

Dataflow



#GlobalPowerPlatformBootcamp

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>



#GlobalPowerPlatformBootcamp

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."

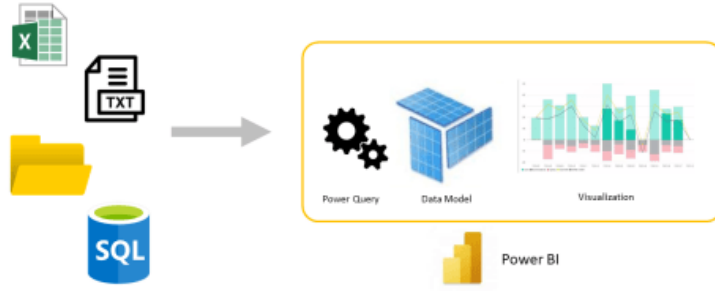
- Gartner



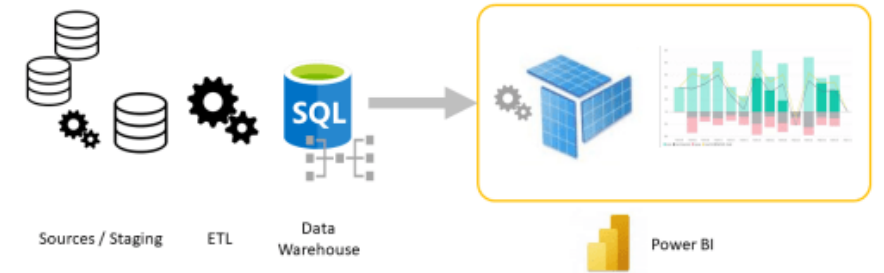
#GlobalPowerPlatformBootcamp

Scenari

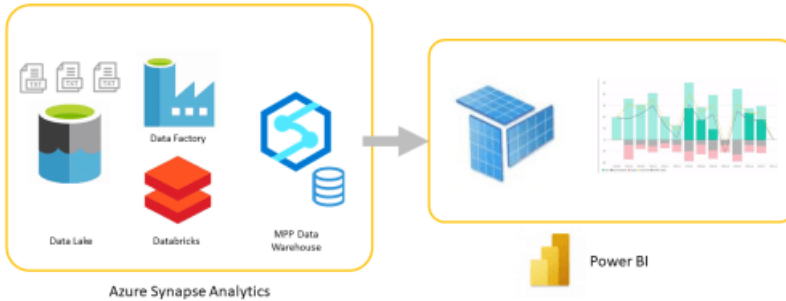
Self-service Data Prep & Analysis



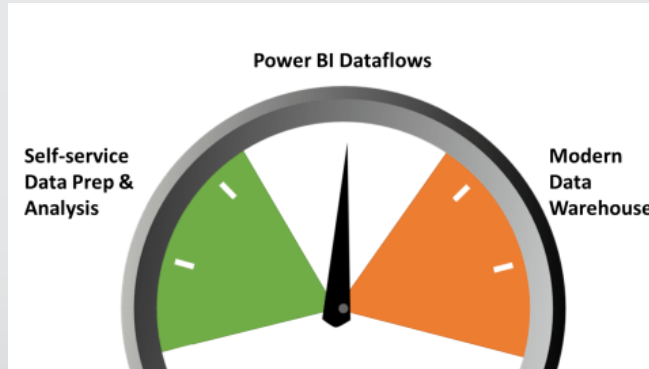
Conventional Business Intelligence



Enterprise Modern Data Warehouse



Modern Data Warehouse



#GlobalPowerPlatformBootcamp

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

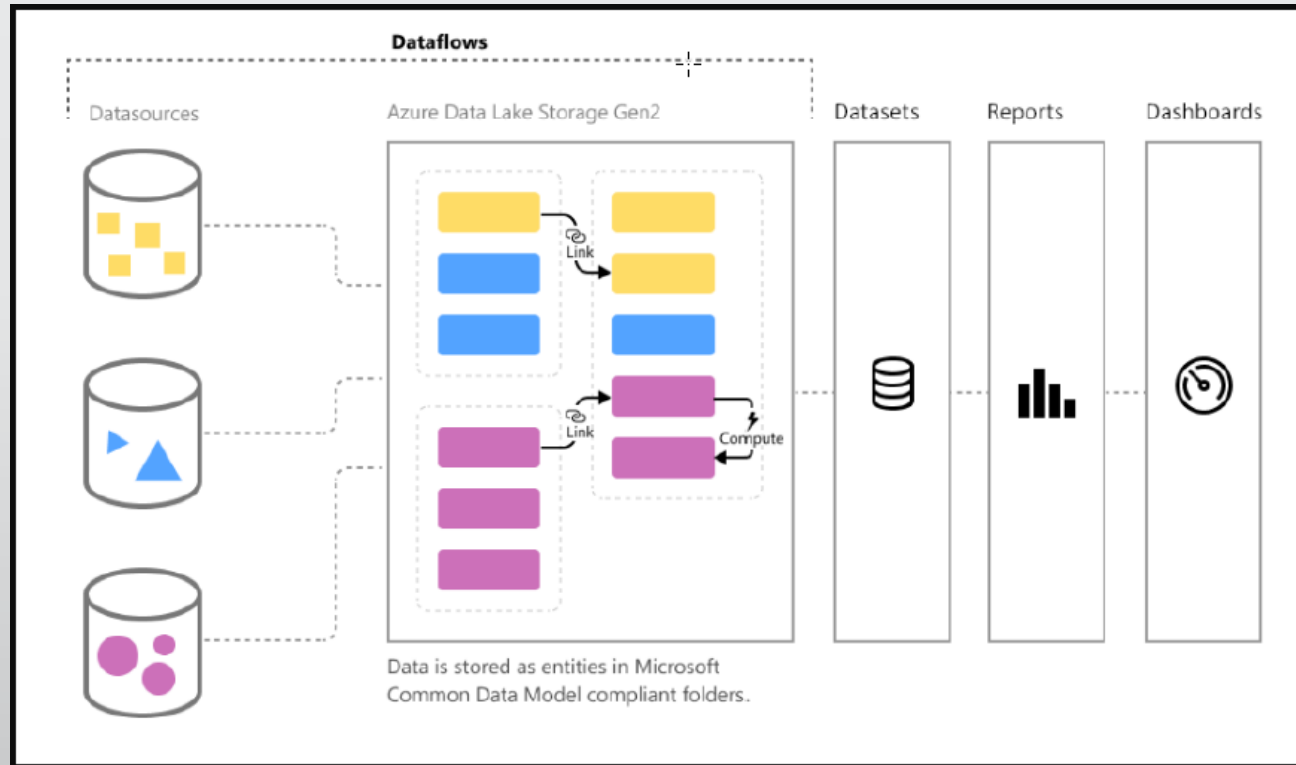


Ridurre il caricamento sui sistemi sorgenti: Ridurre al minimo il numero di query inviate al sistema di origine

Architettura dataflow

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

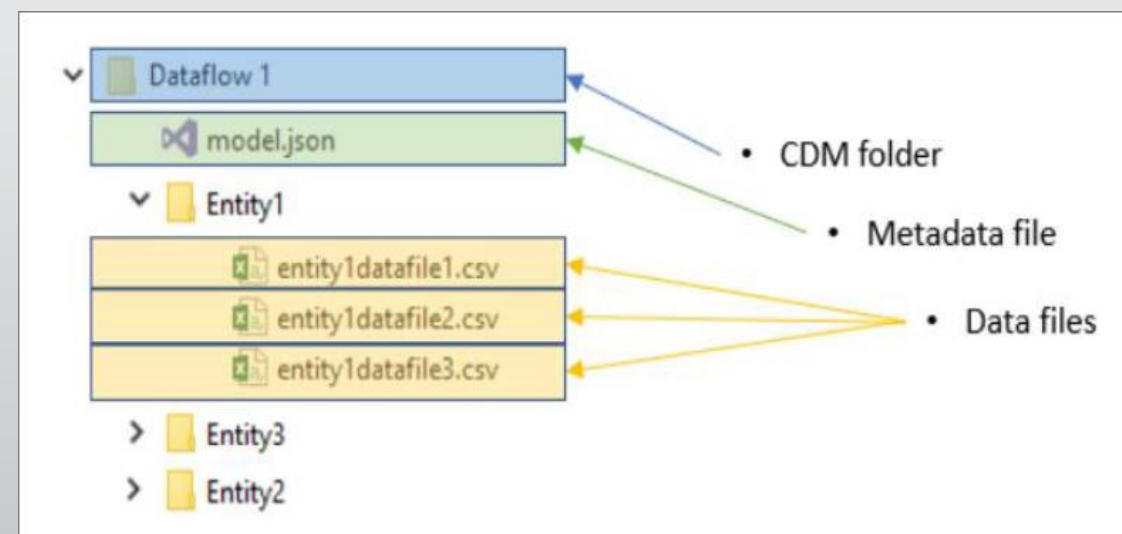
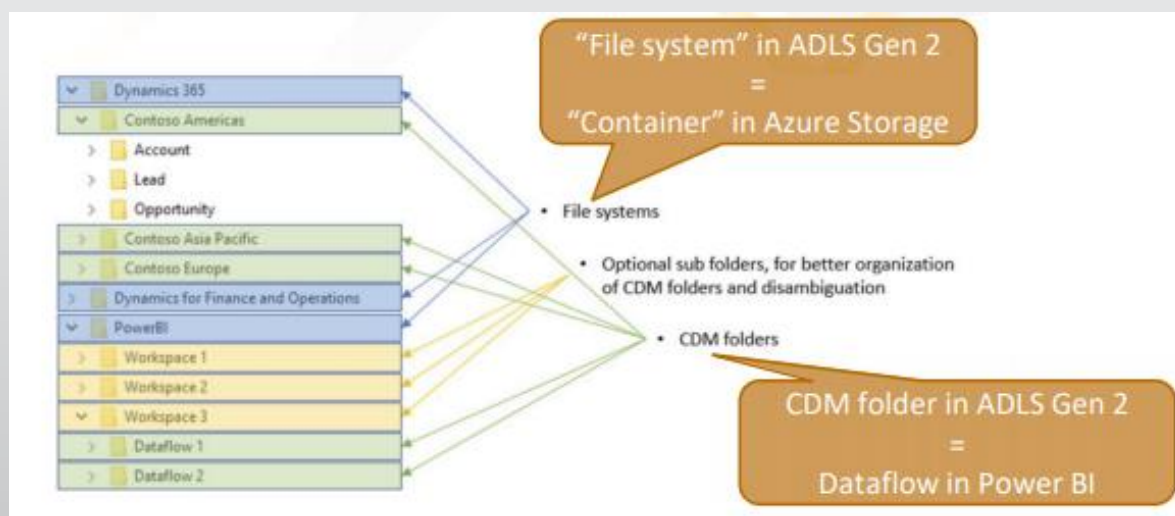
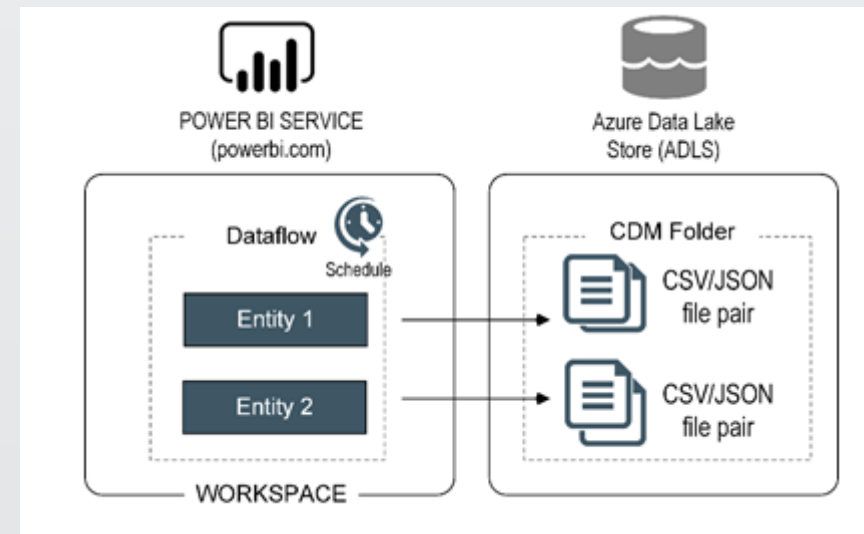
- Archiviazione progettata per soddisfare le esigenze dei big data
- Dataflow serializzati nel formato definito dal Common Data Model (CDM)



Dettagli sullo storage dataflow

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



Common data model metadata

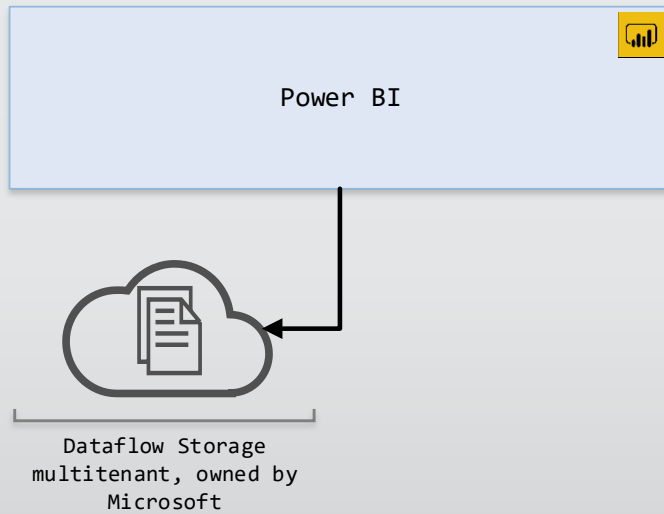
L'output del flusso di dati è memorizzato nel formato CDM

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

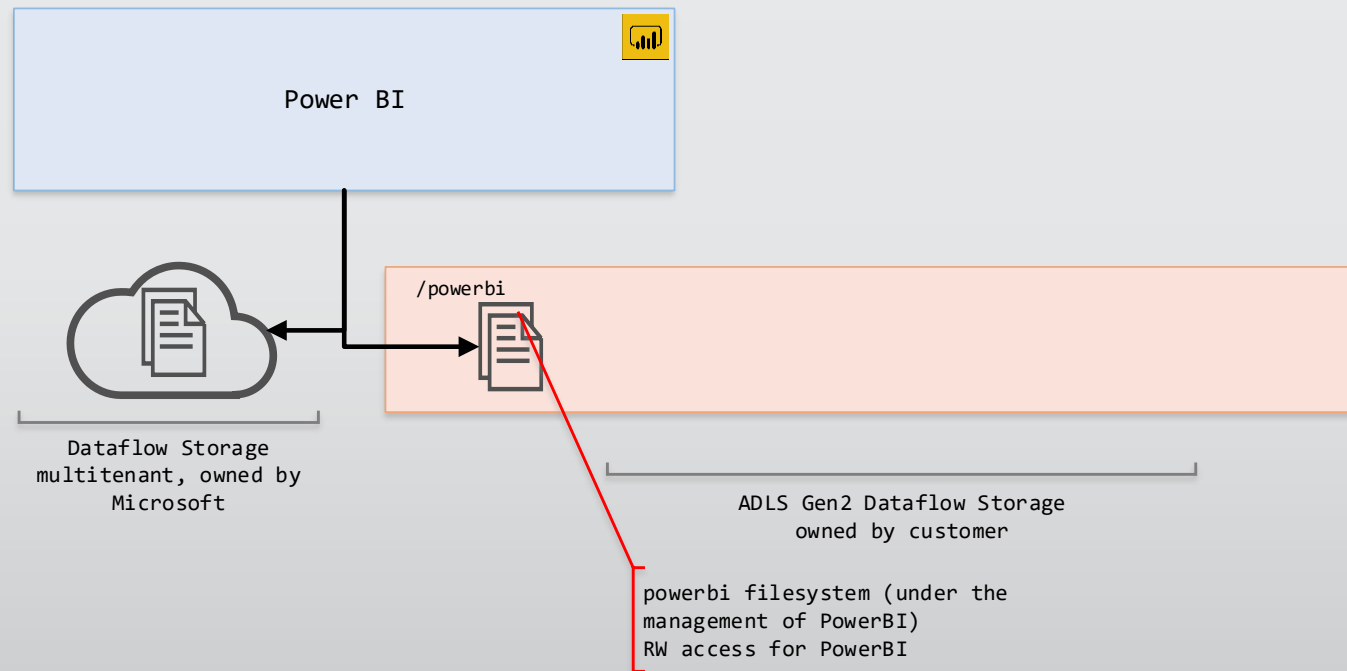
```
{
  "name": "wingtip Sales Dataflow",
  "description": "A sample dataflow",
  "version": "1.0",
  "culture": "en-US",
  "modifiedTime": "2019-10-21T17:54:50.1618626+00:00",
  "pbi:mashup": {
    "fastCombine": false,
    "allowNativeQueries": false,
    "queriesMetadata": {
      "Customers": { "queryId": "58d2a7e0-0298-4d94-8285-7af1f3d54b15", "q",
      "Products": { "queryId": "10577951-df4b-407c-b6fb-c923880bale", "qu",
      "Orders": { "queryId": "ad08816d-be0d-4f6f-b19e-755f23c8fb0f", "quer",
      "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",
    }
  ]
},
```

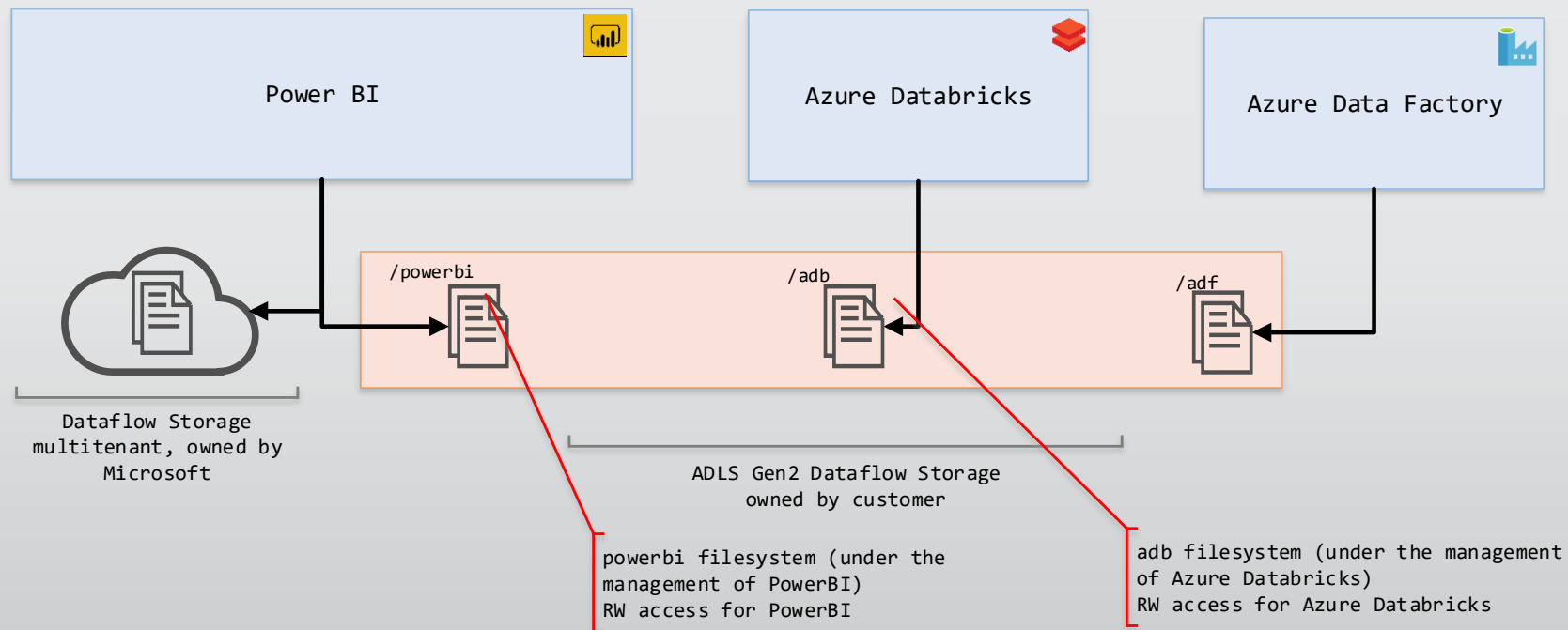

Architettura del common data model



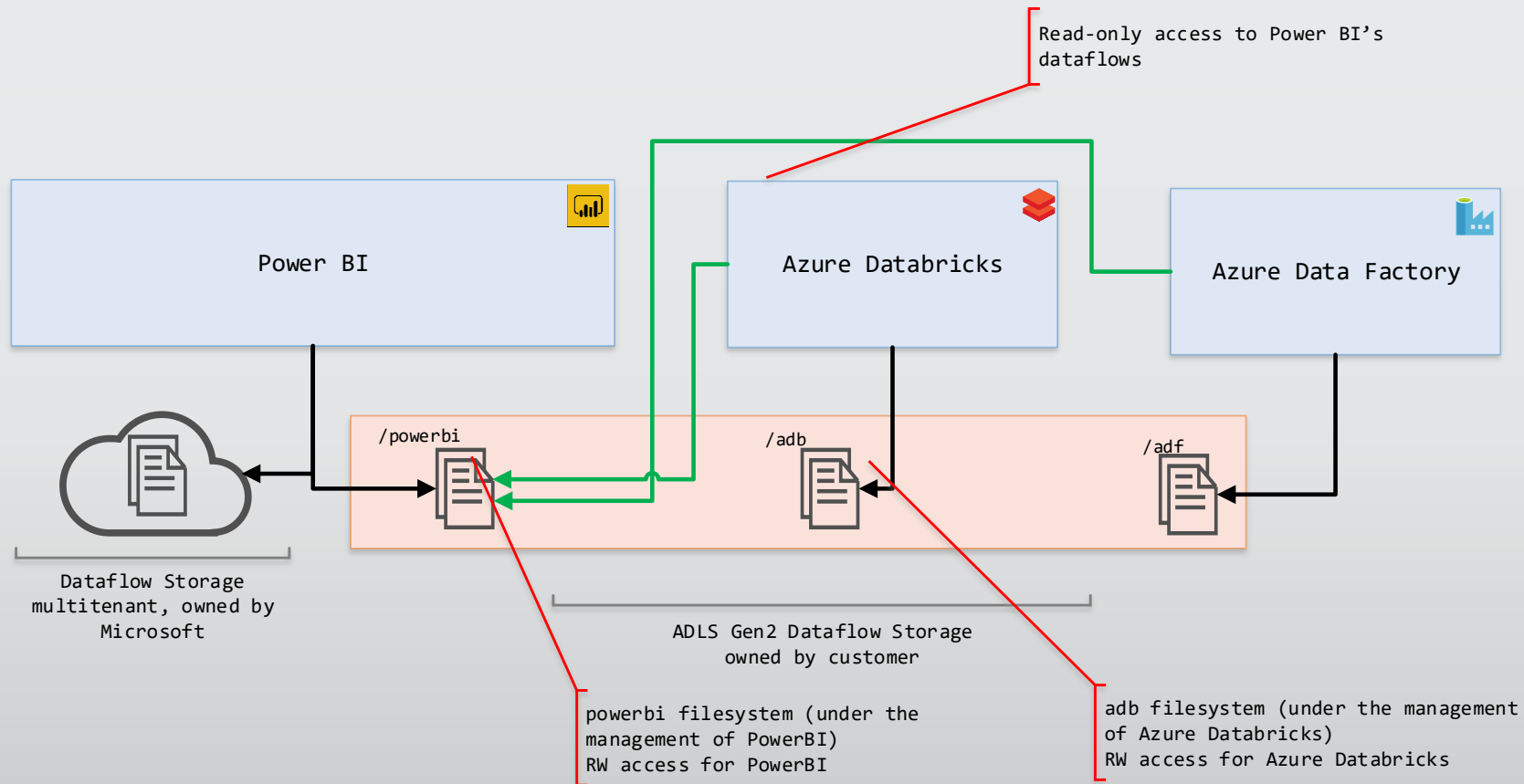
Architettura del common data model



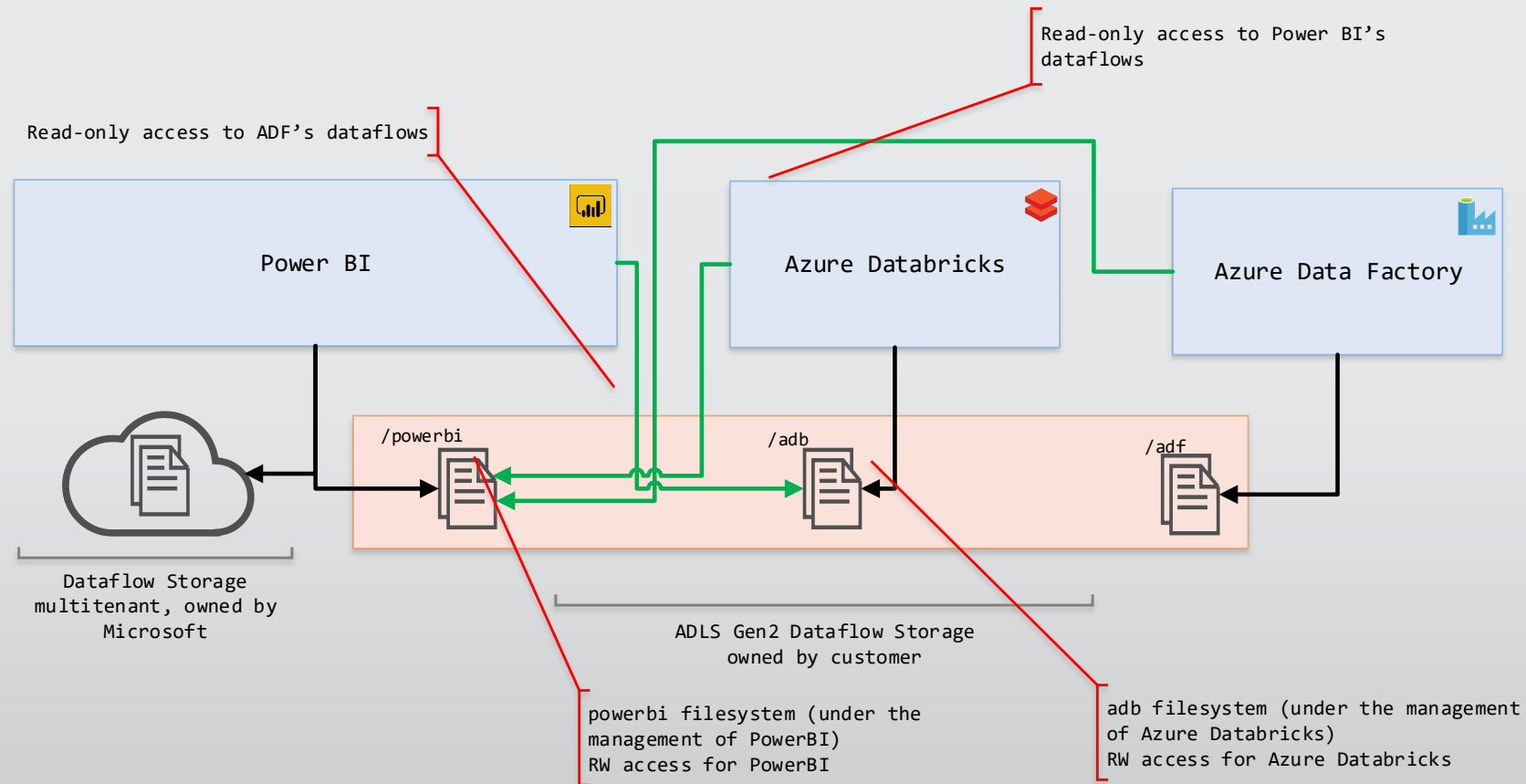
Architettura del common data model



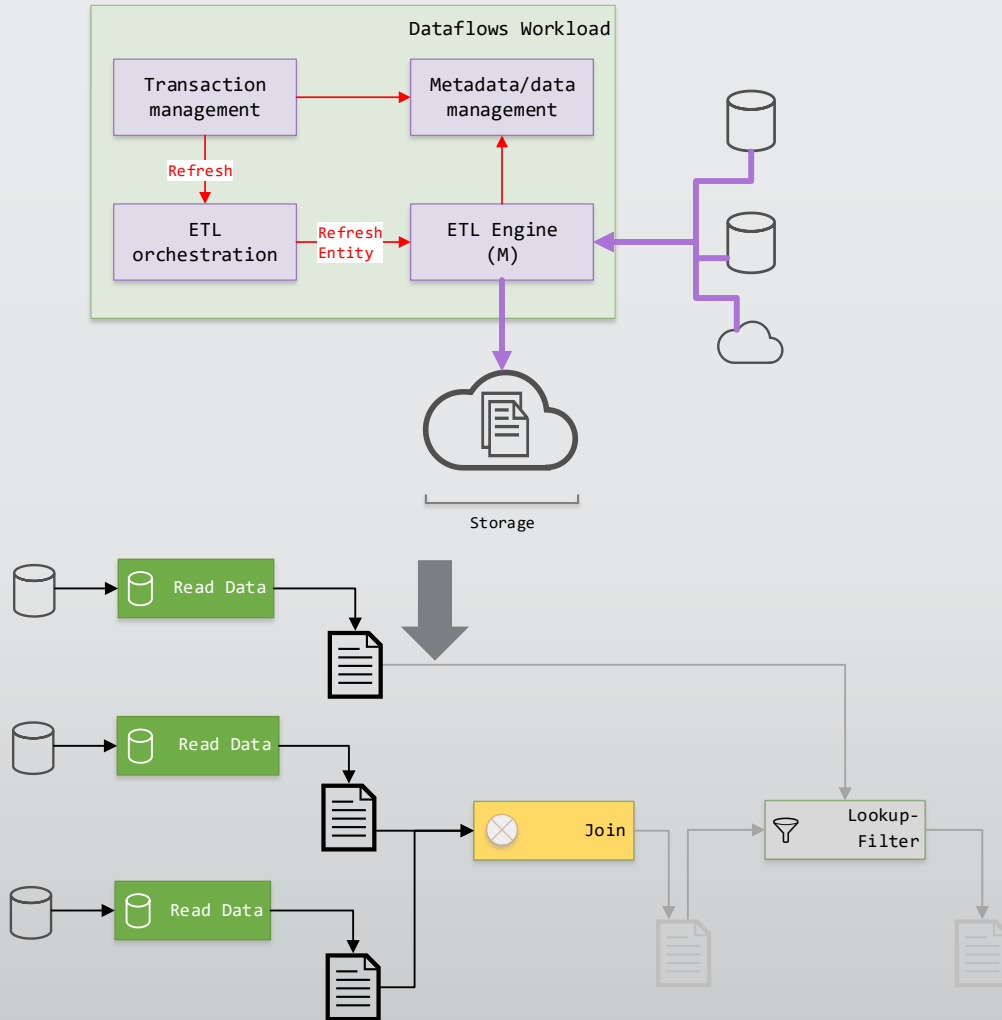
Architettura del common data model



Architettura del common data model



Dettagli sullo storage dataflow con premium capacity



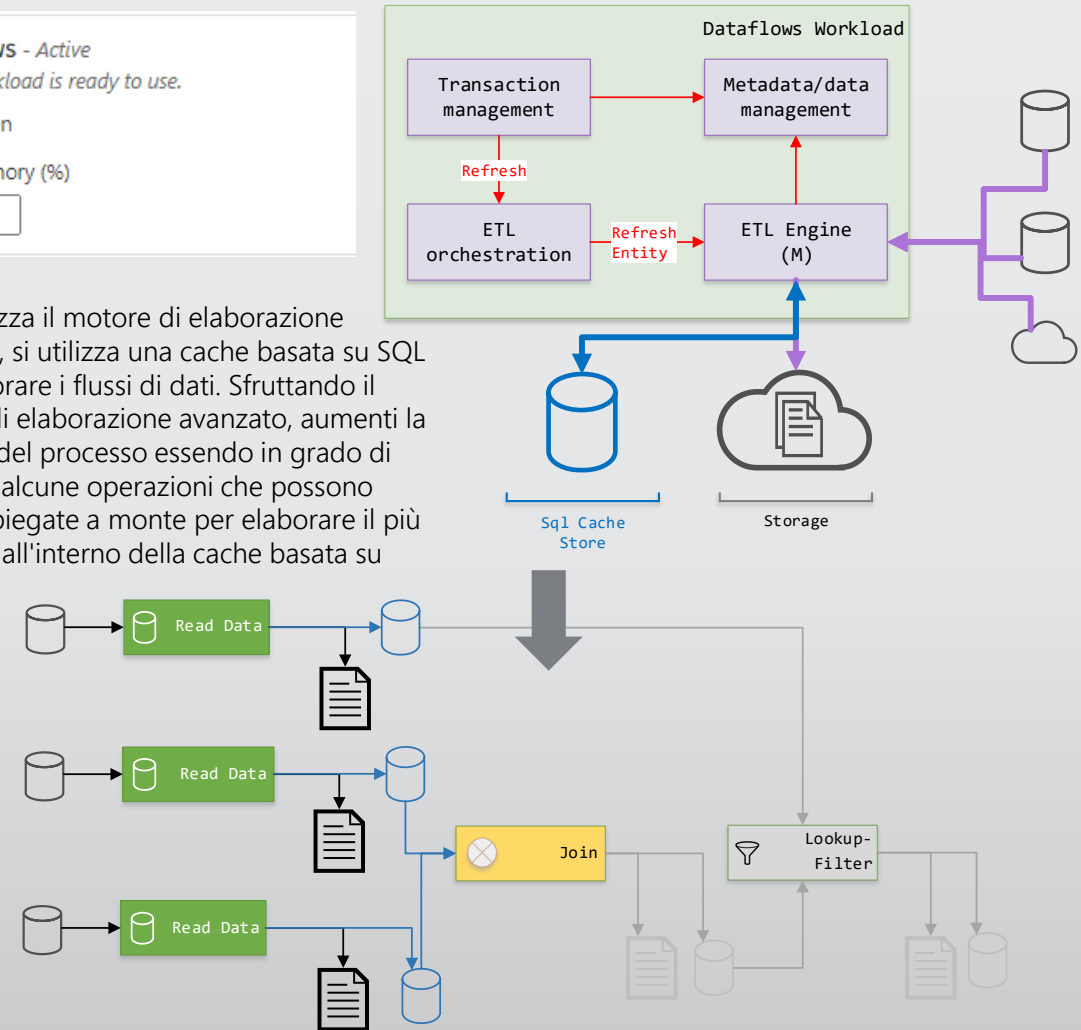
DATAFLOWS - Active
Your workload is ready to use.

☒ On

Max Memory (%)

20

Se si utilizza il motore di elaborazione avanzato, si utilizza una cache basata su SQL per elaborare i flussi di dati. Sfruttando il motore di elaborazione avanzato, aumenti la potenza del processo essendo in grado di spingere alcune operazioni che possono essere ripiegate a monte per elaborare il più possibile all'interno della cache basata su SQL.

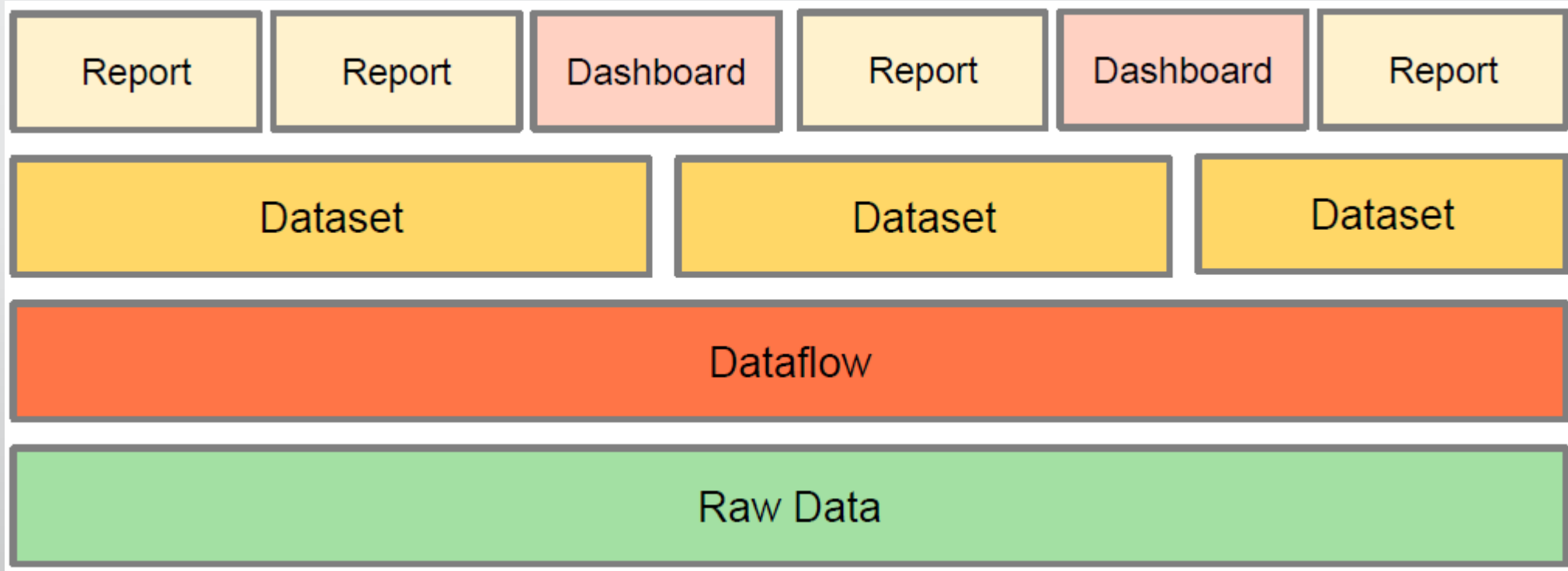


#GlobalPowerPlatformBootcamp

Progetta una soluzione Power BI con dataflow

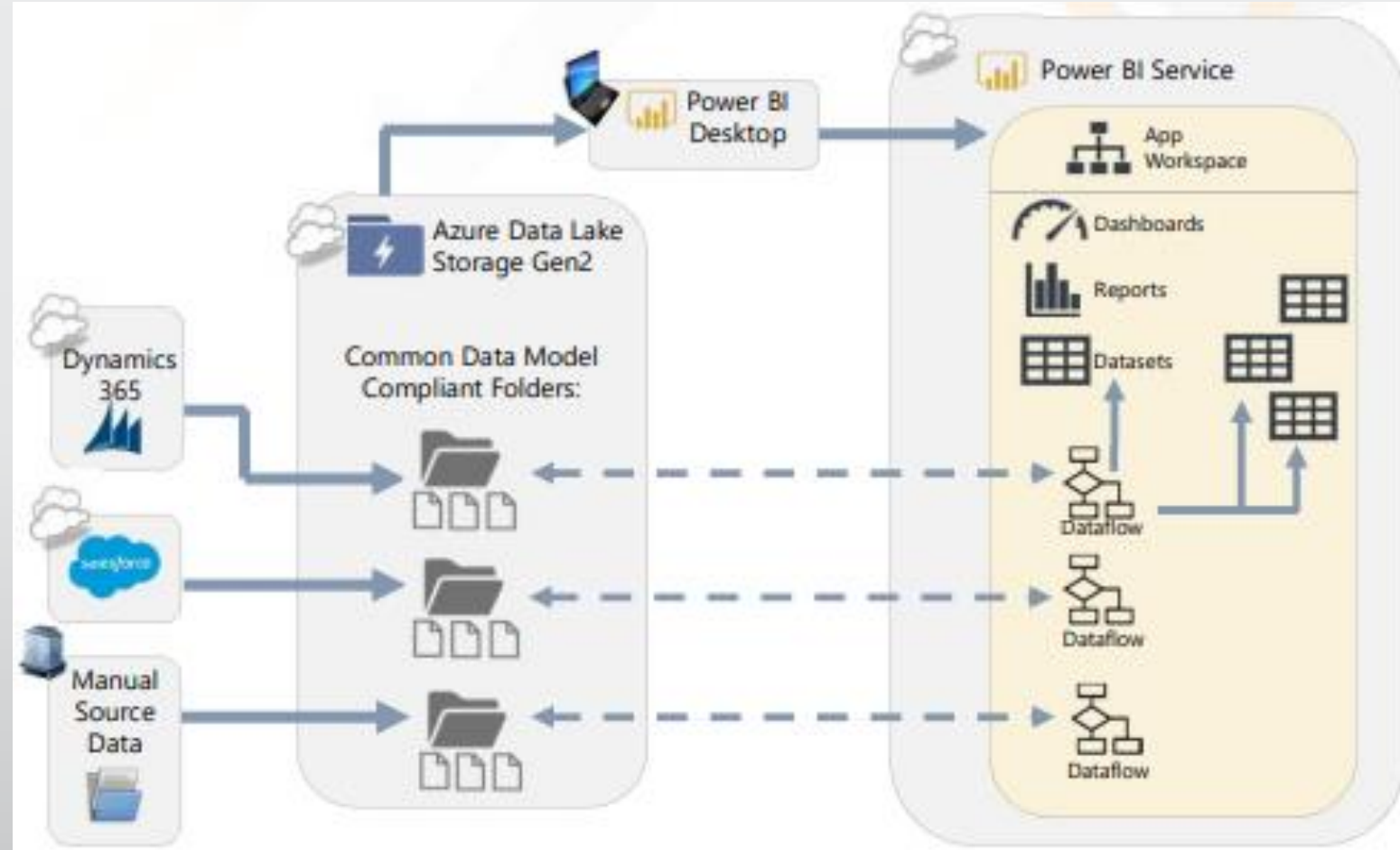
I Dtaflow sono utilizzati per raccogliere tutti i dati

- 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

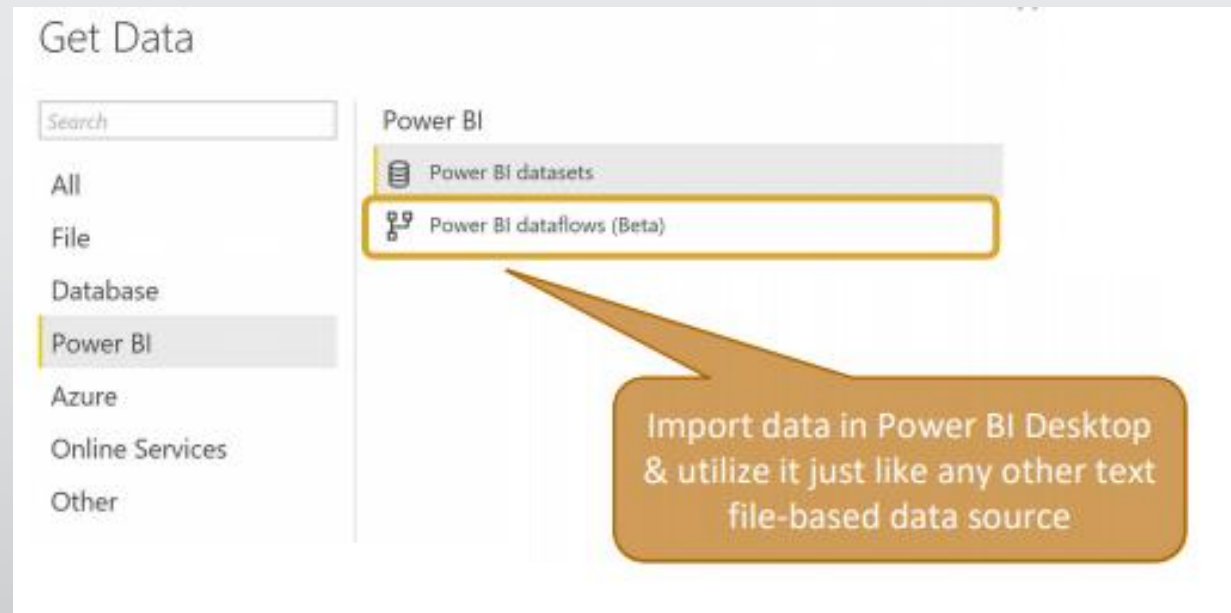
- Un singolo flusso di dati può alimentare i dati di molti dataset
- Un dataset rimane definito e contiene Calcoli, relazioni, RLS e altro
- Ci sono due aggiornamenti separati schedulati uno per il dataset e uno per il dataflow



Progetta una soluzione Power BI con dataflow

Entità del dataflow utilizzate da Power BI Desktop

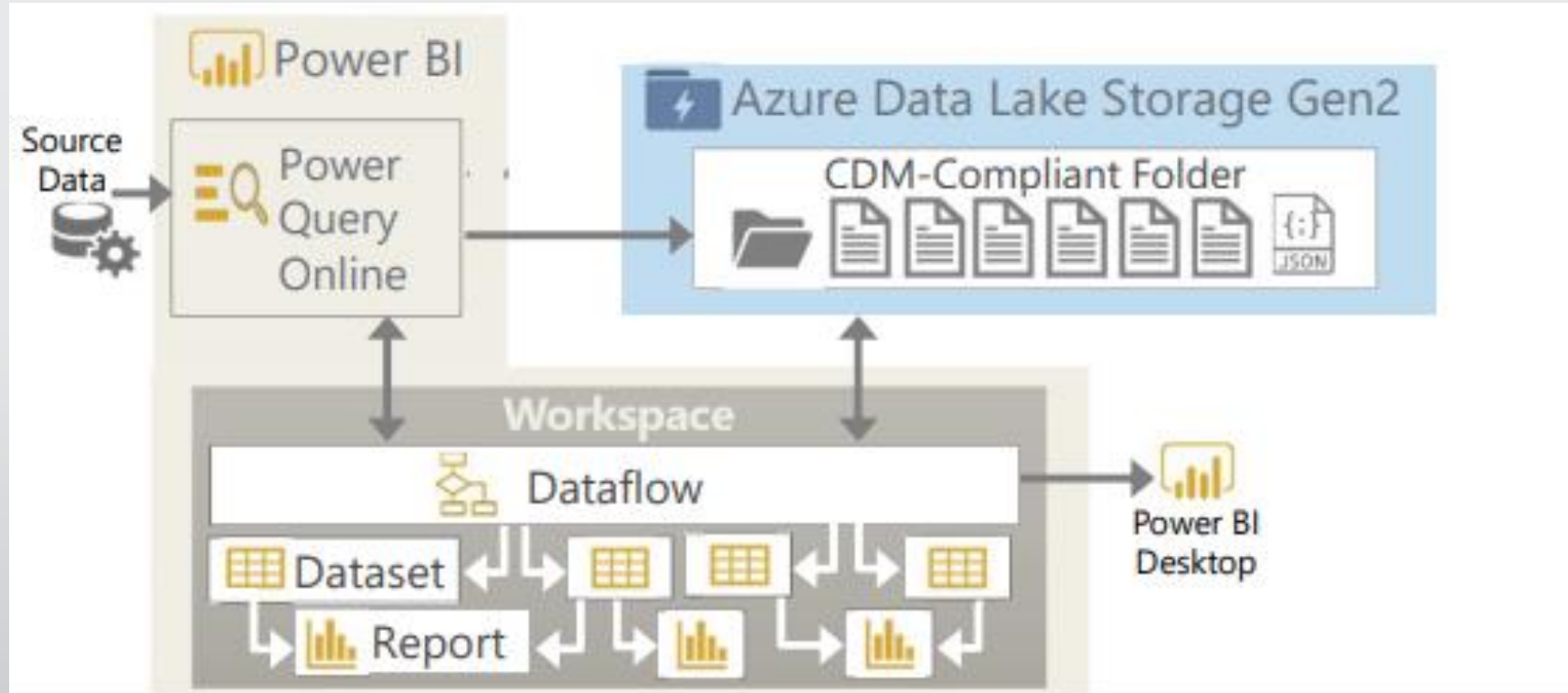
- Utilizzare Power BI Dataflow come sorgente
- Il dataset con il dataflow può essere pubblicato in qualsiasi area di lavoro dell'app



Tre modi di usare i Dataflow: Tipo 1

Struttura dati: Gestita da Power BI

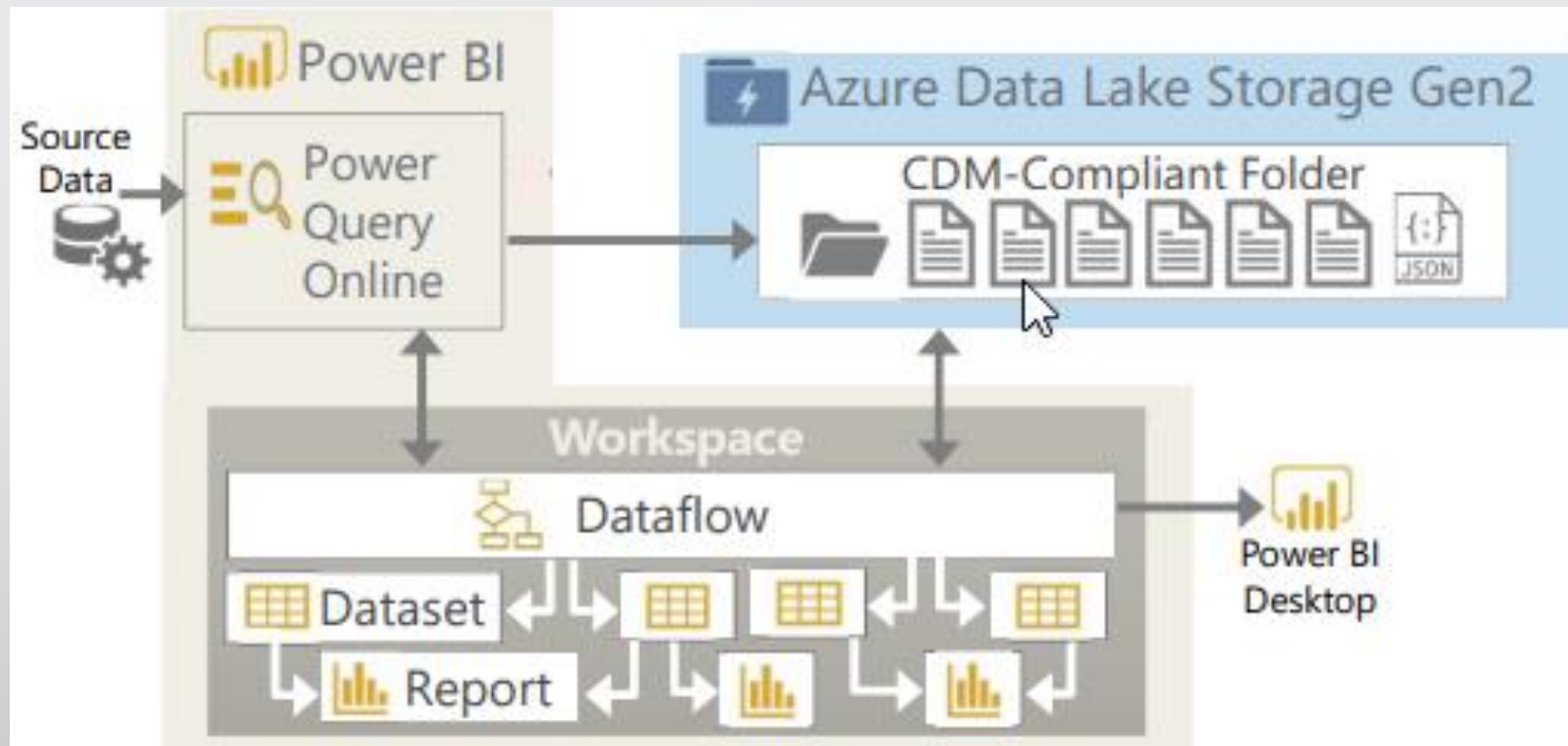
Data lake: I file non sono accessibili da altri tool



Tre modi di usare i Dataflow: Tipo 2

Struttura dati: Gestita da Power BI

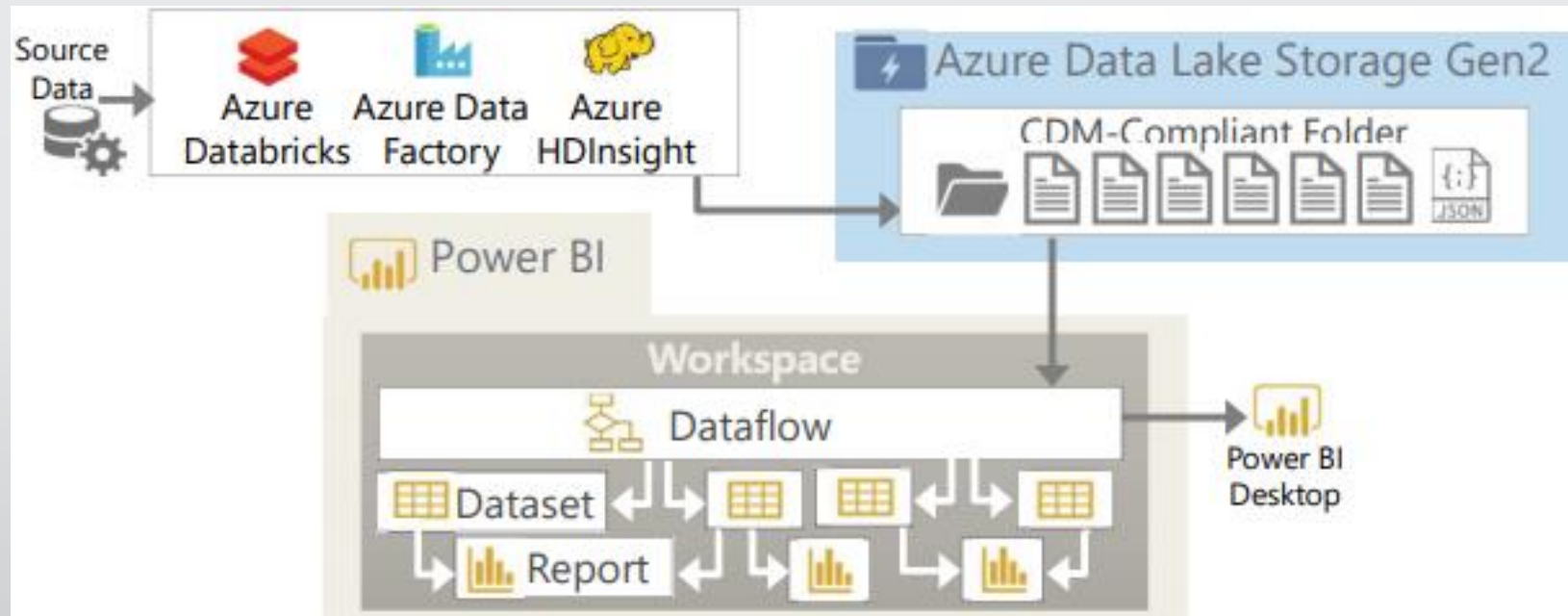
Data lake: lo storage di data lake è associato al tenant di Power BI



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



#GlobalPowerPlatformBootcamp





















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)

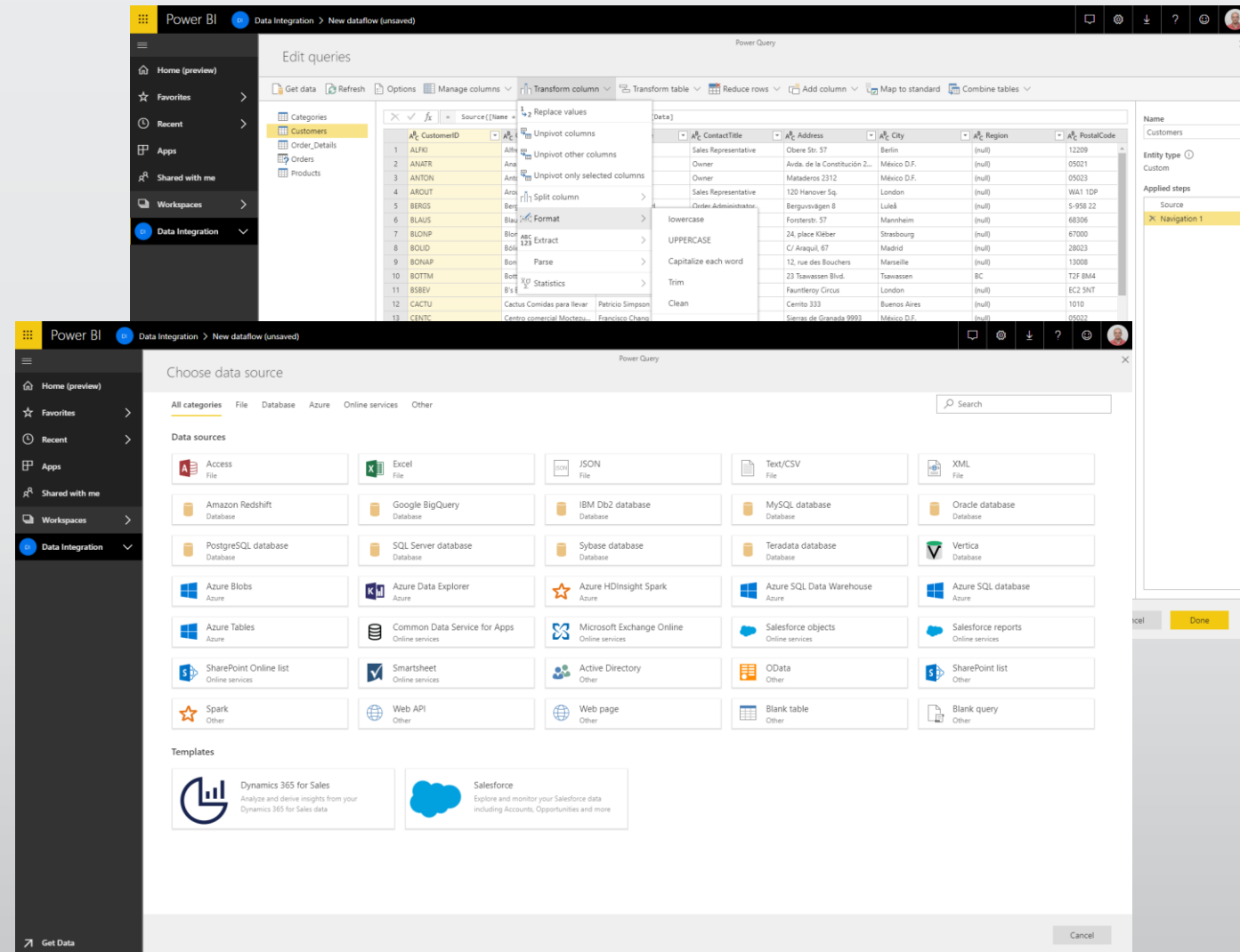
✓ Changes saved Edit entities Add entities | X Close

Entities Machine learning models

ENTITY NAME	ENTITY TYPE	ACTIONS
▶  Customers	Custom	   
▶  Sales	Custom	   
▶  Orders	Custom	   
▶  Products	Custom	   

Dataflows usano Power Query nel browser

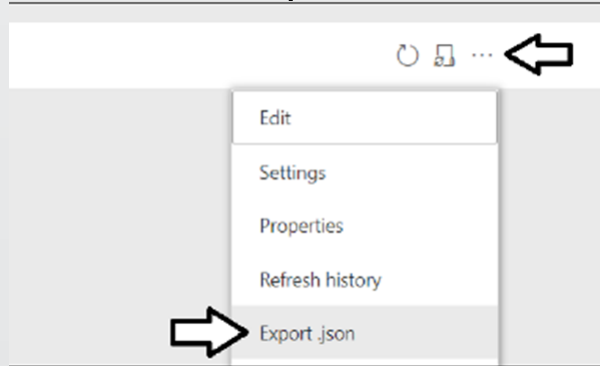
- Esperienza di modifica familiare agli utenti con Power Query
- Power Query è disponibile con un'esperienza utente di preparazione del dato di tipo **web-based self service**
- Supporta lo stesso numero di **300+ transformations** di PBI desktop Power Query (M Engine)
- Correntemente ci sono **~45 connectors**, 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



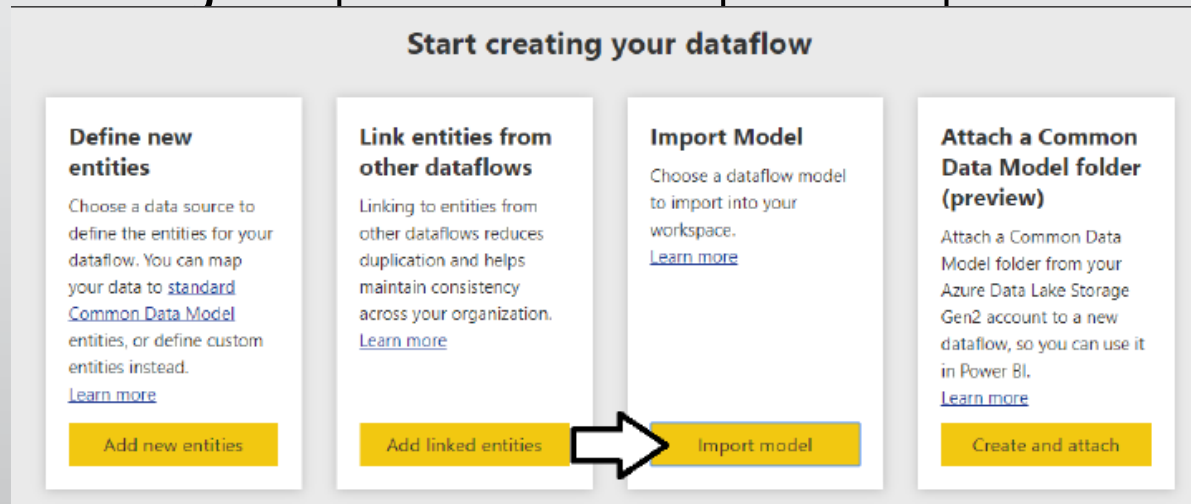
#GlobalPowerPlatformBootcamp

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



#GlobalPowerPlatformBootcamp

Licenze e Funzionalità DataFlow

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

- **PRO** I dataflow non possono essere creati in aree di lavoro personali
- PREMIUM I dataflow possono essere creati in aree di lavoro personali
- PREMIUM Entità collegate (**linked entities**)
- PREMIUM Entità calcolate (**computed entities**)
- PREMIUM Funzionalità di intelligenza artificiale (AI)
- PREMIUM Aggiornamento incrementale
- PREMIUM Esecuzione parallela delle trasformazioni
- PREMIUM **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



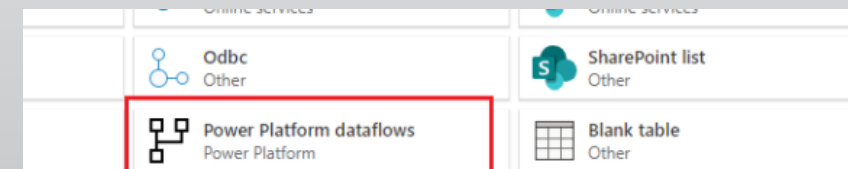
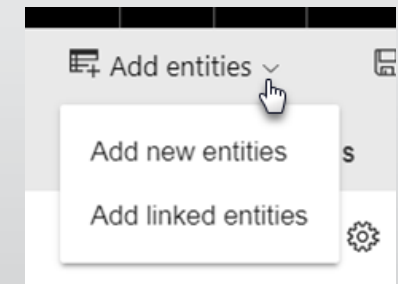
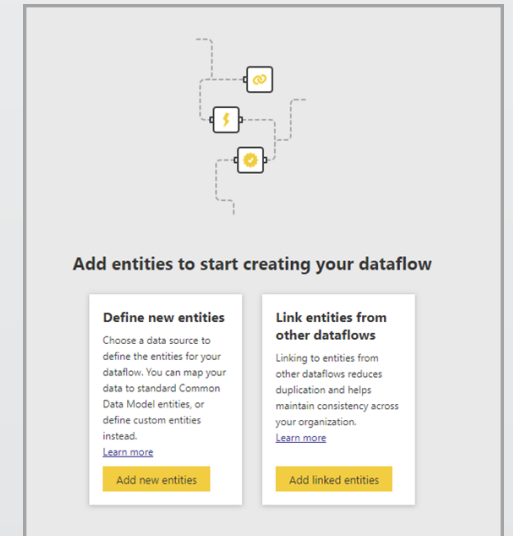
Funzionalità premium del Dataflow

- Entità collegate (**linked entities**)
- Entità calcolate (**computed entities**)
- Funzionalità di intelligenza artificiale (AI)
- Aggiornamento incrementale
- Motore di calcolo avanzato
- **Direct Query su dataflow**



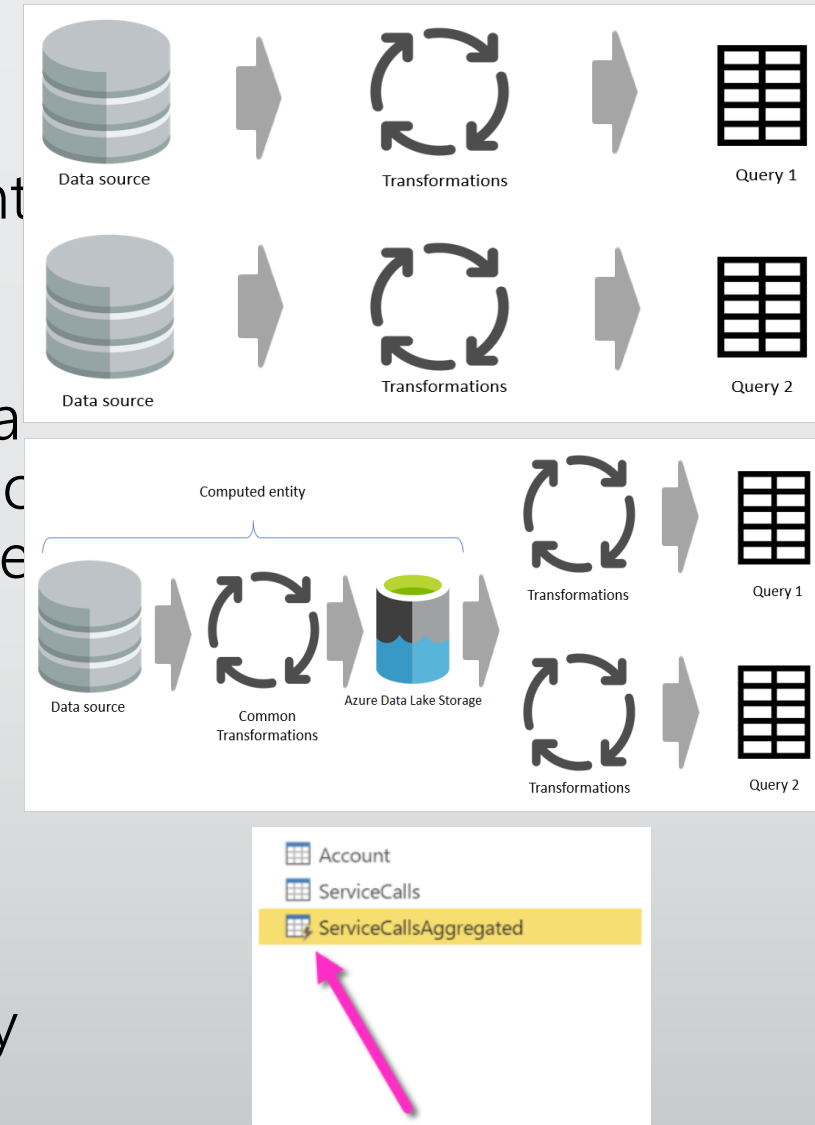
Entità collegate (Linked Entity)

- 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
 - Se viene usata un'entità calcolata, la parte comune (condivisa) della trasformazione verrà elaborata una volta e archiviata in ADLS. Le rimanenti trasformazioni verranno quindi elaborate dall'output della trasformazione comune. Nel complesso, questa elaborazione è molto più veloce.
- Scenari utili
 - Stai creando più entità all'interno di un dataflow dallo stesso dato non elaborato (source) e desideri che i dati non vengano presi dall'origine più di una volta
 - Stai avendo problemi a causa del motore di Power Query che richiede dati più volte in una singola query



Funzionalità di intelligenza artificiale (AI)

E' possibile usare l'intelligenza artificiale (IA) con i dataflow. Ci sono tre modi diversi di utilizzare l'AI:

- Servizi cognitivi
- Machine Learning automatizzato
- Integrazione con Azure Machine Learning



Cognitive Services

- **Servizi cognitivi** in Power BI è possibile applicare diversi algoritmi per arricchire i dati nella preparazione dei dati con i **dataflow**.
- I servizi attualmente supportati sono
 - Analisi dei sentiment,
 - Estrazione chiavi dalle frasi,
 - Rilevamento lingua
 - Assegnazione di tag alle immagini.
- Le trasformazioni vengono eseguite nel servizio Power BI e **non richiedono una sottoscrizione di Servizi cognitivi di Azure**.



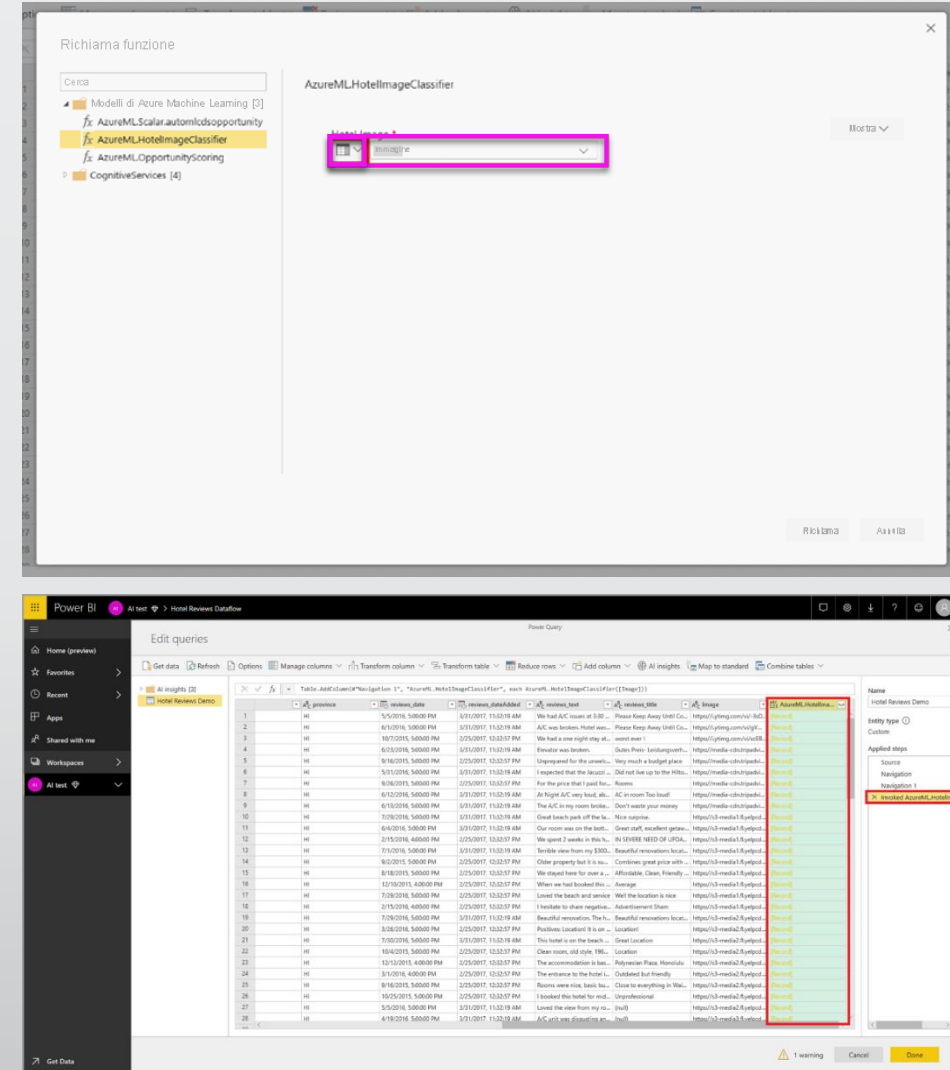
Auto ML

- Machine Learning automatizzato (**AutoML**) per i dataflow consente agli analisti aziendali di eseguire il training, convalidare e richiamare i modelli di Machine Learning (ML) direttamente in Power BI.
- Include una UI semplice per la creazione di un nuovo modello di ML in cui gli analisti possono usare i propri dataflow per specificare i dati di input per il training del modello.
- Il servizio estrae automaticamente le caratteristiche più rilevanti, seleziona un algoritmo appropriato e ottimizza e convalida il modello di ML.
- Dopo aver eseguito il training di un modello, Power BI genera automaticamente un report prestazioni che include i risultati della convalida. Il modello può quindi essere richiamato per tutti i dati nuovi o aggiornati all'interno di un dataflow.



Integrazione con Machine Learning

- Ora con Power BI è facile incorporare le informazioni dettagliate dei modelli ospitati in Azure Machine Learning, usando semplici passi e con pochi click.
- Per usare questa funzionalità, un data scientist può semplicemente concedere all'analista di Power BI l'accesso al modello di Azure Machine Learning usando il portale di Azure.
- All'inizio di ogni sessione, Power Query individua tutti i modelli di Azure Machine Learning a cui l'utente ha accesso e li espone come funzioni dinamiche di Power Query. (semplice chiamata ad una funzione in M)



Aggiornamento Incrementale

- In alcuni casi trasferire una copia completa dei dati in Power BI ad ogni aggiornamento non è pratico. L'**aggiornamento incrementale**, offre i seguenti vantaggi:
 - L'aggiornamento avviene più velocemente: è necessario aggiornare solo i dati modificati. Ad esempio, aggiorna solo gli ultimi cinque giorni di un dataflow di 10 anni.
 - L'**aggiornamento è più affidabile**: ad esempio, non è necessario mantenere connessioni di lunga durata a sistemi di origine volatili.
 - Il **consumo di risorse è ridotto**: meno dati da aggiornare riduce il consumo complessivo di memoria e altre risorse.
 - L'aggiornamento incrementale è disponibile nei dataflow creati in Power BI e nei dataflow creati in Power Apps.

Incremental refresh settings

Appointment

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

createdon

Store rows from the past

5 Years

Refresh rows from the past

10 Days

☐ Detect data changes [Learn more](#)

Only refresh data if the maximum value in this field changes

Choose a field

☐ Only refresh complete [Learn more](#)

When you save these settings, data from the past [storage period] will be loaded to your dataflow storage the next time this dataflow is refreshed. Subsequent refreshes will update only data that's changed in the past [refresh period].

Save Cancel

Motore di calcolo avanzato

- Il motore di calcolo avanzato in Power BI consente ai sottoscrittori di Power BI Premium di usare la propria capacità per ottimizzare l'uso dei dataflow. L'uso del motore di calcolo avanzato offre i vantaggi seguenti:
 - **Riduzione drastica del tempo di aggiornamento necessario per i passaggi ETL** con esecuzione prolungata sulle entità calcolate, ad esempio l'esecuzione di operazioni **join, distinct, filter e group by**
 - Eseguire query DirectQuery sulle entità
- Motore di calcolo avanzato solo con Power BI Premium A3

FLUSSI DI DATI - Attivo

Il carico di lavoro è pronto per l'uso.

☒ Sì

Memoria massima (%)

Motore di calcolo dei flussi di dati avanzato (anteprima)

☒ Sì

Dimensioni del contenitore (MB)

Direct Query su Dataflow (Motore di calcolo avanzato)

- È possibile usare DirectQuery per connettersi direttamente ai dataflow e quindi connettersi direttamente al proprio flusso di dati senza dover importare i dati.
- L'uso di DirectQuery con i dataflow consente i miglioramenti seguenti:
 - **Evita pianificazioni di aggiornamenti separate:** DirectQuery si connette direttamente a un flusso di dati, eliminando la necessità di creare un set di dati importato. Di conseguenza, non sono più necessarie pianificazioni di aggiornamenti separate per il dataflow e il dataset per la sincronizzazione dei dati.
 - **Filtro dei dati:** DirectQuery è utile per lavorare su una visualizzazione filtrata dei dati all'interno di un dataflow.

4 Impostazioni del motore di calcolo avanzato (anteprima)

Configura le impostazioni del motore di calcolo avanzato per questo flusso di dati.

☐ Disabilitato

Disattiva il motore di calcolo avanzato per questo flusso di dati.

☐ Ottimizzato

Il motore di calcolo avanzato verrà attivato solo quando questo flusso di dati sarà collegato a un altro flusso di dati al fine di ottimizzare le prestazioni.

☒ Attivato

Attiva il motore di calcolo avanzato per questo flusso di dati.



Importare da un data lake esterno

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

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



Any Questions?



#GlobalPowerPlatformBootcamp

Please fill out the survey!



<http://bit.ly/GPPBITFORM>

#GlobalPowerPlatformBootcamp



Thank You For Attending

<<speaker contacts>>



#GlobalPowerPlatformBootcamp