A Microsoft Fabric Overview

Multiple Perspectives and Capabilities

Ross McNeely

v0.3 Mar 2024



Summary: A collection of documents on Microsoft Fabric relating to Business Analysis, Data Management, Project Management, and Digital Transformation. Business Analysis and Fabric alignment enables change within an organization. Data Management is the critical data capabilities that Fabric brings to the organization. Project Management brings a key skillset that coordinates the development of multiple Fabric capabilities. A focused Digital Transformation team replaces manual or ineffective processes where Fabric can create or improve business processes and experiences.

I will continue to publish content as it is created and will routinely package content together.

Table of Contents

Change Log 4
Microsoft Fabric Perspectives 5
Overview 5
Building Fabric Teams 6
Organizing by Roles 6
Organizing by Project / Product / Process 7
Organization Example 8
Digital Transformations with Fabric 9
Founder 9
Team & Team of Teams 10
Business Need 11
Capabilities 11
Approaches 11
Capability Analysis – Example 13
Measurable Future State 15
Microsoft Fabric Foundations 23
OneLake Capabilities 24
Fabric Mirroring Capabilities 26
Apache Spark Capabilities 27
Delta Lake Capabilities 28
Notebook UI Overview 29
Data Factory: Pipelines, Dataflows, & Spark 30
Data Factory: Pipeline Capabilities 31
Data Warehouse Capabilities 32
Data Lakehouse Capabilities 34
Real-time Analytic Capabilities 35
Eventstreams in Microsoft Fabric 39
Data Science Capabilities 41
Semantic Link in Data Science 44
Data Wrangler Capabilities 45

MLFlow in Microsoft Fabric 46
Data Activator Capabilities 48
Administration Capabilities 49
Security and Monitoring Capabilities 51
The Fabric Solution Space 52
Comparing Azure SQL DB and Fabric 53
Comparing Azure Synapse and Fabric 54
Comparing Azure Databricks and Fabric 55
Comparing Azure Snowflake and Fabric 56
Logical Data Flow 57
Physical Data Flow 57
Workload Balancing 58
Workloads 58
References 58
Best-of-Suite and Best-of-Breed 59
Evaluating Microsoft Fabric 60
Acceptance and Evaluation Criteria 60
OneLake Balanced Scorecard 61
Data Engineering Balanced Scorecard 62
Data Warehouse Balanced Scorecard 63
Miscellaneous 64
Microsoft Fabric Alignment Example 64
Microsoft Fabric Roadmap Template 65

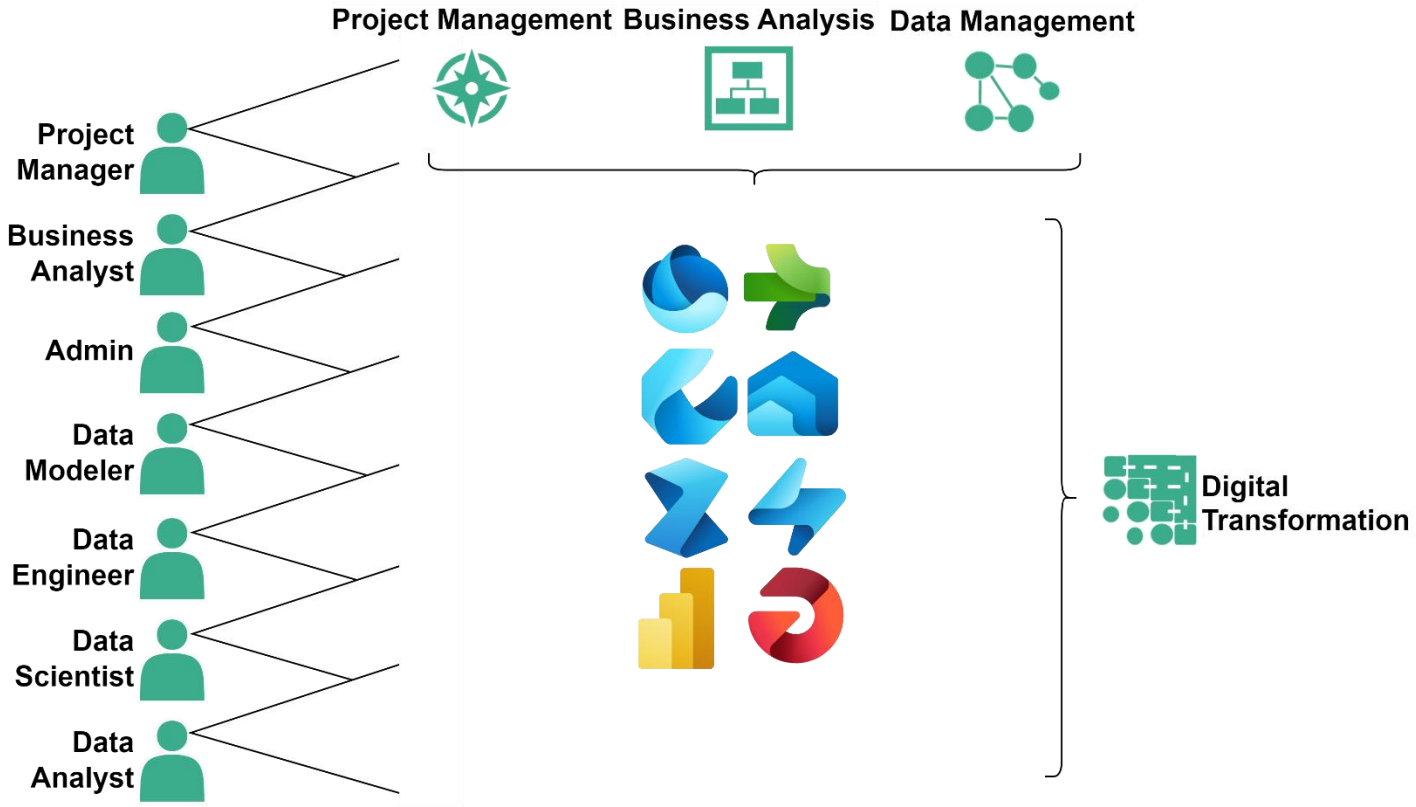
Change Log

2023-12-06	Ross McNeely	Added Data Activator Capability Page
2023-12-07	Ross McNeely	Added Data Wrangler Capability Page
2023-12-10	Ross McNeely	Added MLFlow in Fabric Page
2023-12-13	Ross McNeely	Added Kusto (KQL) examples page
2023-12-13	Ross McNeely	Added Warehouse monitoring page
2023-12-15	Ross McNeely	Updated Delta Lake Capabilities with Spark R
2023-12-15	Ross McNeely	Added Notebook UI Overview
2023-12-18	Ross McNeely	Updated Administration Capabilities
2023-12-18	Ross McNeely	Added Administrator Roles page
2023-12-19	Ross McNeely	Added Fabric Mirroring Capabilities
2024-03-10	Ross McNeely	Added Workload Balancing page
2024-03-11	Ross McNeely	Added Semantic Link in Data Science page

Microsoft Fabric Perspectives

Overview

The purpose of the Microsoft Fabric Perspective diagram is to highlight the persona based technical capabilities, and how they relate to industry standards in Project Management, Business Analysis, Data Management, and Digital Transformation. This document will present you with multiple perspectives of Microsoft Fabric. The key is to acknowledge that different groups will have distinct experiences, needs, and opinions that shape their understanding of Microsoft Fabric. By embracing multiple perspectives, we can drive a deep understanding of the Microsoft Fabric environment and quickly produce better solutions for complex business problems. By recognizing there is not a single "right way" to approach a situation you can provide multiple options each supported by various trade-offs and best practices.

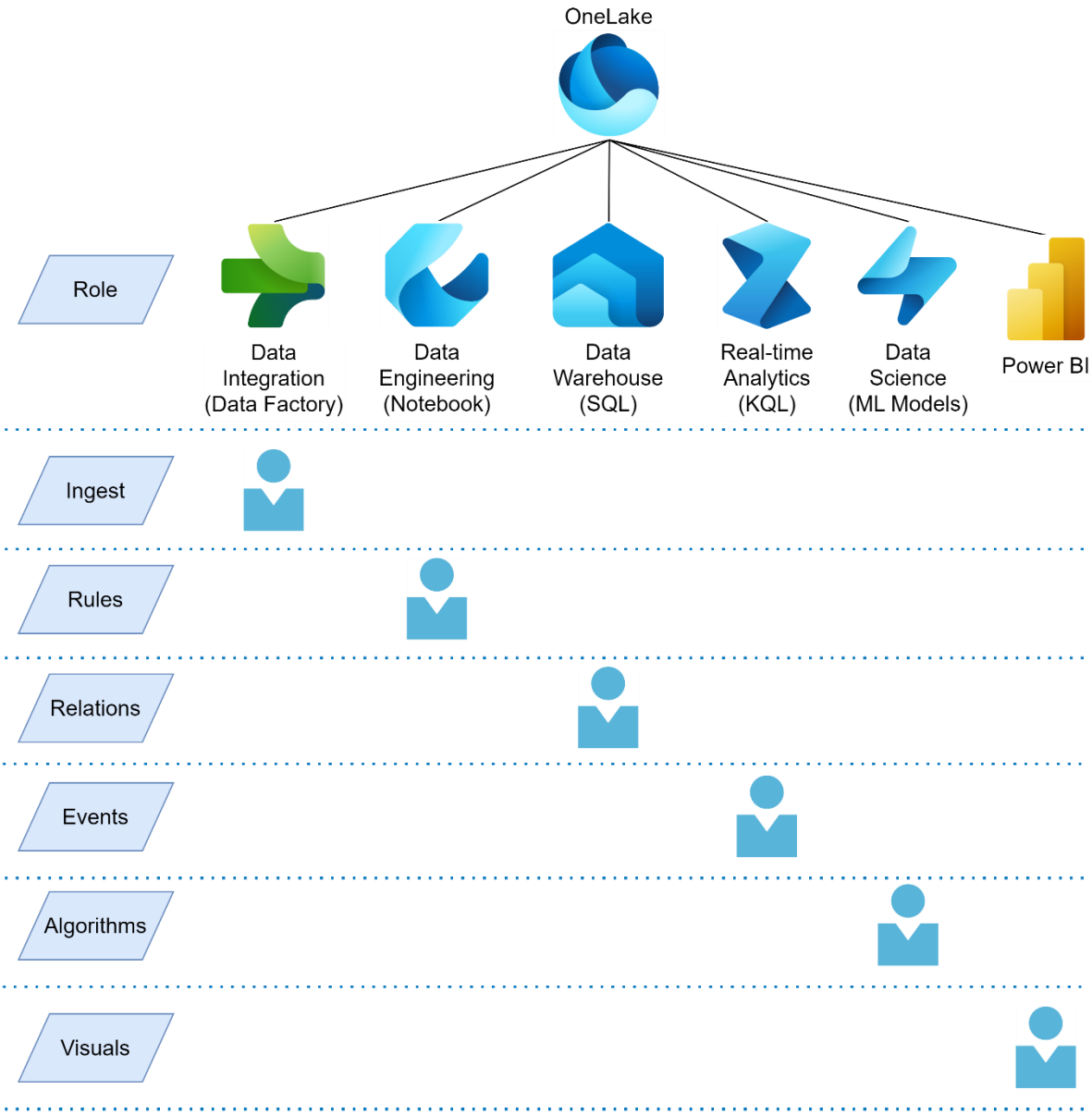


Project Management: Involves the planning, execution, and monitoring to ensure your data project stays on course. **Business Analysis:** Involves documenting the current state and future state of the business process. Including a gap analysis with steps on how to achieve the future state. **Enterprise Data Management:** Encompasses all the data domains necessary to execute a data strategy capable of supporting the business strategy of an organization. Data Governance, Data Storage, Data Integration, Data Architecture and Modeling, Data Security, Data Warehouse, Business Intelligence, Metadata, Data Quality, and Data Lifecycle Management. **Digital Transformation:** Involves the process of using digital platforms to change how an organization fundamentally operates and engages with customers.

Building Fabric Teams

Organizing by Roles

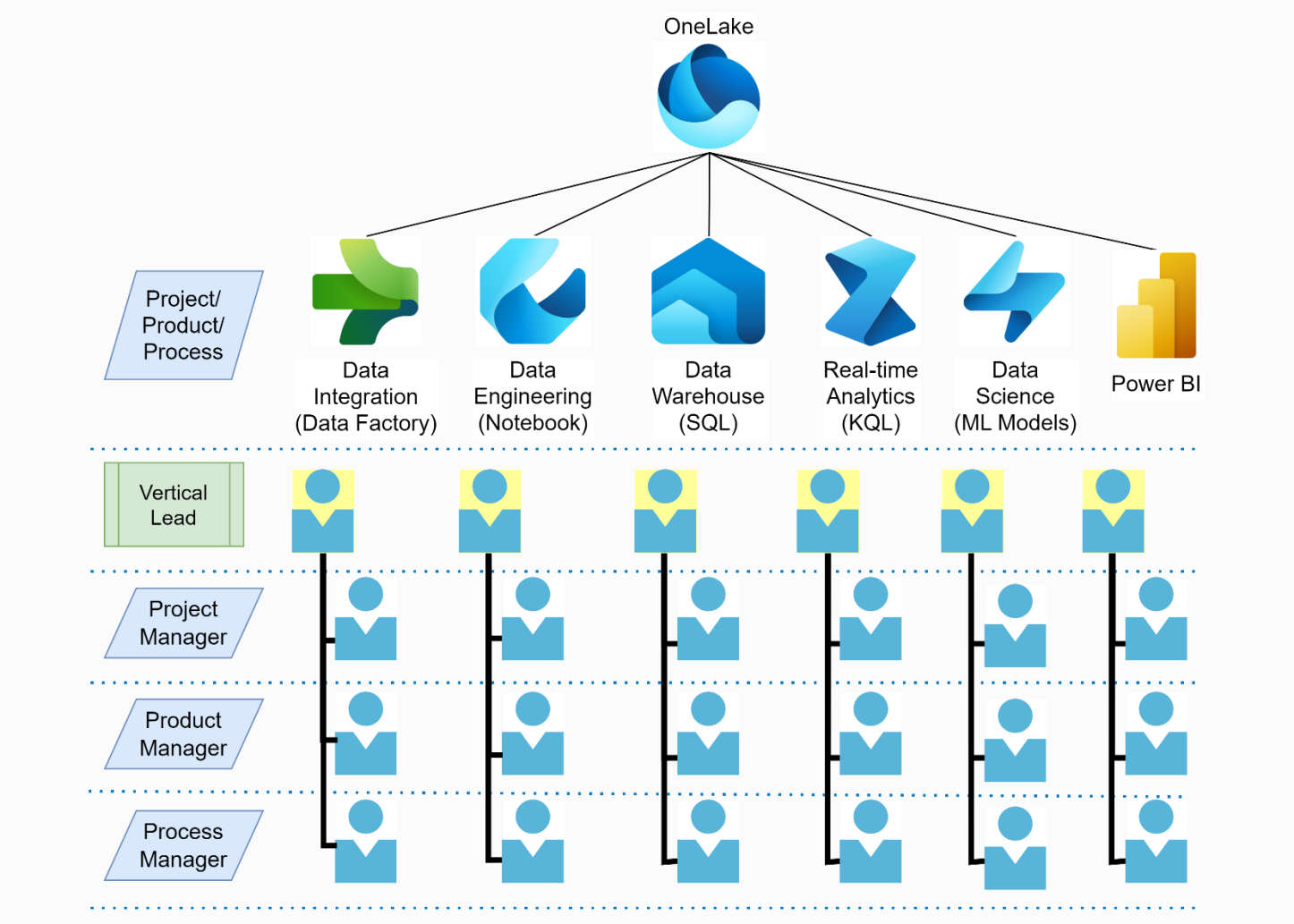
It is common to see a Matrix Org Chart in IT. It can help align a wide variety of technical skills with projects, products, or processes as needed. The challenge with a matrix model has always been the possible conflicting demands between Team Managers, and Project Managers (or Product/Process Manager). Leveraging the DevOps with Sprints helps to focus the priorities. It can still be a challenge when team members work on longer projects and standards start to drift toward the product/project/process. If your team is large, then be sure to use technical leads to help maintain best practices.



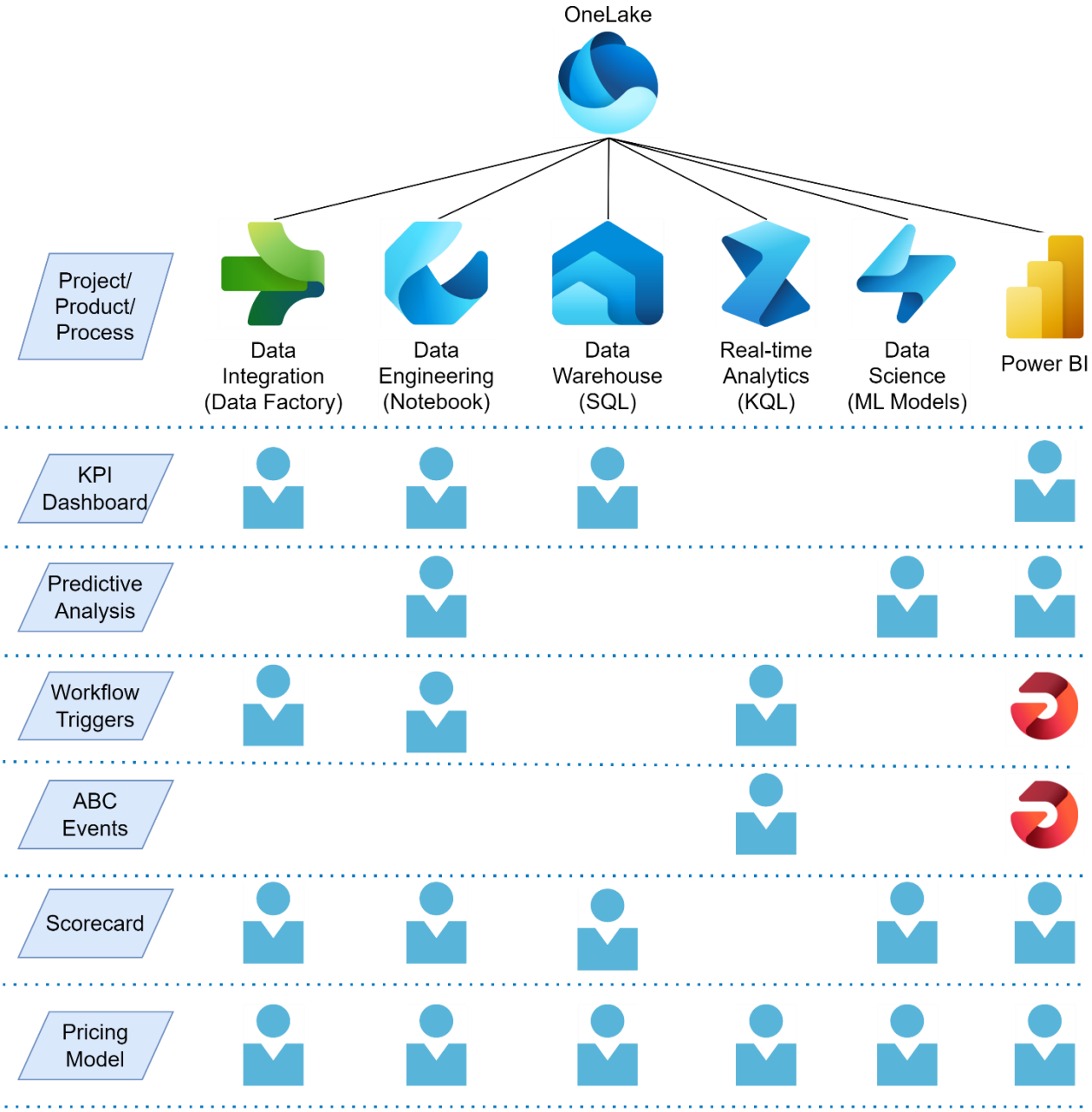
Organizing by Project / Product / Process

There is an overlap with existing Microsoft Azure products and Microsoft Fabric. Most of the Azure products will be needed now and into the future, but other products may be replaced by Fabric capabilities. Your organization will need to assess how to grow and support Fabric.

- Are there cultural changes needed to achieve better alignment with Fabric delivery?
- Do the stakeholders understand the current value-add of their data environment, and how to increase it with Fabric?



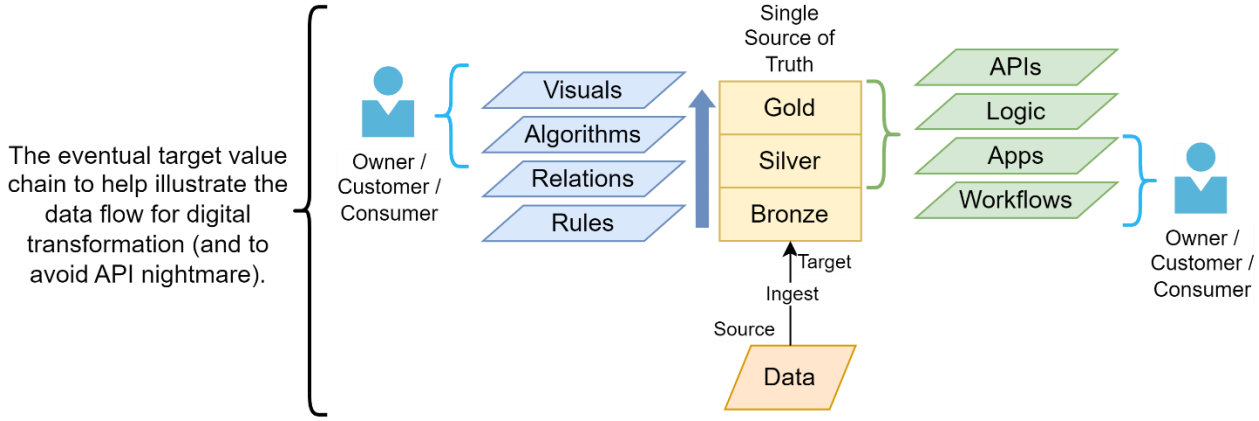
Organization Example



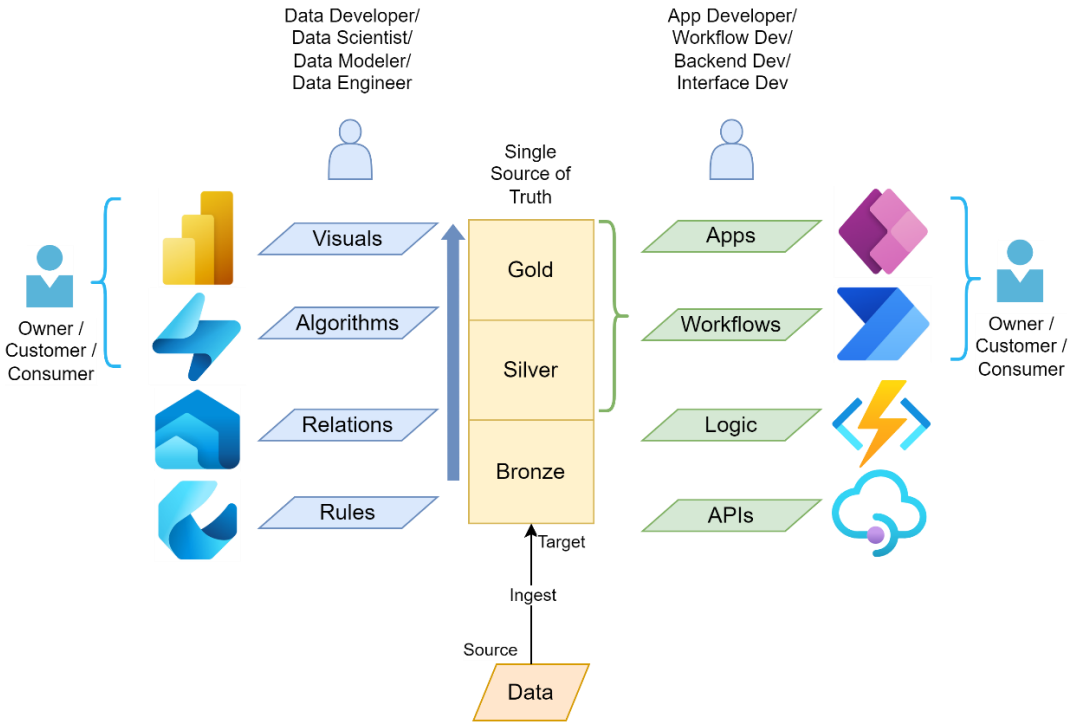
Digital Transformations with Fabric

Founder

Founder: At this early stage it can be 1 to 3 people working to create a hyper-focused digital solution targeting either the owner, customer, or consumer, but ideally you would target all three. The solution can be profit generating or optimization to reduce expenditure in a cost center.



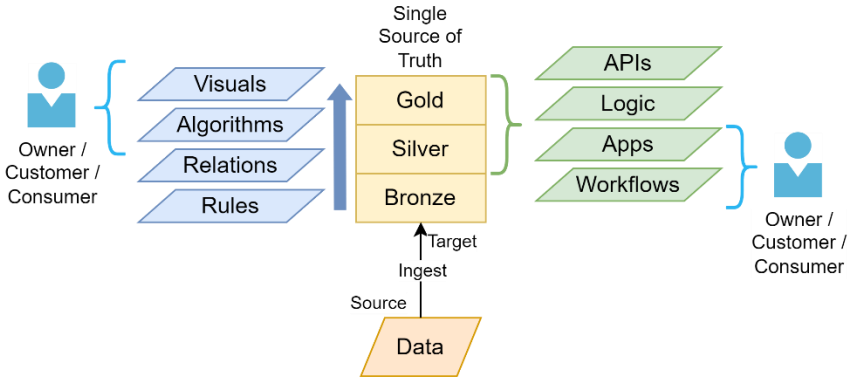
In the early (Founder) stages of a product/project/process there could be 1 to 3 people working to create the solution. They are usually full stack developers able to perform multiple duties. This works well with a narrow scope. If successful and growth is need there will be a point of "crisis" where the development staff will have to grow.

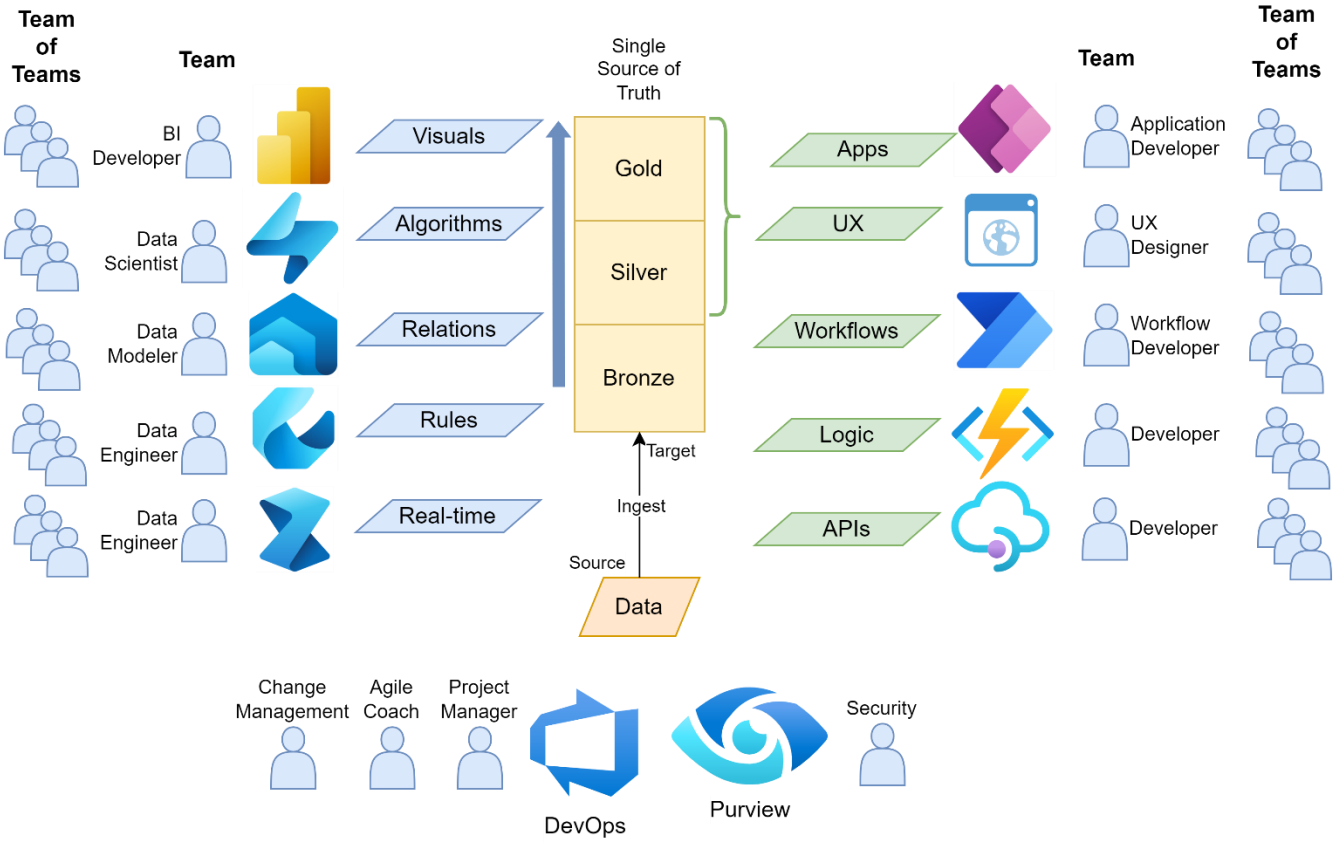


References: Josey A. (2020).The Digital Practitioner Foundation Study Guide. Van Haren Publishing.

Team & Team of Teams

At this stage the solution has grown to need multiple niche resources. The digital solution is still focused on meeting the needs of the owner, customer, or consumer (ideally all 3), but the repeatable nature for growth enables specialization. The additional resources will require a focus on project management to keep a tight alignment with the business need and technical capability. The solution is still profit generating or an optimization to reduce expenditure in a cost center.



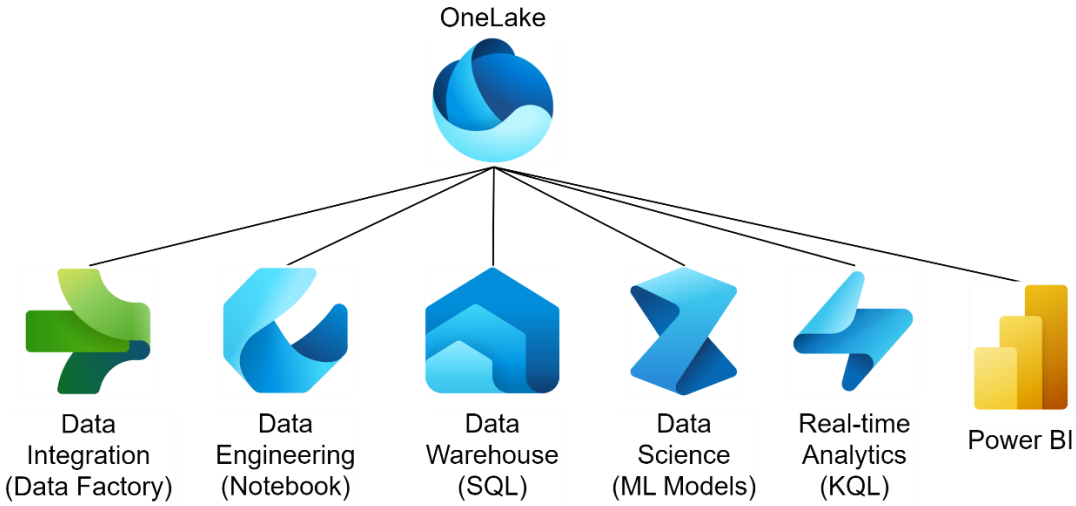


References: Josey A. (2020).The Digital Practitioner Foundation Study Guide. Van Haren Publishing.

Business Need

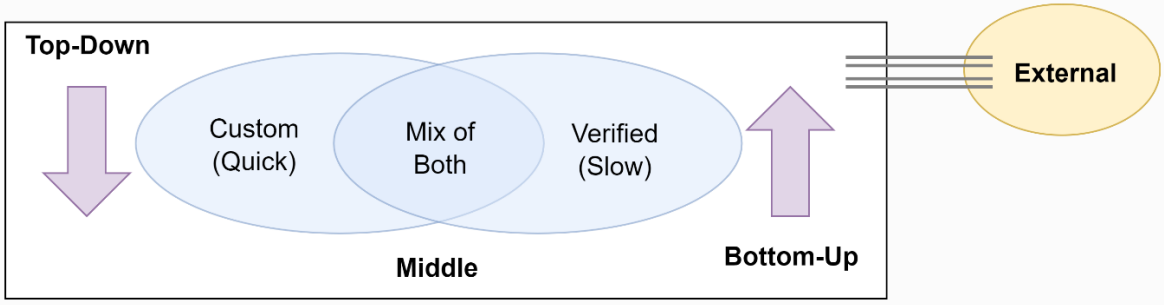
Capabilities

Microsoft Fabric is a suite of technical capabilities able to support anything from a single data project to an enterprise data solution. These technical capabilities are used to meet the needs of the business. It is important to understand how Fabric capabilities can align with the business needs. Aligning the business with technical capabilities usually occurs through Top-Down, Bottom-up, Middle Execution, and External Influence. Below is a diagram showing the various Fabric capabilities for Data Integration, Data Engineering, Data Warehousing, Data Science, Real-time Analytics, and Data Analysis.

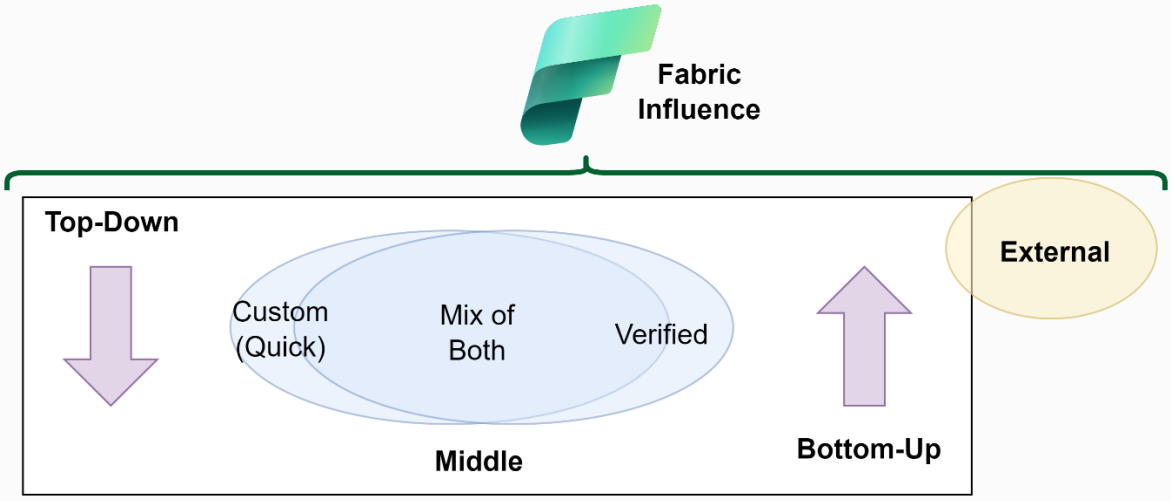


Approaches

- **Top-Down:**
  - Description: There is a need from multiple departments that are enabled to share a single source data they own and utilize without the need of IT intervention.
  - Example: Strategic need to share a single source of extracted data files and tables across departments that can reduce duplication and increase trust in data.
- **Bottom-Up:**
  - Description: IT deploys a new tool that can replace or complement the current data management processes to improve the overall capabilities for business.
  - Example: Departments need to see real-time application telemetry that has potential to correlate to negative customer feedback in order to create an action plan.
- **Middle Execution:**
  - Description: Middle management needs to produce quick insights from a team with a variety of skillsets and will leverage.
  - Example: A team manager needs to provide their team with a data platform capable of supporting the team's skills in Python, SQL, Data Modeling, and Analysis.
- **External Drivers**: Share Fabric Workspaces with guest users, or limit their access to reports, dashboards, datasets, and apps.



The above diagram depicts how a mixture of technical adoption paths often occur within a single organization. A Top-down approach is often driven by the business to quickly implement new capabilities. A bottom-up approach is often used by the IT department to deliver business capabilities with a high degree of quality, and availability. Often you will see a middle approach developed where business and IT departments manage a hybrid solution where stable or certified IT data sources are used with custom departmental data. External partners, vendors, and customers can also influence your organization's data capabilities.



Microsoft Fabric provides a solution at the IT and departmental levels. Using a single platform like Fabric shortens the data value chain and enables both IT and departments to leverage the same capabilities to produce solutions for the business.

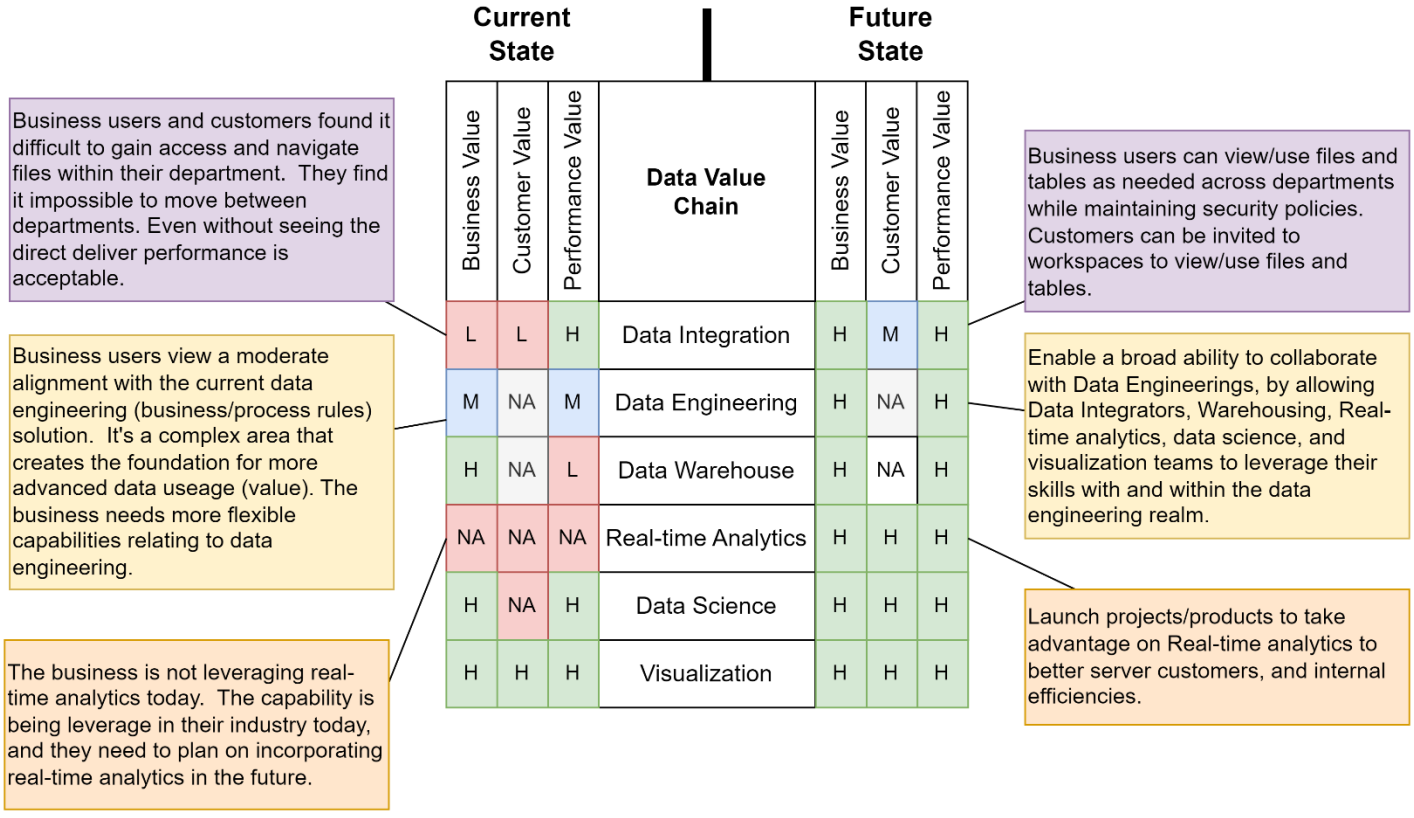
Examples: A mix of IT and business departmental capabilities:

- Able to support data provenance.
- Ability to support multiple skillsets.
- Capable of scale up or down to meet development and consumer demands.
- Suitable solution for the specific data context
- Supports security and reliability features.

Capability Analysis – Example

Business Analysis professionals use the Capability Analysis tool to visually show alignment of the current state and future state. In the example below I used a generic current state, but the future state is intended to be Microsoft Fabric focused. Creating multiple Capability Analysis diagrams may be needed if multiple platforms are being considered for the future state.

There are multiple evaluations occurring for Business, Customer, and Performance Value along the central vertical that represents the Data Value Chain. In my opinion Microsoft Fabric works to shorten the data value chain allowing your organization to benefit from data assets more efficiently. The data value chain provides key data management principles in Data Integration, Data Engineering, Data Warehousing, Real-time Analytics, Data Science and Data Visualization. Analyzing new tools and platforms by the Business, Customer, and Performance provides a deeper impact analysis than just performance along. Multiple points-of-view will help add insights from the business and operations side and highlighting any gaps in any future platform.



References: International Institute of Business Analysis. (2015). A guide to the business analysis body of knowledge (BABOK guide) (3rd ed.). Author.

The example above will start aligning the business needs with MS Fabric. The goal is to identify technical alignment between current state capabilities and future state that will support data management principles and deliver key business capabilities. This is not solely an evaluation of the technical capabilities. The purpose is to help generate a shared understanding of the capability expectations based on the business, customer, and

performance points-of-view. Creating a Capability Analysis chart will require cross-departmental collaboration to achieve a cohesive assessment. An additional benefit of a joint analysis across departments is assessing the dependencies and criticality of any identified risk to your organization's core capabilities.

Here are some questions from Microsoft's ebook titled "Unlocking Transformative Data Value with Microsoft Fabric" that will help drive discovery during your cross-departmental collaborations:

How can we enable innovation with data and analytics for your department?

What innovation can we enable?

How can I make your line of business more efficient?

What immediate initiatives can we collaborate on and how will we measure the impact?

What business value will we deliver over time Cost reduction? Revenue generation?

How can I better enable your line of business to use data in decision-making?

What security and governance concerns should we be aware of?

Is there third-party of IoT data that could enrich your insights?

While you are creating the Capability Analysis chart remember your future state should generally target either a lakehouse, warehouse, machine learning, or real-time analytical solutions. Be sure to create a backlog of other Digital Transformation ideas you hear to help build new use cases for integration.

Measurable Future State

Create a list of business objectives or goals. The objectives need to be specific and measurable to provide a guide for achieving the future state. They do not need to be technical specifications at this point but should be decomposable to provide tasks. Based on the business objectives there will be occasions when risk mitigation can be the primary goal.

Objectives		Risk Mitigation
Specific Goal: Share data across the org securely and act on automation chances. Ensure data asset integrity Measurement: Measure data variety/volume/velocity and shared data assets. Track and confirm security (data) silos.	OneLake	*Data Quality: Reasonability & Completeness Enterprise Ability: Data share policies and data lifecycle mgmt. Technical Skills: BA | DA focus on OneLake data structures Security: Workspace Roles & Item Security
Specific Goal: Increase tracking of data assets that go through an intake process. Measurement: Measure the number of data assets with known and unknown origination processes	Data Ingestion	*Data Quality: Accuracy & Timeliness & Validity Enterprise Ability: Intake processes vs "Take anything process" Technical Skills: Extract & Load operations e.g. APIs, Pipelines, etc Security: Encryption, Landing Zone, Auth, OneLake destination
Specific Goal: Standardize data integration techniques across multiple personas. Deliverable: Best practices, procedures, code review	Data Engineering	*Data Quality: Uniqueness & Validity & Accuracy & Completeness Enterprise Ability: Move all transformation logic to tools/repo Technical Skills: Python & SQL & etc Security: Source | Compute | Destination
Specific Goal: Enable business analysis and decision-making through data domain models. Measurement: Track how many cross-domain models are built/used with additional counters for entities per domain.	Data Warehouse	*Data Quality: Integrity & Uniqueness & Timeliness Enterprise Ability: SMEs wanting to document their data domains Technical Skills: Data modeling skills Security: Domain/Role/Row level
Specific Goal: Drive critical KPI to leverage instant insights from the data streams. Measurement: Track the number of KPIs supported by data streams.	Real-time Analytics	*Data Quality: Timeliness & Accuracy & Reasonability Enterprise Ability: Identify where real-time can be impactful Technical Skills: KQL, streaming, and events Security: Workspace Roles & Item Security
Specific Goal: Generate actionable insights from large data sets. Measurement: Track the number of actionable batch scoring and predicted values in usage.	Data Science	*Data Quality: Reasonability & Accuracy Enterprise Ability: Identify where ML models can be impactful Technical Skills: Feature engineering, model development Security: Workspace Roles & Item Security
Specific Goal: Create dashboards/scorecards/reports based on shared and trusted data sets to drive business analytics Measurement: Track dashboard and business domain alignment along with dashboard and usage alignment.	Power BI	*Data Quality: Culmination Enterprise Ability: Leverage centralized analytical platform Technical Skills: Query, Model, Visualizes Security: Workspace/Visuals/Role/Row level

A chart of measurements based on future state should include the following items: Specific Goals, Measurements, Data Quality Checks, Enterprise Ability, Technical Skills, Security.

Specific goals help to provide a clear direction and internal motivations that will increase the chances of achieving outcomes and measuring progress effectively.

“Without data, your just another person with an opinion.” -W. Edwards Deming. Creating measures to track how your organization implements Fabric capabilities will allow you to identify areas for improvement.

Data quality checks will help reduce the negative impact poor data quality can have in analytics.

The ability of an enterprise to successfully implement new capabilities has a profound impact on its agility and competitiveness in a rapidly evolving business landscape.

Technical skill development or adaptability is critical to build and support the technical capabilities the business needs to execute the organization’s strategy.

Security, Security, Security

OneLake

- Objectives
  - Specific goal: Share data across the organization securely and have the ability act on analytical and automation opportunities that present themselves. Ensure data asset integrity is enforced.
  - Measurement: Measure data variety/volume/velocity and shared data assets. Track and confirm security (data) silos. At the OneLake level we are basically capturing raw data, and do not want to impede the collection of data. We can track from the OneLake environment to understand how well the organization is leveraging data.
- Risk Mitigation
  - Data quality checks
    - "Reasonability asks whether a data pattern meets expectations. For example, whether a distribution of sales across a geographic area makes sense based on what is known about the customers in that area." (DAMA International, 2017)
    - "Consistency can refer to ensuring that data values are consistently represented within a data set and between data sets, and consistently associated across data sets. It can also refer to the size and composition of data sets between systems or across time…" (DAMA International, 2017)
  - Enterprise Ability: Implement data share policies and data lifecycle management. Training on how to use Fabric Workspaces for knowledge workers.
  - Technical Skills: Leverage and Business Analysis and Data Analysis to focus on OneLake data structures created from workspaces.
  - Security: Workspace Roles & Item Security

Objectives

Specific Goal: Share data across the org securely and act on automation chances. Ensure data asset integrity
Measurement: Measure data variety/volume/velocity and shared data assets. Track and confirm security (data) silos.

OneLake

Risk Mitigation

*Data Quality: Reasonability & Completeness
Enterprise Ability: Data share policies and data lifecycle mgmt.
Technical Skills: BA | DA focus on OneLake data structures
Security: Workspace Roles & Item Security

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge Second Edition. Technics.

Data Ingest

- Objectives
  - Specific goal: Increase the tracking of data assets that go through an intake process. Begin to leverage your metadata immediately as you start to apply compute over the data asset. Data elements used in Microsoft Fabric Pipelines or Dataflows gain immediate value simply because compute is being spent to further refine and enrich the data asset.
  - Measurement: You can measure how each data element is processed to help understand the percentage of data being used versus just stored.
- Risk Mitigation
  - Data Quality:
    - "Accuracy refers to the degree that data correctly represents 'real-life entities. Accuracy is difficult to measure, unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy rely on comparison to a data source that has been verified as accurate, such as a system of record or data from a reliable source." [DAMA International, 2017]
    - "…[T]imeliness refers to several characteristics of data. Measures of timeliness need to be understood in terms of expected volatility – how frequently data is likely to change and for what reason." [DAMA International, 2017]
    - "Validity refers to whether data values are consistent with a defined domain of values." [DAMA International, 2017]
  - Enterprise Ability: Intake processes vs "Take everything" process. Giving your organization the ability to see how data is refined and enriched will help them determine areas where they can better leverage data to become data-driven. The quicker Business and Data owners can Your organization will also be able to determine if they are storing data unnecessarily.
  - Technical Skills: Extract & load operations e.g. APIs, Pipelines, Dataflows, etc.
  - Security: Encryption, Landing Zone, Authentication, OneLake Destination

Specific Goal: Increase tracking of data assets that go through an intake process.
Measurement: Measure the number of data assets with known and unknown origination processes

Data Ingestion

*Data Quality: Accuracy & Timeliness & Validity
Enterprise Ability: Intake processes vs "Take anything process"
Technical Skills: Extract & Load operations e.g. APIs, Pipelines, etc
Security: Encryption, Landing Zone, Auth, OneLake destination

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge Second Edition. Technics.

Data Engineering

- Objectives
  - Specific goal: Standardize data integration techniques across multiple personas.
  - Deliverable: Best practices, procedures, code reviews
- Risk Mitigation
  - Data Quality:
    - "Uniqueness states that no entity exists more than once within the data set" [DAMA International, 2017]
    - "Validity refers to whether data values are consistent with a defined domain of values." [DAMA International, 2017]
    - "Accuracy refers to the degree that data correctly represents 'real-life entities. Accuracy is difficult to measure, unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy rely on comparison to a data source that has been verified as accurate, such as a system of record or data from a reliable source." [DAMA International, 2017]
    - "Completeness refers to whether all required data is present. Completeness can be measured at the data set, records, or column level." [DAMA International, 2017]
  - Enterprise Ability: Move business logic into wiki documents and corresponding code repositories. This will help to ensure that the development process remains transparent and enables future collaboration between the business and developer.
  - Technical Skills: Python (PySpark) and SQL (SparkSQL)
  - Security: Data sources, destination

Specific Goal: Standardize data integration techniques across multiple personas.
Deliverable: Best practices, procedures, code review

Data Engineering

*Data Quality: Uniqueness & Validity & Accuracy & Completeness
Enterprise Ability: Move all transformation logic to tools/repo
Technical Skills: Python & SQL & etc
Security: Source | Compute | Destination

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge Second Edition. Technics.

Data Warehousing

- Objectives
  - Specific goal: Represent the business matrix of processes and functions in an analytical model or models. Enable business analysis and decision-making through data domain models.
  - Measure: Track how many cross-domain models are built/used with additional counters for entities per domain.
- Risk Mitigation
  - Data Quality:
    - "Data Integrity (or Coherence) includes indeas associated with completeness, accuracy, and consistency In data, integrity usually refers to either referential integrity (consistency between data ojbects via a ference key contained in both objects) or internal consistency within a setset…" [DAMA International, 2017]
    - "Uniqueness states that no entity exists more than once within the data set" [DAMA International, 2017]
    - "…[T]imeliness refers to several characteristics of data. Measures of timeliness need to be understood in terms of expected volatility – how frequently data is likely to change and for what reason." [DAMA International, 2017]
  - Enterprise Ability: Department heads, and SMEs wanting to create a representation of their business model for analytical purposes and data-driven decisions.
  - Technical Skills: Data modeling skills
  - Security: Domain, Role, Table, Column, Row level security

Specific Goal: Enable business analysis and decision-making through data domain models.
Measurement: Track how many cross-domain models are built/used with additional counters for entities per domain.

Data Warehouse

*Data Quality: Integrity & Uniqueness & Timeliness
Enterprise Ability: SMEs wanting to document their data domains
Technical Skills: Data modeling skills
Security: Domain/Role/Row level

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge Second Edition. Technics.

Real-time Analytics

- Objectives
  - Specific goal: Drive critical KPIs to leverage instant insights from your organization's data streams. Proprietary data stream are critical for driving competitive advantage and should be given priority.
  - Measurement: Track the number of KPIs supported by data streams.
- Risk Mitigation
  - Data Quality:
    - "…[T]imeliness refers to several characteristics of data. Measures of timeliness need to be understood in terms of expected volatility – how frequently data is likely to change and for what reason." [DAMA International, 2017]
    - "Accuracy refers to the degree that data correctly represents 'real-life entities. Accuracy is difficult to measure, unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy rely on comparison to a data source that has been verified as accurate, such as a system of record or data from a reliable source." [DAMA International, 2017]
    - "Reasonability asks whether a data pattern meets expectations. For example, whether a distribution of sales across a geographic area makes sense based on what is known about the customers in that area." [DAMA International, 2017)
  - Enterprise Ability: Identify what current KPIs, trend analysis, pattern monitoring, and anomaly detection can be leveraged with real-time data streams.
  - Technical Skills: KQL, streaming and events
  - Security: Workspace Roles and item security

Specific Goal: Drive critical KPI to leverage instant insights from the data streams.
Measurement: Track the number of KPIs supported by data streams.

Real-time Analytics

*Data Quality: Timeliness & Accuracy & Reasonability
Enterprise Ability: Identify where real-time can be impactful
Technical Skills: KQL, streaming, and events
Security: Workspace Roles & Item Security

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge Second Edition. Technics.

Data Science

- Objectives
  - Specific goal: Generate actionable insights from large data sets. The difficult part is identifying optimal use cases where models will produce reliable results. Ideation sessions with business owners and subject-matter experts help to frame the use cases. This also allows Data Science teams to prioritize efforts on models that will have a larger impact on the business based on the use case identified.
  - Measurement: Track the number of actionable batch scoring and predicted values in usage. This is critical for evaluating model performance, understanding business impact and continued alignment.
- Risk Mitigation
  - Data Quality:
    - "Reasonability asks whether a data pattern meets expectations. For example, whether a distribution of sales across a geographic area makes sense based on what is known about the customers in that area." [DAMA International, 2017)
    - "Accuracy refers to the degree that data correctly represents 'real-life entities. Accuracy is difficult to measure unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy rely on comparison to a data source that has been verified as accurate, such as a system of record or data from a reliable source." [DAMA International, 2017]
  - Enterprise Ability: Identify where ML models can be impactful and ideation session template
  - Technical Skills: Data prep and feature engineering with model development
  - Security: Workspace roles and item security

Specific Goal: Generate actionable insights from large data sets.
Measurement: Track the number of actionable batch scoring and predicted values in usage.

Data Science

*Data Quality: Reasonability & Accuracy
Enterprise Ability: Identify where ML models can be impactful
Technical Skills: Feature engineering, model development
Security: Workspace Roles & Item Security

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge Second Edition. Technics.

Data Visualization

- Objectives
  - Specific goal: Create dashboards/scorecards/reports based on shared and trusted data sets to drive business analytics
- Risk Mitigation
  - Data Quality
    - "Accuracy refers to the degree that data correctly represents 'real-life entities. Accuracy is difficult to measure unless an organization can reproduce data collection or manually confirm accuracy of records. Most measures of accuracy rely on comparison to a data source that has been verified as accurate, such as a system of record or data from a reliable source." [DAMA International, 2017]
  - Enterprise Ability: Track how many dashboards are actively used by the correct departments, teams, and users. Are you leveraging your analytical artifacts?
  - Technical Skills: Data visualization technical and best practices are critical to prevent dashboards and reports that misrepresent the data.
  - Security: Workspace roles and item security along with row and column level security.

Specific Goal: Create dashboards/scorecards/reports based on shared and trusted data sets to drive business analytics
Measurement: Track dashboard and business domain alignment along with dashboard and usage alignment.

Power BI

*Data Quality: Culmination
Enterprise Ability: Leverage centralized analytical platform
Technical Skills: Query, Model, Visualizes
Security: Workspace/Visuals/Role/Row level

DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge Second Edition. Technics.

Microsoft Fabric Foundations

OneLake serves as a comprehensive, integrated data lake solution for your entire organization. Similar to OneDrive, There is only 1 OneLake included in every Microsoft tenant, and it is purpose-built to serve as the central repository for all your data needs in Fabric.

Data Engineering is a key foundational capability allowing users to design and build systems to collect, store, process, and analyze data. **Lakehouses** are comprehensive data architectures designed to enable organizations to consolidate structured and unstructured data into a unified repository. They provide a versatile ecosystem with a wide range of tools and frameworks for data processing and analysis. Including SQL-based queries and PySpark. "**Spark job definitions** are set of instructions that define how to execute a job on a Spark cluster." This includes input, transformations, output, and configuration parameters. **Data Pipelines** are a series of graphical activities that can source, transform, and output data. They are a critical component of data engineering allowing engineers to create new data models and orchestrate large data flows.

Within Microsoft Fabric, the **Data Factory** offers scalable capabilities for data movement and transformations. This will allow you to handle everything from simple to complex ETL challenges. Its primary objective is to deliver a graphical and enterprise-grade data migration experience. There are multiple solutions depending on your ETL requirements. Data Factory **Pipelines** are typically preferred for large data movements leveraging the CopyTo activity and script orchestration. Data Factory **Dataflows** are typically preferred for graphical transformations.

Warehouse

Real-time Analytics

Data Science

Administration

Security and Monitoring

Workspaces

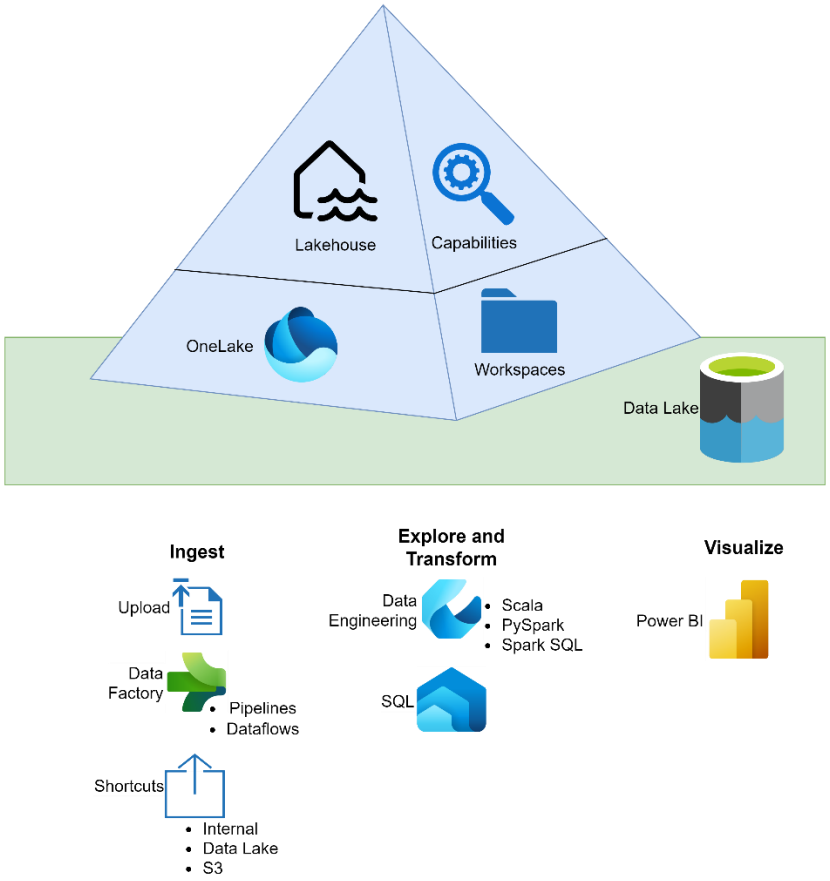
OneLake Capabilities

Capabilities

Lakehouse is designed for schema-on-read. Allowing you to load data without the overhead of modeling. When you read the data you can then defined the necessary schema. The Lakehouse supports ACID transaction with Delta Lake formatted tables for Atomicity, Consistency, Isolation, and Durability. The Lakehouse can be shared without "data copy" overhead. Allowing for a single location for data engineers, data analysts, and data scientists.

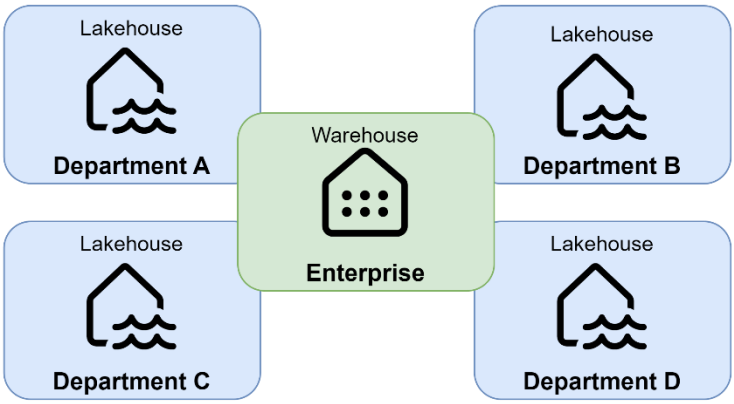
Solution

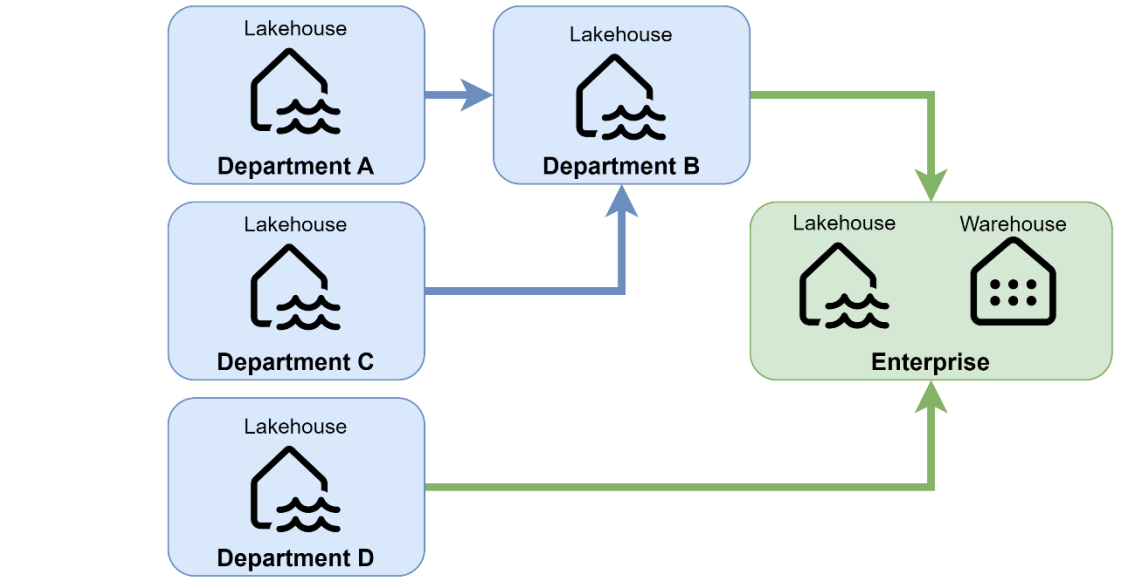
Enable your organization to analyze various data formats across multiple sources by leveraging Microsoft Fabric. Unstructured data from social media, log analytics, or multiple partners combined with structured transactional data. After creating a Lakehouse users can load data from various sources (local files, databases, APIs) Automation can then be applied through Pipelines or Dataflows. You can work with Lakehouse data directly or with a SQL Endpoint in a relational data model. Workspace containers enable organization, collaboration, and a layer of security.

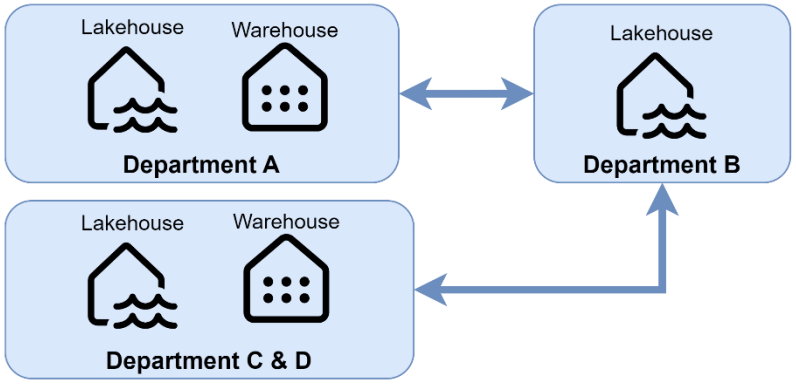


References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

OneLake Workspace Examples

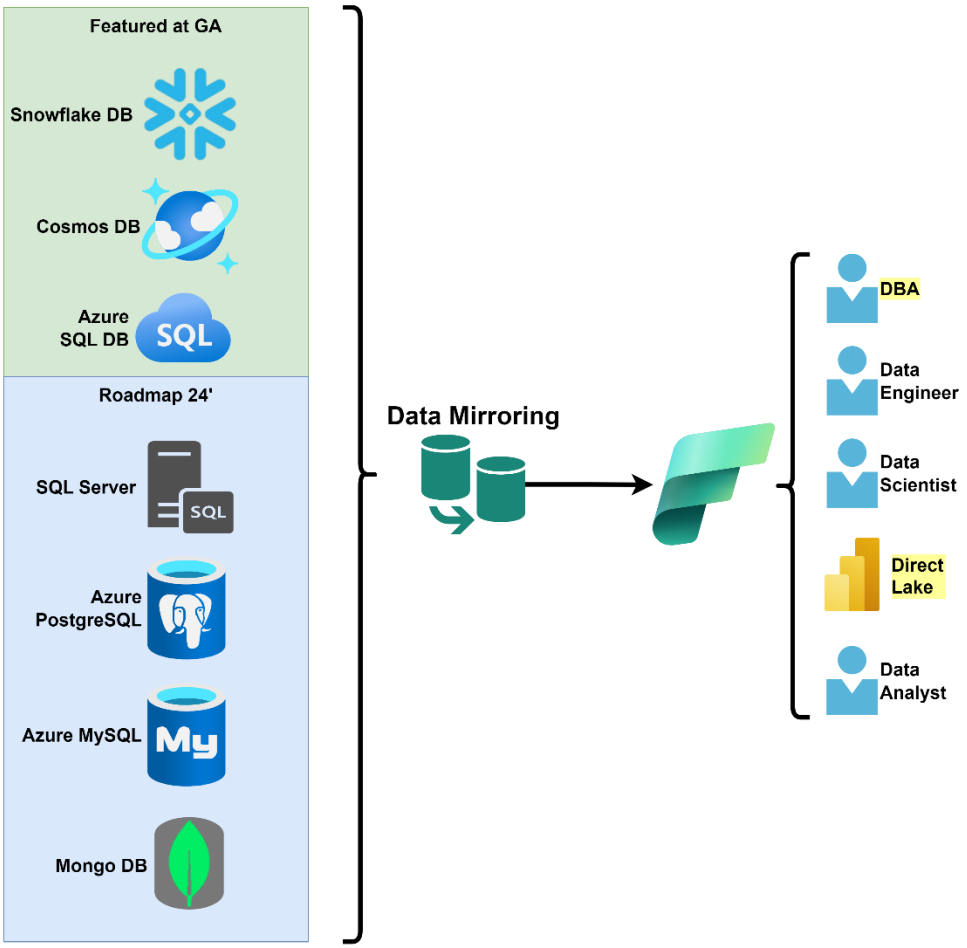






Fabric Mirroring Capabilities

Using Microsoft Fabric Mirroring is a modern way to easily ingest data and allow different workload personas access all within Fabric. This will allow developers or DBAs to access and manage any database centrally within Fabric. Removing complex ETL enables you to sync your data sources in near real-time using change data capture (CDC). Add in Shortcuts and new level of flexibility can be achieved with cross-joining mirrored database, warehouse, and lakehouses. Data Science workloads can use the mirrored databases with Notebooks. Data Analysts can build Power BI reports and leverage the DirectLake mode for fast performance without duplicating data.

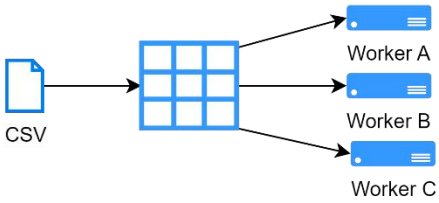


- Access and manage any database or data warehouse
- Real-time data replication
- Data warehousing experiences simplified
- Cross-joining mirrored database, warehouses, lakehouses
- Data Science experiences with Notebooks
- Power BI Direct Lake Mode

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-12-19, from https://blog.fabric.microsoft.com/en-us/blog/introducing-mirroring-in-microsoft-fabric/

Apache Spark Capabilities

Apache Spark distributes data processing across multiple nodes in a cluster. This is a divide and conquer strategy to handle large volumes of data.



You can use a notebook to combine text, images, and code written in multiple languages. This allows you to create an interactive notebook, or a script for automating data loads. If you want to schedule the script all you have to do is create a Spark Job to call the script.

Spark can run a wide range languages:

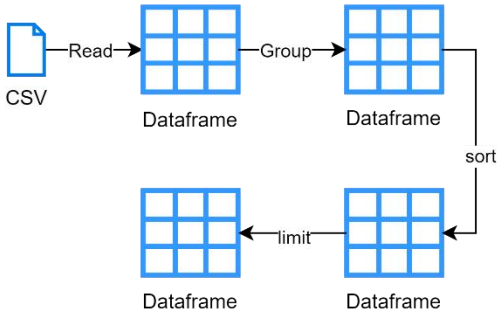
- **Scala**
- **Spark R**
- **Spark SQL**
- **PySpark**

Fabric Workspace Admin

Each workspace can be assigned a Spark cluster. The Spark cluster settings can be managed by an admin in the Data Engineering/Science persona section of the workspace settings.

The Spark dataframe is the most common used data structure and is supplied as part of the Spark SQL Library. You can allow Spark to infer the data type for each column, or provide an explicit schema for the data. You can then write the data to a parquet file to preserve the dataframe results. Parquet is not the only format, but is the recommended.

Example of Dataframe/UI Cell Development



Spark Configuration

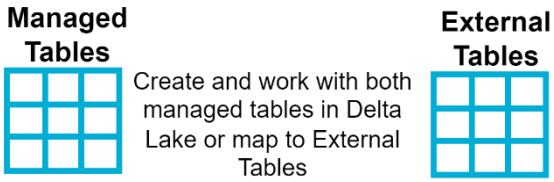
- **Node Family:** Type of virtual machines to be in Spark cluster. Typically, memory optimized nodes is optimal.
- **Runtime Version:** The version of Spark and it subcomponents will run under.
- **Spark Properties:** You can enable or override various spark specific settings.
- **Library Management:** You can include additional code libraries for different tasks needed. Most commonly used libraries are already included.
  - Feed library (.yml)
  - Custom library (.whl, .jar, .tar.gz)

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

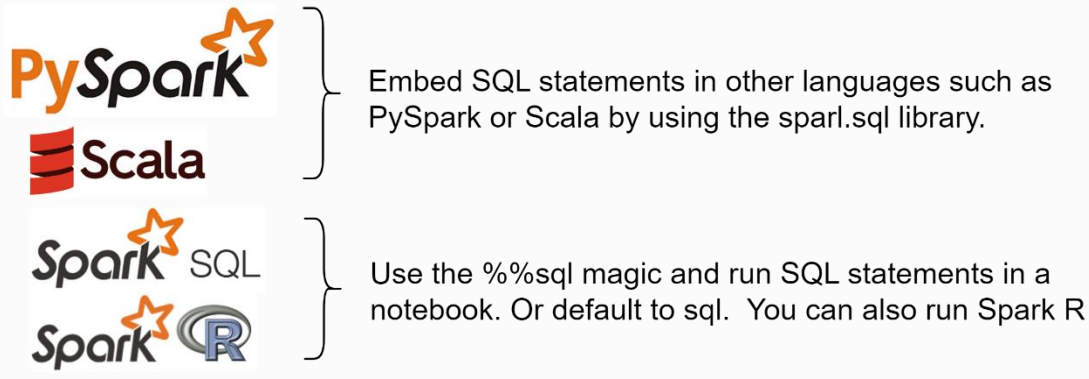
Delta Lake Capabilities

The Fabric Lakehouse is based on the Delta Lake table format. Delta Lake is an open-source layer for Spark to enable relational database capabilities. Delta Lake is able to support batch and streaming solutions.

- **Relational Tables (query & data mods):** Support for select, insert, update, and delete.
- **ACID Transactions:** Atomicity (transaction are a single unit of work), Consistency (transactions leave the database in a consistent state, Isolation (transactions can't interfere with each other), Durability (transactions are persisted)
- **Data Versioning & Time Travel:** All transactions are logged allowing for multiple versions of each table row and even the use of "time travel" to retrieve a previous version of a row.
- **Batch and Steaming Data Solutions:** Structured Spark Streaming API allows Delta Lake tables to be used as both sinks and sources for streaming.
- **Standard Formats & Interoperability:** Delta tables are stored in Parquet format which is commonly used.



1. **Schema + Data**: Save a dataframe as a delta table by creating both the table schema definition in the metastore and the data files in delta format.
2. **Schema Only**: Create a table definition which creates the table schema in the metastore without saving any data files.
3. **Data Only**: Or, save data in delta format without creating a table definition in the metastore, and later overlay a table definition by using the delta format.

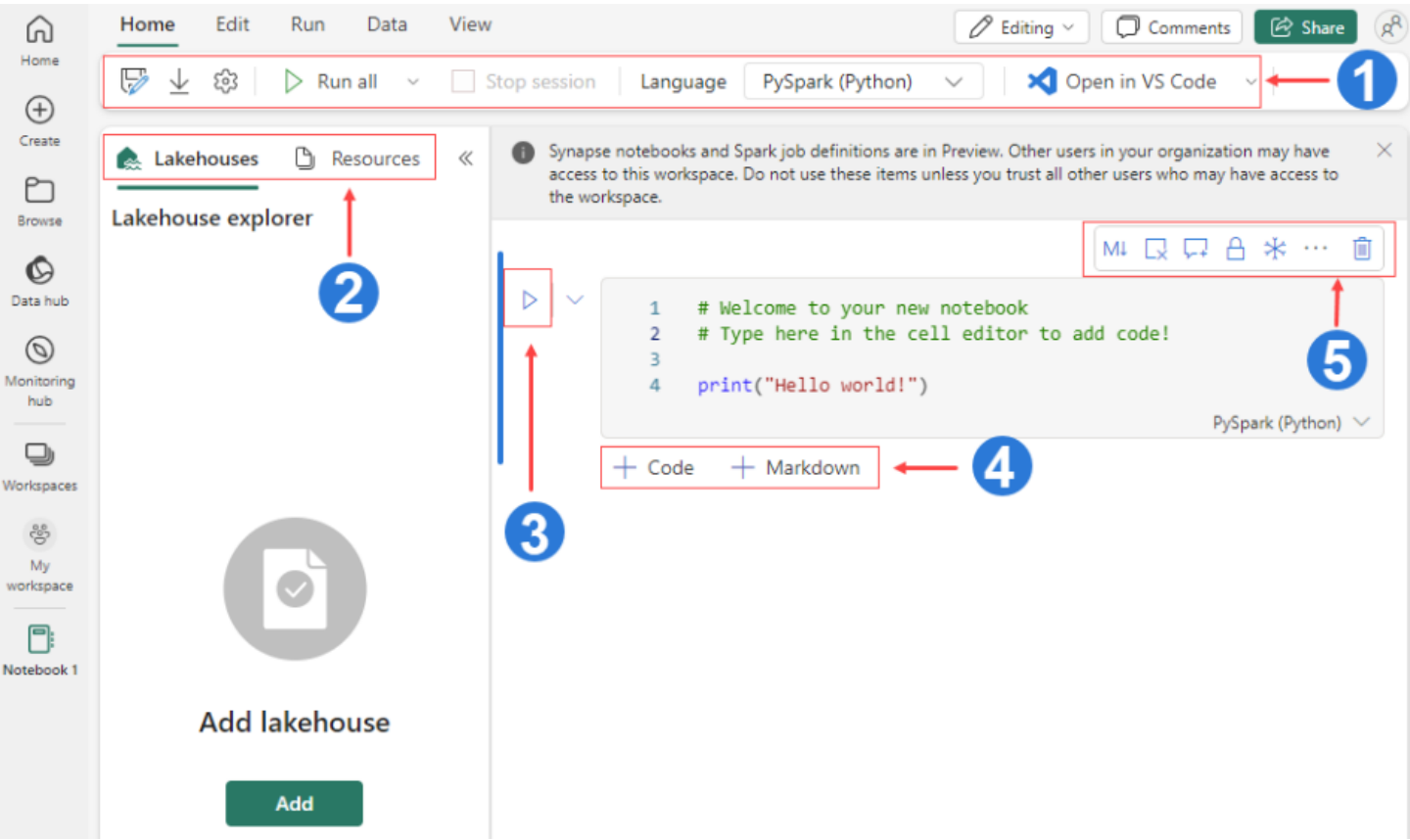


Spark Structured Streaming: Delta tables can be used as a source or a sink for Spark Structured Streaming. This allows you to stream real-time data from IoT devices and directly write to a delta table. Or, you can read from a delta table source. When using delta table streaming source you can only append.

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

Notebook UI Overview

[Explore notebooks - Training | Microsoft Learn]



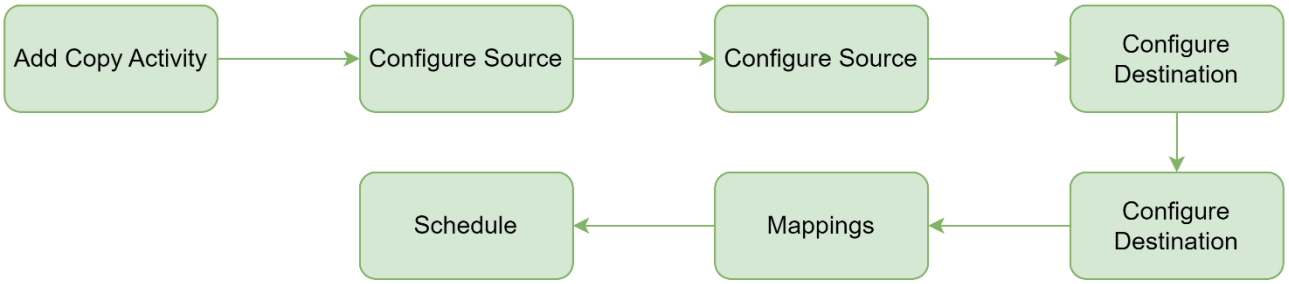
https://learn.microsoft.com/en-us/training/modules/explore-data-for-data-science-microsoft-fabric/2-explore-notebooks

1. Menu Bar: Save | Export | Manage Cells | End Session | Schedule Exe | Data Wrangler | Adjust Layout | etc…
2. Tabs
    a. Lakehouse Tab: Add New Lakehouse | Add Existing Lakehouse
    b. Resources Tab: Small files like code | datasets | images
3. Run Cell: Runs the current cell
    a. Ctrl+Enter to run current cell
    b. Alt+Enter to run current cell and add new cell
4. Add Cell: Code | Markdown
5. Manage Cells: Reposition | Convert to Markdown | Convert to Code | Clear Output | Add Comments | etc…

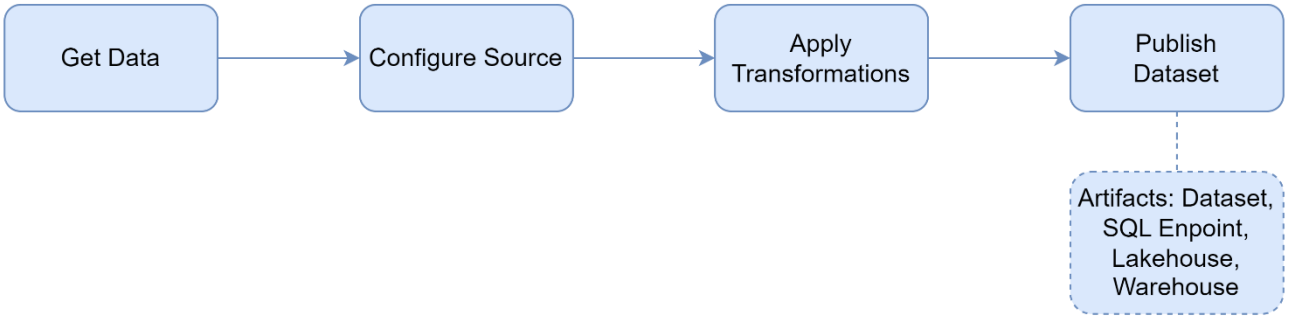
Data Factory: Pipelines, Dataflows, & Spark

There are multiple ways you can leverage Pipelines and Dataflows. Here are a few basic examples. The flexibility will allow your team to ingest, transform, and integrate any data set.

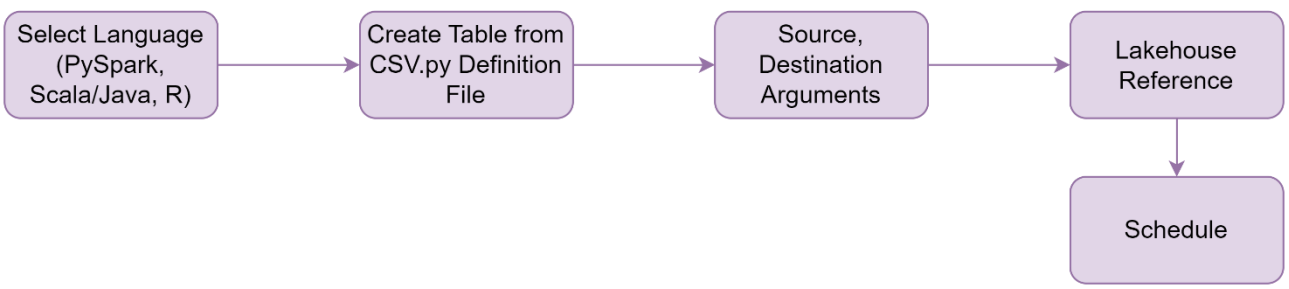
Example 1: Pipeline



Example 2: Dataflow



Example 3: Spark Job Definition



References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

Data Factory: Pipeline Capabilities

Data pipelines define a sequence of activities to control the flow of data. You can create ingest patterns like ELT (Extract, Load Transform) or integration patterns like ETL (Extract, Transform, Load) Today, ELT is the preferred method because it allows the maximum throughput for moving data, and flexibility to apply business rules before user consume the data.



Core Pipeline Concepts

- **Activities:** A pipeline will contain activity sequences which are linked executable tasks.
  - **Data Transformations:** Activities for data transfer operations like "Copy Data" to extract data from a source and load it to a destination. Call Notebooks or Stored Procedures to perform data transformations.
  - **Control Flow:** Activities that enable you to create loops, conditions, or parameter and variable management.
- **Parameters:** Pipelines allow you to parameterize values during the pipeline runtime. This allows you to build loops that can change the source or destination each pass.
- **Pipeline Runs:** A data pipeline run is created each time the pipeline is executed. Each run can be scheduled or run on-demand.



Pipeline Templates: You can pick from a list of predefined pipeline templates.

- Bulk Copy from Database
- Bulk Copy from Files to Database
- Copy data from ADLS Gen2 to Lakehouse File
- Copy Data from ADLS Gen2 to Lakehouse Table
- Copy Data from Azure SQL DB to Lakehouse Table
- Copy Multiple Files Containers between File Stores
- Copy New Files Only By LastModifiedDate
- Delete File Older than 30 Days
- Move Files



View a pipeline history: Start Time | End Time | Duration | Triggered By | Status

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

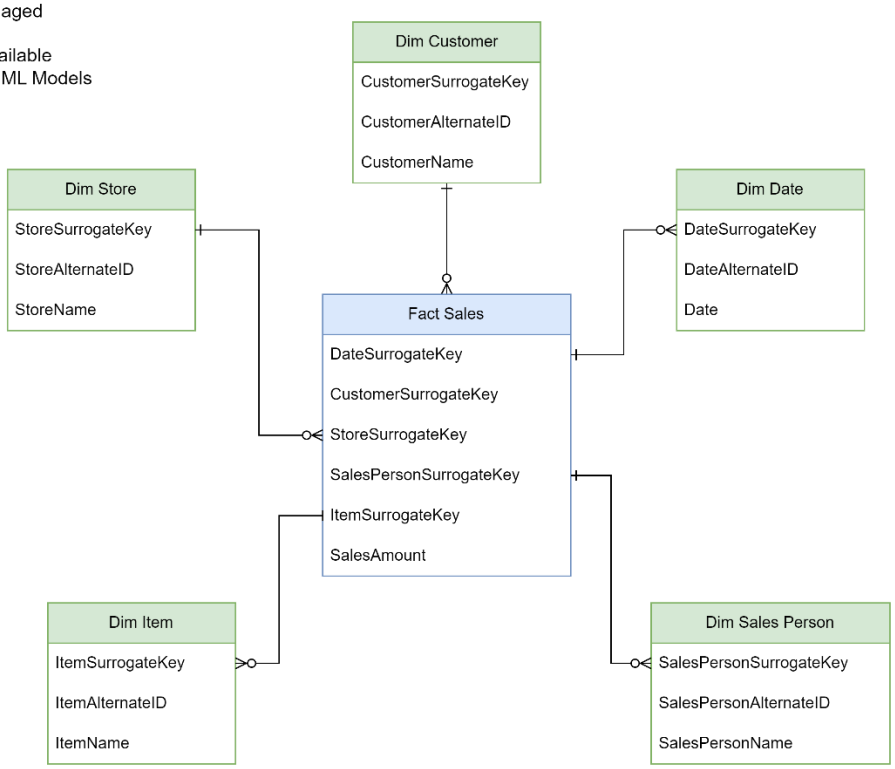
Data Warehouse Capabilities

Denormalized, Multidimensional, Star Schema, Kimball Model, Snowflake, Conformed Dimensions, etc...
All terms used to describe the standard design for a relational data warehouse. Microsoft Fabric warehouse is a modern version of the traditional data warehouse. You can use it to create a single source of truth for analytics across organization's departments. There is the continued full support of SQL semantics with insert, update, and delete. The modern twist comes from the fact it is built on the Fabric Lakehouse which is stored in Delta format. This allows the entire data team use the data solution.

Fact Tables: Used to store metrics for analysis. Fact tables will usually have large number of rows and primarily used for data analytics.
Dimension Tables: Used to store descriptive fields about the fact table metrics. Dimension tables usually have smaller row counts, and are used to provide context.

Key Capabilities

- Full T-SQL Support
- Fully Managed
- Scalable
- Highly Available
- Spark for ML Models



Fabric Data Warehouse Load Sequence

1. Ingest (Extract, Load) to a data lake. Opportunity to apply transformations/data quality checks
2. Load files into staging tables in the warehouse
3. Load the dimension tables from the staging tables with inserts/updates.
4. Load the fact tables from the staging tables with inserts (check for late arriving dims)
5. Optimize post-load with indexes and table distribution statistics.

- Note: you can load the data warehouse with Copy Into by using cross-database querying from the Lakehouse

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

Monitoring a Fabric Data Warehouse

Using the Microsoft Fabric Capacity Metrics App and begin by filtering activity to only show "Warehouse". You can look at historical information about queries that have been run. This will allow you to identify frequently used or long-running queries that may benefit from query tuning.

Query Insights

- queryinsights.exec_requests_history
- queryinsights.long_running_queries
- queryinsights.frequently_run_queries

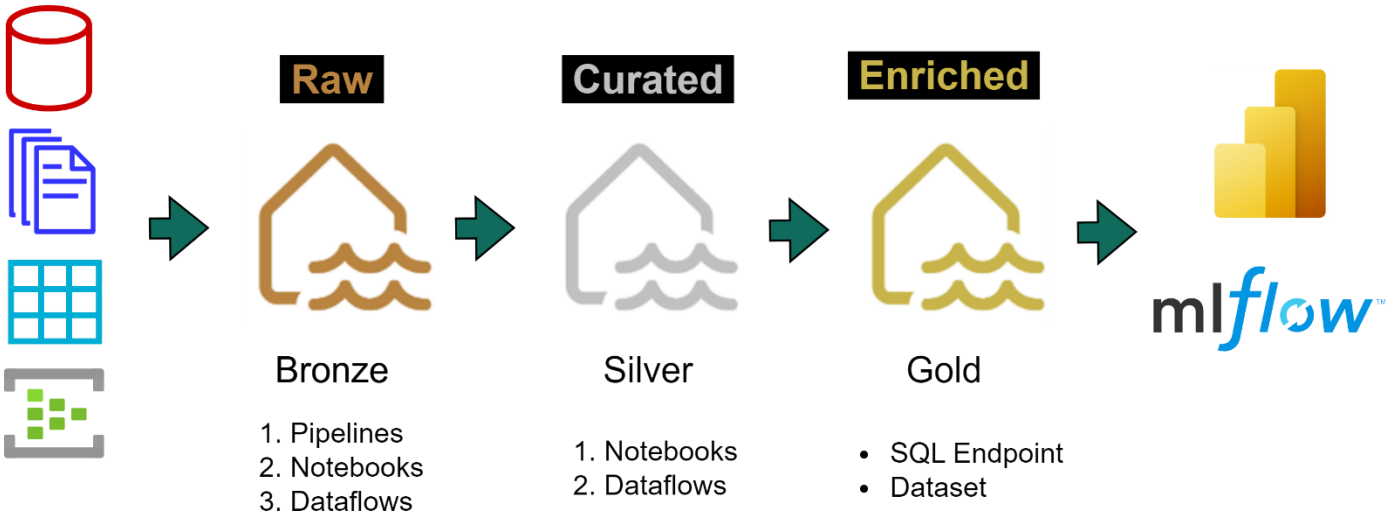
Table Schema

exec_request_history	long_running_queries	frequently_run_queries
Batch_id	Last_dist_statement_id	Avg_total_elapsed_time_ms
Command	Last_run_command	Last_dist_statement_id
Connection_id	Last_run_session_id	Last_run_command
Distributed_statement_id	Last_run_start_time	Last_run_start_time
End_time	Last_run_total_elapsed_time_ms	Last_run_total_elapse_time_ms
Login_name	Median_total_elapsed_time_ms	Max_run_total_elapsed_time_ms
Program_name	Number_of_runs	Min_run_total_elapsed_time_ms
Query_harsh		Number_of_cancled_runs
Root_batch_id		Number_of_failed_runs
Row_count		Number_of_runs
Session_id		Number_of_successful_runs
Start_time		
Status		
Total_elapsed_time_ms		

Data Lakehouse Capabilities

The Lakehouse or Medallion architecture is a design pattern to organize data logically. It typically involves 3 layers called Raw (bronze), Curated (silver), and Enriched (gold). Each layer represents higher and higher level of data quality, and readiness. The Lakehouse architecture is flexible and organizations can implement additional layers as needed. Hint: Platinum, Diamond, Onyx, etc :)

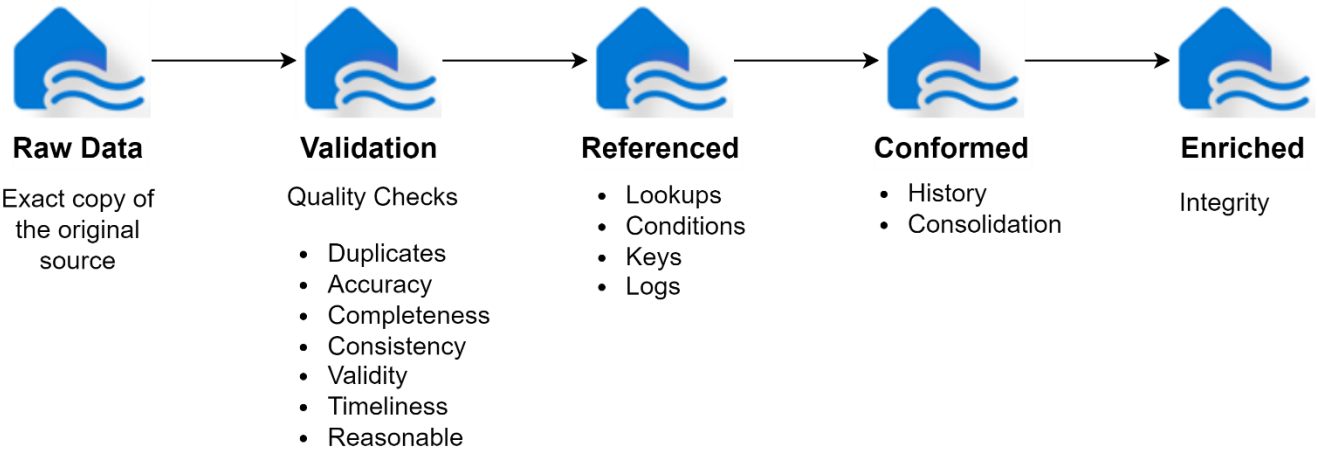
The Raw (Bronze) Layer: This is the landing zone for all data (structured, semi-structured, unstructured)
The Curated (Silver) Layer: This is where the data is refined and quality checks are performed for validation.
The Enriched (Gold) Layer: Here is where data is modeled to align with business and analytical requirements.



Considerations for data moves & transformations across the layers

- What is the volume of data you are working on?
- Are the transformations complex, or straightforward?
- Will data be streamed, or batched?

Example of additional Lakehouse Layers



References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

Real-time Analytic Capabilities

Real-time Analytics in Fabric uses KQL Database for the storage layer and providers KQL (Kusto Query Language) for analysis. A key benefit is the optimization in KQL for time-series analysis from logs or structured data.

- Time-series optimization
- Automatic partitioning and indexing
- Ranging from gigabytes to petabytes

- IoT devices
- Web and App logs
- Social media feeds

- Manufacturing
- Oil and gas
- Automotive
- Etc.

- Structured
- Semi-structured
- Unstructured



Integrate streaming data from various sources. Can land in KQL Database, Lakehouse, or Warehouse, etc.

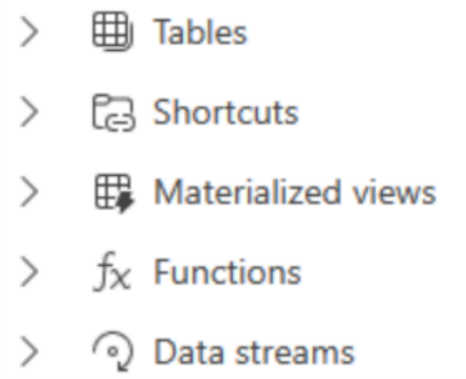


A Kusto database enables you to store multiple tables, stored functions, and materialized views.

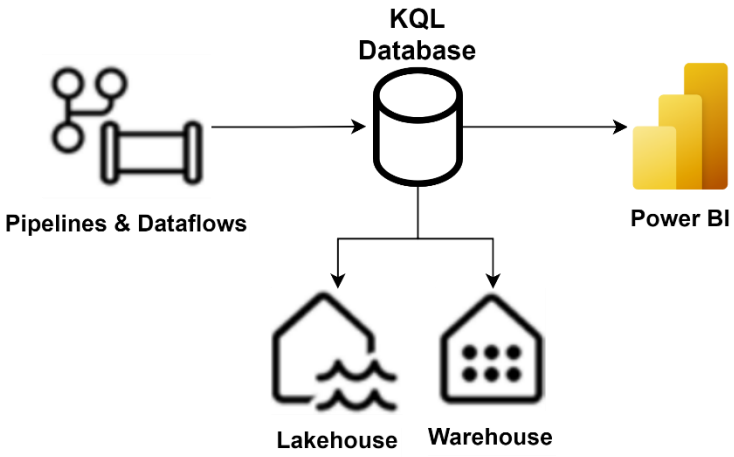


Enables you to create and run queries from Data Explorer.





- Table: A defined schema for columns with data type pairs.
- Materialized Views: Precomputed results of a query stored for faster retrieval. Includes optional list of parameters.
- Functions: A stores a subquery that can be called by KQL queries. Includes optional list of parameters.
- Data Streams: The KQL event stream connected to the KQL database



Kusto Query Language

- Work and visualize various data sources in real-time.
- Create notebooks to encapsulate your code and results.
- Create runbooks that detail troubleshooting steps based on real-time data.
- Portal experience provides guidance to build from scratch.

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

Kusto Query Example The query operates with a pipe ("|") allowing you to filter like a funnel. Starting at the entire data table and filtering rows through each pipe layer until you define the output.

Types of Joins

- Innerunique
- Inner
- Leftouter
- Rightouter
- Fullouter
- Leftsemi
- Leftanti, anti, leftantisemi
- Rightsemi
- Rightanti, rightanitsemi

```
RainMeasures
| where StartTime between (datetime(2023-01-01)..datetime(2023-02-01))
| where State == "MINNESOTA"
| where EventType has "wind"
| top 10 by Measurement
| project EventType, State, County, Measurement

RainMeasures
| where StartTime between (datetime(2023-01-01)..datetime(2023-02-01))
| where State == "MINNESOTA"
| where EventType has "wind"
| project EventType, State, County, StartTime, Measurement
| sort by StartTime desc

RainMeasures
| where StartTime between (datetime(2023-01-01)..datetime(2023-02-01))
| where State == "MINNESOTA"
| where EventType has "wind"
| project EventType, State, County, StartTime, Measurement
| sort by StartTime de
| take 10

RainMeasures
| where StartTime between (datetime(2023-01-01)..datetime(2023-02-01))
| where State == "MINNESOTA"
| where EventType has "wind"
| count
```

```
RainMeasures
| summarize count()

RainMeasures
| summarize MeasureCount = count() by State
| sort by MeasureCount

RainMeasures
| summarize count(),
      EventsWithDamage = countif(Damage > 0),
      Count = dcount(EventType) by State
| sort by Count

RainMeasures
| where State == “MINNESOTA”
| distinct EventType
| sort by EventType asc
```

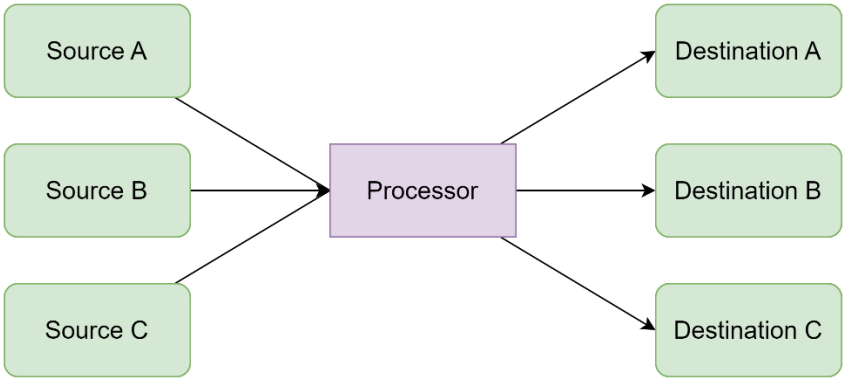
Render Operator: Options include columnchart, barchart, piechart, scatterchart, pivotchart.

```
RainMeasures
| summarize count(),
      EventsWithDamage = countif(Damange > 0),
      dcount(EventType) by State
| sort by count_
| render barchart

RainMeasures
| summarize count() by bin(StartTime, 7d)
| render columnchart

RainMeasures
| extend damange = DamageProperty + DamageCrops
| summarize sum(damage) by bin(StartTime, 7d)
| redner columnchart

RainMeasures
| extend damange = DamageProperty + DamangeCrops
| summarize round(avgif(damage, damage > 0)) by EventType
| sort by avgif_damage
```

```
RainMeasures
let MinDamage = 1; //int
let EventLocation = “MINNESOTA”; //string
RainMeasures
| where State == EventLocation
| where DamageCrops + DamageProperty >= MinDamage
| summarize Damage=round(avg(DamanageProperty + DamageCrops)) by EventType
| sort by Damage

RainMeasures
Let MostFrequentEventType = toscalar(
     StormEvents
     | summarize count() by EventType
     | top 1 by count_
     | project EventType);
StormEvents
| where EventType == MostFrequentEventType
| summarize count() by startofmonth(StartTime)
| render columnchart

TABLE_NAME
| count

TABLE_NAME
| getschema

Customer
| join kind=inner SalesFact on CustomerKey
| take 10

Customers
| join kind=inner SalesFact on CustomerKey
|summarize TotalAmount = round(sum(SalesAmount)) by RegionCountryName
| top 3 by TotalAmount

SalesFact
| lookup Customers on CustomerKey
| summarize TotalSales = count() by RegionCountryName
| order by TotalSales desc
```

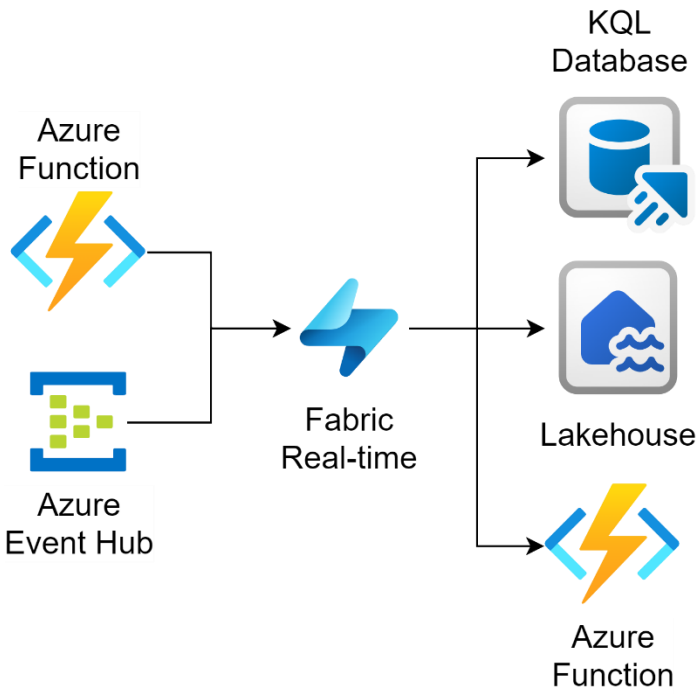
Eventstreams in Microsoft Fabric

Eventstreams allow you to ingest real-time events in a no-code development environment. In the environment you to setup event sources, destinations, and processors for the eventstream. The event processing capabilities include collections, aggregates, filters, groupings, transformations, and more. This will allow you to connect one or more data sources, apply filters, or enrichment operations, and multiple destinations.



Event Stream Components

- **Eventstream:** Created within the portal it contains the event data flow components and endorsement status. In the portal you can view the event data with insights and logs for the data sources and destinations.
- **Source:** The event data source which can be Azure Event Hub or a custom app and includes the data format.
- **Processor:** Once you connect your source and destinations you will have several processing options that you can run before the data is sent to the destination. Aggregate, Expand, Filter, Group By, Field Selection, Union
- **Destination:** The event data sink or destination where the data will end up. The destinations include KQL database or the lakehouse. The includes the table name, and schema mapping.



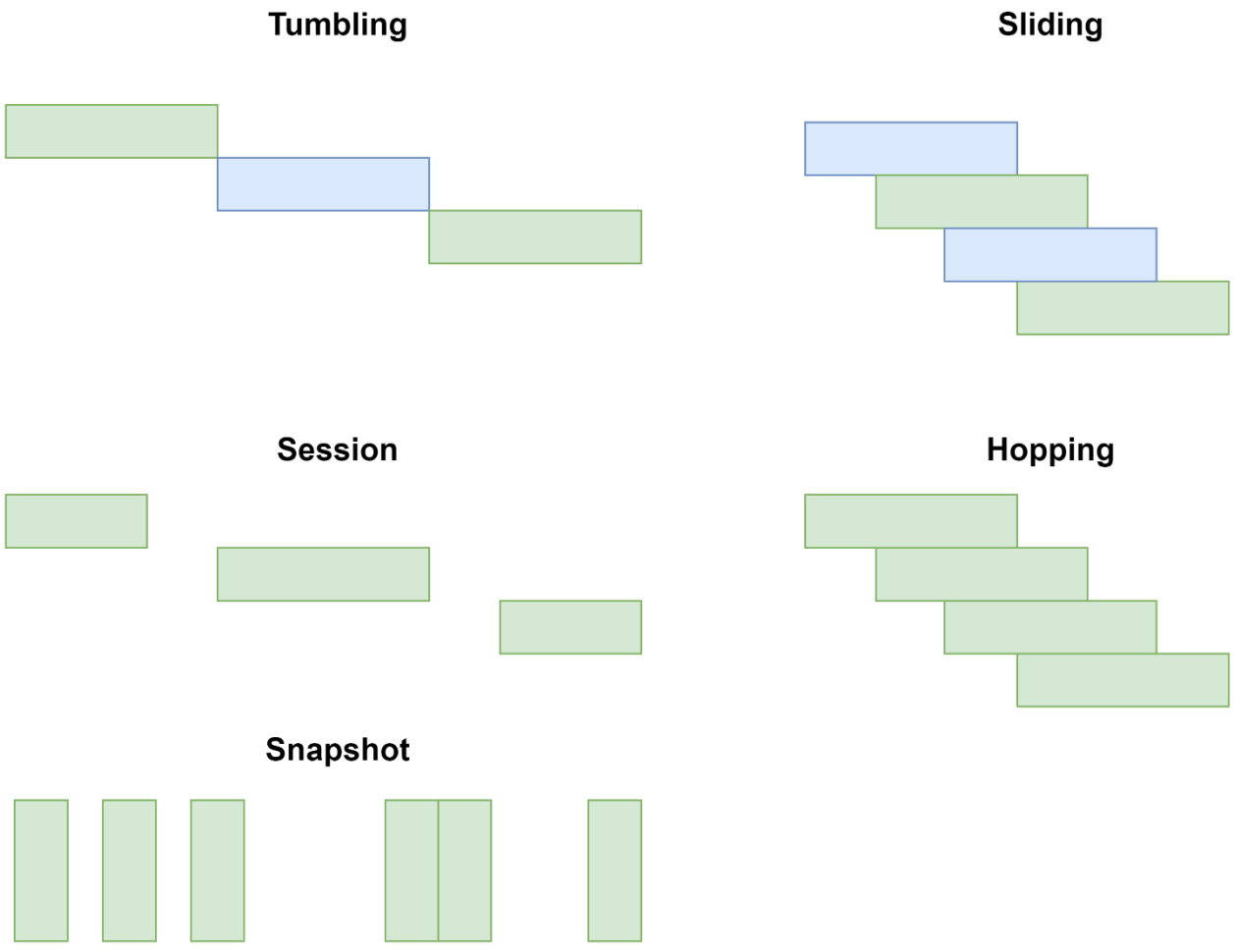
References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-11-10, from https://learn.microsoft.com/en-us/fabric/

Windowing Options in Eventstreams

A **window function** is a way to operate on data within a time-window or temporal window. This enables you to perform various techniques that analyzes data over time. Use cases include IoT sensor readings, or application/web clicks, or any transactions with a timestamp flowing in real-time. Window functions include tumbling, sliding, snapshot, hopping, or session. Beyond picking the type of window you will also set the duration, offset, group key, and aggregate function.

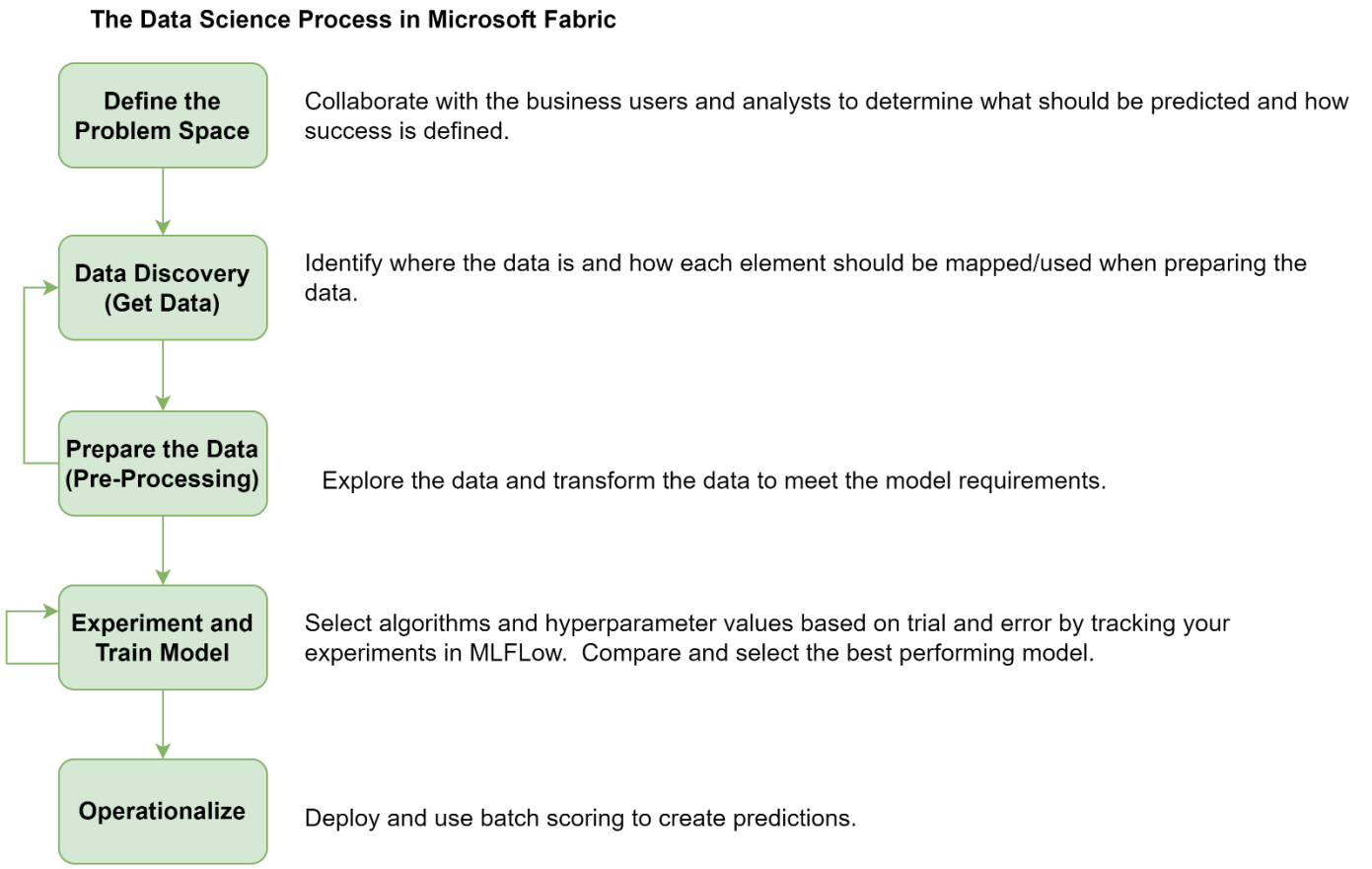
Window Types

- Tumbling: Divides incoming events into fixed and non-overlapping intervals based on arrival time.
- Sliding: Takes incoming events into fixed-continuous and overlapping intervals based on time.
- Session: Divides events into variable and non-overlapping intervals based on a gap of inactivity.
- Hopping: Takes incoming events into multiple fixed and overlapping intervals based on time.
- Snapshot: Simply group events based on the time they occur.



Data Science Capabilities

Covering the basics of Data Science in Microsoft Fabric. In Fabric you are able to train machine learning algorithms in notebooks, and track them with in the industry leading open-source library MLFlow which integrates with Fabric experiments. Machine Learning algorithms or models allow you to generate predictions for action.



Key Features for Data Science Workers

- Notebooks: Write and run code on Apache Spark compute like running one more experiments in a notebook.
- Experiments: Track your workloads for training a machine learning model. An experiment can consist of multiple runs.
- Models: Deploy your ML algorithm as a model in Fabric.

Common Machine Learning Algorithms

- Classification: Predict a categorical value
- Regression: Predict a numerical value
- Forecasting: Predict a future numerical value

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

Importance of Data Distribution

- Exploratory Data Analysis: Dive into the data profile to understand the distribution and patterns.
- Data Preprocessing: Apply transformations to normalize or standardize the data distribution.
- Model Selection: Some models require the data to have a normalized distribution. Be aware of your model selection regarding your data distribution.
- Improve Mode Performance: By normalizing the data distribution, you remove the skewness that could increase the error rate of your models.
- Model Relevance: Track the model drift if the data in production introduces new patterns that create skewed distribution.

Data Distribution

- Normal Distribution
- Binomial Distribution
- Uniform Distribution

Missing Data Distribution

- Identify Missing Data
- Evaluate Nature of Missing Values
    - Missing Completely at Random (MCAR)
    - Missing at Random (MAR)
    - Missing Not at Random (MNAR)
- Handling Missing Data
    - Ignore
    - Remove
    - Impute (fill in with mean median mode or K-Nearest Neighbors)
    - Use it as a new Feature

Simplify Data

- Correlation
- Principal Component Analysis (PCA) Reduce the number of variables while retaining most of the original information

Understand Categorical and Numerical Variables

- Summary Statistics
- Univariate Analysis
- Bivariate and multivariate Analysis

Generate Predictions

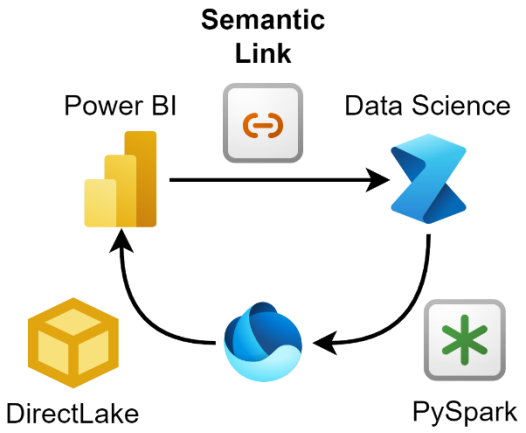
After you trained a model you can use the model to generate predictions. The model will take incoming data as input, and perform the necessary transformations you designed during preprocessing, and out the predicted values. The MLmodel file will stored the expected inputs and outputs definition.

Defining the Model Signature

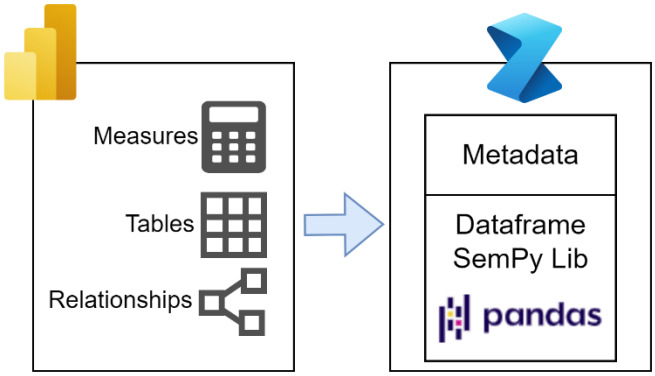
- Select Input Table (column-based or tensor-based)
- Map Input Columns
- Create Output Table
- Map Output Columns
- Configure Notebook

Semantic Link in Data Science

Enable Data Scientists to directly use their organization's most refined data sitting in Power BI semantic models. Using the Fabric Data Science persona you can import a Python library (SemPy) and directly use the columns and measures in a semantic model (dataset) within your notebook. This feature is only supported in Microsoft Fabric. This will help speed up discovery and reduce the risk of duplicating measure logic in multiple places.



Semantic link enables user to connect to the semantic model in a number of ways including the SemPy Python library for pandas users. SemPy allows you to connect and load data from tables, measures, and metadata like relationships.



References

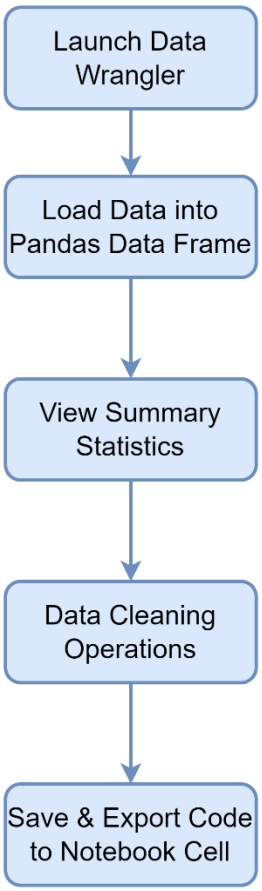
- https://learn.microsoft.com/en-us/fabric/data-science/semantic-link-overview
- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html
- https://learn.microsoft.com/en-us/fabric/data-science/semantic-link-power-bi?tabs=sql#data-augmentation-with-power-bi-measures

Data Wrangler Capabilities

Data Wrangler is a comprehensive tool for preparing data to be used in Machine Learning. Wrangler leverages the Fabric notebooks to perform data exploration and preprocessing work. A key capability to help with Data

Key Features for Observations Include: • Data Display • Dynamic Summary Statistics • Built-in visualizations • Preprocessing Operations Library	Built on Microsoft Fabric Notebooks  All operations can generate reusable code saved back to the notebook.

Explore the data grid and summary information for variable insights. Perform common data operations for cleansing to handle missing values, anomalies, and data type issues. You can create new features based on your understanding of the data before you run your experiments.



Data Wrangler Operators (2023-12):

- **Sort**
- **Filter**
- **One-hot encode**
- **One-hot encode with delimiter**
- **Change column type**
- **Drop column**
- **Select column**
- **Rename column**
- **Drop missing values**
- **Drop duplicate rows**
- **Fill missing values**
- **Find and replace**
- **Group by column and aggregate**
- **Strip whitespace**
- **Split text**
- **Convert text to lowercase**
- **Convert text to uppercase**
- **Scale min/max values**
- **Flash fill**

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-12-07, from https://learn.microsoft.com/en-us/fabric/data-science/data-wrangler

MLFlow in Microsoft Fabric

MLFlow helps Data Scientists to manage their models. Use notebooks to train models and track their work in experiments in Microsoft Fabric.

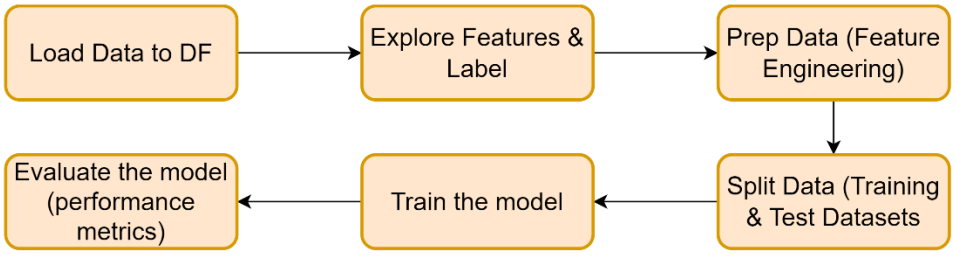
Quick Overview of the ML Frameworks available in Microsoft Fabric



MLFlow helps Data Scientists focus on how to train and track their models. Step 4 below:

1. Define the problem: Ideation with business and analysts what should be predicated and success metrics.
2. Get the data: Identify the data sources to support the process and success metrics.
3. Prepare the data: Preprocessing data is a large portion of the work to clean and transform the data.
4. **Train the model: Select algorithms and their hyperparameters through trail and error. MLFlow will enable you to track your experiment metrics.**
5. Generate predictions: Batch scoring predictions

Steps for the Training the Model



To help manage the results of your model training you can use MLFlow to track the model evaluation (performance metrics). Also, MLFlow can create and track standard performance metrics for you to get started. An added benefit of using MFLow is saving your model in a format for model scoring in a later stage.

MLFlow is installed in Microsoft Fabric notebooks. To track models with MLFlow all you have to do is import the library (`import mlflow`) and you can start logging.



Since MLFlow helps to track your performance metrics through logging. You will be able to reproduce your model results. This allows you to review the metrics and make choices on the best model training approach. MLFlow is an open-source library, and with the Tracking component you can log parameters, metrics, and any artifacts about your model.

- Create an experiment to track work in a notebook. `mlflow.set_experiment`
  - Autologging will track everything.
  - Custom logging allows you to explicityly track any parameters, metrics, artifacts, or models.
    - `mlflow.log_param()`
    - `mlflow.log_metric()`
    - `mlflow.log_artifact()`
    - `mlflow.log_model()`
- Retrieve active metrics from an experiment in a notebook. `mlflow.search_experiments`
- Retrieve specific experiments you can run in a notebook. `mlflow.get_experiment_by_name()`
- Retrieve runs from your notebook. `mlflow.search_runs`

Managing Models with MLFlow

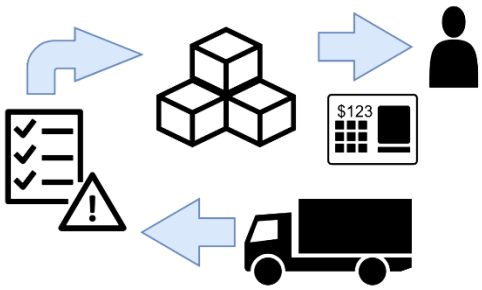
Once you have your model trained and ready scoring and generating new predictions you can use MLflow to store your model. Storing your model allows you to load the model into different environments. The format will depend on the framework you pick PyTorch, TensorFlow, Scikit, SynapseML.

Your model in Fabric will also show you the various versions of the model that have been created (as long as you use the same model name). Your model folder in Microsoft Fabric will typically contain the following files:

`MLmodel`	The models metadata.
`conda.yaml`	The Anaconda environment needed to run the model
`model.pkl`	Contains the trained model.
`python_env.yaml`	Describes the Python environment needed to run the model.
`requirements.txt`	Lists Python packages required to run the model.

Data Activator Capabilities

Data Activator allows you to trigger actions based on conditions you setup against your data. You can evaluate data from multiple sources to create real-time responses. There are 4 core concepts in Data Activator, and they are Events, Objects, Triggers, and Properties. In the Fabric portal you create Data Activator artifacts by creating Reflex items.



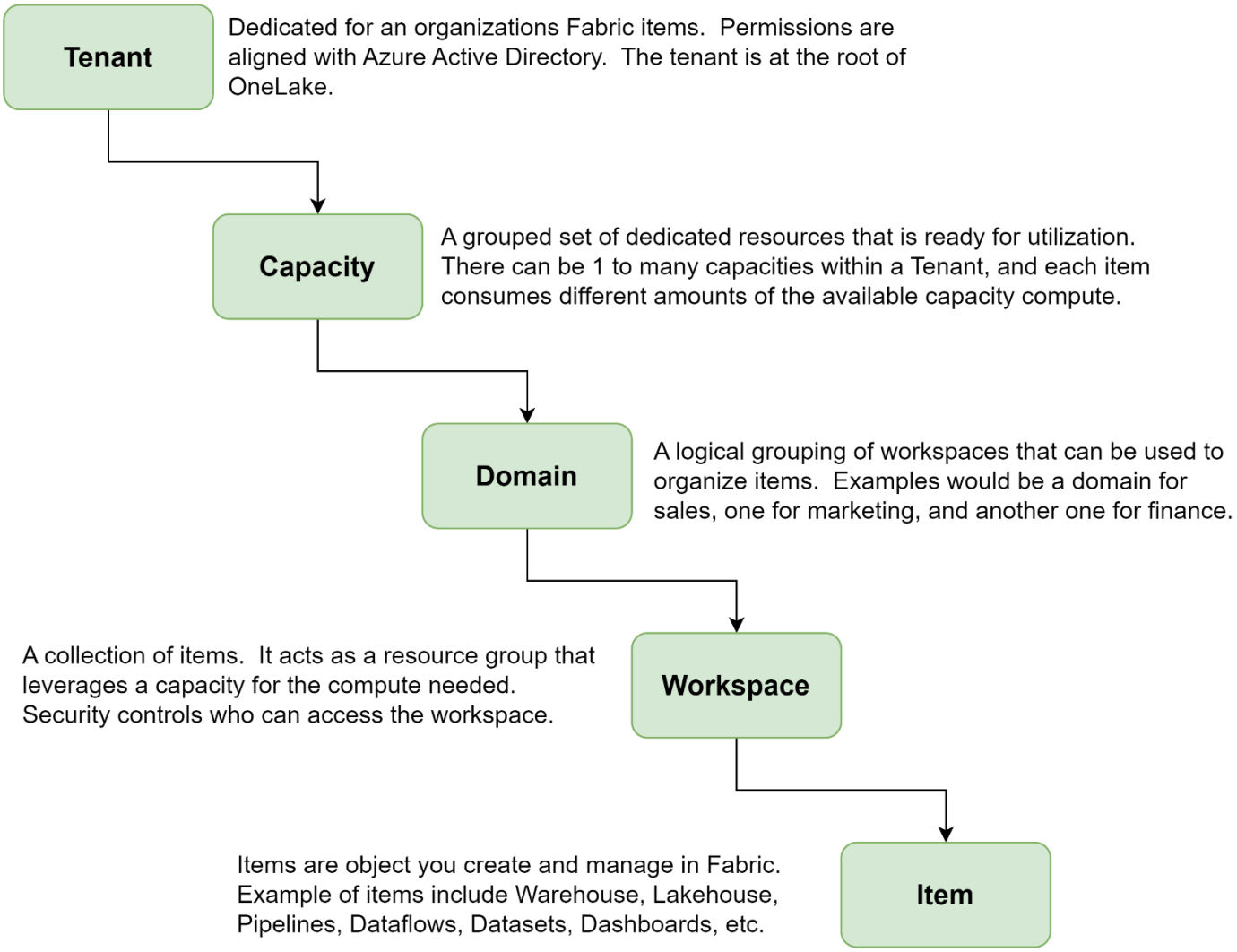
Start Actions: Marketing plans based on sales trend
Start Actions: Operations when equipment failure could affect product
Start Actions: Help Desk when app telemetry customer experience issue
Start Actions: Receiving when tracking updates are missing
Start Actions: IT data quality thresholds are reached in analytics

- **Objects:** The objects in Data Activator are logical representations that you want to monitor.
- **Events:** The event streams are all the data sources that could represent an object. Integration with Fabric allows for easy ingest of EventStreams and Power BI Reports (visuals with both x & y axis).
- **Properties:** From the events you select the object's ID, and various properties about the object.
- **Triggers:** The Data Activator's triggers monitor the events and properties and will start actions based on conditions being met.
    - **Conditions:** The specific circumstances based on events where the trigger is activated.
    - **Actions:** These are tasks that Data Activator executes when conditions are met and include notifications to complex workflows.

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

Administration Capabilities

We will start with the basic concept and that is there can only be one OneLake. Administration tools include Admin Portal, PowerShell cmdlets, Admin APIs and SDKs, Workspace Monitor. The following are Service Management areas.



Additional Service Management Areas (Power BI)

- Embed Codes: Used to control which reports have been published publicly.
- Organizational Visuals: Register and manage organizational visuals.
- Azure Connections: Integrate with Azure services to provide additional functionality.

- References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-12-18, from https://learn.microsoft.com/en-us/power-bi/guidance/fabric-adoption-roadmap-system-oversight
- References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

Fabric Administrator Roles

Fabric Administrator — Admin over tenant level settings. This is a high privilege role able to impact settings relating to user experience, full security access, personal workspace access, and metadata.

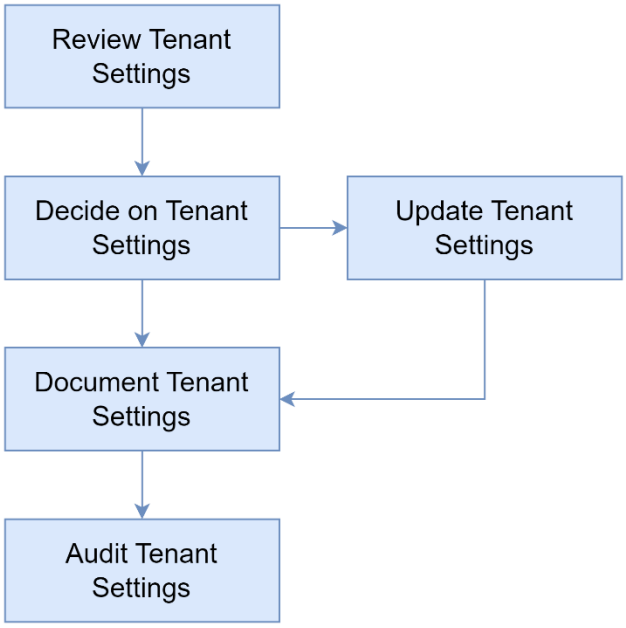
Capacity Administrator — Admin over workspaces, workloads, and monitoring. (1:N capacities as assigned)

Gateway Administrator — Admin over data gateway data source configuration, credentials, and user assignments. Coordinates gateway software updates. (1:N gateways as assigned)

Workspace Administrator — Manages workspace settings and access. (1:N workspaces as assigned)

Managing Tenant Level Settings as Fabric Administrator: You can leverage the API call to document the tenant settings for monthly or quarterly review. Also, recommend setting up the API to running weekly. If a change is made incorrectly having a weekly log of the settings will make it easier to reset.

```
GET https://api.fabric.microsoft.com/v1/admin/tenantsettings
```



Admin Tasks

- Security and access control: Manage Users & Manage Items & Sharing
- Data governance: Endorse Content & Sensitive data scans & Data Lineage
- Customization and configuration
- Monitoring and optimization

Security and Monitoring Capabilities

Security Features

- RBAC Role-Based Access Control to the warehouse and data.
- AAD Azure Active Directory access control to the warehouse with Multi-Factor Authentication.
- SSL Secure Socket Layer to encrypt communication between the client app and the Fabric Warehouse.
- ASSE Azure Storage Service Encryption protects the data in transit and at rest.

Workspace Permissions

Fabric organizes data into workspaces. Users and Security Groups control access to the workspaces. Workspace roles include Admin, Member, Contributor, or Viewer.

- **Admin:** Administrator of the workspace
- **Member:** Add members and manage permissions
- **Contribute:** Publish content to the workspace
- **Viewer:** Read-only access to workspace content

Item Permissions

Workspace roles apply to all items within a workspace. Item permissions can be grant access to individual warehouses. Use T-SQL scripts or the portal for Read, ReadData, or ReadAll.

- **Read:** Users can connect via the connection string.
- **ReadData:** Users can read data from any table or view within the warehouse.
- **ReadAll:** Users can read data from the raw parquet files in OneLake with Spark.

Monitoring

Use these monitoring DMVs (dynamic management views) to manage connections, sessions, and status of live SQL queries.

- **sys.dm_exec_connections:** Returns information on each connection between the warehouse and the engine. SELECT * FROM sys.dm_exec_connections;
- **sys.dm_exec_sessions:** Returns information on each session authenticated between the item and the engine. SELECT * FROM sys.dm_exec_sessions;
- **sys.dem_exe_requests:** Returns information on each active session request. SELECT * FROM sys.dm_exec_requests;

References: Microsoft. (n.d.). Microsoft Fabric. Retrieved 2023-09-10, from https://learn.microsoft.com/en-us/fabric/

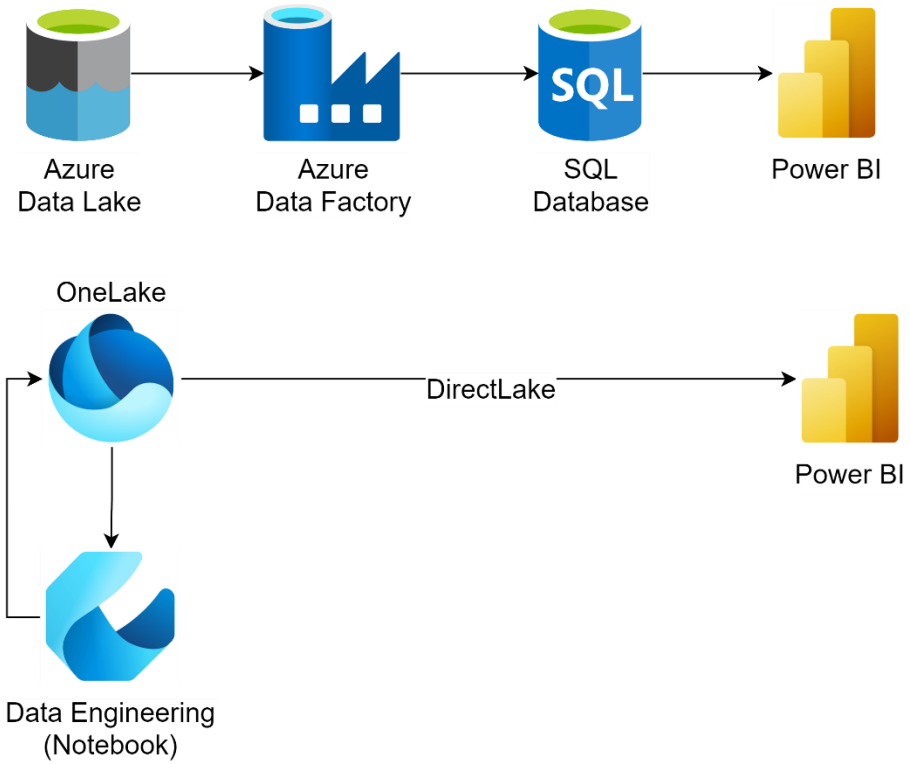
The Fabric Solution Space

Microsoft Fabric is built with well-established pre-existing technical solutions designed to support established data methodologies. The solution leverages Delta Lake created by the founders of Databricks and supported as open source through the Linux Foundation Projects.

The data value chain is a general framework that visualizes the process of increasing the value data assets. The general stages include: **Source > Data Ingest > Transformation > Model > Data Science > Visualization** Tools can help lower the barrier between each stage, and help shorten the value chain by reducing unnecessary data movements.

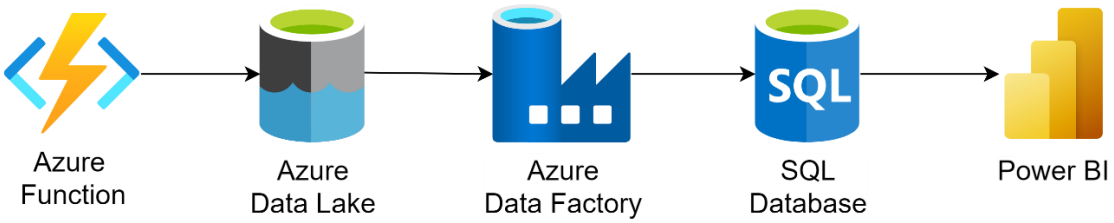
There are multiple solutions to meet your business needs. Often, there can be solution overlap where one tool could meet the needs of multiple business needs. It is important to understand the Solution(s) Space and how best to leverage Multi-persona vs Niche Solutions.

Basic Comparison Azure and Fabric

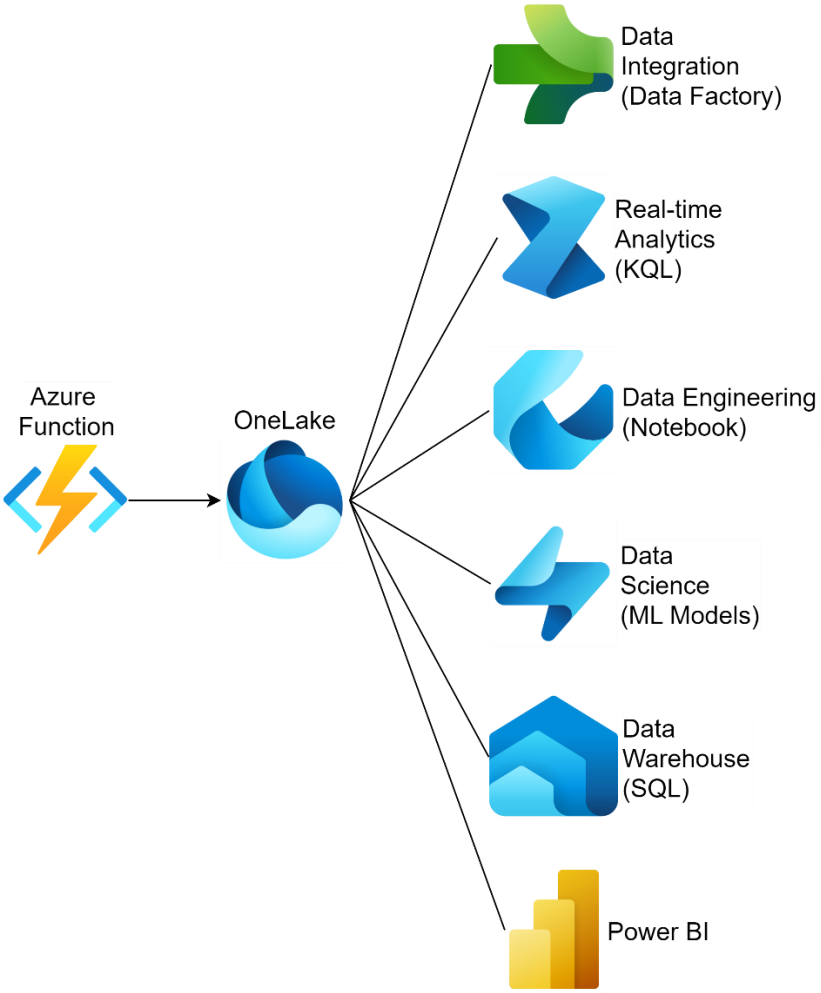


Comparing Azure SQL DB and Fabric

Azure

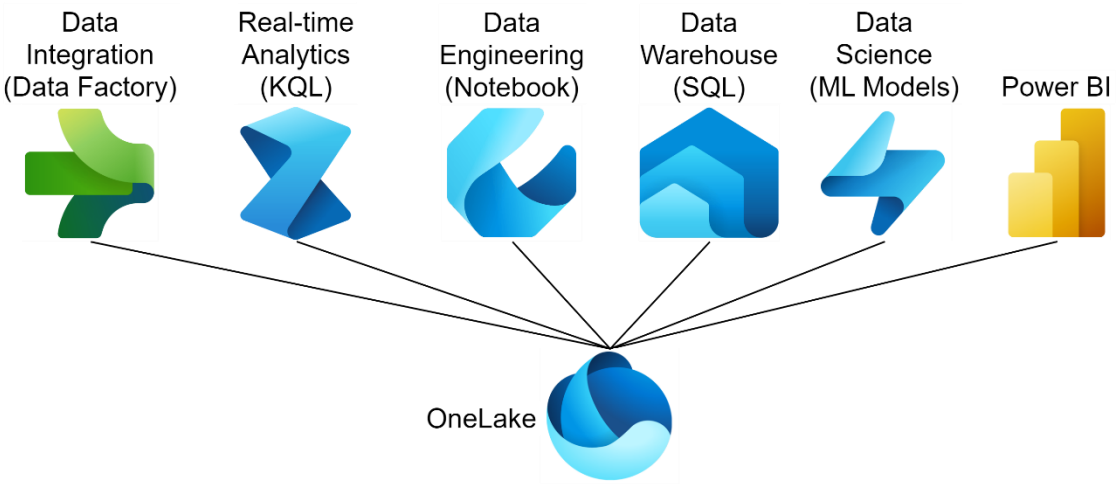


Fabric

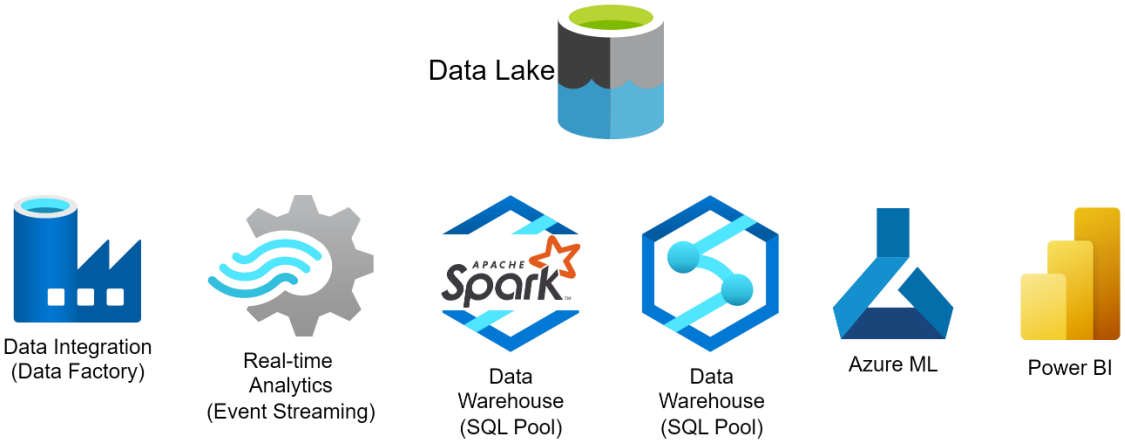


Comparing Azure Synapse and Fabric

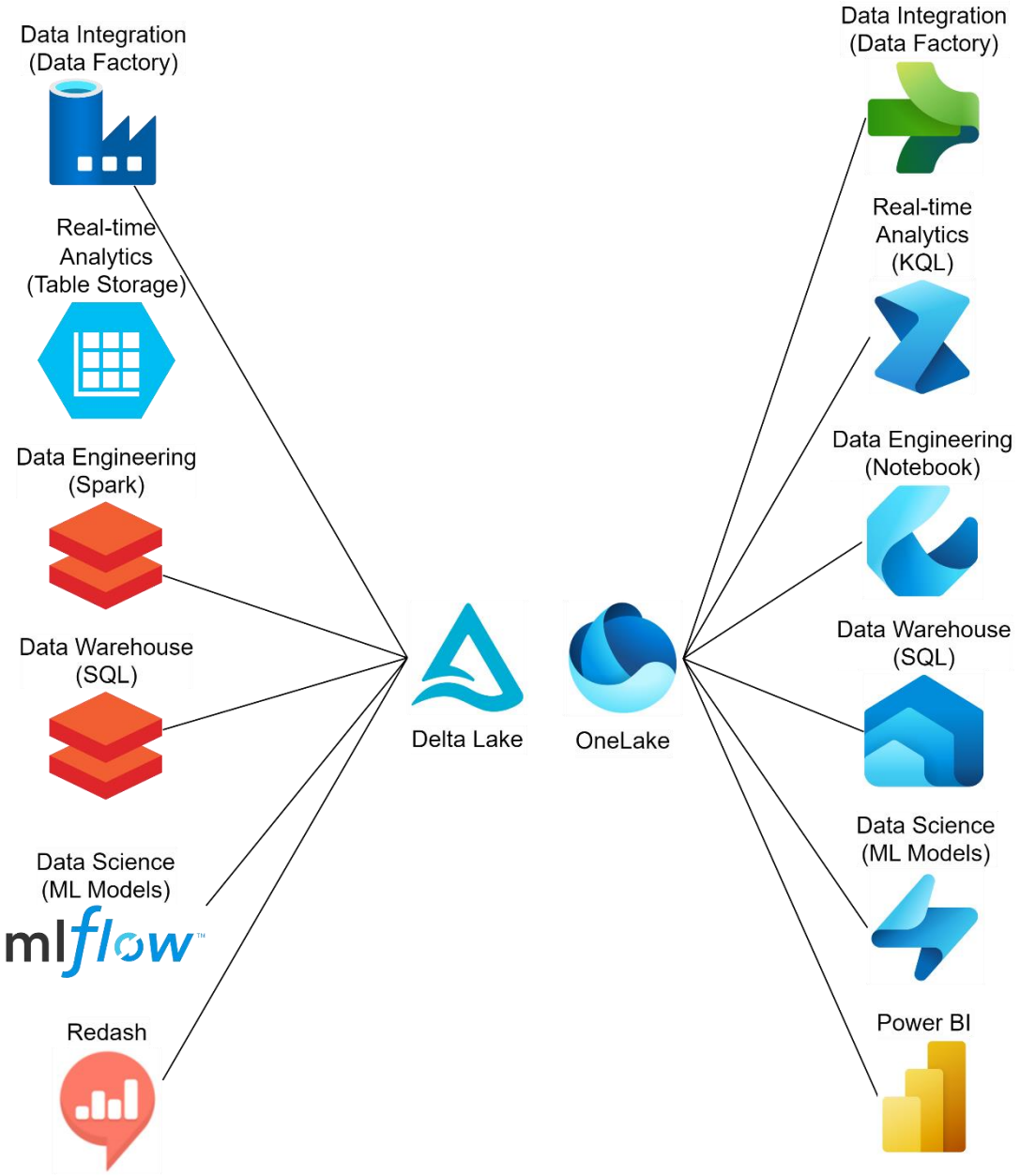
Microsoft Fabric



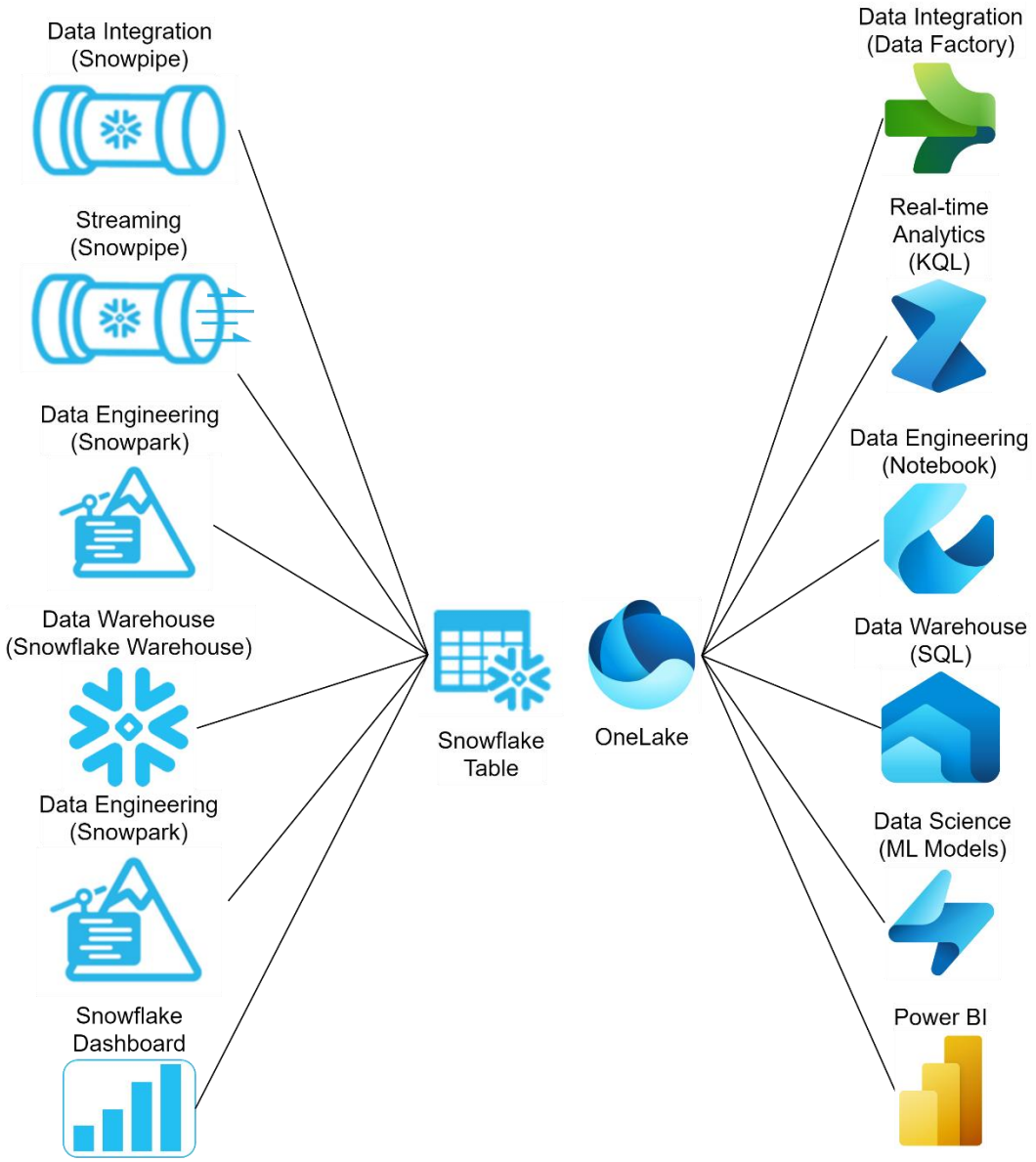
Azure Synapse



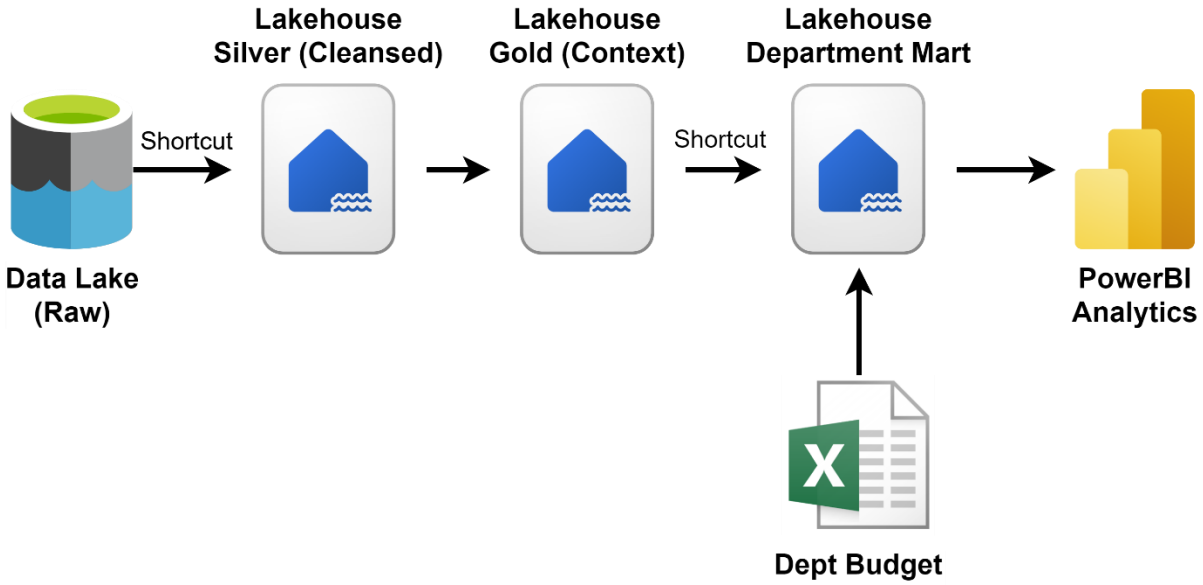
Comparing Azure Databricks and Fabric



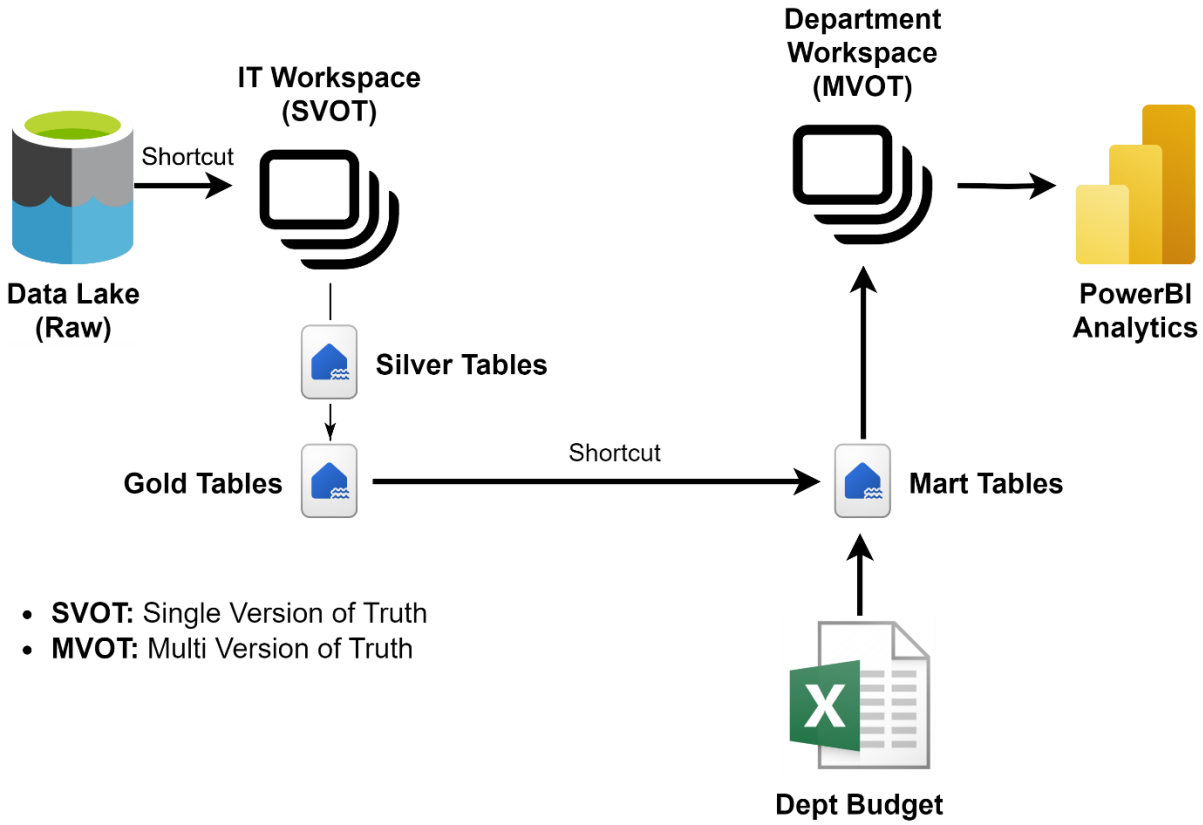
Comparing Azure Snowflake and Fabric



Logical Data Flow



Physical Data Flow



- **SVOT:** Single Version of Truth
- **MVOT:** Multi Version of Truth

Workload Balancing

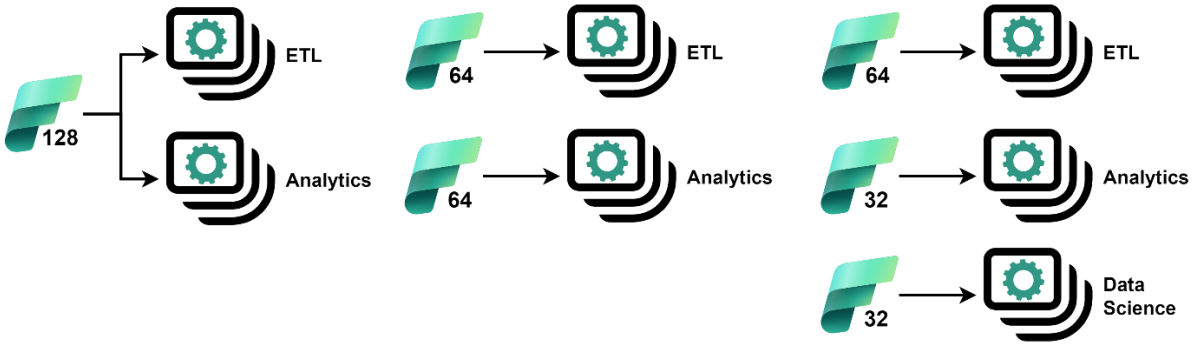
Managing the Fabric capacity is critical for all workload experiences. The experiences include both interactive user jobs, and scheduled background jobs. Capacity administrators play a vital role.

- Identify contributors to peak activity and work to optimize their queries or distribute to a new capacity.
- Maximize the capacity by pushing full utilization until throttling occurs, and scaling capacity as needed.
- Set alerts to be notified before a capacity is throttled to warn users or scale as needed.

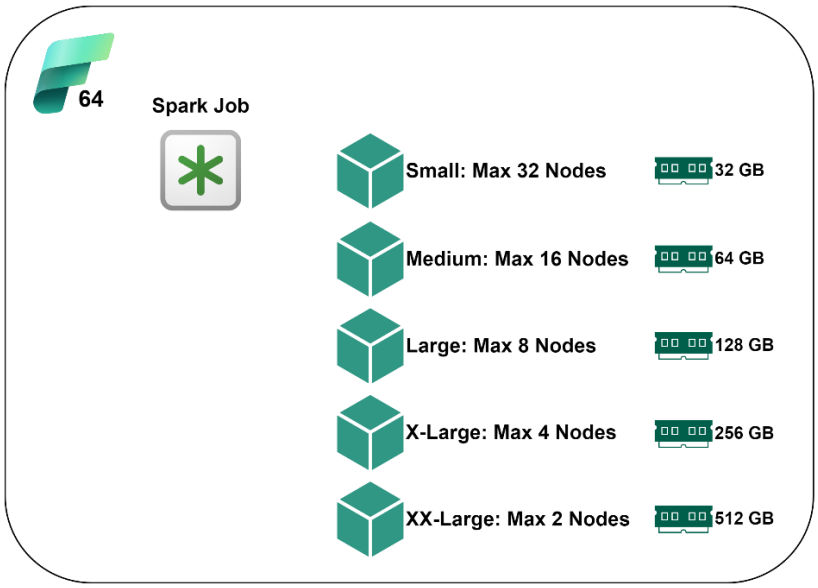
Workloads

Ingestion Isolation and Analytical Queries. Load operations have resource isolation from analytical workloads in the backend compute pool of Warehouses. Beyond isolation there are 2 architecture layers to help with workload management.

Sample capacity deployment options.



Sample spark pool configuration options.

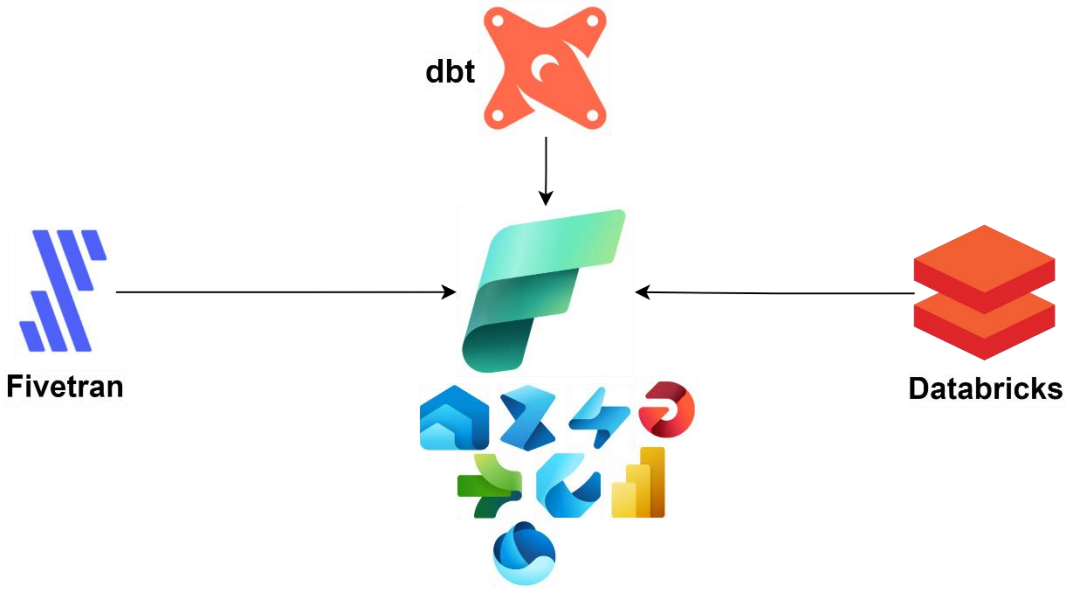


References

https://learn.microsoft.com/en-us/fabric/data-warehouse/compute-capacity-smoothing-throttling
https://learn.microsoft.com/en-us/fabric/data-engineering/create-custom-spark-pools
https://learn.microsoft.com/en-us/fabric/data-warehouse/workload-management

Best-of-Suite and Best-of-Breed

It may sound counterintuitive, but effectively utilizing Microsoft Fabric does not mean your data strategy is 100% Fabric based. Another way to say it is, we will leverage Microsoft Fabric as our organization's best-of-suite data platform while utilizing best-of-class components where needed in Azure (and beyond). Consider for example, using Fabric Pipelines to load data from SQL Server located on a VM into OneLake. Verses, leveraging Fivetran's SAP ERP connector.



Evaluating Microsoft Fabric

Acceptance and Evaluation Criteria

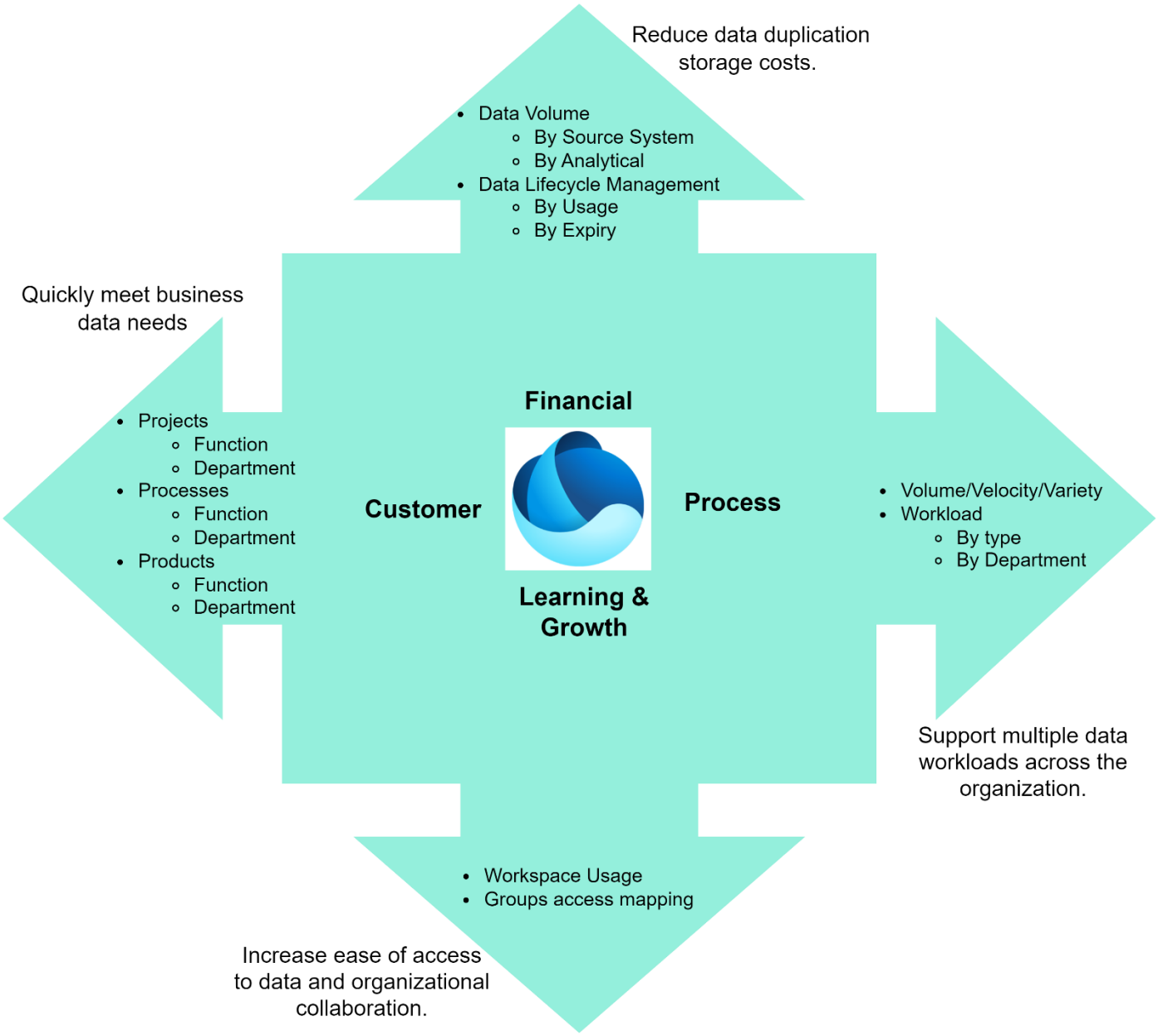
Regardless if you evaluate a single solution or multiple solutions it is key for all stakeholder to understand how the attributes were defined. The definition should include whether or not it is a pass/fail or value attribute, and a description to provide context (not shown). Below is a sample to help you get started.

Attribute	Definition	Data Integration	Real-time Analytics	Data Engineering	Data Warehouse	Data Science	Dashboard
Cost							
Ingres/Egress	Value						
Storage	Value						
Compute	Value						
Performance							
Write	Value						
Read	Value						
Usability							
Developer	Pass/Fail						
Analysts	Pass/Fail						
User	Pass/Fail						
Functionality	Pass/Fail						
Data Variety							
Dev Lang							
Features							
Security	Pass/Fail						
Scalability	Pass/Fail						
Reliability	Pass/Fail						
Etc.	Pass/Fail						

References: International Institute of Business Analysis. (2015). A guide to the business analysis body of knowledge (BABOK guide) (3rd ed.). Author.

OneLake Balanced Scorecard

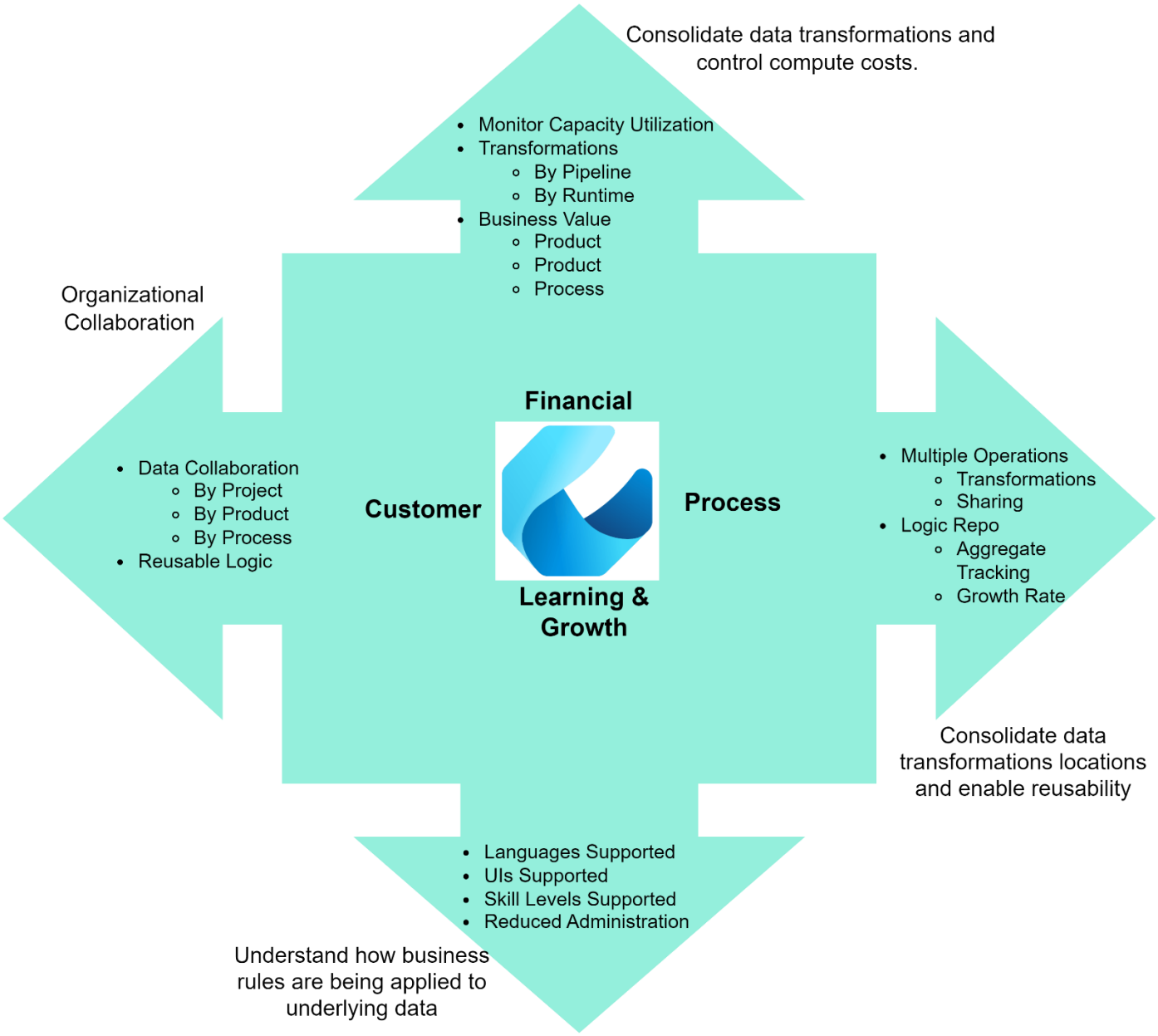
Ideas for targeting measures in the future state regarding business model, organizational structure, or process/product/project. This will help with a balanced view beyond financial impact. Each organization should identify their potential measures.



References: International Institute of Business Analysis. (2015). A guide to the business analysis body of knowledge (BABOK guide) (3rd ed.). Author.

Data Engineering Balanced Scorecard

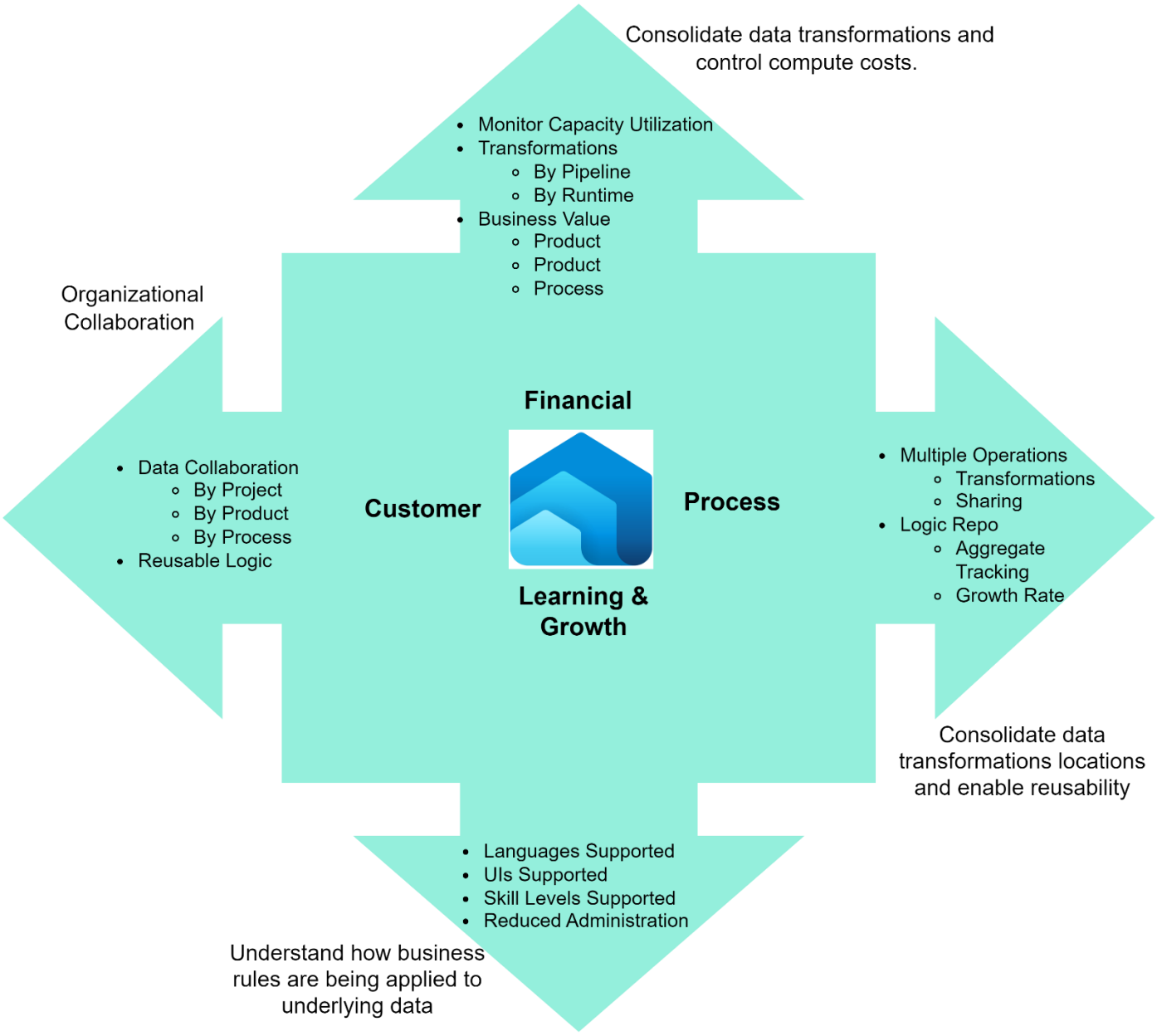
Ideas for targeting measures in the future state regarding business model, organizational structure, or process/product/project. This will help with a balanced view beyond financial impact. Each organization should identify their potential measures.



References: International Institute of Business Analysis. (2015). A guide to the business analysis body of knowledge (BABOK guide) (3rd ed.). Author.

Data Warehouse Balanced Scorecard

Ideas for targeting measures in the future state regarding business model, organizational structure, or process/product/project. This will help with a balanced view beyond financial impact. Each organization should identify their potential measures.



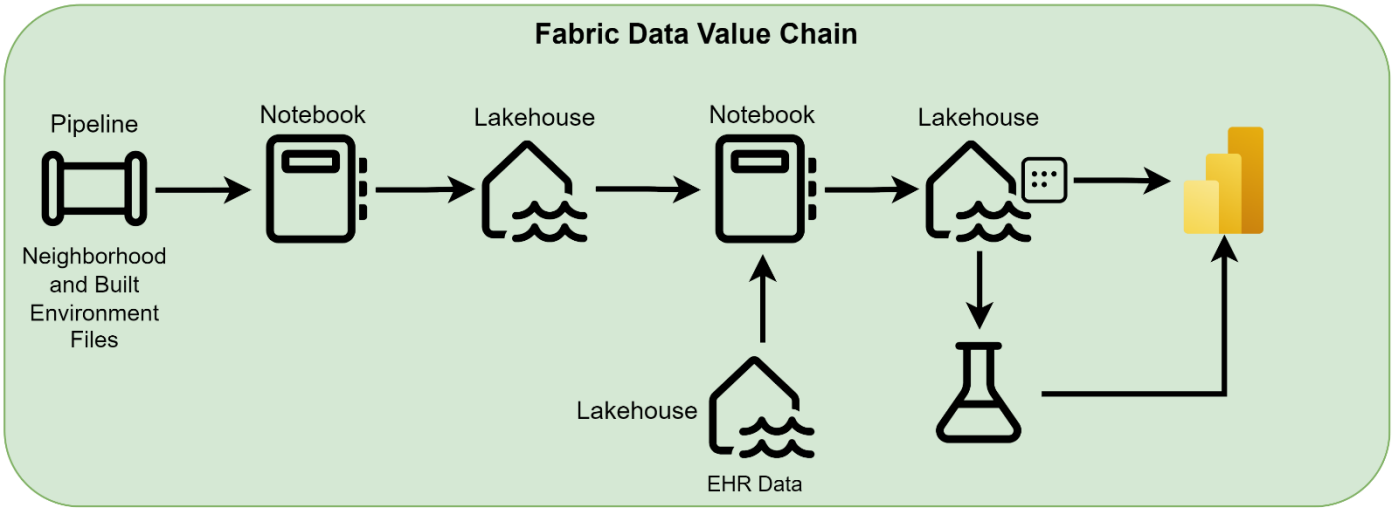
References: International Institute of Business Analysis. (2015). A guide to the business analysis body of knowledge (BABOK guide) (3rd ed.). Author.

Miscellaneous

Microsoft Fabric Alignment Example

Portfolio: Social Determinants of Health (SDOH) focused on the 5 domains of Economic Stability, Education Access and Quality, Health Care Access and Quality, Neighborhood and Built Environment, Social and Community Context. The purpose of this portfolio solution is to incorporate the SDOH data sources with internal EHR data for enhanced analytics to identify opportunities to increase quality, efficiency, and outcomes.

Project: Neighborhood and Built Environment Where people live has a big influence on their health and overall well-being. Identify where outreach programs would be the most effective. Objectives can include programs with local schools to promote health and safety. Promote walking and biking to work based on greenways within a hospital's patient population. Partner with a mobile grocery organization to target patient populations with limit access to healthy food.



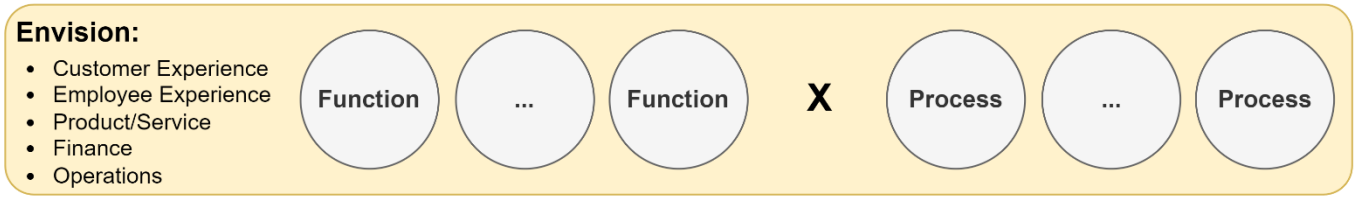
Dependencies: EHR Data Warehouse in Fabric
The project to load EHR data into Fabric is needed so we can align based on relationships for analysis context. Relational context should be at the lowest granularity as possible.

Resources: Data Engineer, Visualization Developer

Stakeholders:

Microsoft Fabric Roadmap Template

Understand the motivations, challenges, and experiences of your employees, customers, partners, suppliers, and competitors.



Prioritize and build prototypes to quickly experiment and develop internal standards and best practices.

