



DATABRICKS

A brief scroll — Think less “treatise,” more “tavern chat with a diagram.”

ABSTRACT

A pragmatic overview of the Databricks architecture, illustrating how it unifies business intelligence (BI) and machine learning (ML) on a single, scalable platform. Grounded in real-world use, it outlines a layer-by-layer pipeline—from raw ERP exports to curated insights—framed through the Medallion architecture (Bronze, Silver, Gold). Key technologies like Delta Lake, MLflow, and Snowflake integration are discussed, with a narrative that bridges data engineering and business value. This guide is designed to support practitioners seeking clarity, accountability, and delivery confidence in modern data platforms.

Iain Hamilton Toolin

The intellectual property for the Databricks platform and its architectural framework is owned by Databricks, Inc., the creators of the Unified Analytics Platform and key contributors to Apache Spark. The narrative and illustrative interpretations presented here are the result of independent research by Iain Toolin, developed to support learning and engagement through accessible, context-rich storytelling. This work is not affiliated with or officially endorsed by Databricks, Inc.

1.2 Architecture of Pipeline (Databricks Blackbox)

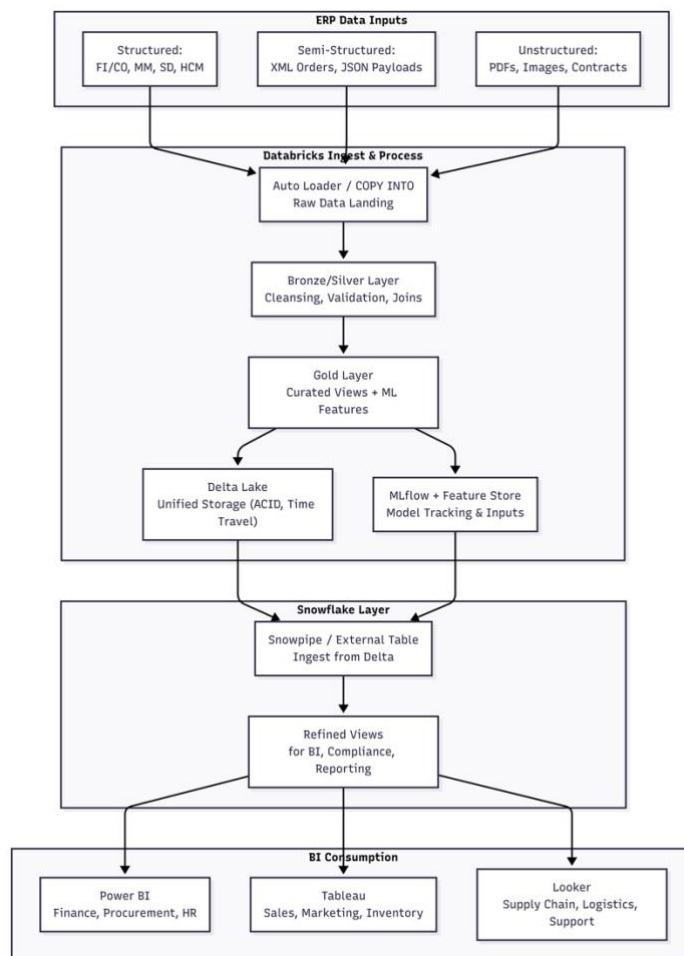





Figure 1: Databricks pipeline architecture

1.3 Narrative: Layer-by-Layer Breakdown

 Layer	 What It Is	 Pragmatic narrative
ERP Data Inputs	Real-time or batch exports from your ERP modules—SAP, Oracle, Dynamics, etc.	“This is the mess at the edge—clean, messy, and wild data all showing up at once. XML from procurement, structured GL entries, scanned contracts—it all starts here.”
Databricks Ingestion	Handles all formats via Auto Loader or COPY INTO, dropping them into Bronze Layer.	“Think of Bronze as the raw archive. No tampering. You land it, you tag it, and you step away. Good for audits, replays, and future regret fixes.”
Bronze/Silver Layer	Apply validation rules, deduping, schema alignment, joins with lookup data.	“Silver is where you stop apologising for your data. Business logic starts to take root. If you're filtering duplicates or stitching org hierarchies, it's here.”

The intellectual property for the Databricks platform and its architectural framework is owned by Databricks, Inc., the creators of the Unified Analytics Platform and key contributors to Apache Spark. The narrative and illustrative interpretations presented here are the result of independent research by Iain Toolin, developed to support learning and engagement through accessible, context-rich storytelling. This work is not affiliated with or officially endorsed by Databricks, Inc

Layer	What It Is	Pragmatic narrative
Gold Layer + MLflow	Curated business views and/or feature sets ready for ML or dashboards.	“Gold is what the business actually sees. This is where KPIs live, forecasts start, and you’ll be held accountable for numbers.”
Delta Lake	ACID-compliant storage format with versioning, schema enforcement, and rollback.	“This is the version control system your data never had. Without it, change means chaos. With it, you can roll back, branch, or ‘git diff’ your ledger.”
Snowpipe & Views	Push clean outputs to Snowflake—either streaming or batched.	“Hand-off to Snowflake keeps your finance team happy—clean, governed, and performant. Great for compliance and standardised reports.”
BI Tools	Serve the right dashboards to the right audiences.	“Power BI is for ledger-heads, Tableau is for the storytellers, and Looker’s for the process wonks. Pick your tool based on what story the data needs to tell.”

Table 1: Narrative on architectural layers

1.4 Architecture of Databricks (Databricks White Box)

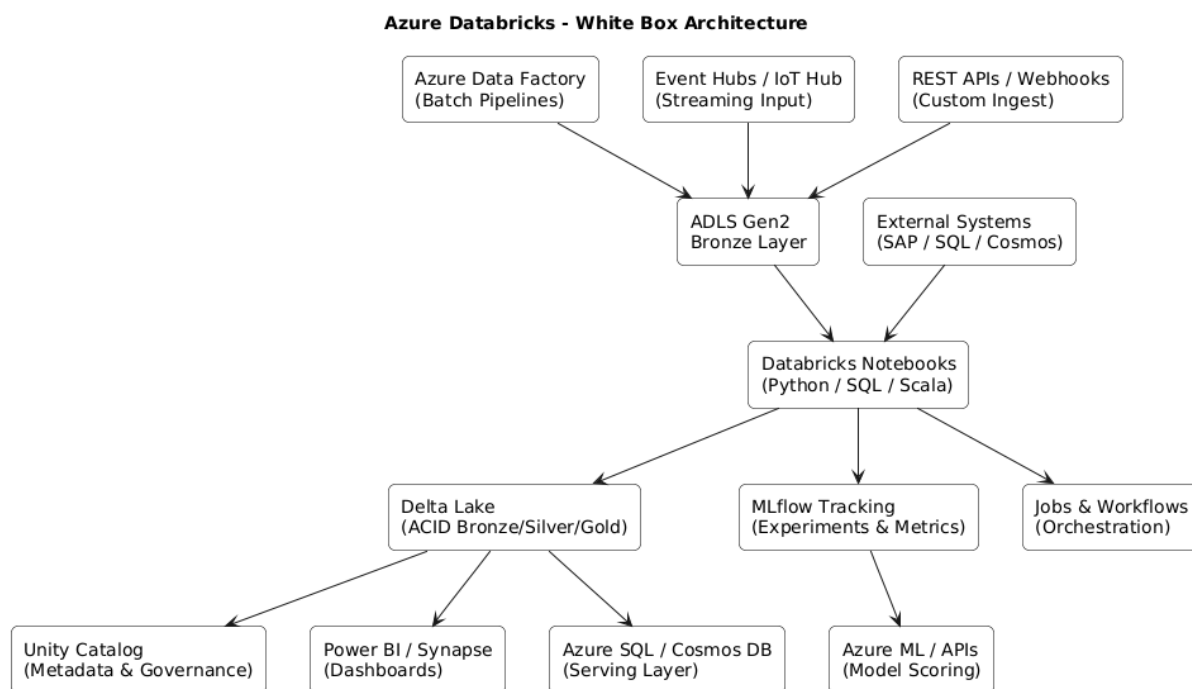


Figure 2: Databricks internal services

The intellectual property for the Databricks platform and its architectural framework is owned by Databricks, Inc., the creators of the Unified Analytics Platform and key contributors to Apache Spark. The narrative and illustrative interpretations presented here are the result of independent research by Iain Toolin, developed to support learning and engagement through accessible, context-rich storytelling. This work is not affiliated with or officially endorsed by Databricks, Inc

1.5 Narrative Table – Azure Databricks Services & Roles



Layer / Component	 What It Is	 Pragmatic Narrative
Azure Data Factory	Batch ingestion service	“Moves structured data like invoices or GL entries from your ERP into your data lake—timed and dependable.”
Event Hubs / IoT Hub	Streaming pipeline inputs	“Handles live stuff: meters, devices, and sensors. Streams it in like telemetry on autopilot.”
ADLS Gen2	Primary storage layer	“Everything lands here first. It’s your Bronze vault—raw but retrievable.”
Databricks Notebooks	Interactive dev & analysis	“Where engineers sketch pipelines, analysts try joins, and everyone shares debug hacks.”
Delta Lake	ACID storage format	“Your structured layer cake: Bronze (raw), Silver (clean), Gold (insight). Plus versioning and rollback.”
MLflow Tracking	Model versioning and metrics	“Where ML experiments go to be tracked—not lost on someone’s laptop.”
Jobs & Workflows	Scheduled or triggered tasks	“This is your automation muscle. When data moves itself, this is who’s pushing.”
Unity Catalog	Metadata and lineage governance	“Gives you the audit trail. Who touched what, and where the data’s been.”
Power BI / Synapse	Dashboarding and insight delivery	“Your final product for stakeholders. It’s not real until someone sees a chart.”
Azure SQL / Cosmos DB	Externalised data store	“Use when your clean data needs to be shared with apps or APIs in real time.”
Azure ML / API Layer	Scored predictions & endpoints	“Where your model gets deployed, called, and earns its keep in the real world.”

Table 2: Narrative on services and roles.

The intellectual property for the Databricks platform and its architectural framework is owned by Databricks, Inc., the creators of the Unified Analytics Platform and key contributors to Apache Spark. The narrative and illustrative interpretations presented here are the result of independent research by Iain Toolin, developed to support learning and engagement through accessible, context-rich storytelling. This work is not affiliated with or officially endorsed by Databricks, Inc