DATA SATURDAYS MADRID

Fabric End-To-End

Miguel Félix

Sponsors:

alesonITC data thinking · NTT DATA · Verne TECH · FOUR9S · V-Valley enhancing your business · TD SYNNEX · waibbo · Namasdata

#DataSatMadrid

1

¡Gracias sponsors!

Platinum

alesonITC data thinking · NTT DATA · FOUR9S · Verne TECH

Gold

V-Valley enhancing your business · TD SYNNEX

Silver

waibbo · Namasdata

#DataSatMadrid

2

Miguel Félix

Lead BI Architect

- Microsoft Data Platform + Fabric MVP
- Microsoft MCT
- Super User at Fabric Community
- Fabric Power BI Portugal meetup

https://pbiportugal.com/

https://www.linkedin.com/in/miguel-a-s-felix/

@MFelixPBI

#DataSatMadrid

9

Session Goal

- Implement a Fabric solution Beginning to end
- Learn the main concepts in Fabric
- Produce a analytics solution using Fabric capabilities

#DataSatMadrid

10

Agenda

- Fabric main concept
- Starting Fabric Trial Capacity
- Creation of a lakehouse
- ELT process
  - Dataflow
  - Notebook
  - Pipeline
- Creation of a semantic
- Creation of a report
- Maintenance of semantic model using notebooks

11

Requirements / Resources

Requirements

Computer: Windows, Mac, Linux

Fabric Trial capacity: Subscription steps will be presented during workshop

Power BI Service account created (can be a free account)

Use this link if you don't have an account: Start Session in Power BI

Minimum 1 year of Power BI knowledge

Understanding of basic data concepts and structures

Understanding of fundamentals in modeling and DAX

Basic SQL (not mandatory)

Not covered in workshop

Star schema, aggregations, DAX context, and others won't be covered in detail

Advance capabilities of Scala, Python, R won't be covered on the workshop

Advanced Data visualizations

Resources and files

https://github.com/MASFelixPBI/DataSaturdaysMadrid24

SCAN ME

12

FABRIC

Main concepts

#DataSatMadrid

13

Fabric

The data platform for the era of AI

Complete Analytics Platform	Lake centric and open	Empower Every Business User	AI Powered
Everything, unified	OneLake	Familiar and intuitive	Copilot accelerated
SaaS-ified	One Copy	Built into Microsoft 365	ChatGPT on your data
Secured and governed	Open at every tier	Insight to action	AI driven insights

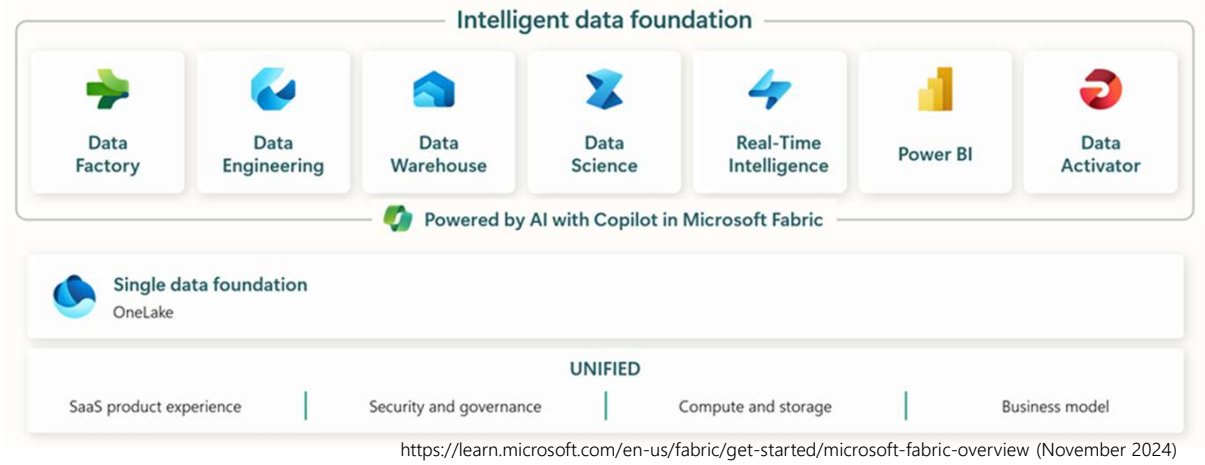
#DataSatMadrid

14

Fabric

The data platform for the era of AI

An end-to-end analytics platform that brings together all the data and analytics tools that organizations need to go from the data lake to the business user



https://learn.microsoft.com/en-us/fabric/get-started/microsoft-fabric-overview (November 2024)

15

OneLake for all Data

"The OneDrive for Data"

- A single SaaS lake for the whole organization
- Provisioned automatically with the tenant
- All workloads automatically store their data in the OneLake workspace folders
- All the data is organized in an intuitive hierarchical namespace
- The data in OneLake is automatically indexed for discovery, MIP labels, lineage, PII scans, sharing, governance and compliance

OneDrive

OneLake

OneLake provides a data lake as a service without you needing to build it

16

OneLake for all Data

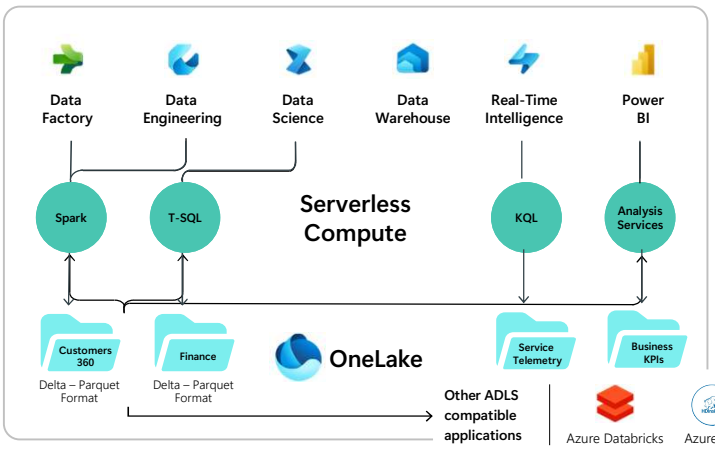
"The OneDrive for Data"

All the compute engines store their data automatically in oneLake as data items.

The data is stored in a single common format.

Delta – Parquet, an open standards format, and it is the storage format for all tabular data in Fabric.

All the compute engines have been fully optimized to work with Delta Parquet as their native format.

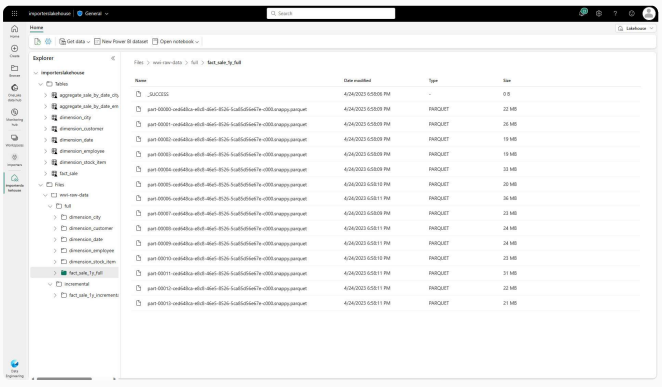


Lakehouse

Store, manage and analyze all your data in a single location & easily share across the entire enterprise

Key Capabilities:

- Flexible and scalable solution that enables organizations to handle large data volumes of all types and sizes
- Built-in SQL endpoint unlocks data warehouse capabilities on top of your Lakehouse with no data movement
- Use 'direct lake' mode to build reports in seconds directly on top of the data lake with blazing fast performance
- Easily ingest data into the Lakehouse through a variety of methods
- Share your Lakehouse as a data product with consumers



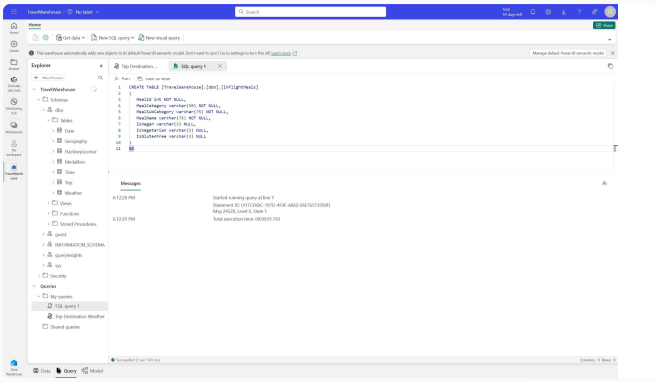
Warehouse

Enterprise scale data warehouse with open standard format

No knobs performance with minimal set-up and deployment, no configuration of compute or storage needed

Key Capabilities:

- Lake-centric warehouse stores data in OneLake in open Delta format with easy data recovery and management
- Use Fabric Mirroring for Zero-ETL integration of data from Azure SQL, Snowflake, or Azure Cosmos DB
- Data loading and transforms at scale, with full multi-table transactional guarantees provided by the SQL engine
- Virtual warehouses with cross-database querying and a fully integrated semantic layer
- Flexibility to build data warehouse or data mesh based on organizational needs and choice of no-code, low-code, or T-SQL for transformations



Warehouse vs Lakehouse

	Warehouse	Lakehouse
Data volume	Unlimited	Unlimited
Type of data	Structured	Unstructured, semi-structured, structured
Primary developer persona	Data warehouse developer, SQL engineer	Data engineer, data scientist
Primary dev skill	SQL	Spark(Scala, PySpark, Spark SQL, R)
Data organized by	Databases, schemas, and tables	Folders and files, databases, and tables
Read operations	T-SQL, Spark*	Spark, T-SQL
Write operations	T-SQL	Spark(Scala, PySpark, Spark SQL, R)
Multi-table transactions	Yes	No
Primary development interface	SQL scripts	Spark notebooks,Spark job definitions
Security	Object level, RLS, CLS, DDL/DML, dynamic data masking	RLS, CLS**, table level (T-SQL), none for Spark
Access data via shortcuts	Yes	Yes
Can be a source for shortcuts	Yes (tables)	Yes (files and tables)
Query across items	Yes	Yes
Advanced analytics	Interface for large-scale data processing, built-in data parallelism and fault tolerance	Interface for large-scale data processing, built-in data parallelism and fault tolerance
Advanced formatting support	Tables defined using PARQUET, CSV, AVRO, JSON, and any Apache Hive compatible file format	Tables defined using PARQUET, CSV, AVRO, JSON, and any Apache Hive compatible file format
Ingestion latency	Available instantly for querying	Available instantly for querying

Dataflow Gen2

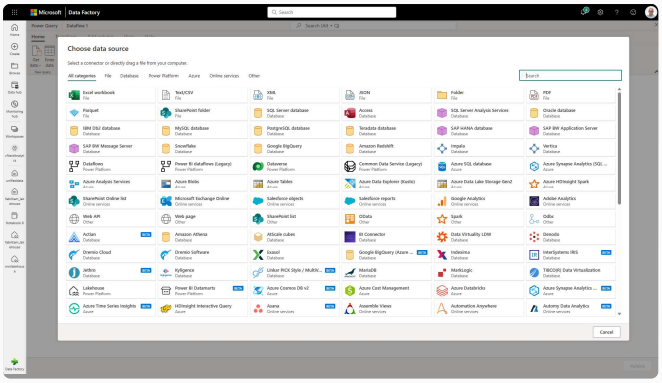
Dataflow provides a low-code interface for ingesting data from hundreds of data sources

Dataflow quickly and easily unify disparate data sources, establish a more collaborative analytics approach, and promote more informed, agile decision making.

Key Capabilities:

- Accelerate data transformation with code-free data flows
- Scale out using Fabric compute and Data Factory fast copy
- Load results of data transformations into multiple destinations (Azure SQL Databases, Lakehouse, etc.)

Simply write into a Lakehouse from a Dataflow



Dataflow Gen2 vs Gen1

Dataflow provides a low-code interface for ingesting data from hundreds of data sources

Dataflow quickly and easily unify disparate data sources, establish a more collaborative analytics approach, and promote more informed, agile decision making.

Key Capabilities:

- Accelerate data transformation with code-free data flows
- Scale out using Fabric compute and Data Factory fast copy
- Load results of data transformations into multiple destinations (Azure SQL Databases, Lakehouse, etc.)

Simply write into a Lakehouse from a Dataflow

Feature	Dataflow Gen2	Dataflow Gen1
Author dataflows with Power Query	✓	✓
Shorter authoring flow	✓	
Auto-Save and background publishing	✓	
Data destinations	✓	
Improved monitoring and refresh history	✓	
Integration with data pipelines	✓	
High-scale compute	✓	
Get Data via Dataflows connector	✓	✓
Direct Query via Dataflows connector		✓
Incremental refresh		✓
AI Insights support		✓

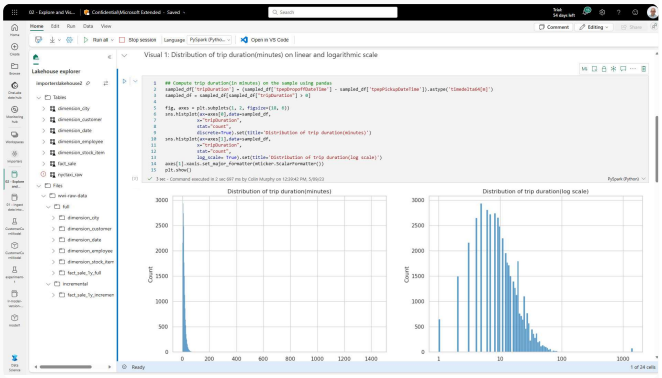
Notebooks

Immersive authoring experience for data developers

Rich notebook capabilities including native Lakehouse integration, real-time collaboration with commenting support, auto-save support, lightweight scheduling and pipeline integration

Key Capabilities:

- Manage your Python and R libraries through in-line installs using commands like %pip install
- Advanced notebook development support with ability to reference notebooks in notebooks, and snapshots for easy troubleshooting
- In context monitoring complete with real time advice and error analysis
- Streamline data prep without giving up the power of reproducibility of Python



Data Pipelines

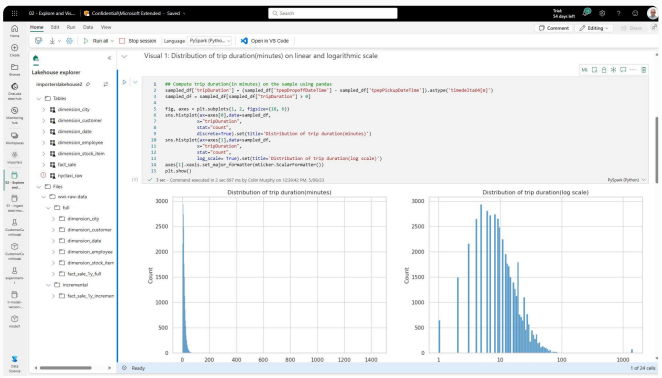
Data Pipelines enable powerful workflow capabilities at cloud-scale like building complex workflows, moving PB-size data, and defining sophisticated control flow pipelines

Data pipelines can be used to build complex ETL and data factory workflows that can perform a number of different tasks at scale.

Sample Datasets helps new users get started quickly, building out their ELT processes using Data Pipelines

Simply copying data to a Lakehouse with **copy assist capabilities** within the Data Pipeline.

Template help reduce development time by providing an easy way to create pipeline for common data integration scenarios



Data Pipelines

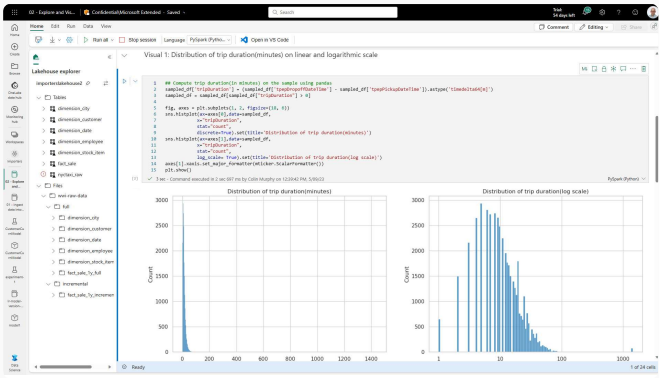
Data Pipelines enable powerful workflow capabilities at cloud-scale like building complex workflows, moving PB-size data, and defining sophisticated control flow pipelines

Data pipelines can be used to build complex ETL and data factory workflows that can perform a number of different tasks at scale.

Sample Datasets helps new users get started quickly, building out their ELT processes using Data Pipelines

Simply copying data to a Lakehouse with **copy assist capabilities** within the Data Pipeline.

Template help reduce development time by providing an easy way to create pipeline for common data integration scenarios



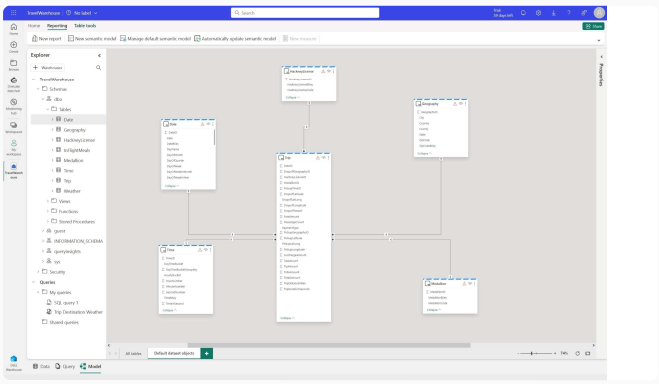
Semantic Model

Reduce integration and gain insight from your data in seconds

Built-in Power BI enables everyone to visualize their data in seconds.

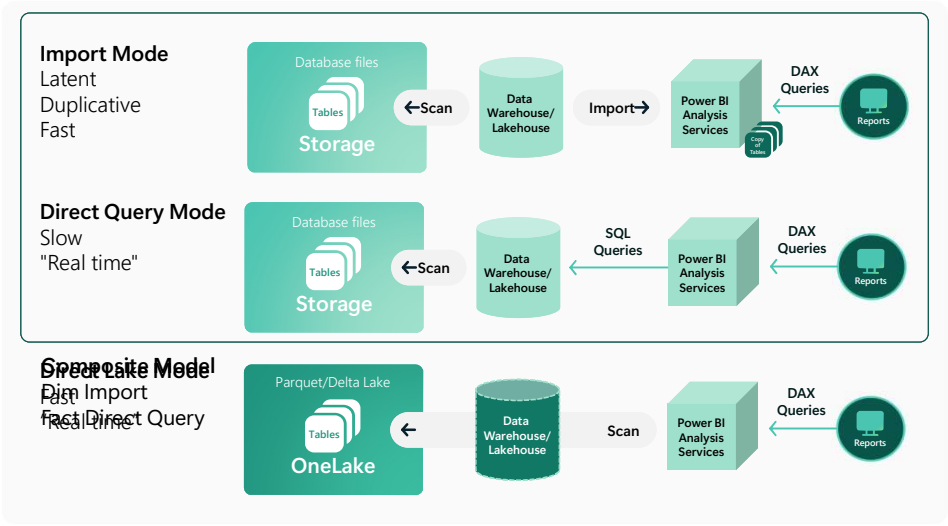
Key Capabilities:

- Auto-generated semantic models always in sync
- Default dataset in Direct Lake mode but automatically switches to Direct Query or Import mode as security or performance needs change
- Flexibility to add/remove tables to dataset
- Full web authoring experience for creating measures



Direct Lake

Best of Import and DirectQuery for High Scale, Performant, Real Time Reporting



Composite Model
Dim Import
Fact Direct Query

#DataSatMadrid

27

Direct Lake

Best of Import and DirectQuery for High Scale, Performant, Real Time Reporting

- Semantic model capability for analysing very large data volumes.
- Based on loading parquet-formatted files directly from a data lake without having to query a Lakehouse or Warehouse,
- Eliminates the import requirement by loading the data directly from OneLake. (removes duplication of data)
- Fast-path to load the data from the lake straight into the Power BI engine, ready for analysis.
- Unlike DirectQuery, there is no translation from DAX or MDX to other query languages or query execution on other database systems, yielding performance

- Avalilable on Lakeahouse or Warehouse
- SQL Endpoint for querying and a default model with all the tables
- XMLA endpoint read-write support using tools like SSSM, Tabular Editor and DAX Studio, **Power BI Desktop**
- Database object Compatibility level 1604 or higher
- Fallback occurs when exceeds the limits for the SKU or features that don't support Direct Lake (settings available)
- Data changes are automatically reflected (settings available)

#DataSatMadrid

29

Direct Lake

Fallback

Fabric/Power BI SKUs	Parquet files per table	Row groups per table	Rows per table (millions)	Max model size on disk/OneLake[1] (GB)	Max memory (GB)
F2	1,000	1,000	300	10	3
F4	1,000	1,000	300	10	3
F8	1,000	1,000	300	10	3
F16	1,000	1,000	300	20	5
F32	1,000	1,000	300	40	10
F64/FT1/P1	5,000	5,000	1,500	Unlimited	25
F128/P2	5,000	5,000	3,000	Unlimited	50
F256/P3	5,000	5,000	6,000	Unlimited	100
F512/P4	10,000	10,000	12,000	Unlimited	200
F1024/P5	10,000	10,000	24,000	Unlimited	400
F2048	10,000	10,000	24,000	Unlimited	400

Direct Lake semantic models read delta tables directly from OneLake.

A query can fall back to DirectQuery mode when:
- DAX query exceeds limits for the SKU,
- Usage of features that don't support Direct Lake mode (example SQL views in a Warehouse)

DirectQuery mode use SQL to retrieve the results from the SQL endpoint of the Lakehouse or Warehouse, which can have an impact on query performance.

Guardrails define resource limits for Direct Lake mode beyond which a fallback to DirectQuery mode is necessary to process DAX queries.

Max Memory represents the upper memory resource limit for how much data can be paged in.

30

¿Preguntas?

31

DATA SATURDAYS MADRID

Thank you

https://pbiportugal.com/

https://www.linkedin.com/in/miguel-a-s-felix/

@MFelixPBI

Sponsors:

alesonITC data thinking · NTT DATA · Verne TECH · FOUR9S · V-Valley enhancing your business · TD SYNNEX · waibbo · Namasdata

#DataSatMadrid

32

Q&A

#DataSatMadrid

33