

Dataflow

Chi sono

- Consulente e formatore in ambito business intelligence, business analytics e data mining
- Dal 2017 mi occupo della modern data warehouse con prodotti Azure: Synapse, Azure Data Factory, Stream Analytics, Data Lake
- Dal 2002 le attività principali sono legate alla progettazione di data warehouse relazionale e alla progettazione multidimensionale con strumenti Microsoft.
- Docente all'Università di Pordenone nel corso Architetture Big Data e DWH: Tecniche di modellazione del dato
- Community Lead di 1nn0va (www.innovazionefvg.net)
- MCP, MCSA, MCSE, MCT SQL Server
- dal 2014 MVP per SQL Server e relatore in diverse conferenze sul tema.
 - info@marcopozzan.it
 - [@marcopozzan.it](https://www.instagram.com/marcopozzan)
 - www.marcopozzan.it
 - <http://www.scoop.it/u/marco-pozzan>
 - <http://paper.li/marcopozzan/1422524394>



Microsoft
CERTIFIED
Trainer

Microsoft
CERTIFIED
Professional

Microsoft
CERTIFIED
Solutions Associate
SQL Server 2012/2014

Perchè concentrarsi tanto sulla preparazione dei dati

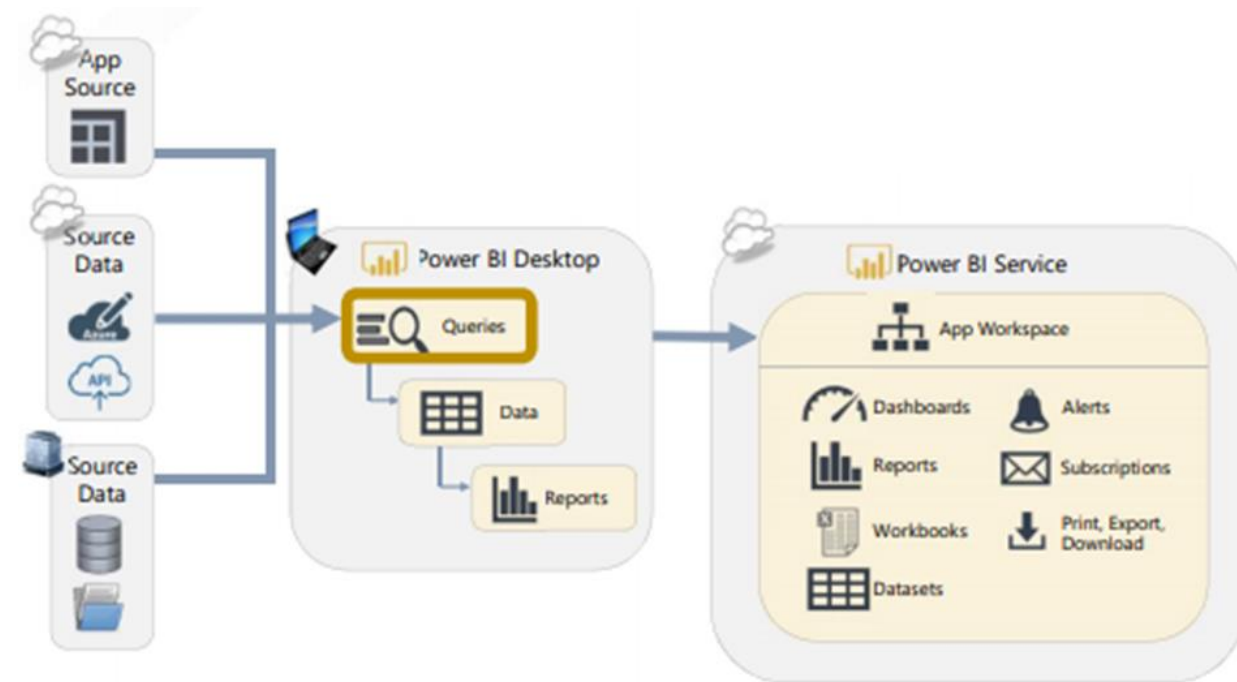
"Analysts spend up to 80% of their time on data preparation delaying the time to analysis and decision making."

- Gartner

Evoluzione del caricamento dati

Stato prima dei dataflow: Self-Service BI senza dataflow

- **Pro:**
 - Si ha una completa libertà nella self-service BI
 - SME (Small and Medium Size Enterprises) si gestiscono la pulizia dei dati
- **Limitazioni:**
 - Query Editor è confinato ad un singolo PBIX File
 - Altri tool non possono accedere all'output del query editor



Evoluzione del caricamento dati

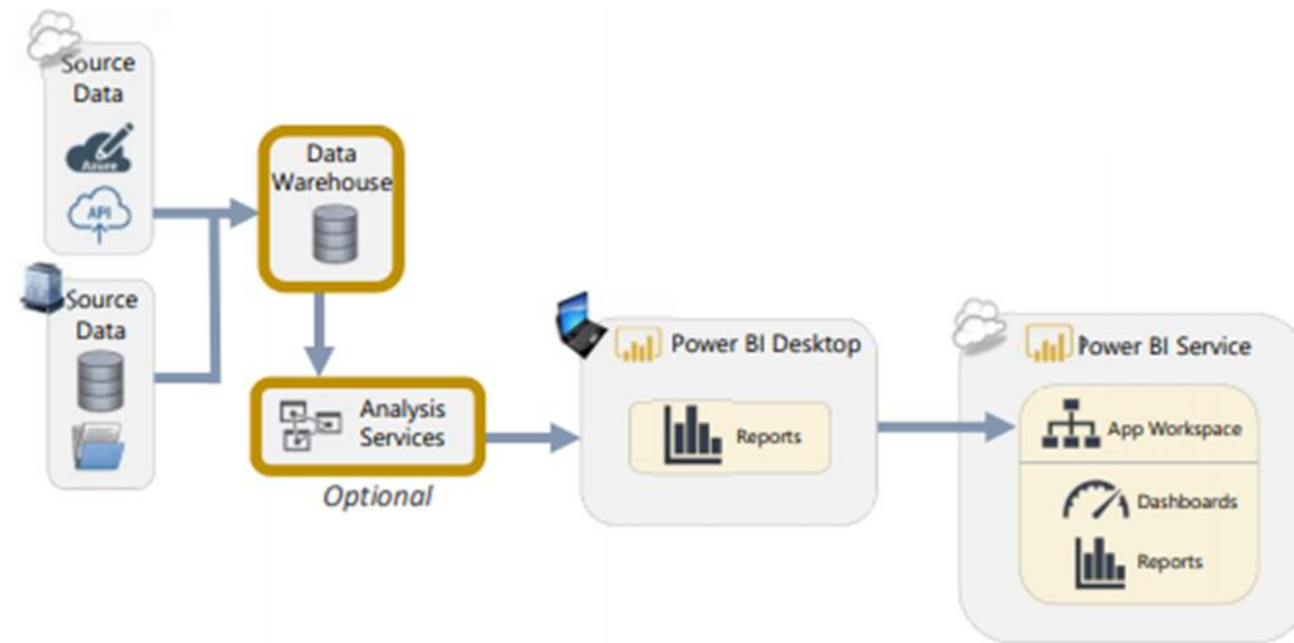
Stato prima dei dataflow: Corporate BI senza dataflow

- **Pro:**

- Si ha la centralizzazione della distribuzione del dato per la corporate BI

- **Limitazioni:**

- DWH e livello semantico non incontrano tutti i bisogni
- I grandi DWH non possono reagire velocemente alle esigenze
- Richiede dei data engineering

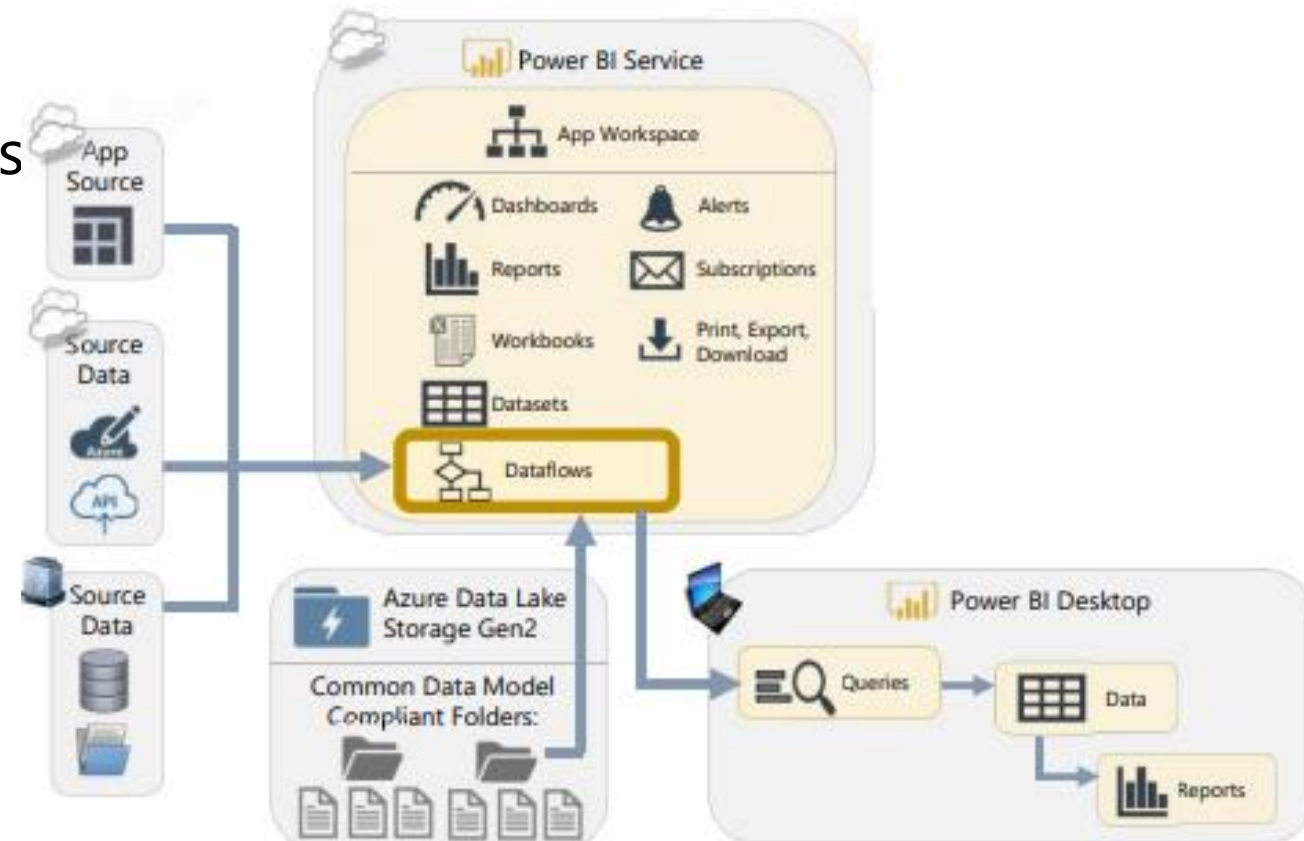


Evoluzione del caricamento dati

Stato prima dei dataflow: Corporate BI con dataflow

- **Pro:**

- La preparazione del dato è familiare attraverso un tool facile per il business
- La preparazione dei dati è messa a disposizione per molti dataset



Caratteristiche dei dataflow

Compatibilità:

- Preparazione dei dati self-service che può essere riutilizzata
- Riutilizzo dei dataflow da diversi dataset

Fatto per:

- Analisti responsabili per l'acquisizione dei dati e della loro pulizia che vogliono mantenere uniche queste attività
- Analisti che vogliono usare dati preparati da altri colleghi

Scopi:

- Promuovono la consistenza del dato
- Riducono i costi, il tempo e gli expertise richieste

Scenari dei dataflow



Standardizzazione e riusabilità dei dati: I dati sono abbastanza preziosi da avere molti casi d'uso per molti dataset, diversi tipi di analisi, molti tipi di app



Pre-Processamento: Elaborare set di dati più grandi che superano le risorse disponibili laptop locale o Power BI Desktop



Stage dei dati: Fornire dati per i modellisti di dati di Power BI per completare la preparazione dei modelli

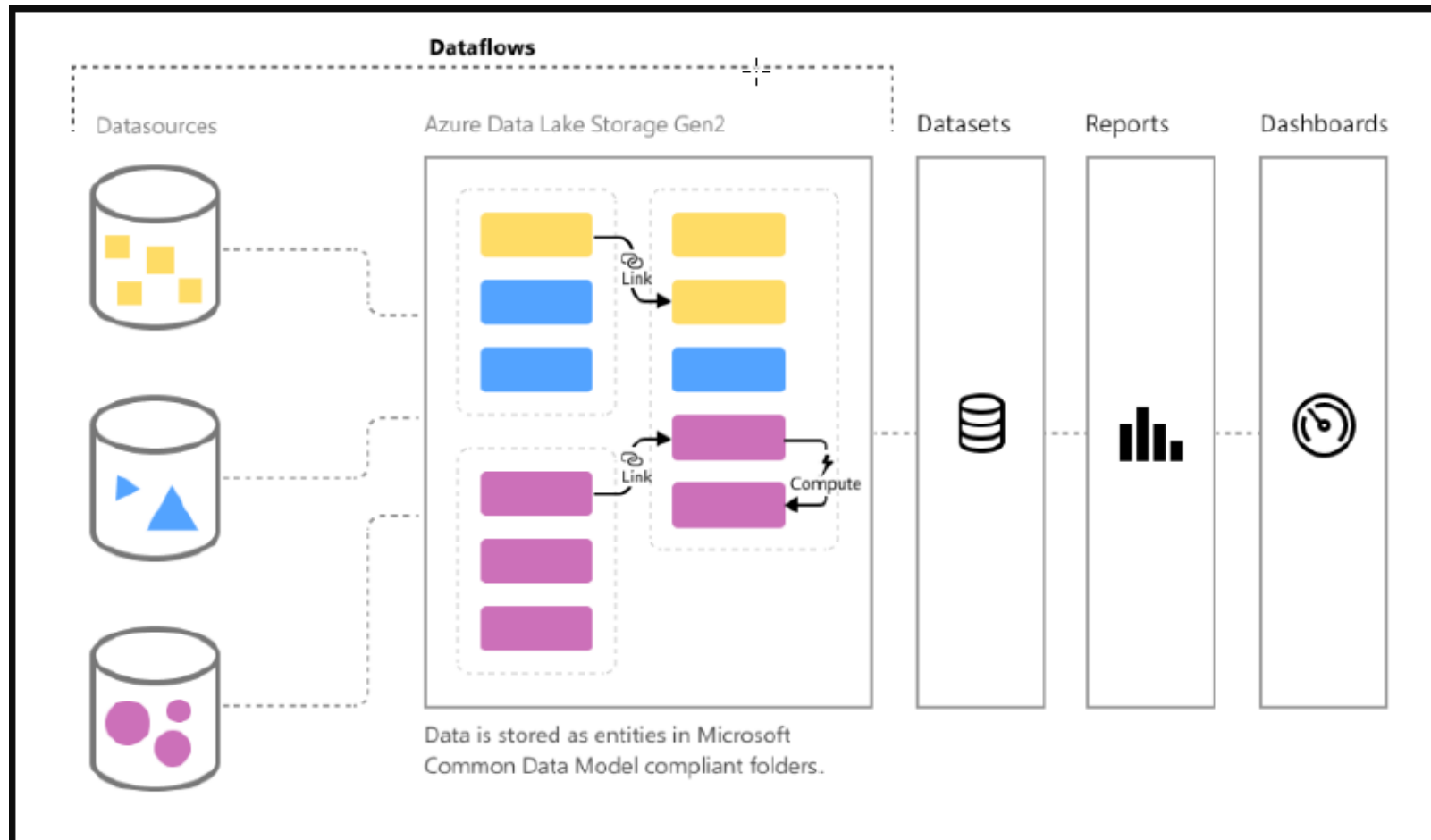


Ridurre il caricamento sui sistemi sorgenti: Ridurre al minimo il numero di query inviate al sistema di origine

Architettura dataflow

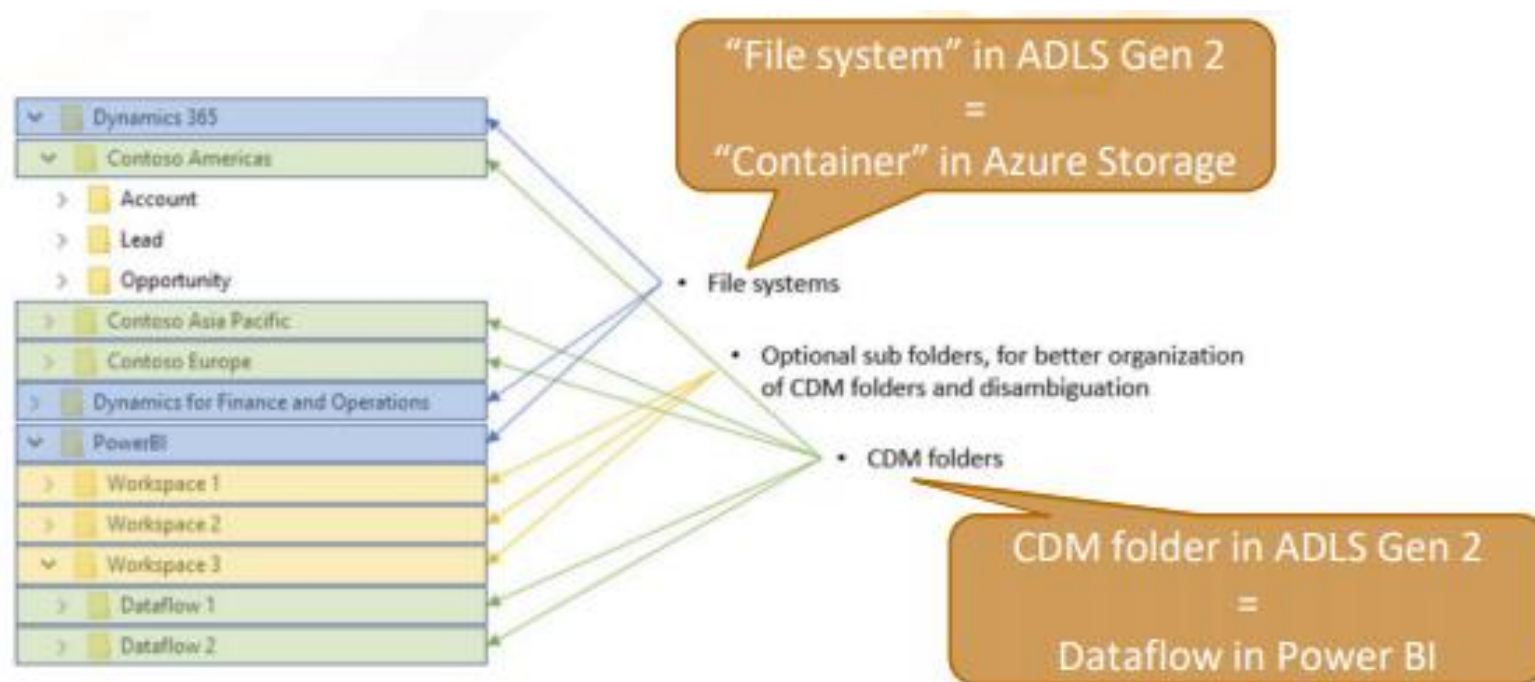
I flussi di dati usano l'archiviazione di Azure Data Lake Gen2

- Archiviazione progettata per soddisfare le esigenze dei big data
- Dataflow serializzati nel formato definito dal Common Data Model (CDM)



Dettagli sullo storage dataflow

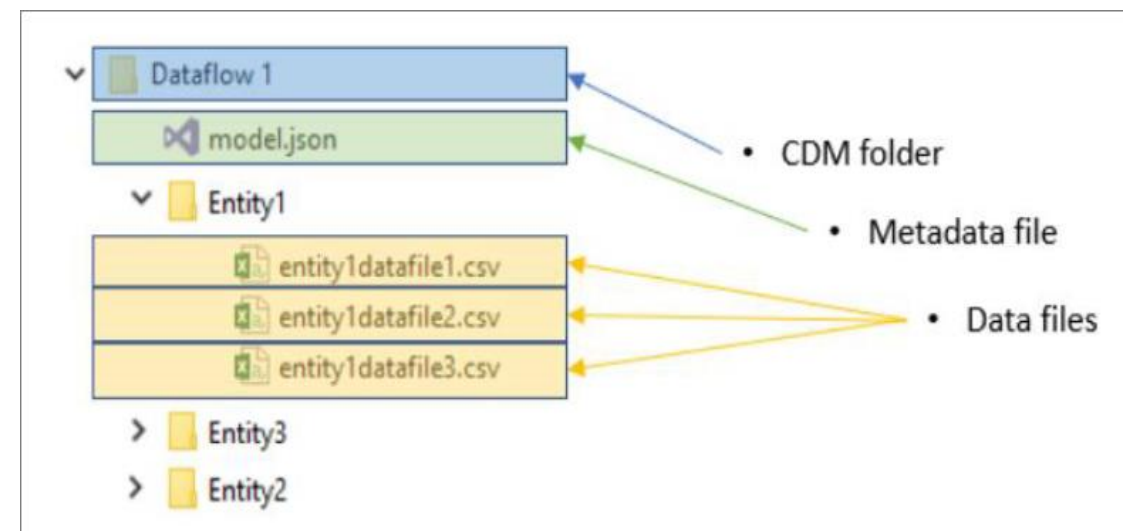
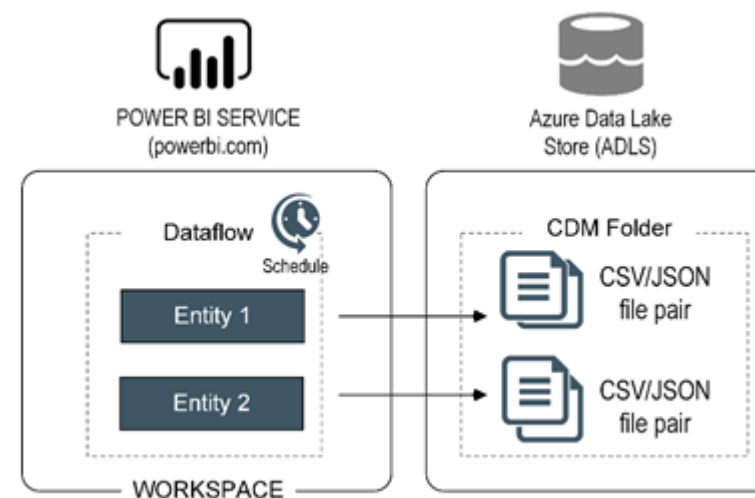
Ci sarà una cartella nel data lake conforme alle specifiche **CDM**, ben definito e standardizzato con strutture di metadati autodescrittive



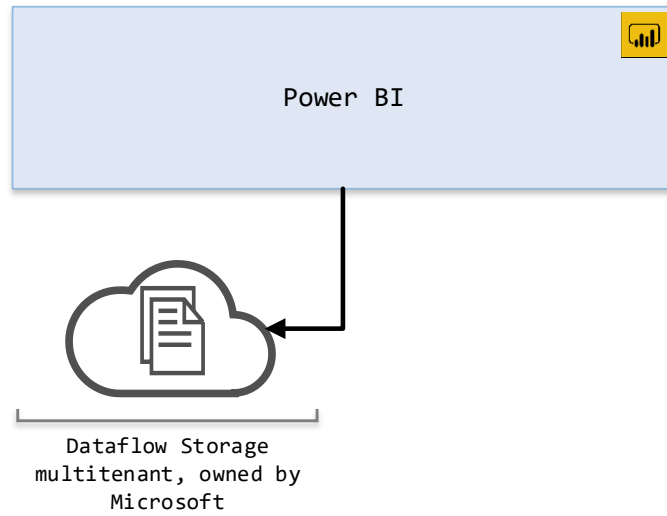
Dettagli sullo storage dataflow

Il formato di serializzazione è definito dalle specifiche del Common Data Model

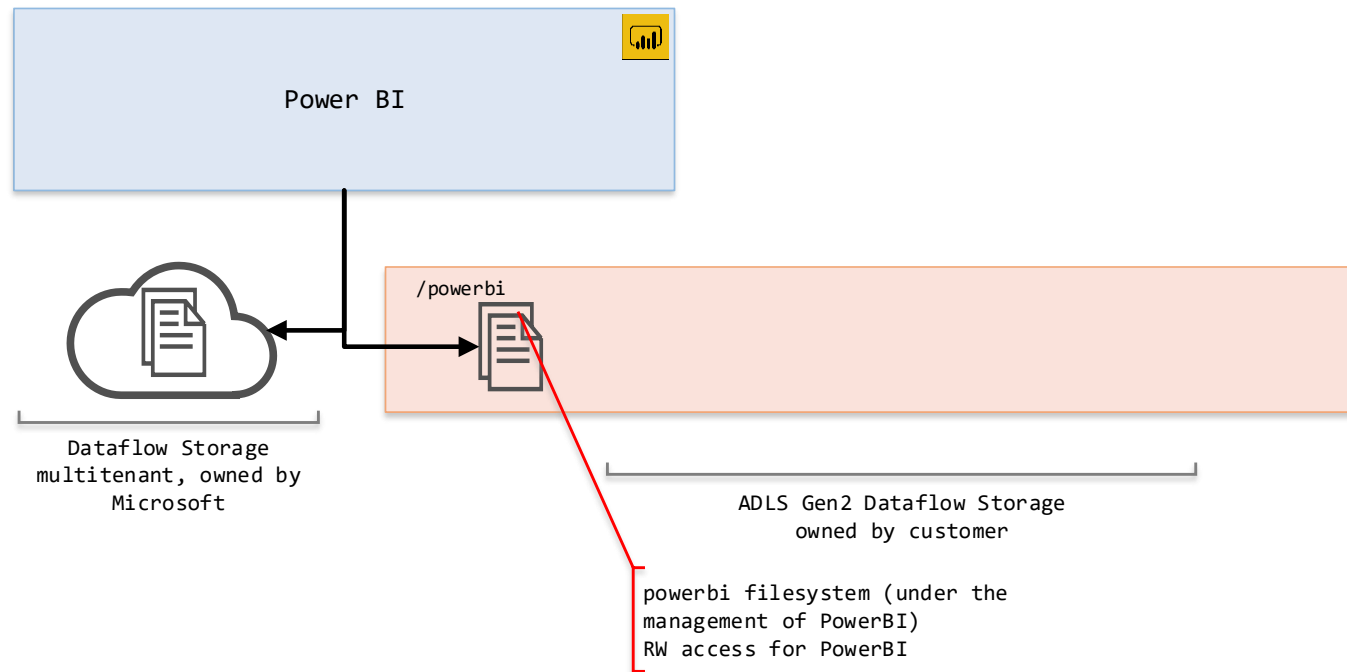
- Metadati del dataflow archiviati nel file **model.json**
- Righe di dati dei dataflow sono archiviate in file CSV
- Per impostazione predefinita, Power BI gestisce l'archiviazione dei dataflow dietro le quinte



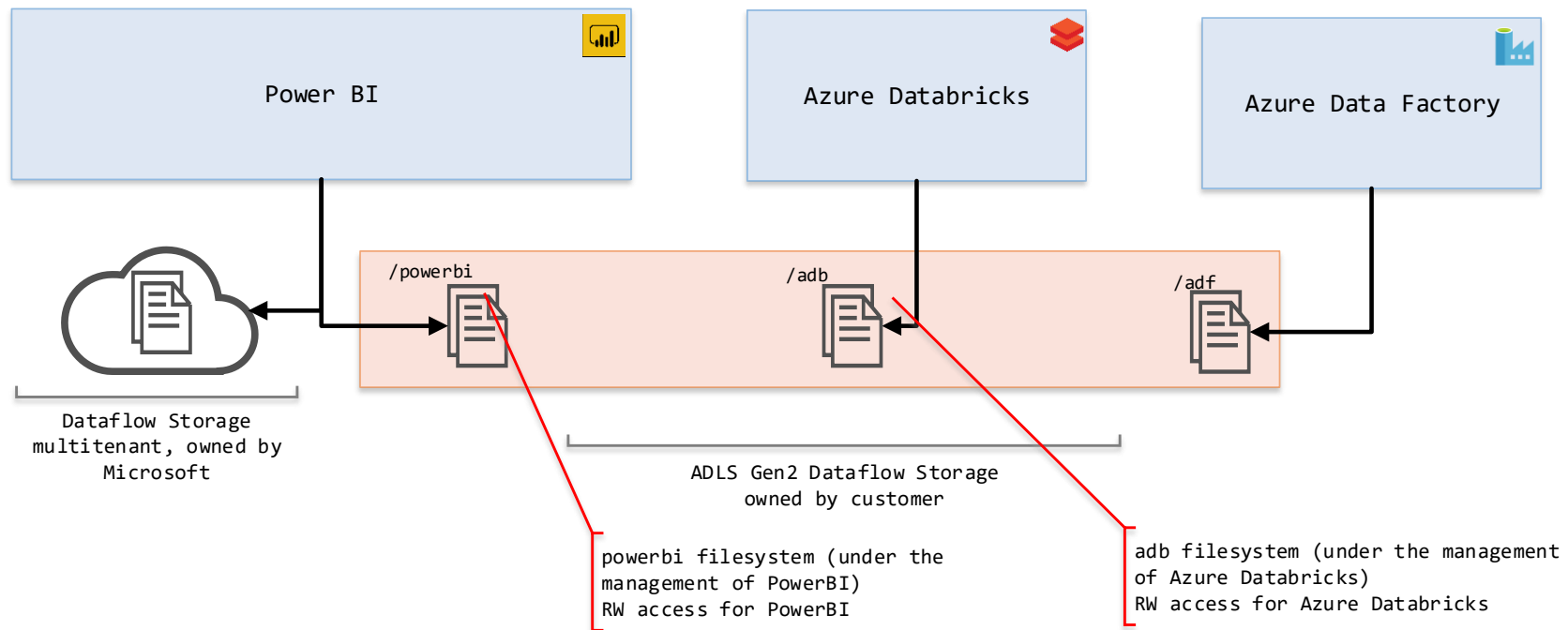
Architettura del common data model



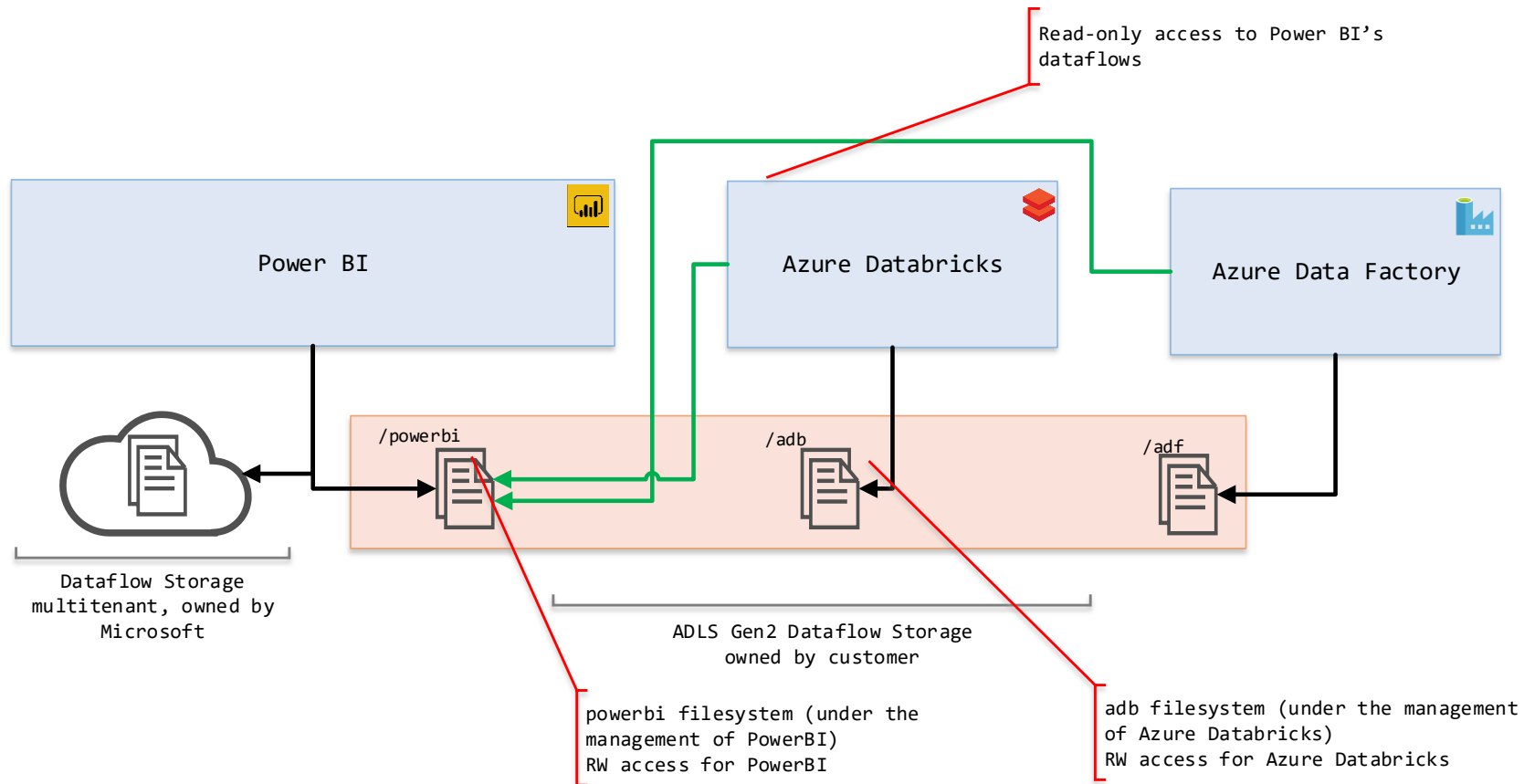
Architettura del common data model



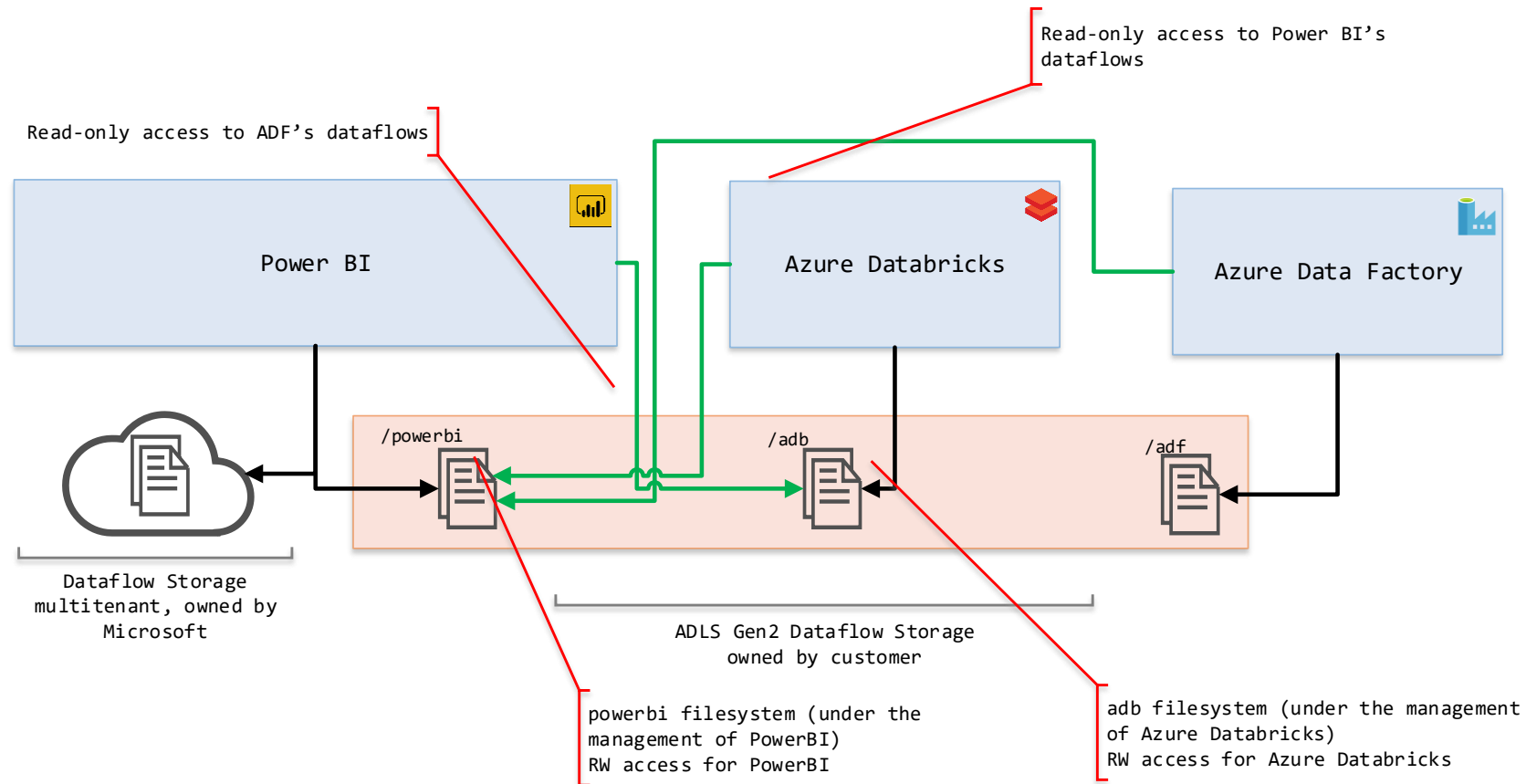
Architettura del common data model



Architettura del common data model



Architettura del common data model



Common data model metadata

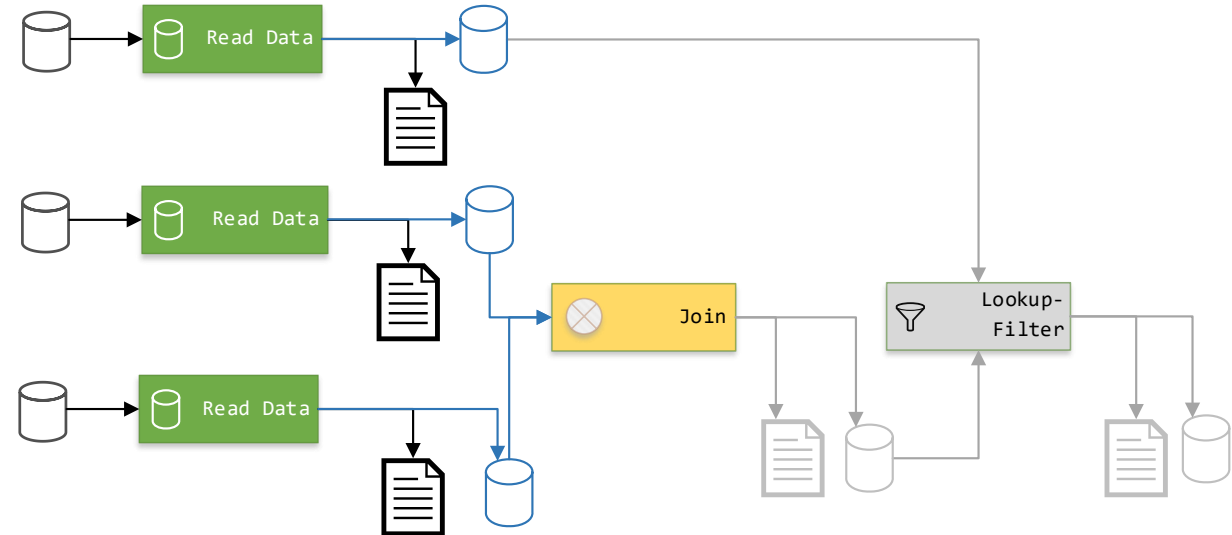
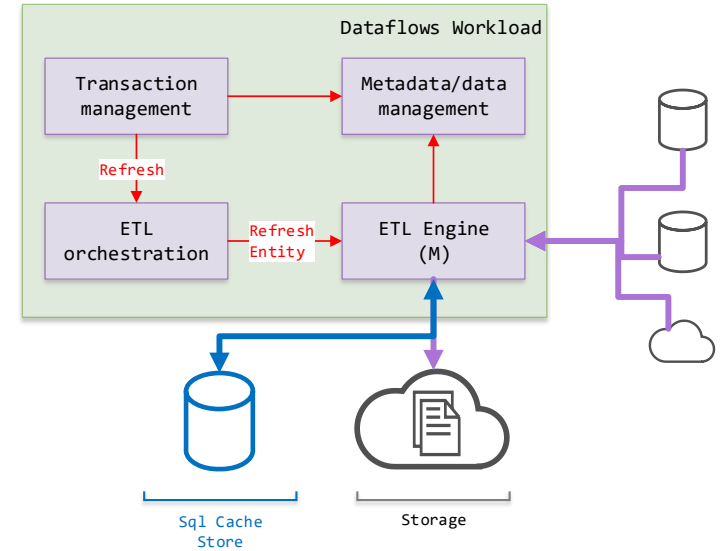
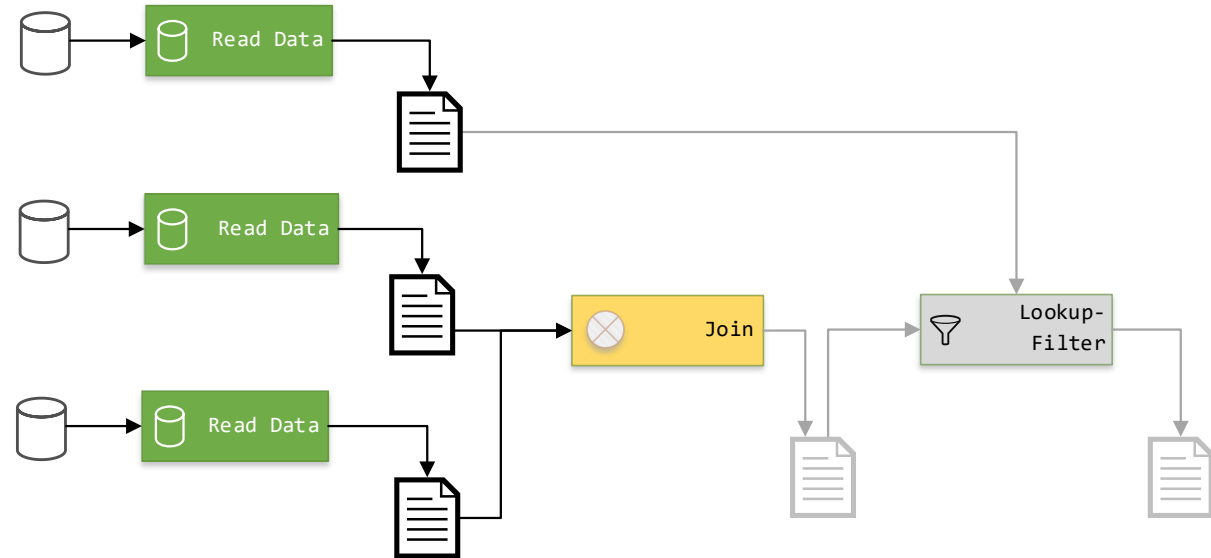
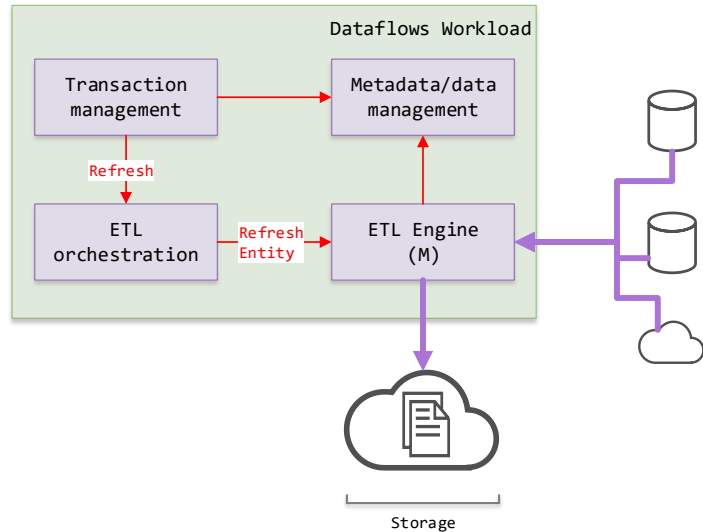
L'output del flusso di dati è memorizzato nel formato CDM

- Il file model.json contiene metadati relativi alle entità
- Il file model.json contiene il codice M per le query

```
{
  "name": "wingtip Sales Dataflow",
  "description": "A sample dataflow",
  "version": "1.0",
  "culture": "en-US",
  "modifiedTime": "2019-10-21T17:54:50.1618626+00:00",
  "pbi:mashup": {
    "fastCombine": false,
    "allowNativeQueries": false,
    "queriesMetadata": {
      "Customers": { "queryId": "58d2a7e0-0298-4d94-8285-7af1f3d54b15", "query": "" },
      "Products": { "queryId": "10577951-df4b-407c-b6fb-c923880bale", "query": "" },
      "Orders": { "queryId": "ad08816d-be0d-4f6f-b19e-755f23c8fb0f", "query": "" },
      "Sales": { "queryId": "4613190e-da33-4a3a-af5d-a567cbde4dd2", "query": "" }
    },
    "document": "section Section1;\r\nshared Customers = let\r\n  Source = s
  },
  "entities": [
    {
      "$type": "LocalEntity",
      "name": "Customers",
```

```
{ "$type": "LocalEntity", "name": "Products", "description": "",
  "pbi:refreshPolicy": { "$type": "FullRefreshPolicy", "location": "Products.csv" },
  "attributes": [
    { "name": "ProductId", "dataType": "int64" },
    { "name": "Product", "dataType": "string" },
    { "name": "Description", "dataType": "string" },
    { "name": "Category", "dataType": "string" },
    { "name": "Subcategory", "dataType": "string" },
    { "name": "UnitCost", "dataType": "decimal" },
    { "name": "ListPrice", "dataType": "decimal" },
    { "name": "Product Image", "dataType": "string" }
  ],
  "partitions": [
    {
      "name": "Part001",
      "refreshTime": "2019-10-21T17:59:55.5031318+00:00",
      "location": "https://wabieus2cdsap1.blob.core.windows.net:443/913b7aae-5
    }
  ]
},
```

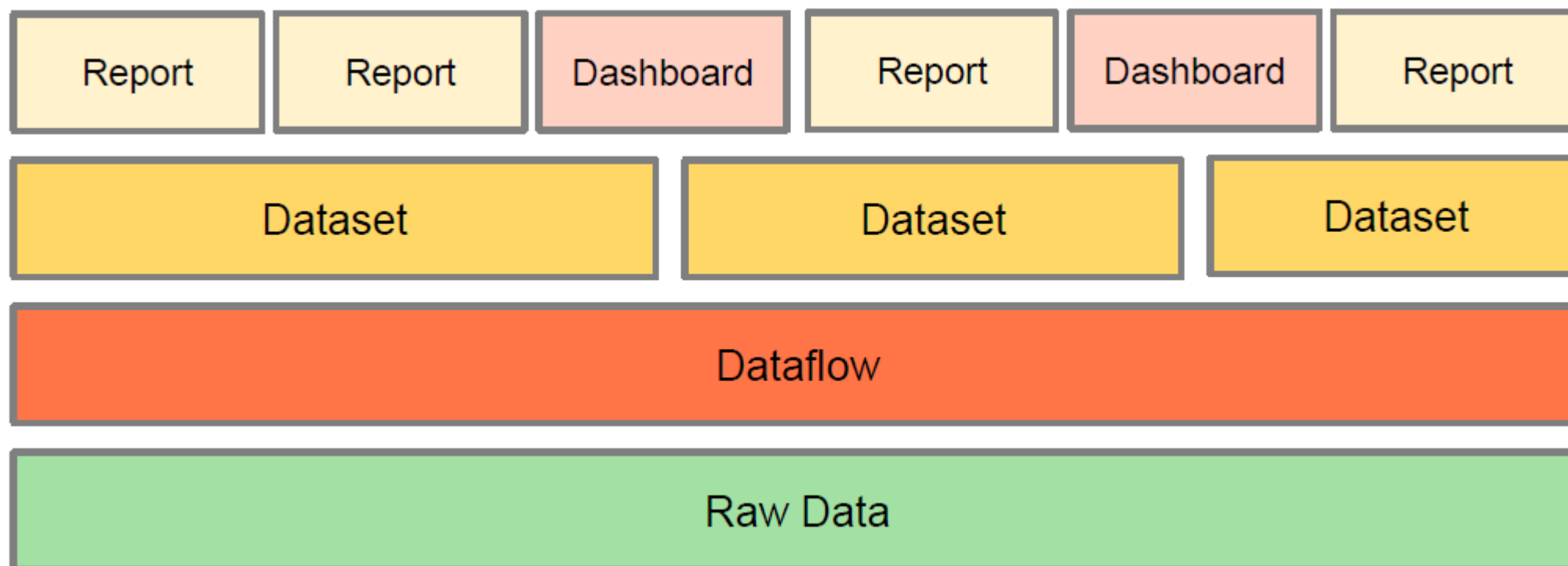
Dettagli sullo storage dataflow (miglioramenti)



Progetta una soluzione Power BI con dataflow

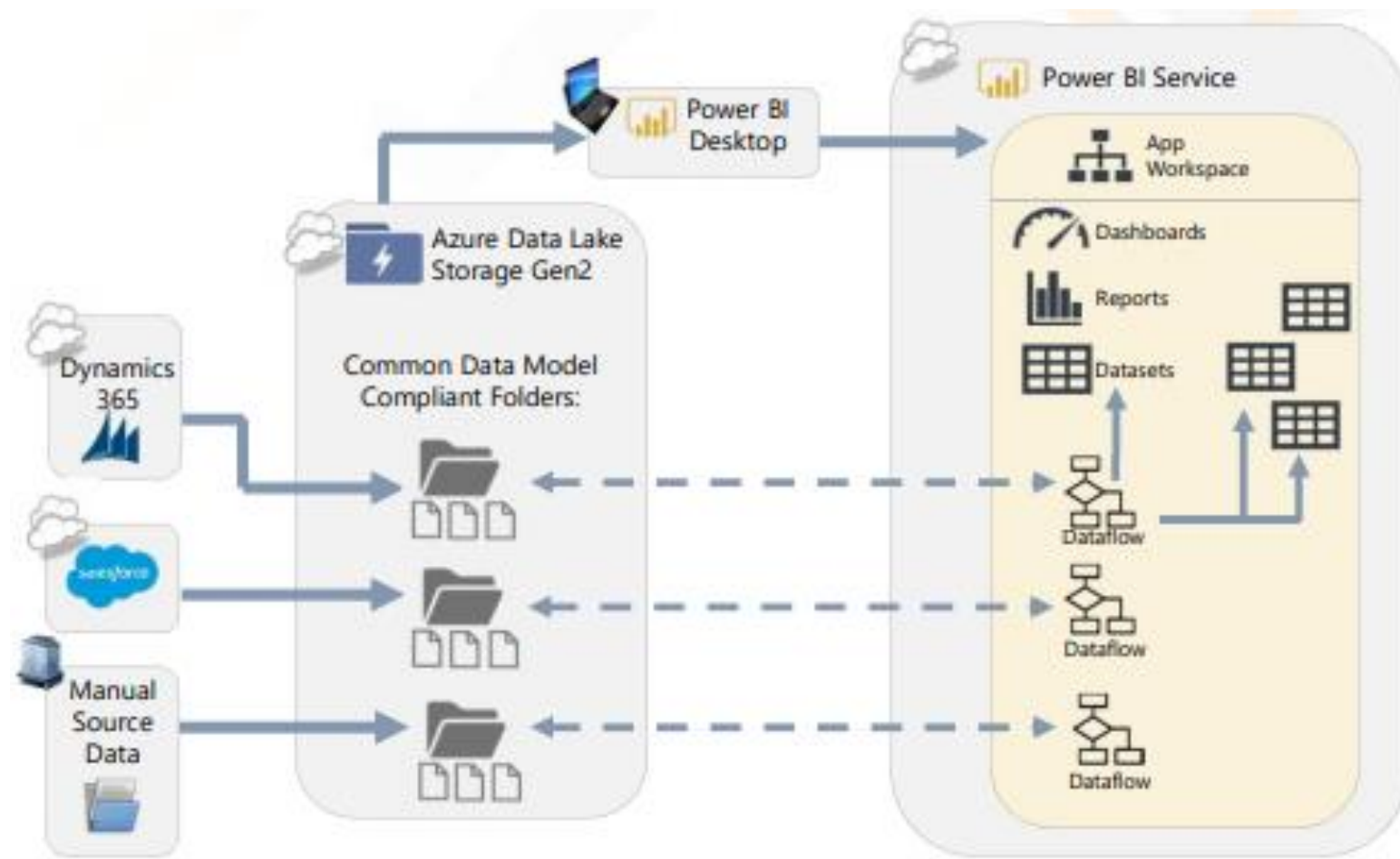
I Dataflow sono utilizzati per raccogliere tutti i dati

- I progetti di Power BI Desktop importeranno dati dai dataflow
- Il lavoro ETL non è più necessario nei progetti di Power BI Desktop



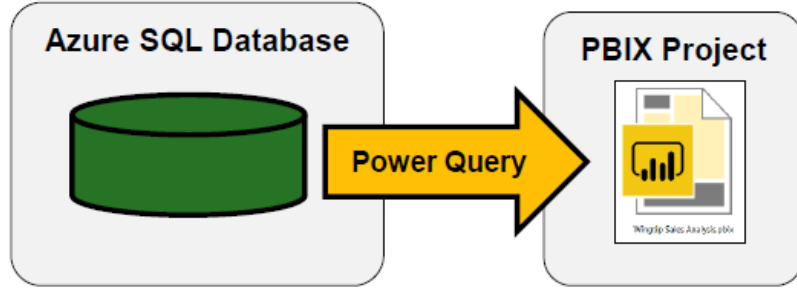
Progetta una soluzione Power BI con dataflow

- Un singolo flusso di dati può alimentare i dati di molti dataset
- Un dataset rimane definito e contiene Calcoli, relazioni, RLS e altro
- Ci sono due aggiornamenti separati schedulati uno per il dataset e uno per il dataflow



Progetta una soluzione Power BI con dataflow

- Progetto Power BI Desktop senza Dataflow



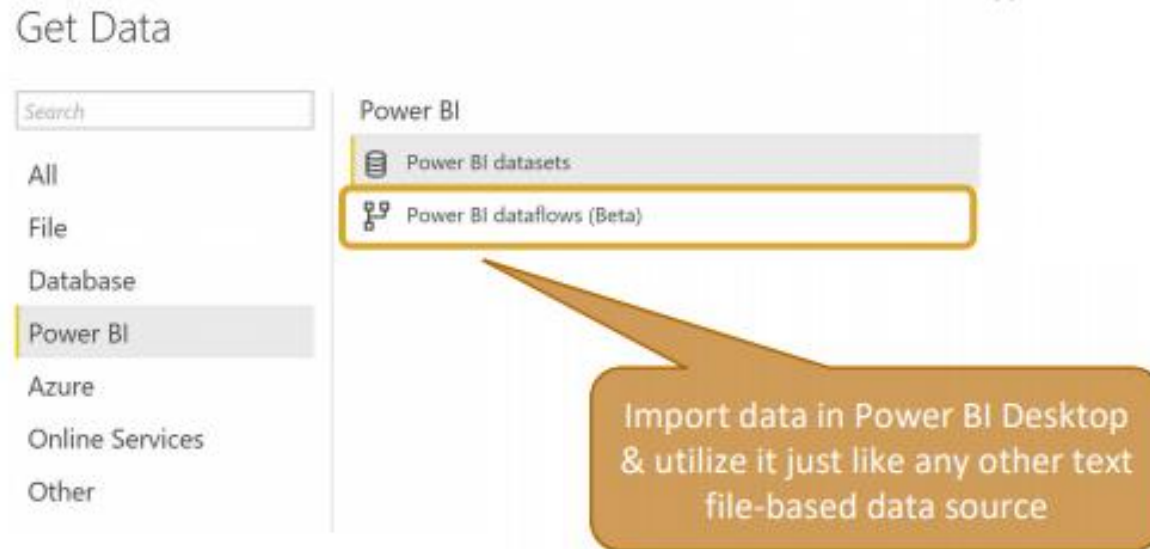
- Progetto Power BI Desktop con Dataflow



Progetta una soluzione Power BI con dataflow

Entità del dataflow utilizzate da Power BI Desktop

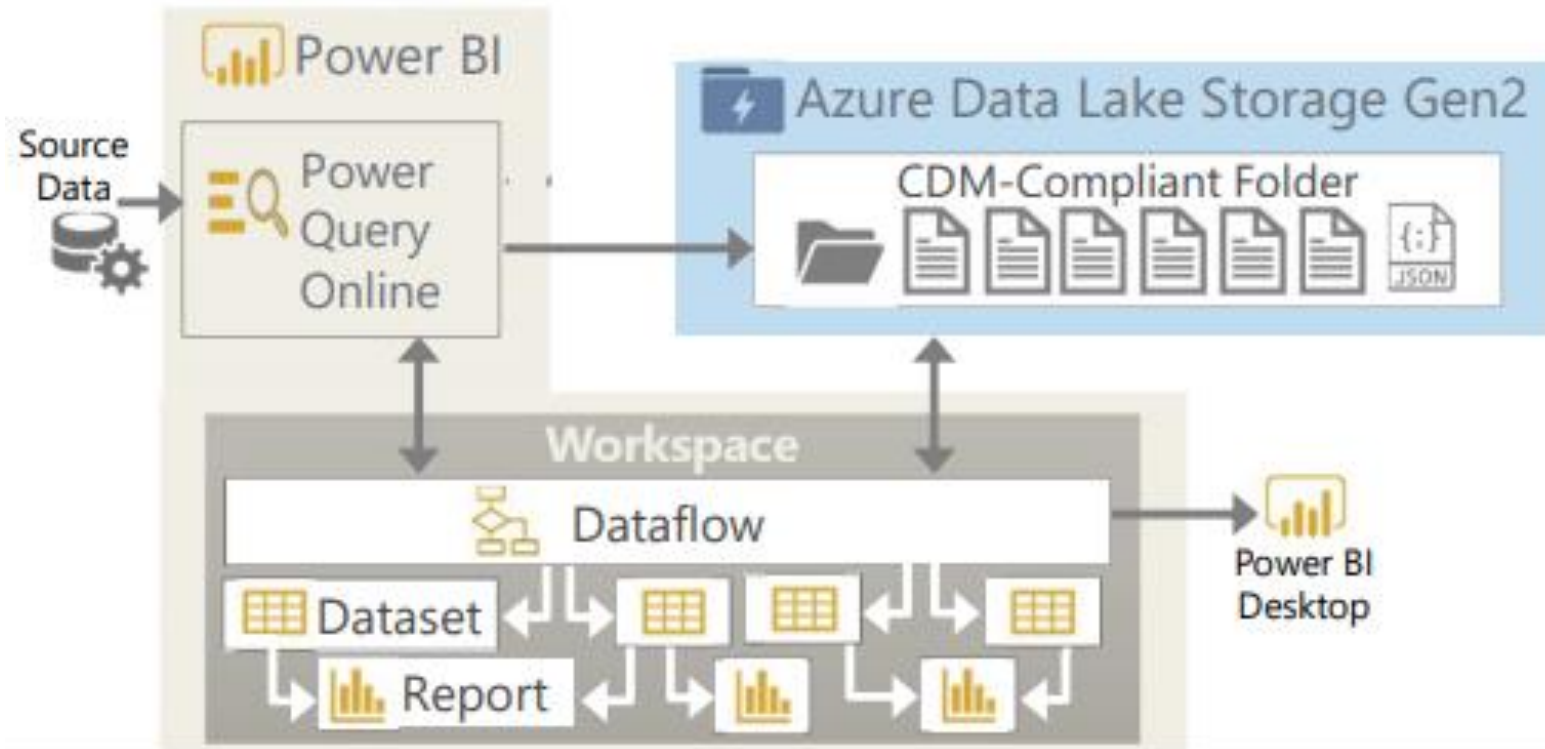
- Utilizzare Power BI Dataflow come sorgente
- Il dataset con il dataflow può essere pubblicato in qualsiasi area di lavoro dell'app



Tre modi di usare i Dataflow: Tipo 1

Struttura dati: Gestita da Power BI

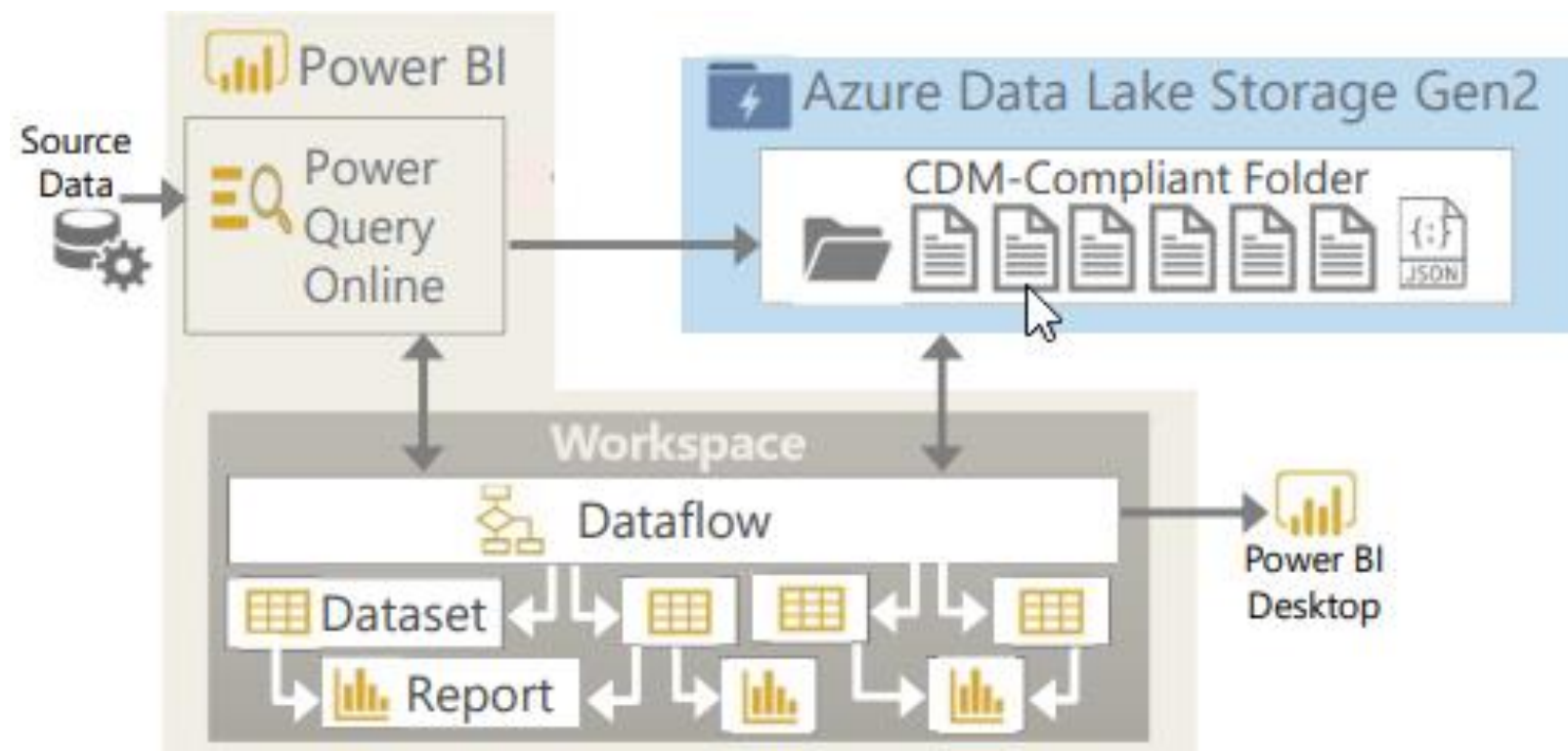
Data lake: I file non sono accessibili da altri tool



Tre modi di usare i Dataflow: Tipo 2

Struttura dati: Gestita da Power BI

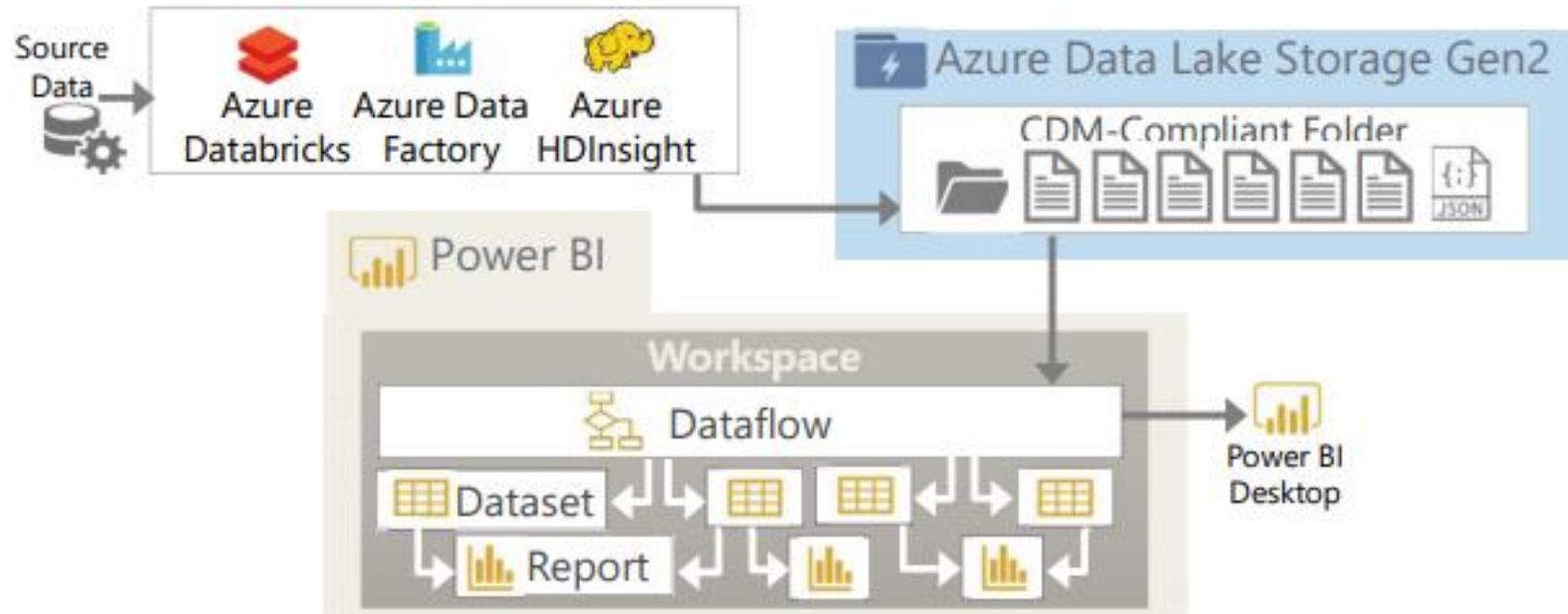
Data lake: lo storage di data lake è associato al tenant di Power BI



Tre modi di usare i Dataflow: Tipo 3

Struttura dati: Gestito da altri tool

Data lake: lo storage di data lake è associato al tenant di Power BI























Demo 1: Creazione Dataflow Tipo 1 e 2

Entità Dataflow

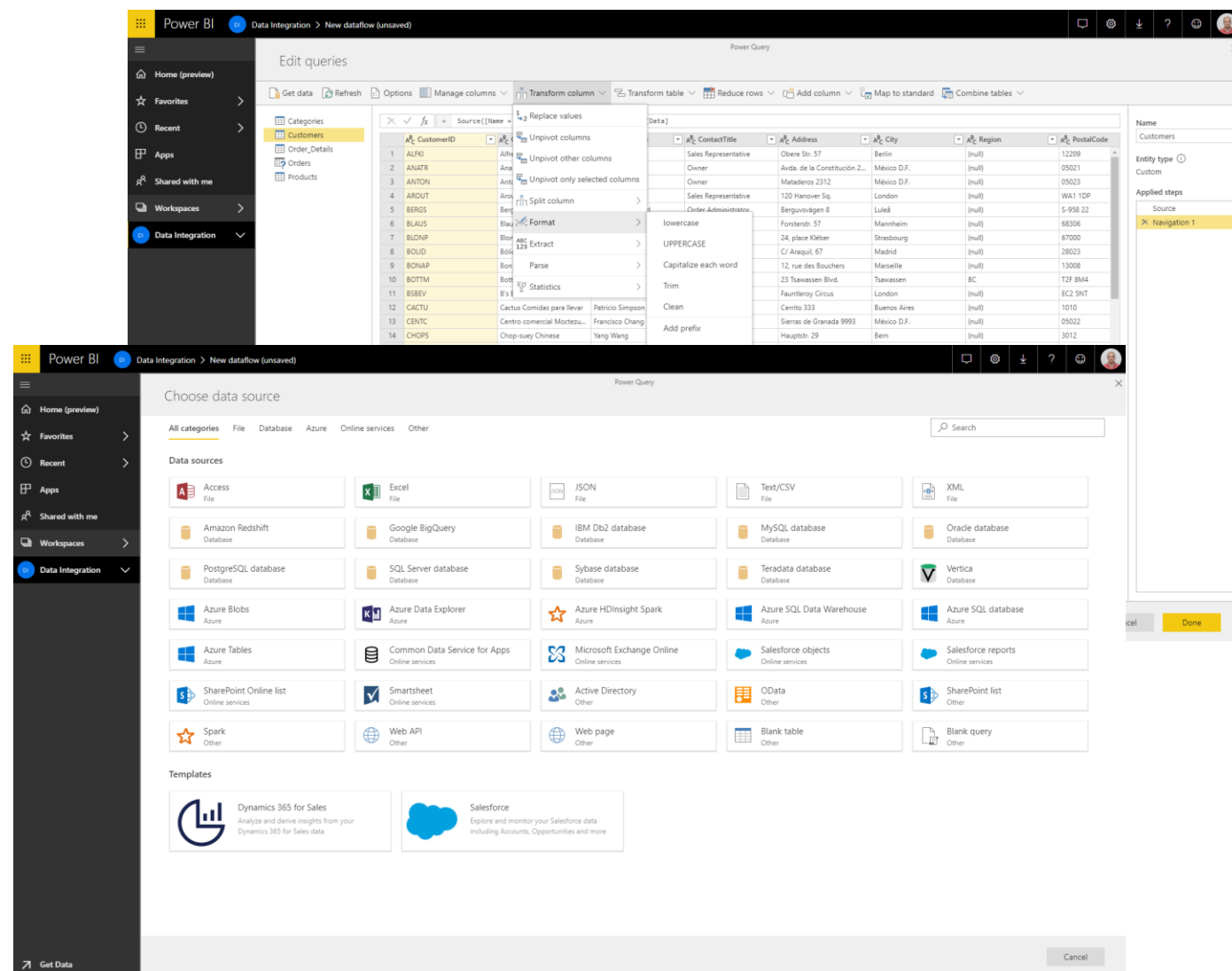
Esiste uno o più dataflow all'interno di un'area di lavoro (workspace)

- Il dataflow contiene una o più entità
- L'Entity è una tabella con schema ben definito
- L'Entity è popolata eseguendo una query (codice M)

Entities Machine learning models		✓ Changes saved	Edit entities	Add entities	Close
ENTITY NAME	ENTITY TYPE	ACTIONS			
▶  Customers	Custom				
▶  Sales	Custom				
▶  Orders	Custom				
▶  Products	Custom				

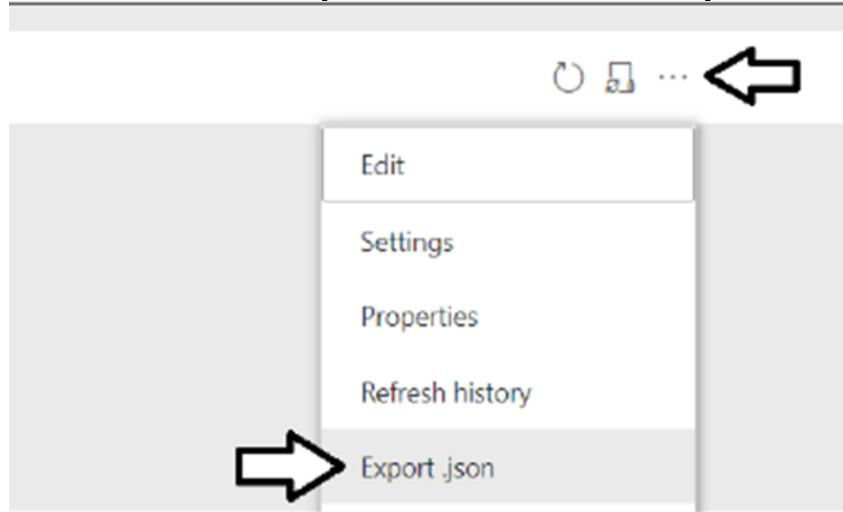
Dataflows usano Power Query nel browser

- Esperienza di modifica familiare agli utenti con Power Query
- Power Query è disponibile con un'esperienza utente di preparazione del dato di tipo **web-based self service**
- Supporta lo stesso numero di **300+ transformations** di PBI desktop Power Query (M Engine)
- Correntemente ci sono **~45 connectors**, incluse connessioni a sorgenti cloud & on-prem via **On-premises data gateway**
- Usa la potenza del cloud per elaborare grandi volumi di dati in Power BI
- Sfrutta il calcolo di Power BI per trasformare i dati in modo semplice e rapido

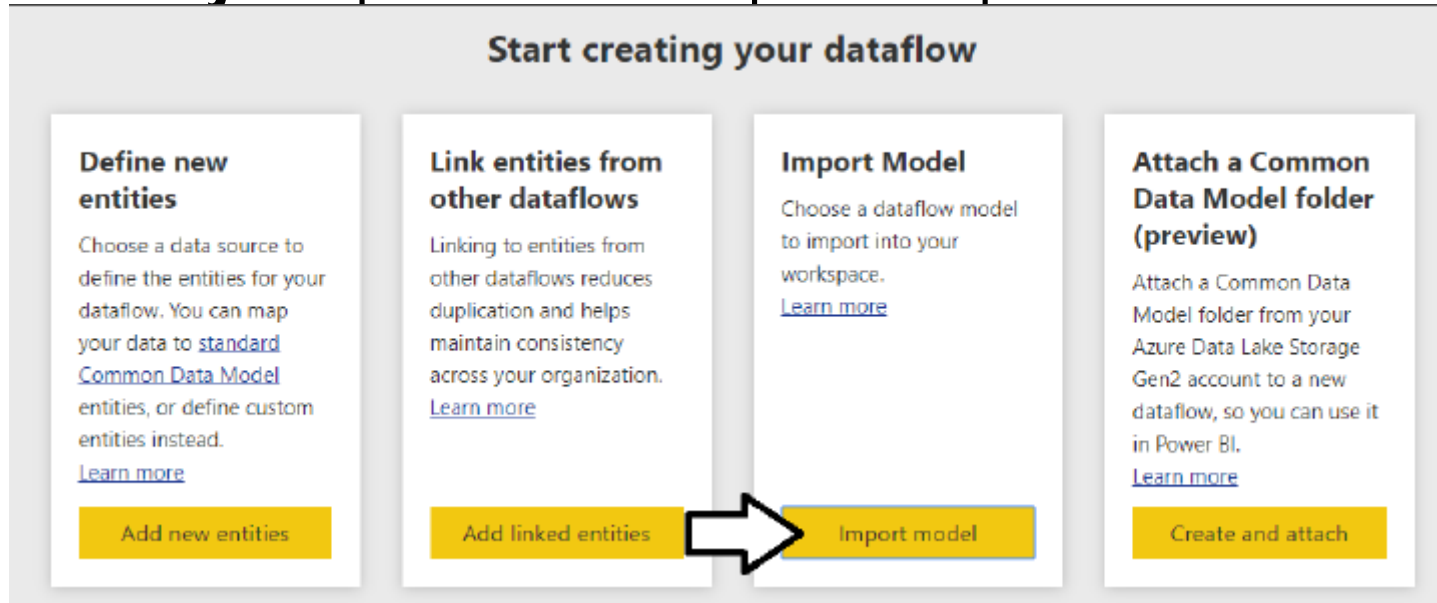


Importare ed esportare i dataflow

- Il dataflow può essere esportato come **model.json**



- model.json** può essere importato per creare un nuovo dataflow



Demo 2: Dataflow

Licenze per i dataflow

La creazione del dataflow richiede **Power BI Pro**

- I dataflow non possono essere creati in aree di lavoro personali

La capacità dedicata (**Premium**) aggiunge funzionalità extra per il flusso di dati

- Capacità di gestire volumi di dati più grandi
- Migliori prestazioni di aggiornamento
- Entità **collegate** e **calcolate**
- Funzionalità AI per trasformare i dati

Funzionalità premium del Dataflow

- Entità collegate (**linked entities**)
- Entità calcolate (**computed entities**)
- Funzionalità di intelligenza artificiale (AI)
- Aggiornamento incrementale
- Esecuzione parallela delle trasformazioni
- **Direct Query su dataflow**

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

Entità collegate (Linked Entity)

- Le entità collegate(**linked entity**) consentono di condividere i dati tra:
 - Dataflow diversi nello stesso workspace
 - Dataflow diversi in diversi workspace
- La creazione di una **linked entity** non duplica i dati di origine
 - È possibile utilizzare un'entità esistente in un altro workspace come origine
 - Utilizza lo stesso codice M utilizzato dal dataset per ottenere dati da un'entità
 - Le entità collegate sono di sola lettura
 - Se vuoi ulteriori trasformazioni crei un'entità calcolata
 - La vista a diagramma semplifica la visualizzazione dell'utilizzo di entità collegate

Entità calcolate (Computed Entity)

- Entità calcolate (**Computed Entity**) basate su di altre entità
 - Consente alle entità di utilizzare altre entità come sorgenti all'interno di un dataflow
- Scenari utili
 - Stai creando più entità all'interno di un dataflow dallo stesso dato non elaborato (source) e non desideri ottenere dati dall'origine dati originale più di una volta
 - Stai avendo problemi a causa del motore di Power Query che ha l'abitudine di richiedere dati più volte durante una singola query

Importare da un data lake esterno

- Importare i dati da un modello che si trova su data lake

The image displays four vertical panels, each representing a different method to import data from an external data lake. Each panel has a title, a descriptive paragraph, a link for more information, and a yellow action button at the bottom.

- Definisci nuove entità**
Scegliere un'origine dati per definire le entità per il flusso di dati. È possibile eseguire il mapping dei dati a entità [Common Data Model standard](#) o definire entità personalizzate.
[Altre informazioni](#)
Aggiungi nuove entità
- Collega entità da altri flussi di dati**
Il collegamento alle entità da altri flussi di dati consente di ridurre la duplicazione e contribuisce a garantire la coerenza nell'intera organizzazione.
[Altre informazioni](#)
Aggiungi entità collegate
- Importa il modello**
Scegliere un modello di flusso di dati da importare nell'area di lavoro.
[Altre informazioni](#)
Importa il modello
- Collega una cartella Common Data Model (anteprima)**
Collegare una cartella Common Data Model dell'account Azure Data Lake Storage Gen2 a un nuovo flusso di dati in modo da poterla usare in Power BI.
[Altre informazioni](#)
Crea e collega

Benefici dei dataflow

- Sostituisce altri strumenti ETL (ad es. Azure Data Factory, Power Automate)
- Disaccoppia il lavoro degli ETL dai set di dati nei progetti PBIX
- Abilita la condivisione di tabelle provenienti dalla sorgente tra set di dati
- Riduce il numero di query sulle origini dati live
- Elimina la necessità di connettere i computer degli utenti direttamente all'origine dati
- Centralizza gli sforzi per pulire e preparare i dati
- Condividi le tabelle che non hanno origine (esempio tabelle del calendario)

Svantaggi dei dataflow

- Aggiunge ulteriore complessità
- I dati devono essere **aggiornati in 2 fasi separate**
- **Non** supporta le funzionalità di **modellazione dei dati di DAX**
- Alcune funzionalità dei dataflow richiedono capacità Premium



Afishapa

Sung Tan Chuk Ha
Chuc Mung Giang Sinh

Feliz Natal

Buon Natale Sretan Bozic
Bara Din Mubarrak Ho

Srecen Bozic

Sretam Bozic, Hristos se Rodi
Mo'adim Lesimkha Feliz Navidad

Feliz Navidad

Sing dan fuy loc Bon Nadâl Sawasdee Pee Mai
Merii Kurisumasu Natal Mubarak

Buon Natale

God Jul Zelig Kerstfeest Merry Christmas
Joyeux Noël

Fröhliche Weihnachten

Vesele Vanoce Gleðileg Jól Merii Kurisumasu
Cestitamo Bozic Gleðileg Jól Milad Mubarak

Merry Christmas

2020

Boas Festas

Frohliche Weihnachten

Bon Nadâl

Nollaig Shona Dhuit

God Jul