# Extracting and Preparing Data using Dataflows

Agenda

- Understanding Dataflow Architecture

- Creating and Consuming Dataflows

- Creating a Star Schema

- Importing and Exporting Dataflows

- Using Premium Dataflow Features

Motivation for Dataflows

- Dataflows provide self-service ETL that scales
  - Intuitive and familiar authoring using Power Query
  - Seamless integration with Power BI Desktop
  - Dataflow storage designed for interoperability & big data
 - Data storage format based on Common Data Model
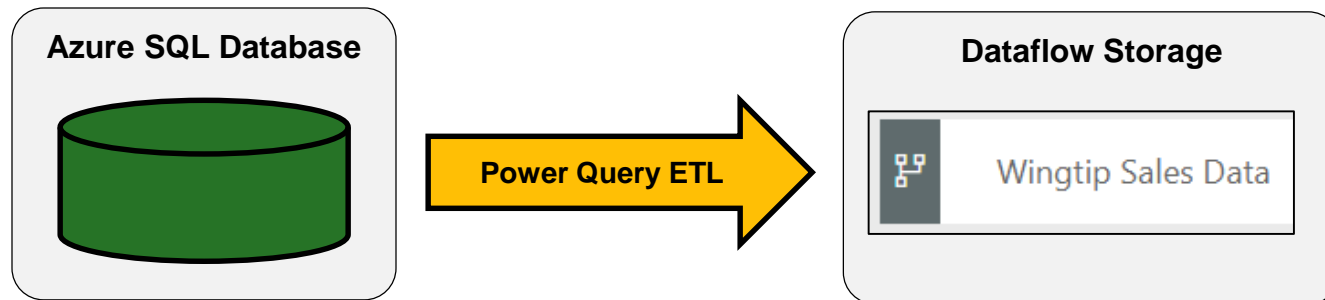
Dataflow Architecture

- Dataflows use Azure Data Lake Gen2 Storage
  - Storage designed to meet requirements of big data
  - Dataflows serialized in format defined by Common Data Model (CDM)
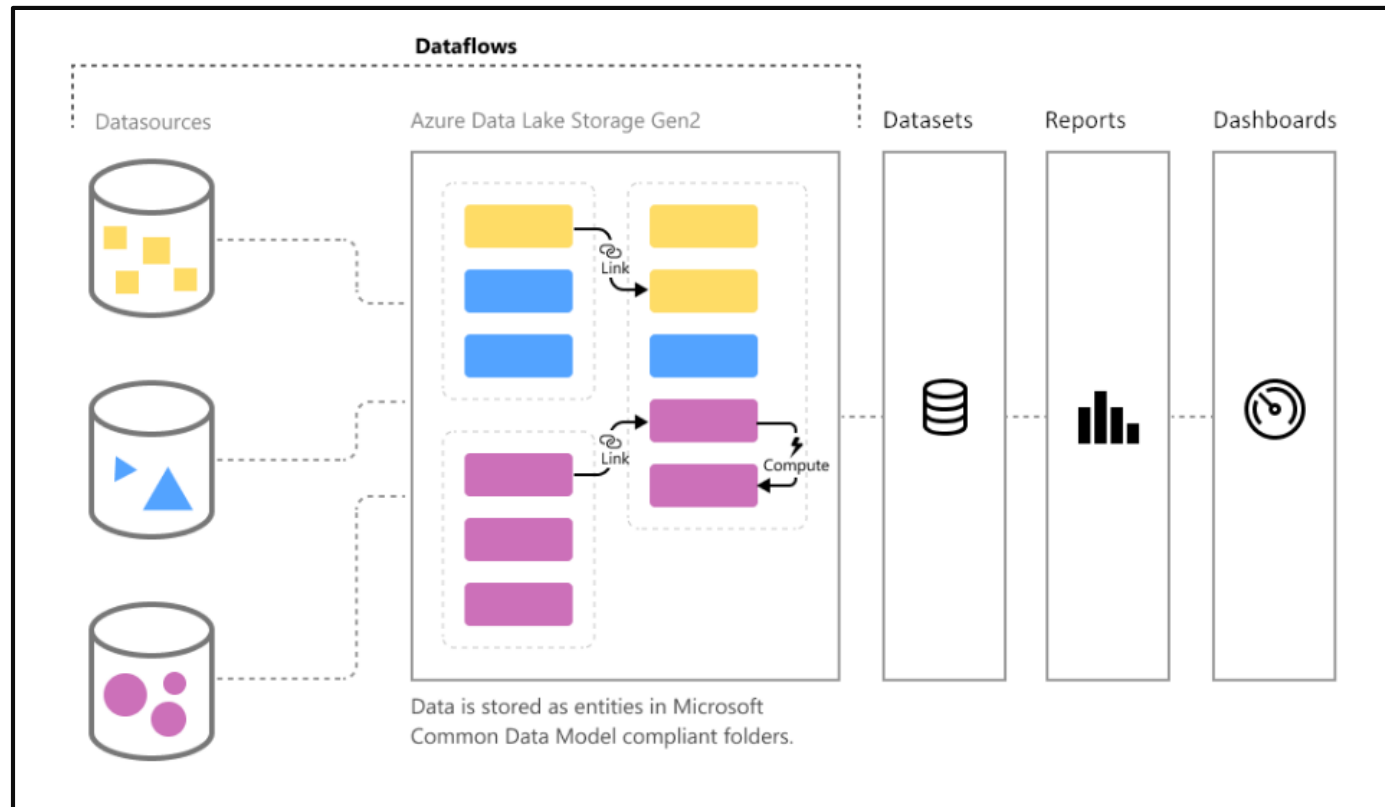
Dataflow Storage Details

- Serialization format is defined by the Common Data Model specification
  - Dataflow metadata stored in **model.json** file
  - Dataflow data rows stored in CSV files
  - By default, Power BI manages dataflow storage behind the scenes
- Dataflows can be configured to write data to Azure storage account
  - Allows access to dataflows by service other than Power BI
  - Allows with an organization to store larger data volumes
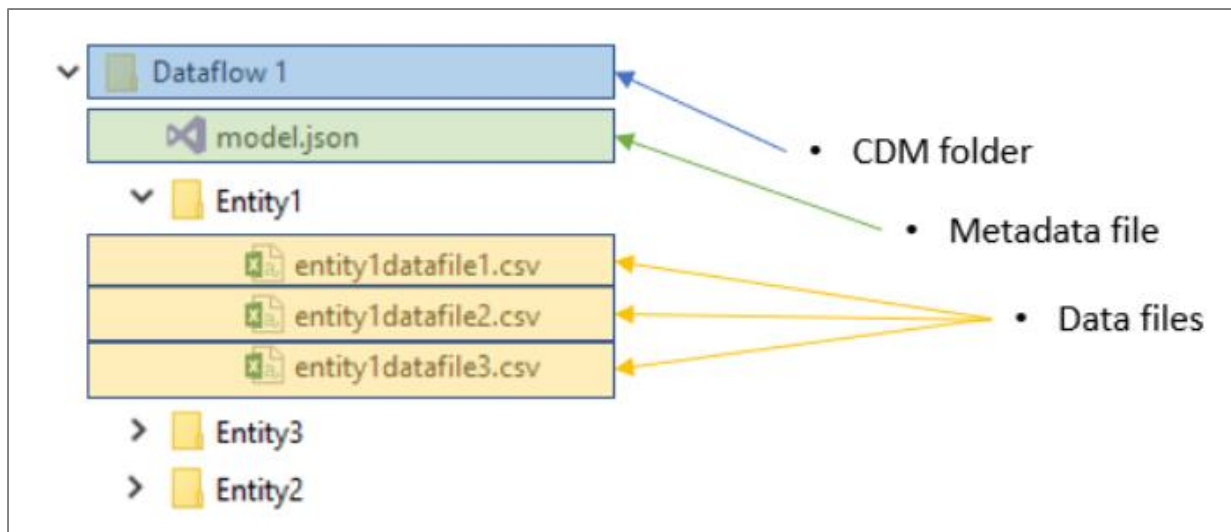
Common Data Model Metadata

- Dataflow output is stored in CDM format
  - model.json file contains metadata about entities
  - model.json file contains M code for queries

Designing Power BI Solutions with Dataflows

- Dataflows used to collect all data
  - Power BI Desktop projects import data from dataflows
  - Complex ETL work no longer required in Power BI Desktop projects

Report	Report	Dashboard	Report	Dashboard	Report

Dataset	Dataset	Dataset

Dataflow

Raw Data
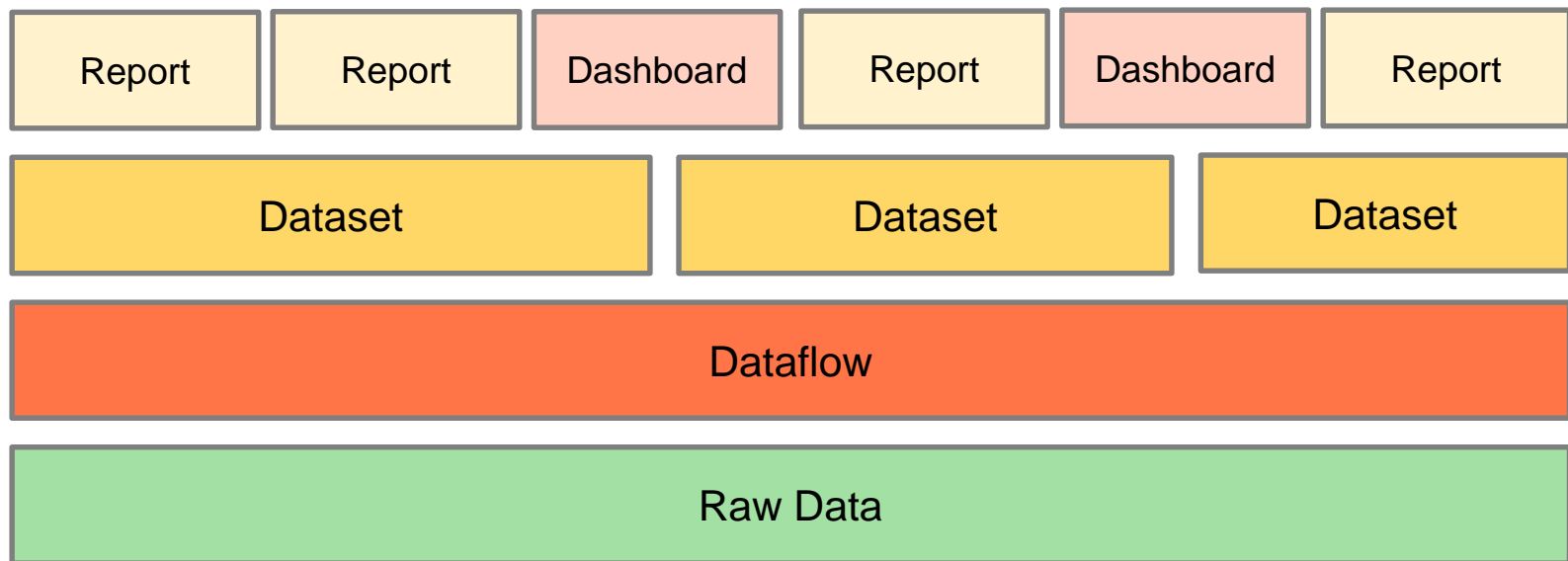
Dataflow Benefits

- Replaces other ETL tools *(e.g. Azure Data Factory, Power Automate)*
- Decouples ETL work from datasets in PBIX projects
- Enable sharing of source tables of data between datasets
- Reduces number of queries on live data sources
- Eliminates need to connect user computers directly to data source
- Centralizes efforts to clean and prepare data
- Share tables that have no source (calendar tables)

Dataflow Disadvantages

- Adds extra complexity

- Data must be refreshed in 2 seperate phases

- Does not support data modelling features of DAX

- Some dataflow features require dedicated capacities

Licensing with Dataflows

- Dataflow creation requires Power BI Pro
  - Dataflows can only be created in app workspaces
  - Dataflows cannot be created in personal workspaces

- Dedicated capacity adds extra dataflow features
  - Ability to handle larger data volumes
  - Better refresh performance
  - Linked and computed entities
  - AI features to transform data
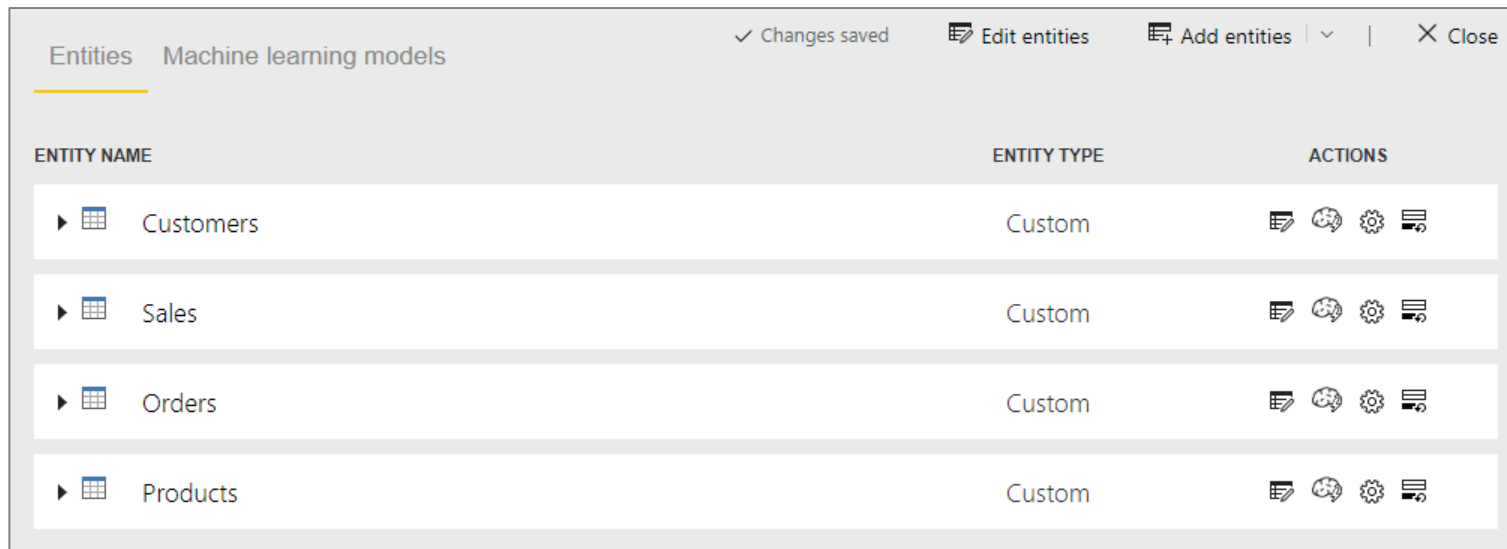
Agenda

- ✓ Understanding Dataflow Architecture
- ➢ Creating and Consuming Dataflows
- • Creating a Star Schema
- • Importing and Exporting Dataflows
- • Using Premium Dataflow Features

Dataflow Entities

- A **dataflow** exists within a workspace
  - Dataflow contains one or more **entities**
  - Entity is a table with well-defined schema
  - Entity data populated by running a query (M code)

Power Query in the Browser

- Editing experience familiar to Power Query users

Refreshing a Dataflow

- Dataflow can be refreshed on demand or scheduled

- Dataflow displays last refresh time

Dataflows and Power BI Desktop

- ## Power BI Desktop project without dataflow
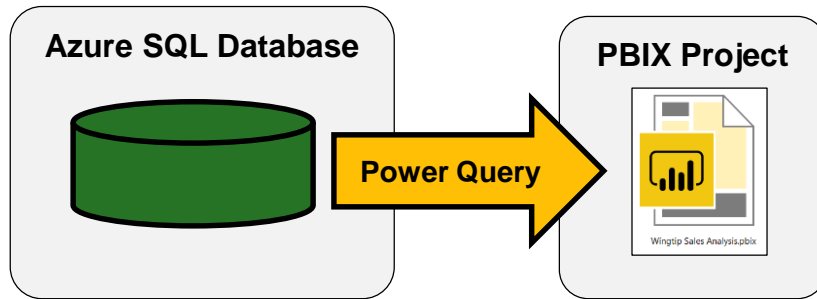
- ## Power BI Desktop project with dataflow
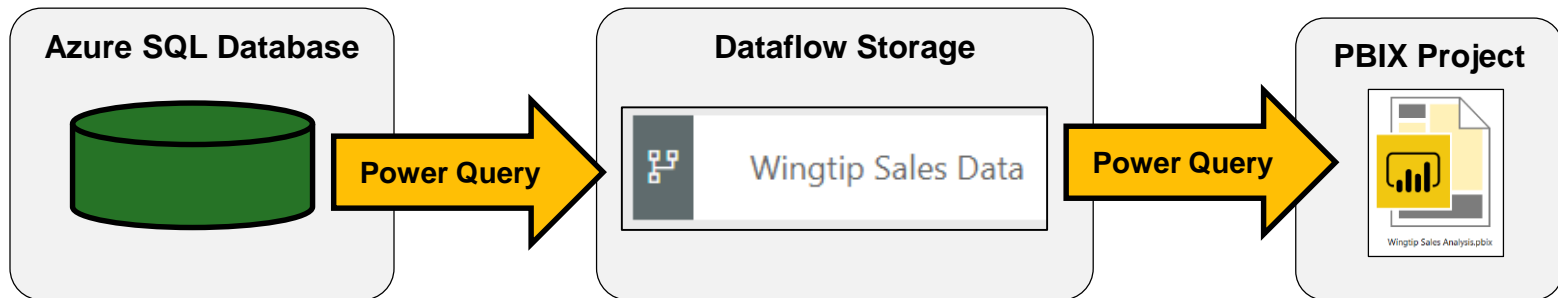
Consuming Dataflows

- Dataflow entities consumed by Power BI Desktop
  - Use **Power BI dataflows** data source
  - Dataset with dataflow can be published to any app workspace
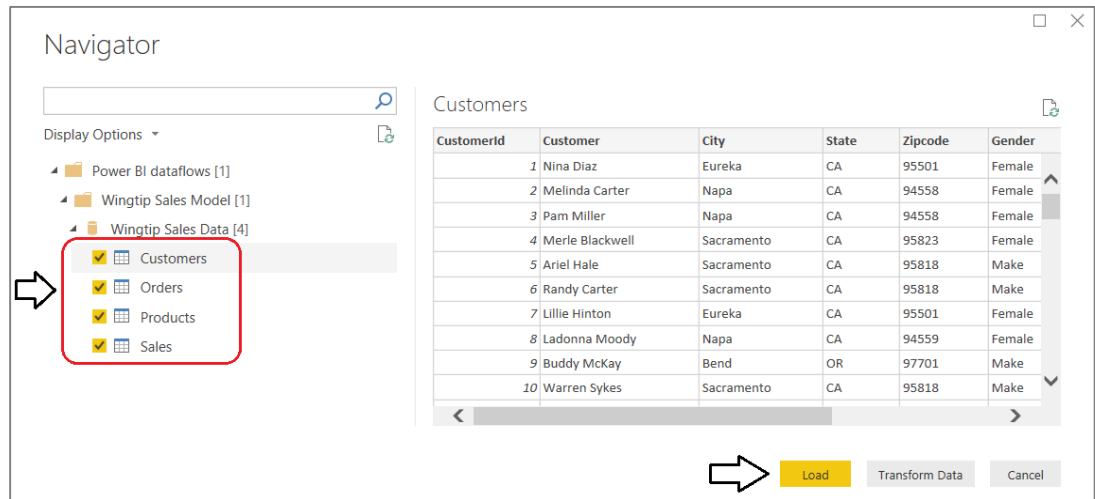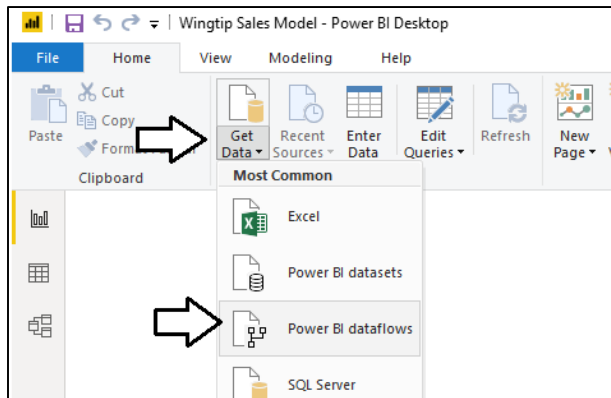
Dataflow Provides Starting Point for Data Model
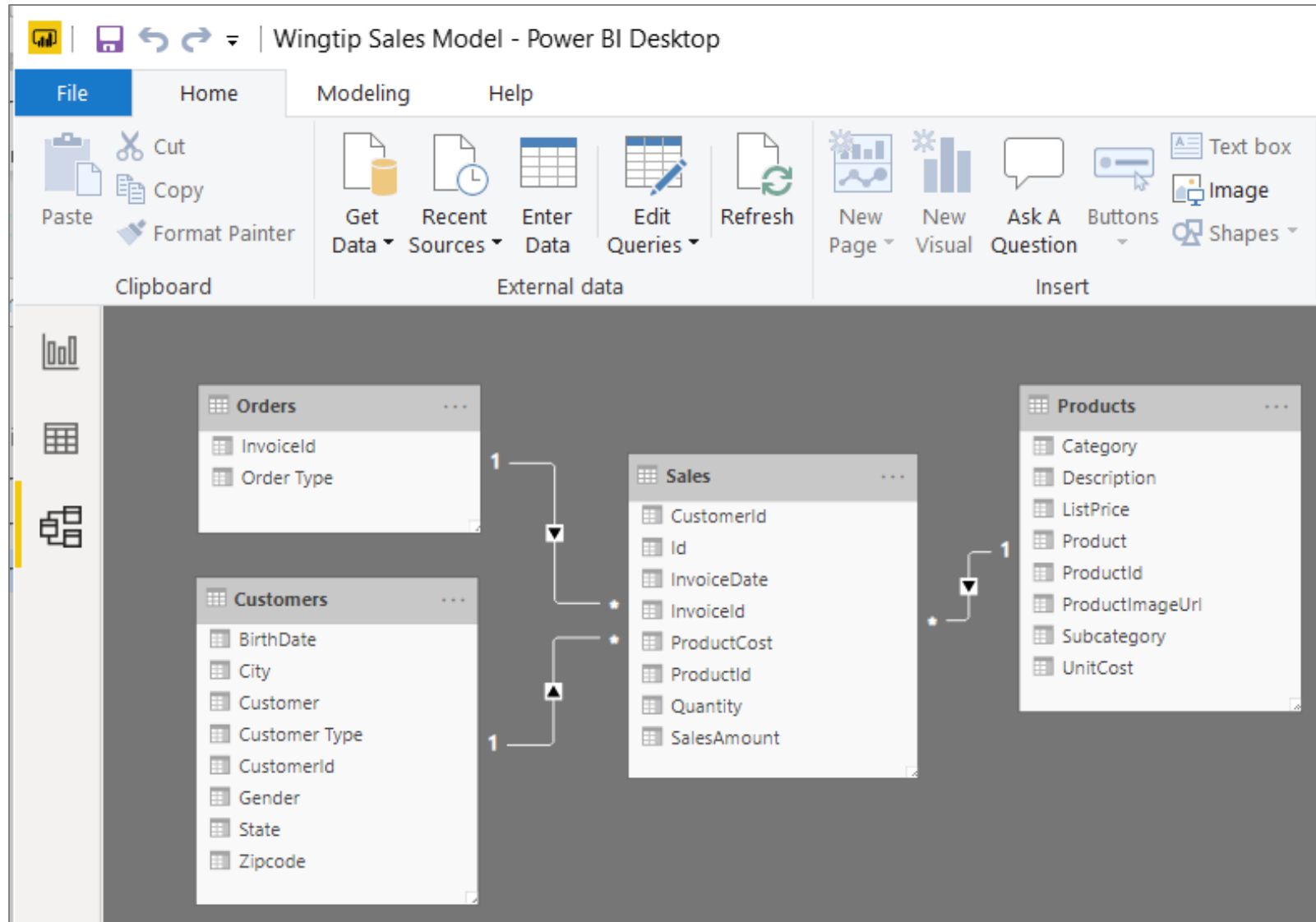
Agenda

- ✓ Understanding Dataflow Architecture
- ✓ Creating and Consuming Dataflows
- ➢ Creating a Star Schema
- • Importing and Exporting Dataflows
- • Using Premium Dataflow Features

Sample OLTP Database: WingtipSalesDB

- Online Transaction Processing (OLTP) System
  - Used for real-time data access and transaction-based data entry
  - Optimized for faster transactions (e.g. inserts, updates & deletes)
  - Tables normalized to reduce/eliminate redundancies
  - Table schemas can be hard for business users to understand
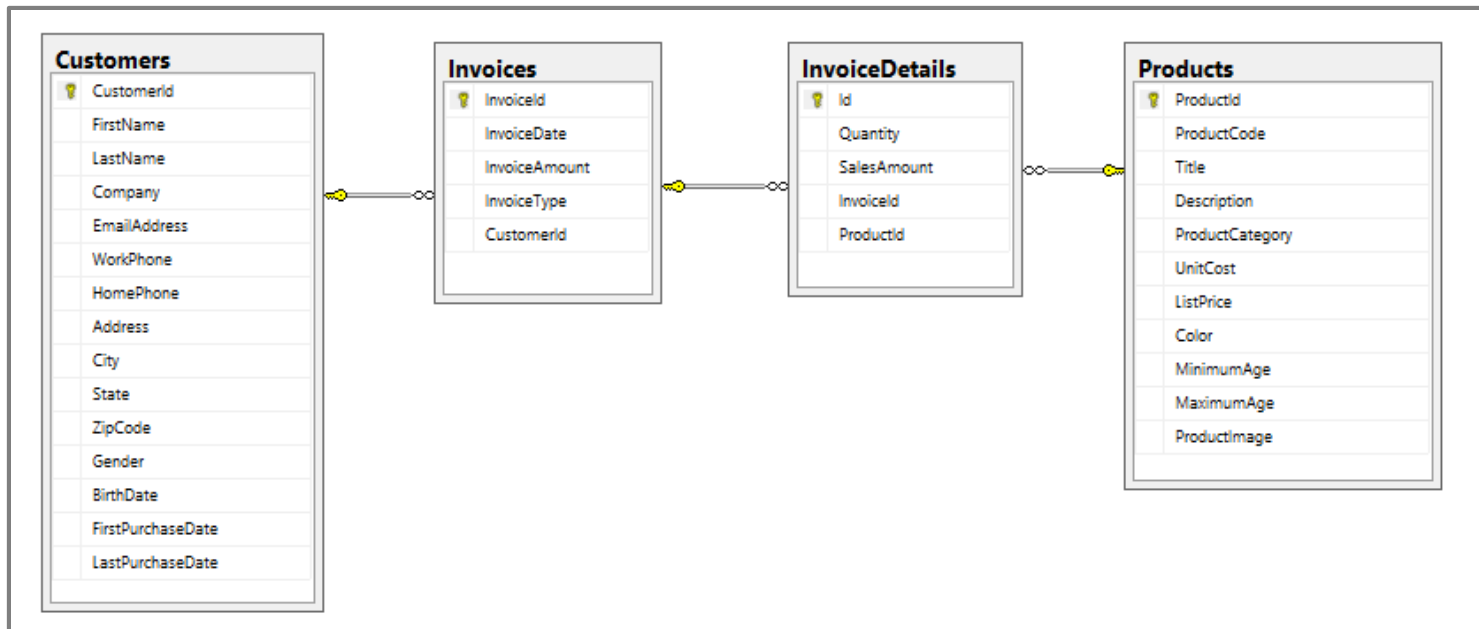
Data Modeling using a Star Schema

- OLAP Modeling often based on Star Schema
  - Tables defined as fact tables or dimension tables
  - Fact tables related to dimension table using 1-to-many relationships
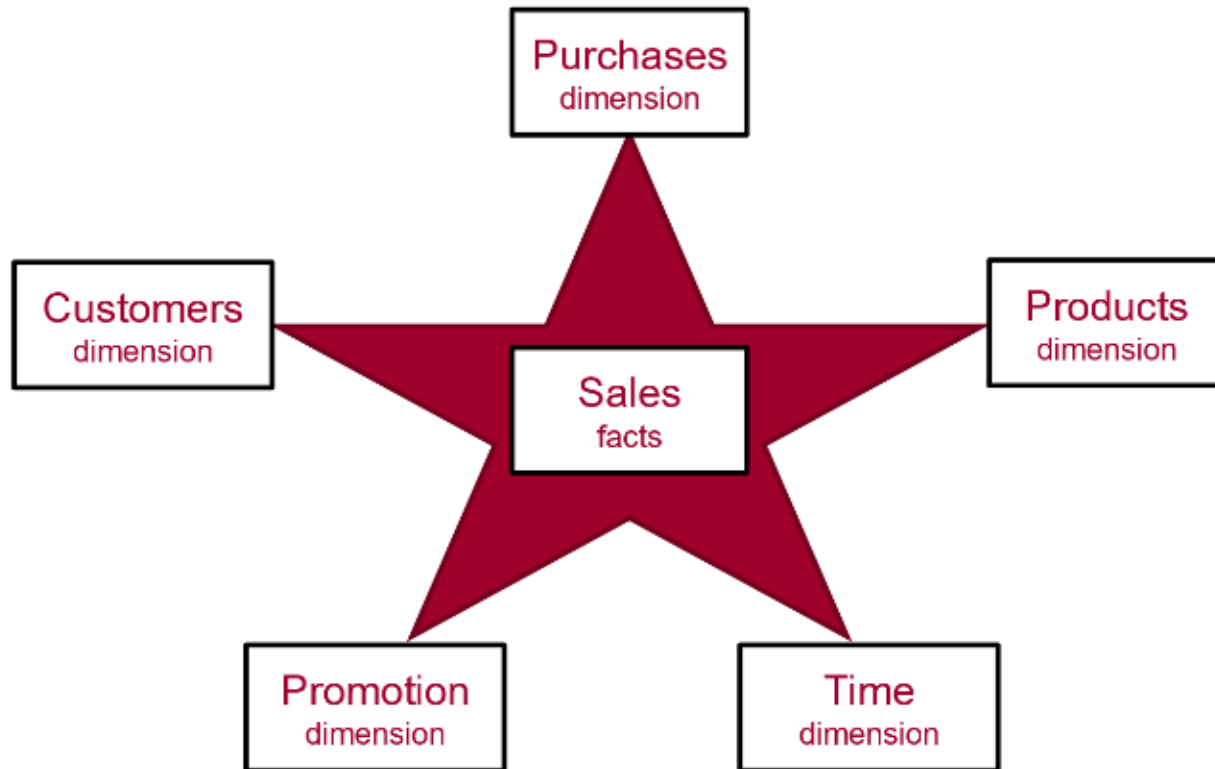
Designing Queries to Build a Star Schema

- Converts OLTP Data Model to OLAP Data Model
  - Sales table is modeled as a OLAP Fact Table
  - Other tables are modeled as OLAP Dimension tables
  - Requires pulling CustomerId column into Sales table
  - All dimension tables should be directly related to fact table
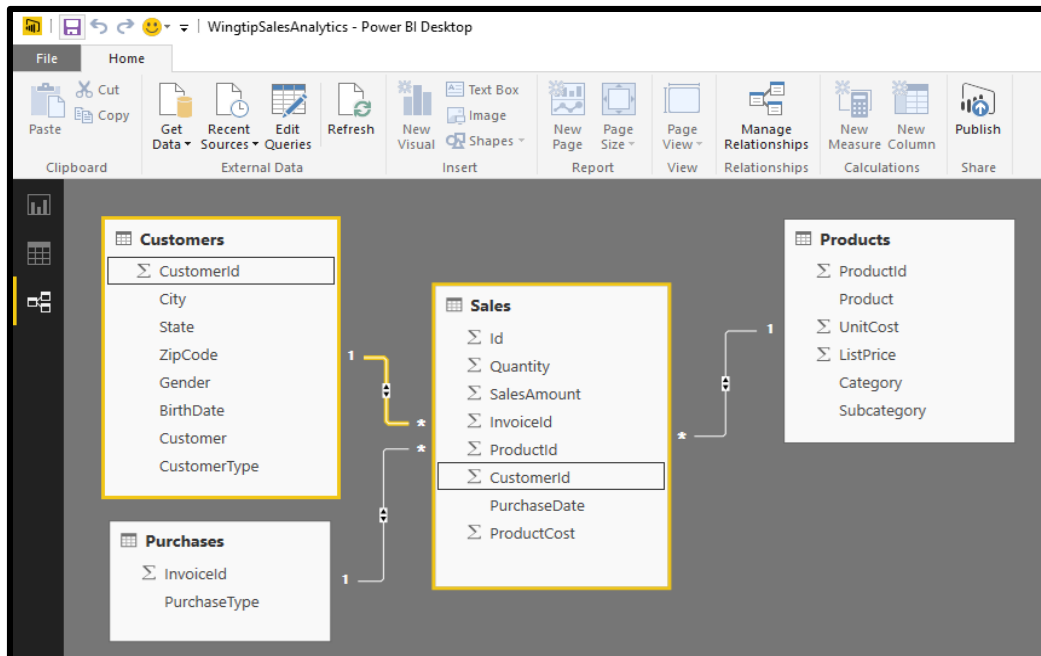
Agenda

- ✓ Understanding Dataflow Architecture
- ✓ Creating and Consuming Dataflows
- ✓ Creating a Star Schema
- ➢ Importing and Exporting Dataflows
- • Using Premium Dataflow Features

Importing and Exporting Dataflows

- Dataflow can be exported as model.json
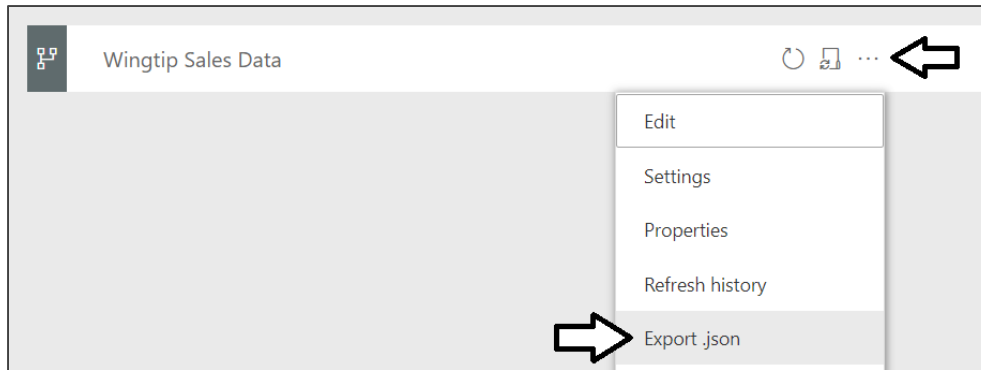
- model.json can be imported to create new dataflow
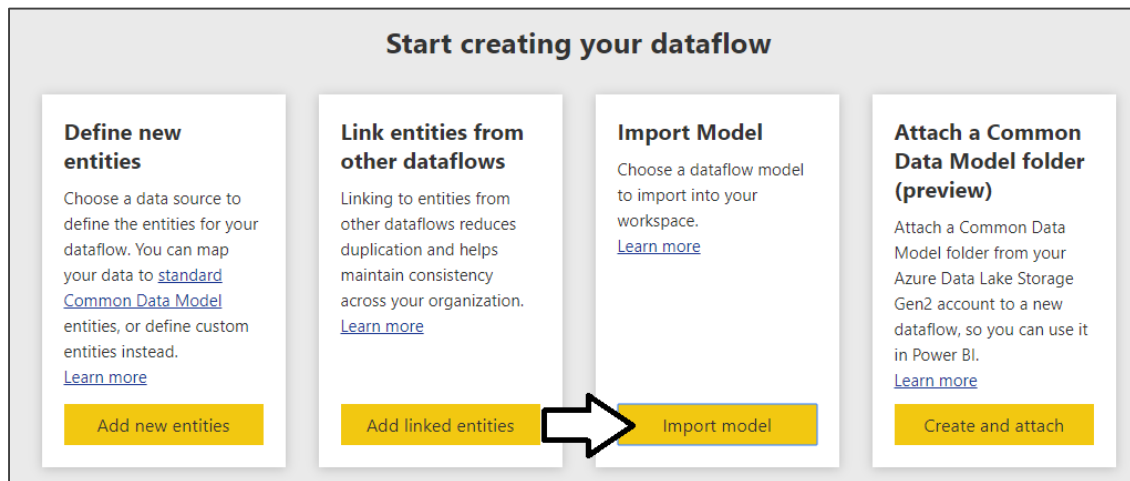
Agenda

- ✓ Understanding Dataflow Architecture
- ✓ Creating and Consuming Dataflows
- ✓ Creating a Star Schema
- ✓ Importing and Exporting Dataflows
- ➤ Using Premium Dataflow Features

Dataflow Premium Features

- Linked entities

- Computed entities

- Artificial intelligence (AI) features

- Incremental refresh

- Parallel execution of transforms

Linked Entities

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

Computed Entities

- Computed entities built on to of other entities
  - Allows entities use other entities as data sources
  - Computed entity uses the persisted output of the source entity
- Useful scenarios
  - 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

AI Features

- Dataflows in Power BI Premium includes AI features
  - Call Cognitive Services functions
  - Detect sentiment, language, tag images
  - Extract key phrases

Incremental Refresh

- Incremental refresh can be configured for entities
  - Very helpful to speed up refresh time for large tables

Incremental refresh settings

Sales

Incremental refresh updates only data that's changed, to speed refresh, reduce capacity usage, and store historic data. Learn more

On

Choose a DateTime field to filter by

Timestamp

Store rows from the past

| 4 | Years |

Refresh rows from the past

| 7 | Days |

Summary

- Understanding Dataflow Architecture

- Creating and Consuming Dataflows

- Creating a Star Schema

- Importing and Exporting Dataflows

- Using Premium Dataflow Features