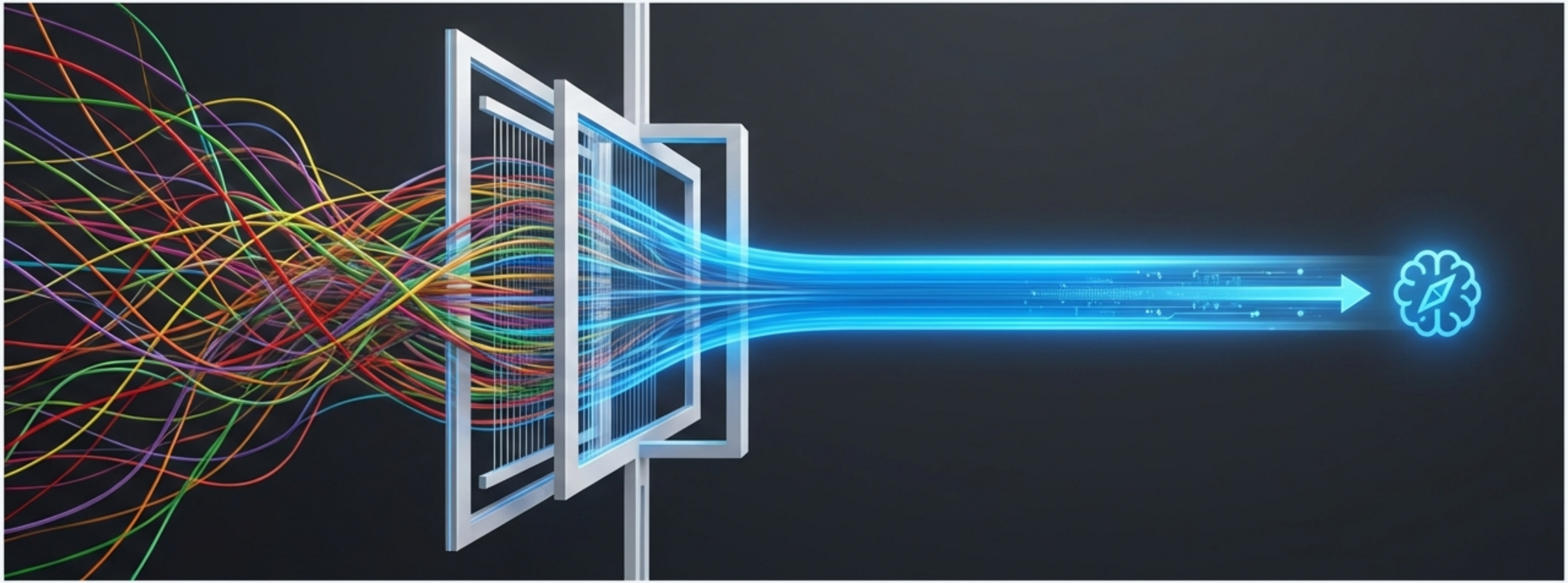
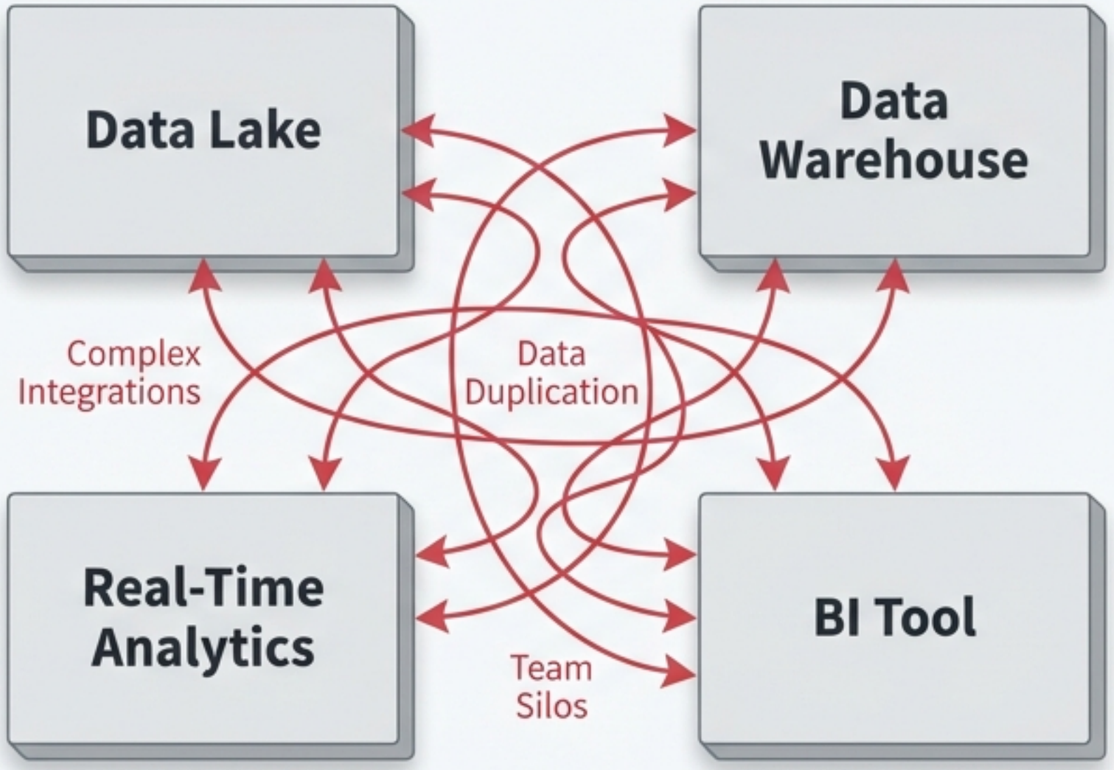
Microsoft Fabric: An Architectural Journey

From Raw Data to Actionable Insight



The Fragmentation Challenge

Today’s data platforms are often a collection of disconnected silos, creating complexity, inefficiency, and hindering collaboration.



- **Data Duplication**: Multiple copies of data across systems lead to inconsistency and ballooning storage costs.
- **Integration Complexity**: Stitching together disparate tools with custom code is fragile, time-consuming, and difficult to maintain.
- **Siloed Personas**: Engineers, scientists, and analysts work in separate environments, preventing seamless collaboration and creating friction.

One Platform to Unify Them All

Fabric is an integrated SaaS platform that brings together all data experiences—from engineering to business intelligence—into a single environment built on a shared foundation.

One Product: A single, integrated SaaS experience.

One Lake: A single, logical storage space eliminates data silos.

One Copy: Different compute engines (Spark, SQL, Power BI) work on the same copy of data.

One Security Model: Unified governance and security across all experiences.

Building on a Foundation of Openness and Reliability

Fabric’s power begins with its storage layer, which provides a single source of truth using an open format with enterprise-grade reliability features.

OneLake: The “OneDrive for Data.”

A single, logical, and coherent storage location for the entire organization, accessible from all tools.

Delta-Parquet: The Gold Standard.

It combines the high performance of Parquet’s compressed columnar format with the transactional reliability of Delta Lake.

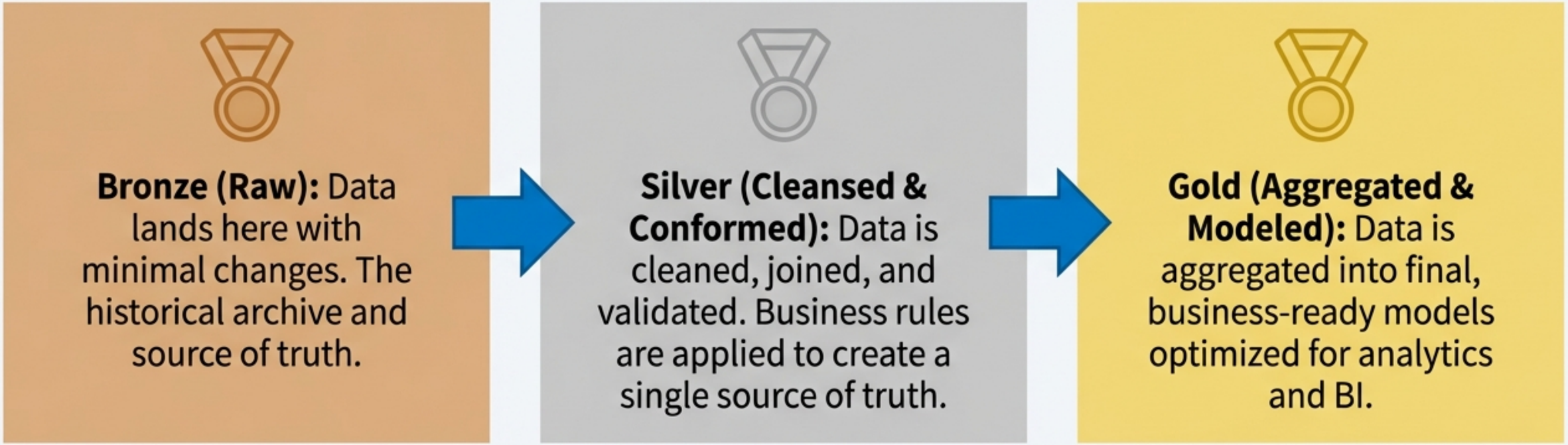
ACID Transactions
Guarantees consistency (Atomicity, Consistency, Isolation, Durability), preventing data corruption from concurrent write operations.

Time Travel
Query, audit, and roll back to previous versions of the data, as all changes are captured in the Delta log.

Schema Enforcement & Evolution
Protects data quality by rejecting invalid data, while allowing the schema to evolve by adding new columns as needed.

A Journey of Progressive Refinement

Key idea: The Medallion Architecture is the standard pattern for transforming raw data into high-quality, business-ready assets within a Fabric Lakehouse.



This architecture ensures data is reliable, easy to understand, and performant at every stage, combining the flexibility of a data lake with the power of a data warehouse.

Stage 1 (Bronze): Orchestrating Data Ingestion

Key Idea: Data Factory Pipelines are the orchestrators that define the workflow for moving data from various source systems into the Bronze layer of the Lakehouse.

Pipeline Responsibilities

- **Orchestration:** Defines the *what, when, and in what order* of data movement and transformation activities. Contains activities like Dataflows, notebooks, and SQL scripts.
- **Data Movement:** Uses 100+ native connectors to copy data efficiently from sources.
- **Scheduling & Triggering**: Automates workflows to run based on a schedule or in response to specific events.
- **Parameterization**: Creates dynamic, reusable workflows. (e.g., A parameter 'FolderName' can direct a pipeline to process data for `Clientes/2024-01-15` or `Clientes/2024-01-16` without duplicating the pipeline logic).

Stage 2 (Silver): Refining Raw Data into Business-Ready Assets

Key Idea: Fabric provides powerful and distinct tools for every skill set to clean, shape, and enrich data, moving it from the Bronze to the Silver layer.

Dataflow Gen2 (Low-Code Transformation)

Visually transform data with the familiar Power Query Online experience. Encapsulates reusable ETL logic with over 300 transformations without code.

Ideal for: Analysts and citizen developers.

Keywords: Visual UI, Reusability, Joins/Merge, Filtering, Calculated Columns.

Spark Notebooks (Code-First Transformation)

Leverage the full power of Apache Spark with Python, Scala, or SQL for complex, large-scale transformations. Built for data engineers and scientists.

Ideal for: Engineers and Data Scientists.

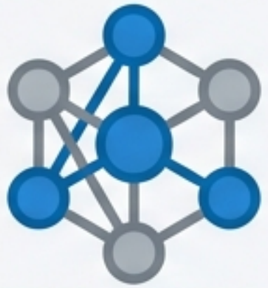
Keywords: Distributed Computing, DataFrame API, Custom Libraries, Scalability.

High-Performance, Managed Spark

Fabric manages the complexity of Spark clusters, allowing you to focus on logic, not infrastructure, while optimizing for cost and performance.

Dynamic Scaling

Pools automatically scale nodes up or down based on workload demand (Autoscale). This optimizes cost, especially for variable workloads.



Spark Pool

Custom Environments

Define a consistent runtime by pre-loading custom libraries (Python packages like pandas, JARs like JDBC drivers) for all notebook sessions.

Efficient Code Patterns

```
# Best Practice: Filter and Select early to reduce I/O and memory usage
(df.filter(df.year == 2024) \
    .select("product_id", "revenue") \
    .groupBy("product_id") \
    .agg(sum("revenue").alias("total_revenue")))
```

Stage 3 (Gold): Delivering Actionable Insights

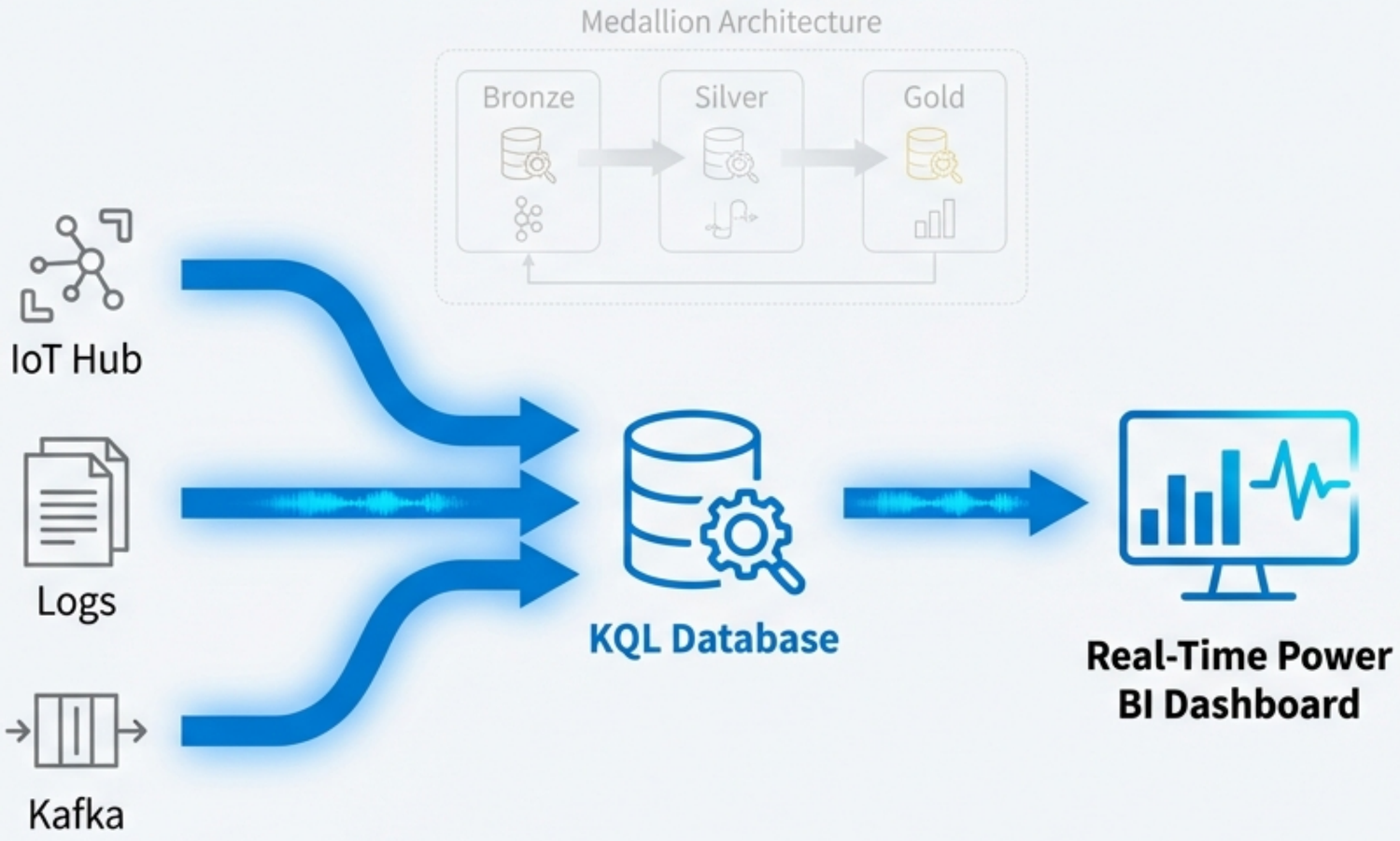
The Gold layer contains business-level models optimized for direct consumption by Power BI and other analytics tools, eliminating data movement and import delays.

Key Concepts

- **Optimized Models:** Contains aggregated data, star schemas, and business-specific KPIs ready for reporting.
- **SQL Endpoint:** Every Lakehouse is automatically queryable via a standard T-SQL endpoint, enabling connectivity from any SQL-compatible tool.
- **DirectLake Mode:** A revolutionary feature. Power BI directly queries the Delta-Parquet files in OneLake with warehouse-level performance. This eliminates the need to import or duplicate data into a separate Power BI dataset.

Beyond Batch: The Real-Time Stream

For high-volume, high-velocity data like logs, IoT, and telemetry, Fabric provides a dedicated, high-performance real-time analytics path powered by Kusto Query Language (KQL).



Core Strengths

- **Kusto Query Language (KQL):** A powerful and intuitive language optimized for time-series, text, and log data, capable of querying terabytes in seconds.
- **Near Real-Time Ingestion:** Ingest and query data with seconds of latency using continuous connectors for sources like Kafka or Kinesis.
- **High-Performance Dashboards:** Power real-time operational monitoring and interactive dashboards without performance degradation.

Optimizing Real-Time Queries for Speed and Reusability

KQL provides advanced features to make real-time analytics more efficient, maintainable, and incredibly fast for recurring queries.

Stored Functions

Encapsulate complex KQL logic into a function that can be reused across multiple queries. This ensures consistency and simplifies maintenance.

Materialized Views

Pre-calculate and store aggregated results from a source table. Delivers millisecond query performance for dashboards by querying the much smaller, pre-computed data. The view is incrementally updated as new data arrives.

Key Takeaway: Use Materialized Views for recurring, high-performance dashboard needs; reserve direct queries against the full table for ad-hoc exploration.

A Unified Governance and Operations Framework

Fabric's integrated design provides centralized control, security, and visibility across the entire data lifecycle.

Access Control

Workspaces provide granular security boundaries. Best practice is to use separate workspaces for Bronze, Silver, and Gold layers to control access by team and responsibility.

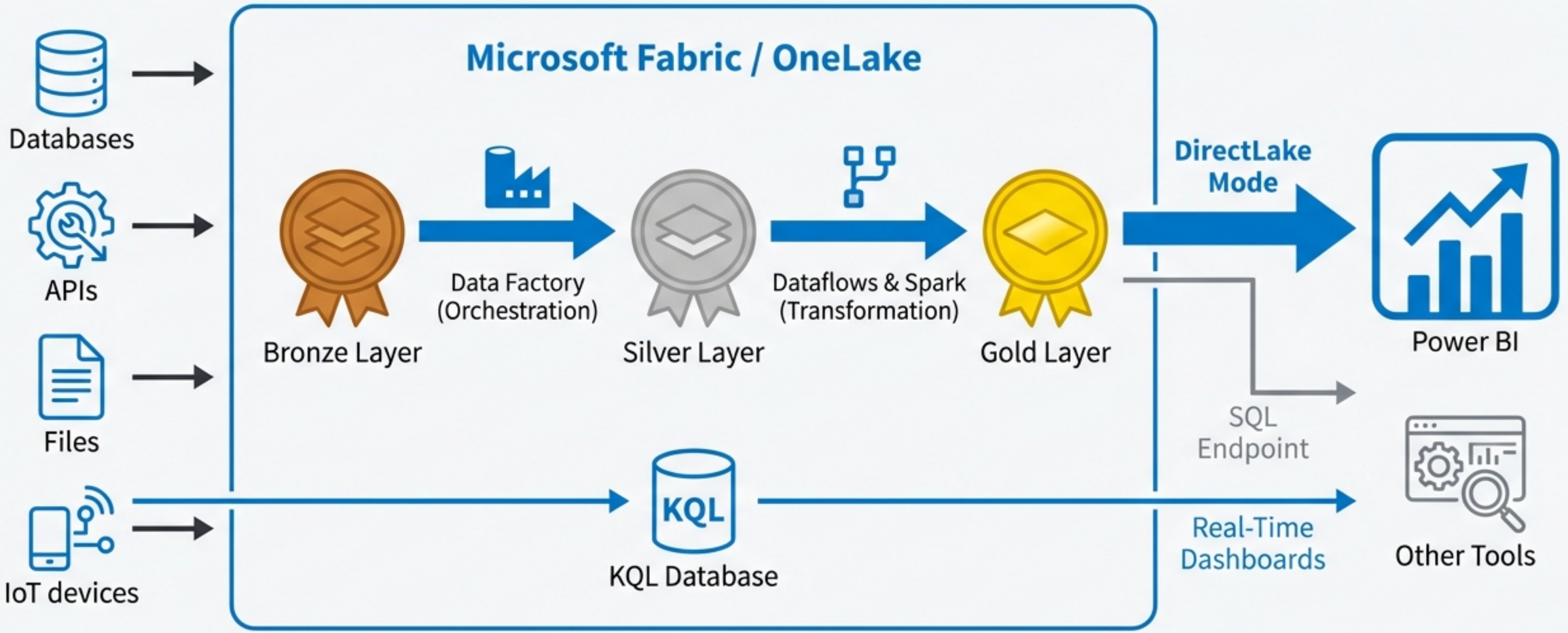
Data Management

Distinguish between **Managed Tables** (data files deleted when table is dropped) and **External Tables** (only metadata is dropped; data files remain). This enables flexible data sharing patterns.

Monitoring

The centralized **Monitoring Hub** provides detailed run history, status, duration, and error messages for all pipelines, dataflows, and other activities, critical for debugging and meeting SLAs.

From Chaos to Clarity: The Complete Fabric Architecture



A single, unified platform for every data need—from batch to real-time, from engineer to business analyst.

The Core Principles of a Unified Data Platform

OneLake + Delta Lake.

Eliminates data duplication and provides a single, reliable source of truth with enterprise-grade features like ACID transactions and time travel.

Choose the Right Tool for the Job.

Seamlessly use Spark, T-SQL, Power Query, and KQL on the same copy of data without moving or duplicating it.

Break Down Silos.

A single SaaS platform where engineers, analysts, and data scientists can collaborate effectively in one shared environment.

One Fabric. One Lake. One Copy.