Microsoft Fabric in Action

Data Engineering Case Study

Vinodh Kumar TD

Azure Data Architect, Presidio

About Me…

- Microsoft Certified Trainer x3
- C#SharpCorner MVP x3
- Global Speaker
- Data Blogger → vinsdata.wordpress.com

  Vinodh Kumar (c-sharpcorner.com)

What is Microsoft Fabric?

Microsoft Fabric is a cloud-based data platform that provides a range of services for data engineering, data science, and business intelligence.

- It is a unified platform
- brings together Power BI, Data Factory, Big data processing, and the Data Lake, on a new generation of the Synapse data infrastructure
- Computing power is measured in SKU capacities
- Available on both pay-as-you-go and reservation models

What is a Capacity?

- A Capacity is a pool of resources (which can be measured by CU: Capacity Unit) for dedicated use. These resources can be included but not limited to CPU, Memory, etc., on the Microsoft Azure environment, which gives you the computing power to process the Fabric services.

SKU	Capacity Units (CU)	HOURLY $	MONTHLY $
F2	2	$0.36	$262.80
F4	4	$0.72	$525.60
F8	8	$1.44	$1,051.20
F16	16	$2.88	$2,102.40
F32	32	$5.76	$4,204.80
F64	64	$11.52	$8,409.60
F128	128	$23.04	$16,819.20
F256	256	$46.08	$33,638.40
F512	512	$92.16	$67,276.80
F1024	1024	$184.32	$134,553.60
F2048	2048	$368.64	$269,107.20

Tenant

- In the context of Microsoft Fabric, a tenant is a logical container that holds all the resources and services required for a specific organization. It is tied to a specific Domain Name System (DNS) and can have multiple capacities active at the same time.

Workspace

- On the other hand, a workspace is a container for Microsoft Fabric items and resides within a capacity. Each Microsoft Fabric user has a personal workspace known as My Workspace, and more workspaces can be created to enable collaboration.

Fabric Items

All-in-one

Store data

- Lakehouse
- Warehouse
- KQL Database

Prep & query data

- Notebook
- Dataflow gen2
- Data pipeline
- KQL Queryset
- Spark Job Definition

Train and use ML models

- Experiment
- Model

Present data

- Scorecard
- Report
- Dashboard
- Paginated Report
- Realtime Dashboard

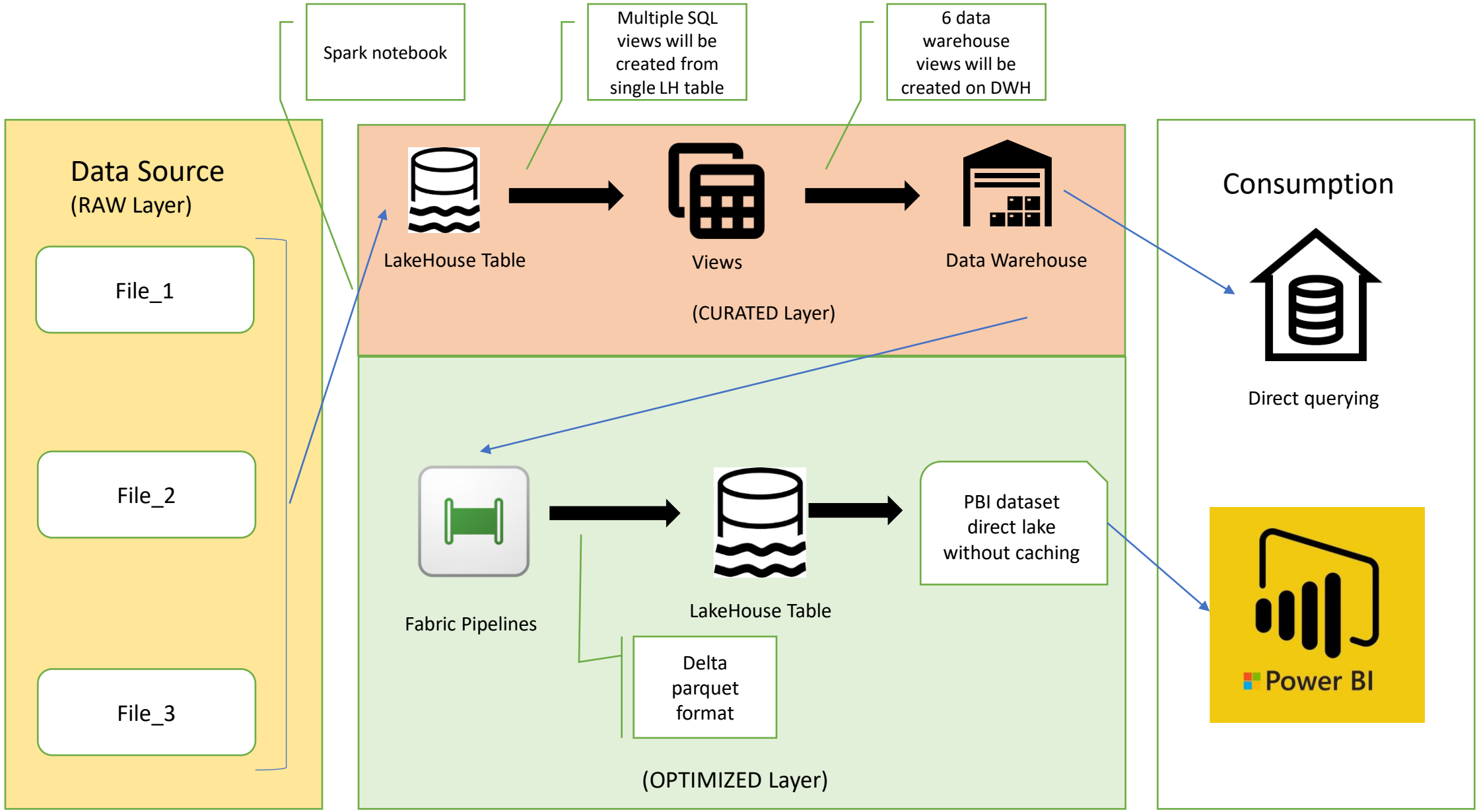
Manage real-time data

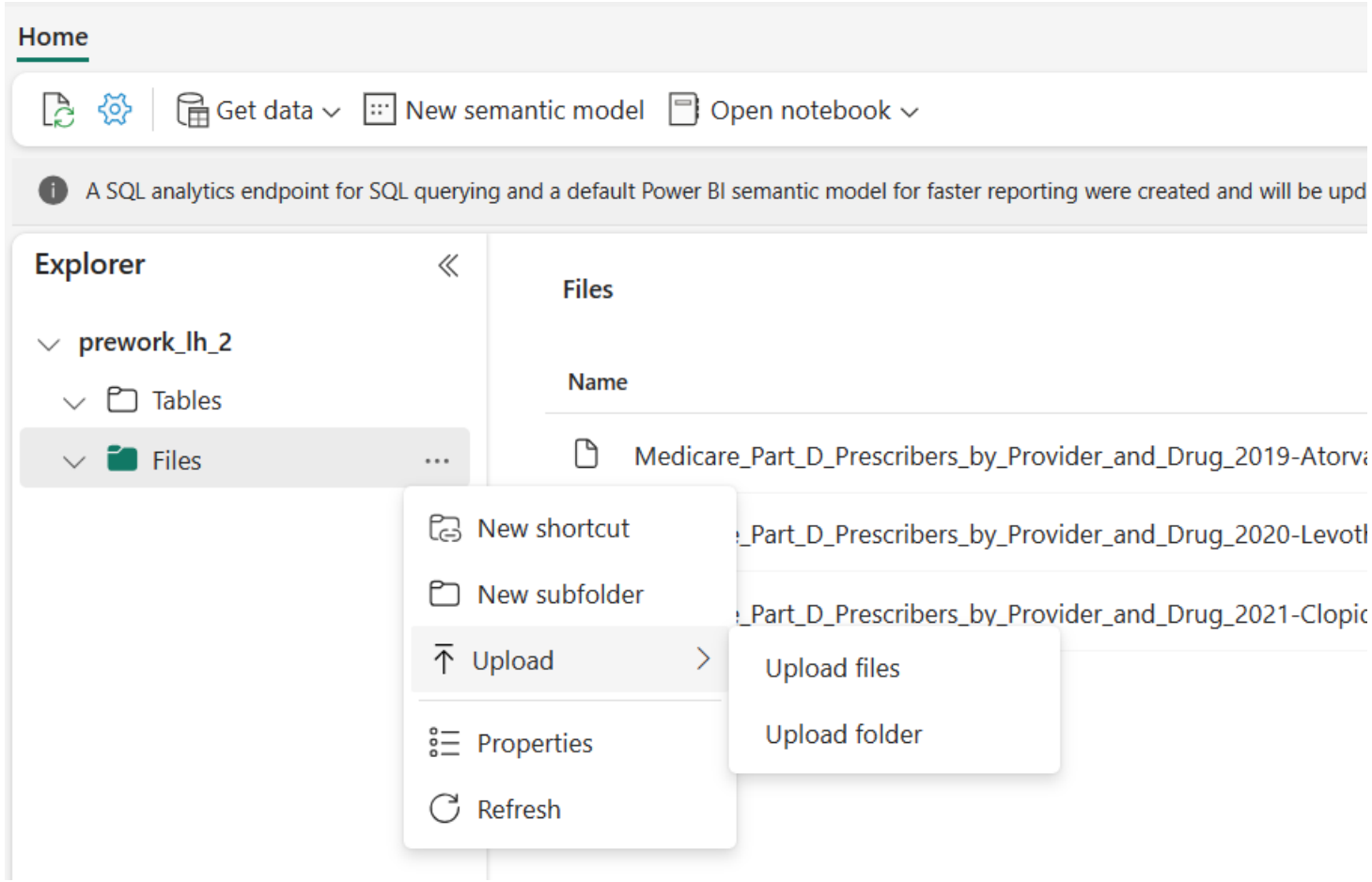
- Eventstream
- Streaming Dataflow
- Streaming Dataset

@cmaneu

Agenda for today's use case:

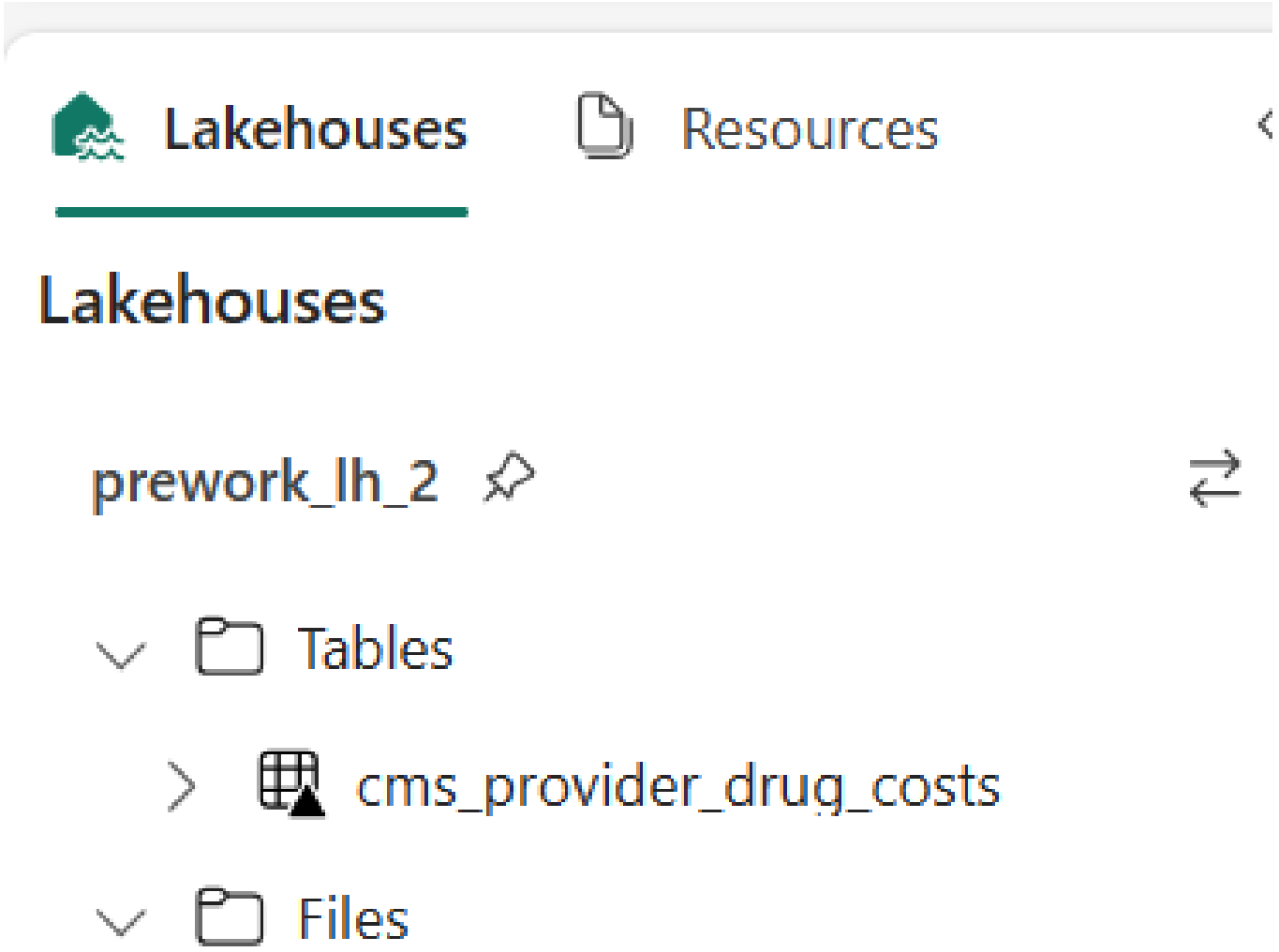
- Join multiple files from Azure storage to form a single LakeHouse Table
- Create many views from the flattened LakeHouse table using SQL/Spark notebook
- Create Fact & Dim tables in the Lakehouse from the SQL views using the Fabric Pipeline activity
- Create a new semantic model to be used in the PBI report for the following tables
  - cms_provider_dim_drug
  - cms_provider_dim_geography
  - cms_provider_dim_provider
  - cms_provider_dim_year
  - cms_provider_drug_costs_star
- Create a PBI dataset to be saved within the WS by creating a relationship between fact and dim tables
- Rename the columns into user-friendly names
- Add DAX measures
- Create a new PBI report calling the saved dataset/semantic model


Spark notebook
Multiple SQL views will be created from single LH table
6 data warehouse views will be created on DWH
Data Source
(RAW Layer)
File_1
File_2
File_3
LakeHouse Table
Views
Data Warehouse
(CURATED Layer)
Fabric Pipelines
LakeHouse Table
PBI dataset direct lake without caching
Delta parquet format
(OPTIMIZED Layer)
Consumption
Direct querying
Power BI



Creating LakeHouse from workspace dashboard and uploading files

- Once after successfully running the spark file, a table will be created



Questions??