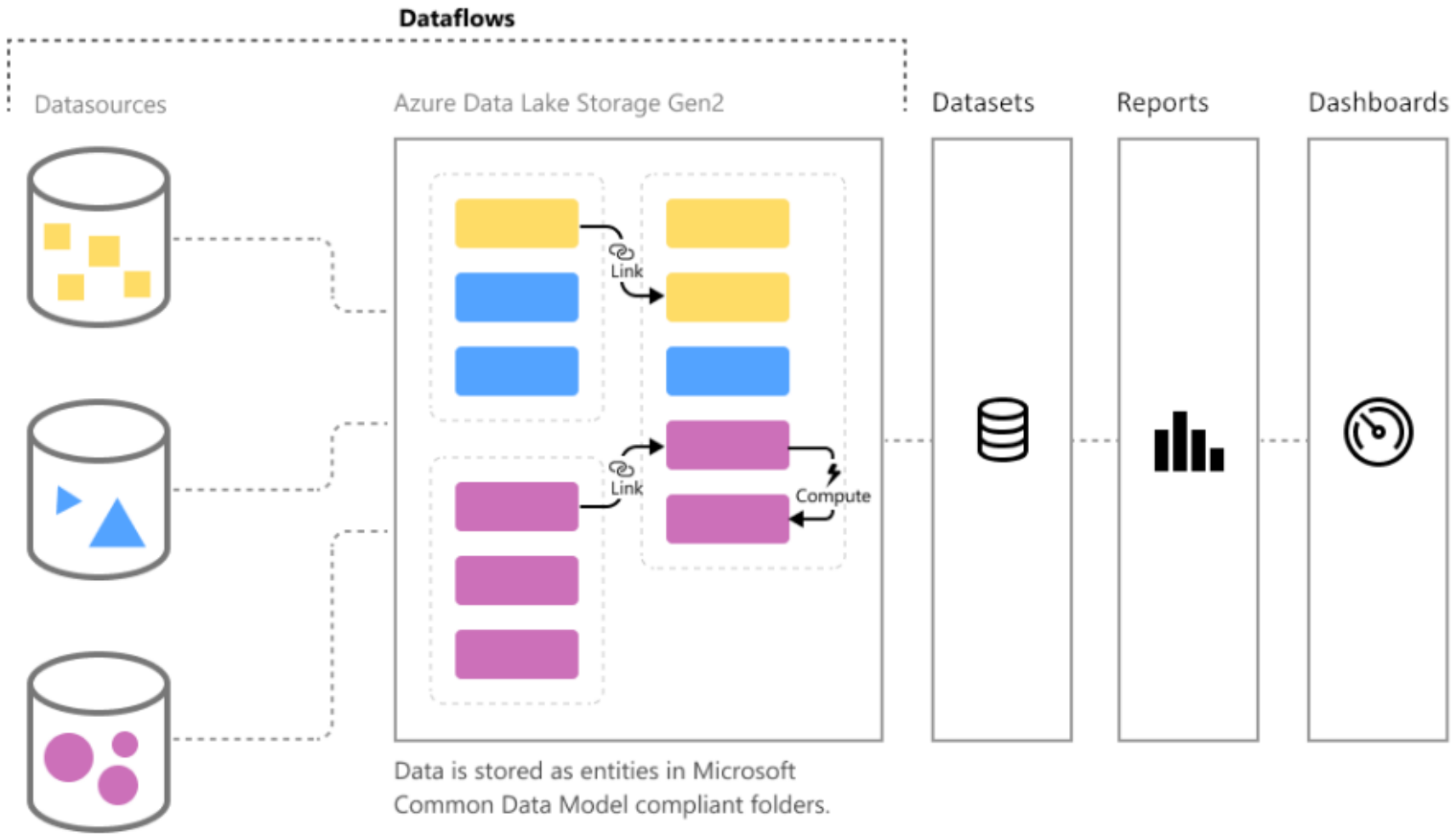
Agenda

- What are dataflows?
- Configuring, creating and consuming dataflows
- Why use dataflows?
- Linked and computed entities
- AI features in dataflows
- Dataflows and the Common Data Model
- Roadmap

What are dataflows?

- Self-service ETL tool in the Power BI cloud service
  - Not Azure Data Factory
  - Not Microsoft Flow
- Power Query in the cloud/in the browser plus a lot more
- They split the extraction/preparation away from datasets
- They allow for sharing of tables of data between datasets
- They **do not** allow sharing of dataset features like relationships and DAX calculations

The Big Picture



Licensing

- Not available in My Workspace or for Power BI Free users
- Most features available for Power BI Pro users in an app workspace
- Premium gives you:
  - Ability to handle larger data volumes
  - Better refresh performance
  - Linked and computed entities
  - Ability to use AI features to transform data (coming soon)

Concepts

- A **dataflow** exists within a workspace
- One dataflow contains one or more **entities**
- An entity is:
  - A table of data plus some metadata
  - The output of a Power Query M query or some other process

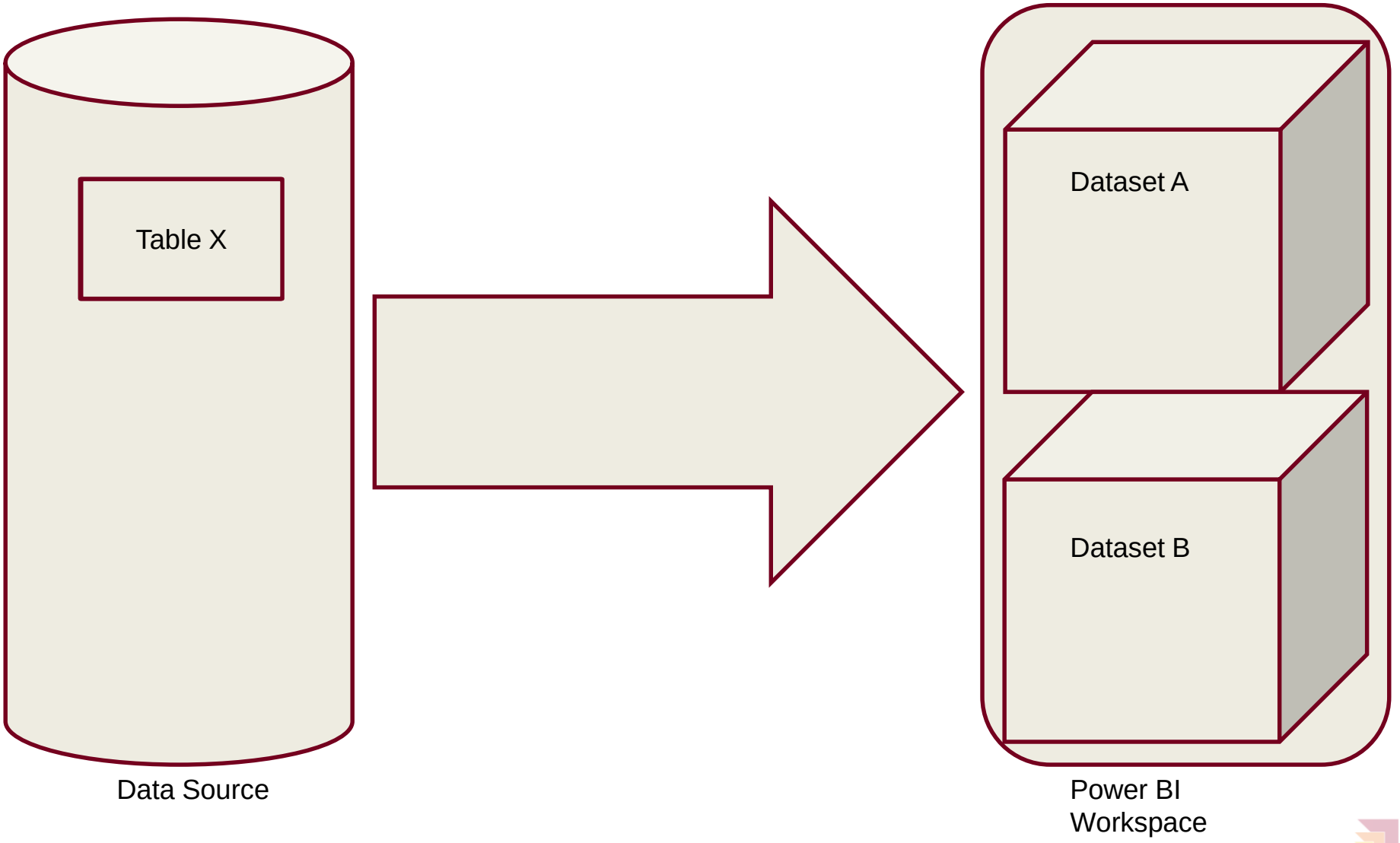


Table X
Data Source
Dataset A
Dataset B
Power BI
Workspace

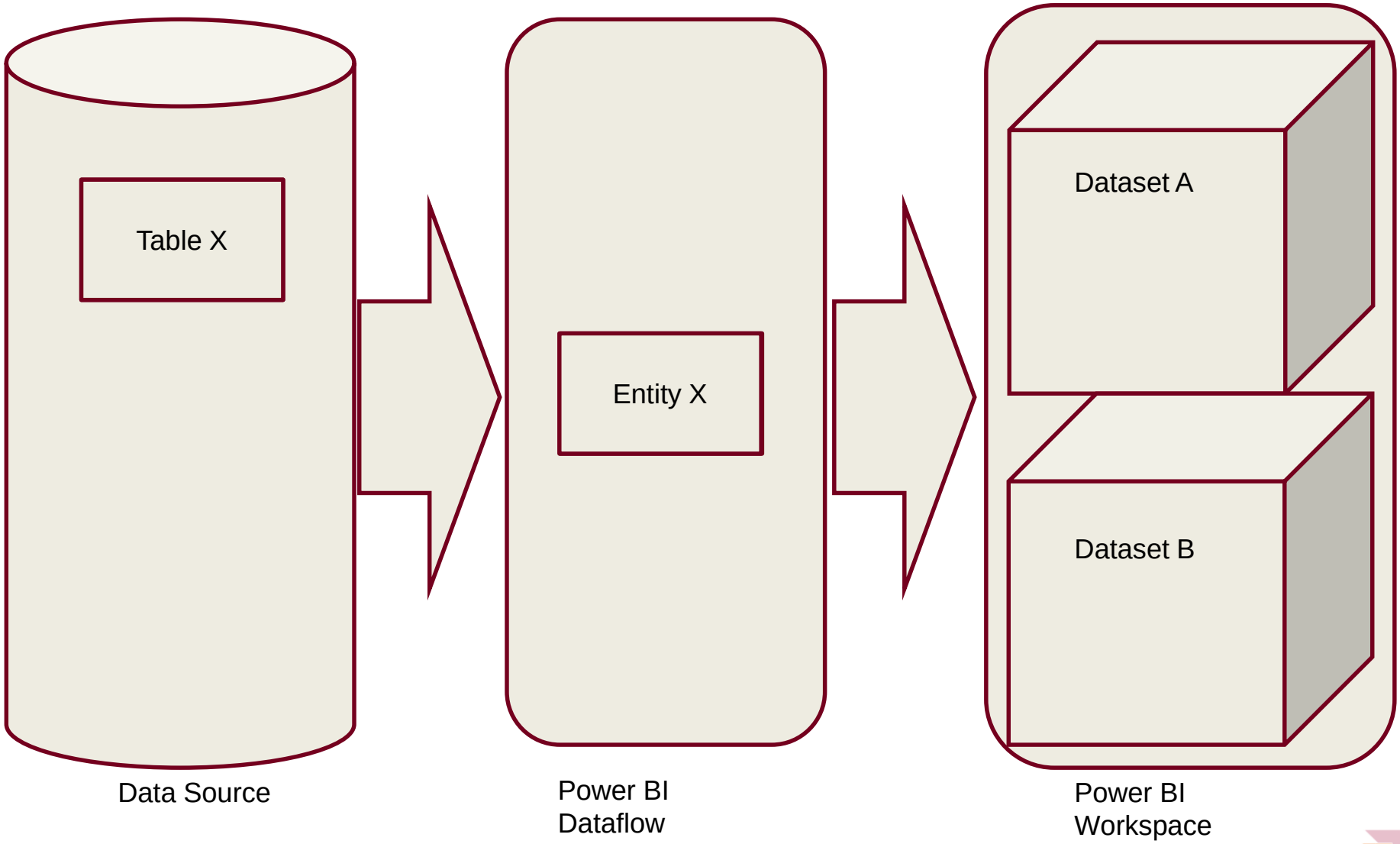


Table X
Entity X
Dataset A
Dataset B
Data Source
Power BI
Dataflow
Power BI
Workspace

Configuring dataflows

- Enable “Create and use dataflows” in the Admin Portal
- Dataflows are a workload in a Premium capacity
  - Therefore also needs to be enabled in the Capacity Settings section of the Admin Portal before you can use them on a Premium capacity
  - How much memory do you want to allocate?
- Link to bring-your-own Azure storage also needs to be configured in the Admin Portal
- Some features are only available in “New” workspaces

Creating dataflows

- Create a dataflow in a workspace
- Create entities in the dataflow using Power Query Online
  - Copying and pasting M code from Power BI Desktop into a new blank query may be a better idea for now
- Save and close
- Refresh

Consuming dataflows

- Entities can be consumed in Power BI Desktop by using the “Power BI dataflows” source
- You can make your own transformations as normal in the Power Query Editor in Power BI Desktop
- Datasets that use entities as a data source can be published to **any** workspace
  - Not just the workspace that contains the dataflow

Why use dataflows?

- **Not** a replacement for a data warehouse, but useful when:
  - There is no data warehouse in your organisation
  - Data warehouse does not contain the data you need
- Reduces overall data refresh time:
  - Extracting once and re-using multiple times means you only pay the performance price for the initial slow extract once
  - Reading data from a dataflow is fast, probably much faster than extracting data from the original source
  - Computed entities may be faster than referencing

Why use dataflows?

- Reduces load on/number of calls to source system
  - Eg when refresh could affect the performance of a line-of-business database
  - Eg when there is a limit on the number of calls to an API
- More consistency between datasets – less chance that different users will make different decisions when preparing data
- Share complex M queries that some users would not be able to write
- Share tables that have no source, eg Date dimensions generated in M

Linked entities (Premium only)

- Linked entities let you share data between
  - Different dataflows in the same Workspace
  - Different dataflows in different Workspaces
- **Do not** duplicate data from the source entity
- Specifically: use an existing entity in another workspace as a source
  - Uses the same M code a dataset uses to get data from an entity
- Linked entities are read-only
  - If you want to do further transformations you must create a computed entity
- Diagram view makes it easy to see usage of linked entities

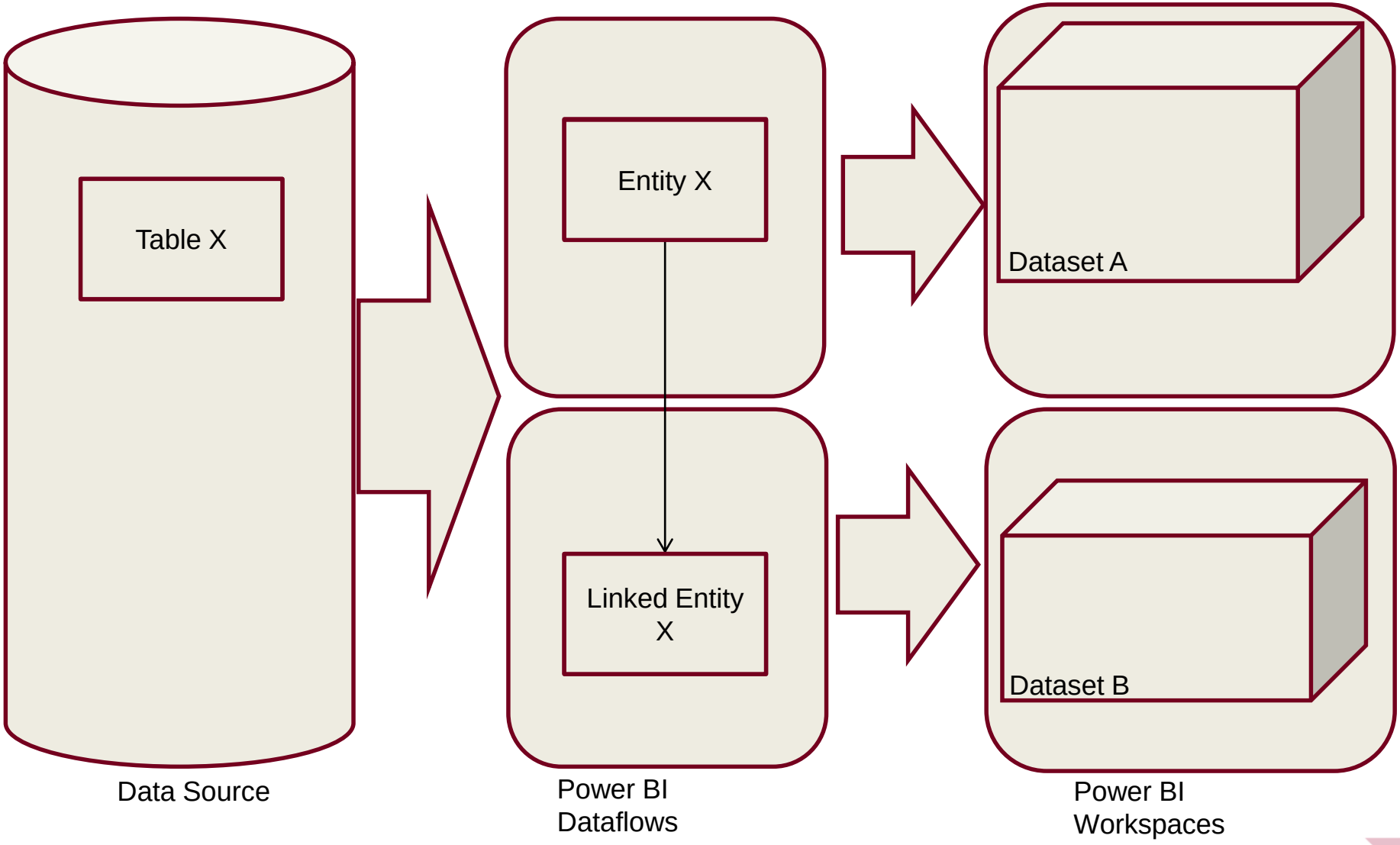


Table X
Data Source
Entity X
Linked Entity X
Power BI Dataflows
Dataset A
Dataset B
Power BI Workspaces

Computed Entities

- Computed entities let entities use other entities as data sources
- A bit like referencing – but uses the persisted output of the source entity as the input
  - Referencing is still possible between Entities in the same Dataflow – just turn off the “Enable load” option
- Useful if:
  - You are creating multiple entities within a dataflow from the same raw data, and don’t want to get data from the original data source more than once
  - You are hitting problems as a result of the Power Query engine’s habit of requesting data multiple times during a single query execution

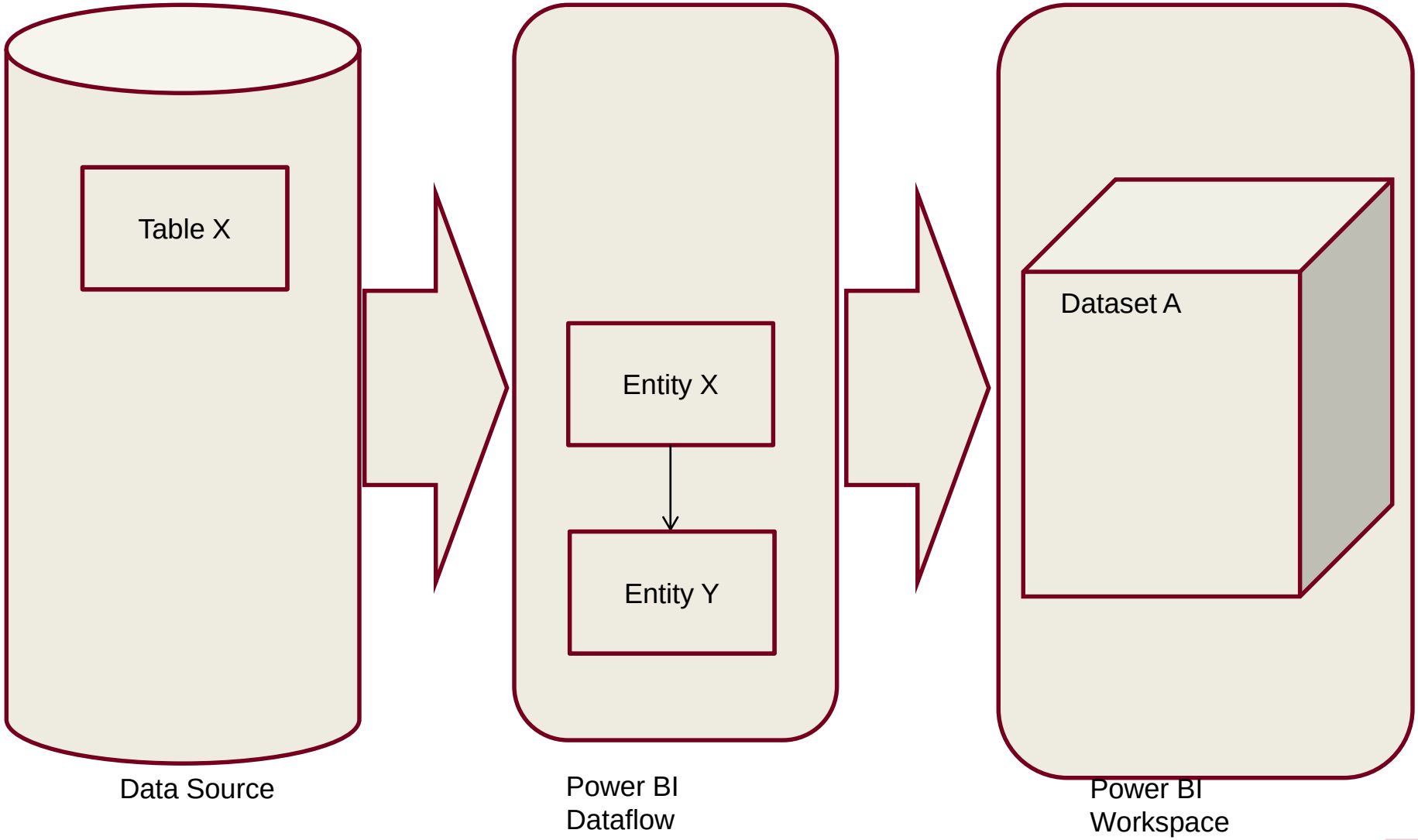
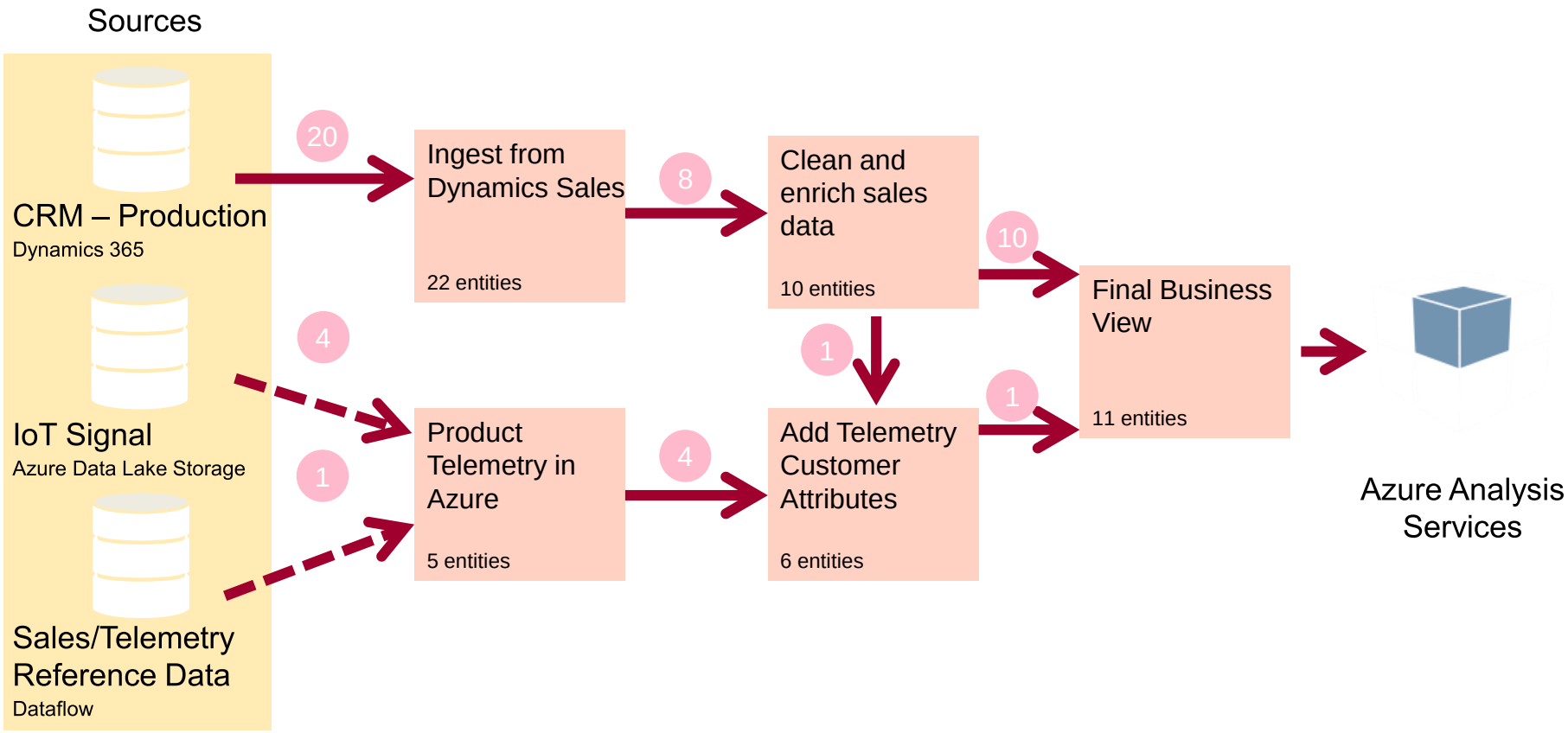


Table X
Data Source
Entity X
Entity Y
Power BI
Dataflow
Dataset A
Power BI
Workspace

Refresh

- Dataflows can be refreshed manually or on a schedule
- Entities in dataflows stored in a Premium capacity can use incremental refresh
  - Great for data warehouse style refresh scenarios
  - But how do I just add new data onto my existing data…?
- Linked entities in the same workspace are automatically refreshed when the source dataflow is refreshed
- Linked entities in different workspaces are not automatically refreshed when the source dataflow is refreshed

Sample Production Dataflow Design

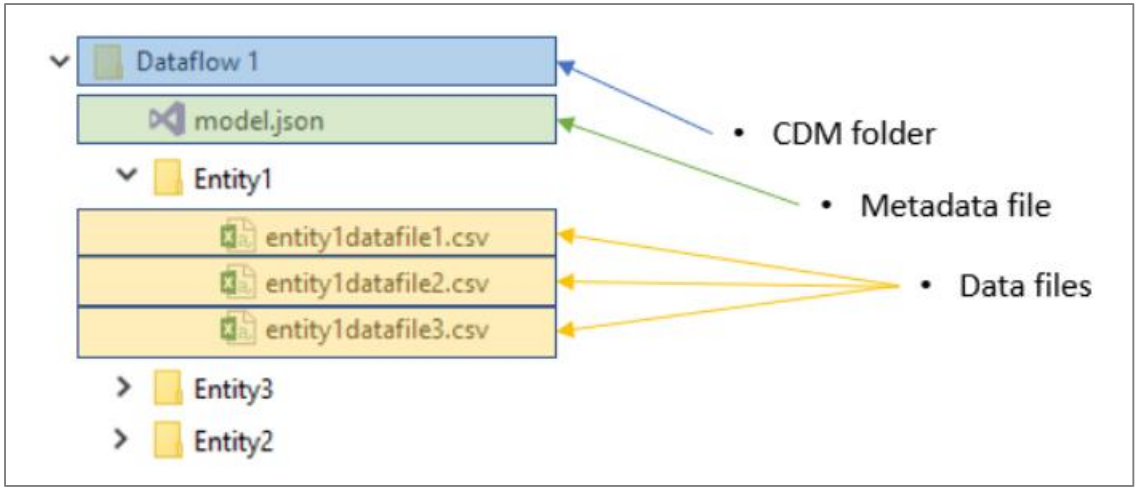


AI Features

- Power BI Premium users will soon get extra AI/machine learning features for data preparation
- Another separate workload inside your Premium capacity
- These include:
  - Call Cognitive Services functions (eg detect sentiment, language, tag images) as steps in a query
  - Creation and training of simple machine learning models (Binary Prediction, Classification, Regression, Forecasting)
  - Call machine learning models built in Azure Machine Learning

Storage

- Entities are stored in one or more CSV files
- There is also JSON metadata file for the dataflow
- The format is defined by the Common Data Model specification
- Everything is stored in Azure Data Lake Gen2 Storage
  - By default this is owned and managed by Power BI, so you don't see it
- You can also bring your own Azure storage, which allows for:
  - Larger data volumes
  - Access to this data by other services



Common Data Model integration

- The Common Data Model provides a:
  - Simple, consistent way of describing data
  - Common, extensible schema for business entities
- Dataflow output is stored in CDM format
- Dataflow entities can be mapped to CDM Entities
- Joint initiative between Microsoft, Adobe and SAP
- See https://github.com/Microsoft/CDM for more information

Common Data Model Integration

- Aim is make it easy to share data between multiple services
- Other applications and services are able to:
  - Read data from entities created by Power BI Dataflows
  - Output data to CDM folders which can then be attached as Dataflows by Power BI users
- Integration planned with Azure Data Factory, Azure SQL DW, Azure Machine Learning and Azure Databricks

User education

- Dataflows are for end-users, or at least power-users
  - But will these users plan ahead enough in order to use them?
- User education will be key:
  - Create some dataflows for them to use and learn from
  - Look at existing datasets for duplicated tables and help users to refactor to use dataflows
  - Look at dataset refresh times and see if dataflows might help
- Restricting access to data sources might encourage more use of dataflows?

Roadmap

- Power Query Online (as used by dataflows) will be functionally equivalent to Power Query in Power BI Desktop
- Performance and scalability enhancements for working with large data volumes
- Ability to refresh dataset after a dataflow it depends on has refreshed
- Ability to create dataflows though the API
- Lots of interesting features in the CDM specification…?