



ON, TORONTO, WATERPARK PLACE

88 Queens Quay West, Suite 2500, Toronto, M5j 0B8

T: 506-496-3964

Email: info@datakrypton.ai

MEDALLION ARCHITECTURE ON AZURE: RECOMMENDED TECH STACK BY BUSINESS GOALS





INTRODUCTION

In today's data-driven world, organizations need robust and flexible data platforms that can scale with their business goals. Whether you are operating in highly regulated industries, driving innovation through artificial intelligence (AI), enabling business teams to self-serve analytics, or delivering real-time operational insights, the way you architect your data matters.

Medallion Architecture on Azure offers a structured, layered approach to managing data—from raw collection to trusted business insightstailored to meet diverse business needs. This guide breaks down how you can implement this architecture aligned with your priorities, helping you understand the components, their purpose, and how they come together to deliver value.

1. COMPLIANCE-HEAVY ENVIRONMENTS

1. **Controlled Data Collection:**

Bring in data from multiple sources, including transactional systems, devices, and external feeds, using tools that provide visibility and control over what data is ingested and when. Schedule batch data transfers to reduce errors and enable auditing, while supporting streaming for near real-time needs where appropriate.

2. **Organize Data in Clear Layers:**

Establish a Bronze layer where raw data lands exactly as it comes, untouched and immutable. This preserves an audit trail for regulatory review. Develop a Silver layer where data is cleaned, validated, and enriched, preparing it for business use. Build a Gold layer of curated, high-quality data approved for reporting and decision-making.

3. **Ensure Secure and Compliant Processing:**

Use environments hardened to meet compliance standards (such as HIPAA or PCI) where data is processed and transformed. Enforce strict security policies, including encryption, network isolation, and access controls, to prevent unauthorized data access during processing.



4. **Orchestrate with Transparency:**

Manage data workflows with centralized tools that log every action—when data moves, who triggered jobs, and whether any errors occurred. Provide audit-ready reports showing compliance with policies and procedures.

5. **Embed Data Quality Controls:**

Build automated checks to identify missing, inconsistent, or out-of-range data before it impacts reports. Quarantine problematic data for investigation without blocking the entire pipeline.

6. **Centralize Metadata and Data Cataloging:**

Maintain a business-friendly data catalog describing what each dataset contains, its sensitivity, and its owner. Enable users to search and discover approved data assets easily, reducing risk of unauthorized or incorrect data use.

7. **Track Data Lineage End-to-End:**

Show how data flows from origin through every transformation and report, so auditors and business users can trust the data's journey. Use lineage information to quickly investigate and resolve data issues or compliance questions.

8. Enforce Layered Access Controls:

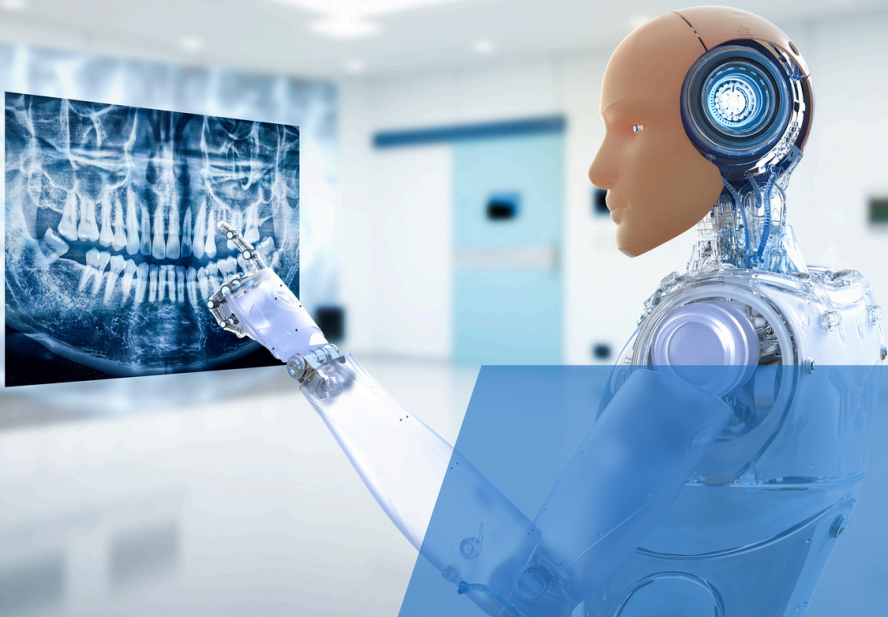
Control access at every stage, from raw files to reports, ensuring users see only data they're authorized to view. Implement role-based permissions with detailed logging of data access.

9. Deliver Governed Business Insights:

Provide business intelligence tools connected only to certified data, ensuring reports are accurate, compliant, and consistent. Apply security measures such as row-level security to protect sensitive segments of data even within reports.



2. AI AND MACHINE LEARNING READY USE CASES



Artificial intelligence and machine learning offer transformative potential but require a solid foundation of reliable, well-governed data. Data scientists need ready access to comprehensive, high-quality datasets, while organizations must maintain control and transparency throughout the ML lifecycle.

1. Ingest Diverse Data Sources Flexibly:

Capture structured data like sales and finance records alongside unstructured or semi-structured data such as logs, images, or social media feeds. Support batch and streaming ingestion to feed training and operational models with fresh, relevant data.

2. **Structure Data for Machine Learning:**

Maintain raw data in the Bronze layer for reproducibility and auditing. Clean, join, and transform data in the Silver layer to create consistent datasets. Curate feature-rich tables in the Gold layer designed specifically to feed ML models efficiently.

3. **Foster Collaboration Between Teams:**

Provide data engineers and data scientists a shared, secure environment to develop feature pipelines and models collaboratively. Leverage tools that track experiments, parameters, and results automatically to ensure reproducibility and governance.



4. **Automate ML Workflows:**

Use orchestration to chain data preparation, model training, validation, and deployment steps seamlessly. Enable rapid retraining triggered by new data or performance drift.



5. **Ensure Data Quality for Models:**

Validate data distributions, detect anomalies, and monitor for data drift to protect model accuracy. Prevent poor data from reaching training or inference stages through automated quality gates.

6. **Maintain a Comprehensive Metadata Repository:**

Document each dataset and feature with clear descriptions, lineage, and ownership to facilitate reuse and transparency. Register models and track their provenance alongside data assets.

7. **Provide Full Lineage Visibility:**

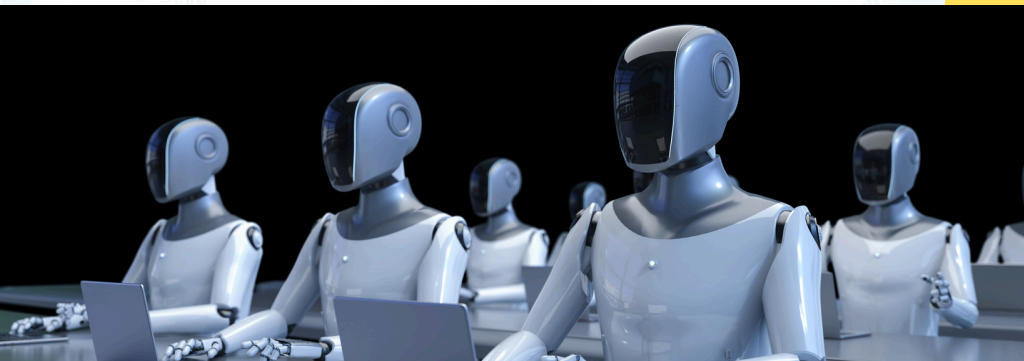
Trace every ML outcome back through the features and raw data that produced it. Empower auditing and troubleshooting across the entire AI pipeline.

8. **Balance Access and Security:**

Allow data scientists the freedom to experiment while restricting sensitive data exposure. Manage permissions at granular levels, including masking or tokenizing personally identifiable information.

9. **Present AI Insights Clearly:**

Use business-friendly dashboards that translate model outputs into actionable metrics and categories. Ensure security and privacy controls remain enforced through reporting layers.



3. SELF-SERVICE BUSINESS INTELLIGENCE

Empowering business users to generate insights independently accelerates decision-making and innovation. However, this requires a well-governed data environment that ensures consistent, trusted data is used while enabling flexibility

1. **Support Both Centralized and Self-Service Data Ingestion:**
IT manages ingestion of core enterprise data for broad use. Business teams can bring in their own datasets through user-friendly, governed tools, with appropriate oversight.



2. **Organize Data for Easy Discovery and Use:**
Maintain clear data layers (Bronze, Silver, Gold) that users can navigate confidently. Store data in accessible formats optimized for fast querying and analysis.

3. Prepare Data with Business Context:

Data engineers create refined tables with business-friendly entities and consistent definitions. Enable reusability by standardizing dimensions like customers or products.

4. Schedule Regular Updates with User Flexibility:

Automate nightly or hourly data refreshes to keep information current. Allow controlled on-demand refreshes for business users who need immediate updates.

5. Embed Data Quality and Certification:

Apply rules and validations to catch errors early. Mark datasets as certified to build trust and encourage use of approved data.

6. Centralize Metadata and Business Glossaries:

Provide clear definitions and ownership information so users understand data context. Link data assets to business terms to reduce misunderstandings.

7. Capture and Share Data Lineage:

Show how data flows from source systems to reports, helping users verify data accuracy. Facilitate impact analysis before changes are made.

8. Apply Appropriate Access Controls:

Use role- and row-level security to protect sensitive data while maximizing accessibility. Audit data usage to monitor compliance and usage patterns.

9. Deliver Rich, User-Friendly Reporting Tools:

Build semantic models with common business measures to ensure consistency. Enable users to focus on analysis, not data preparation.

4. REAL-TIME ANALYTICS

The ability to react instantly to business events is a competitive advantage. Real-time analytics enable operational visibility, personalized customer experiences, and rapid response to changing conditions.

1. Continuously Capture Event Data:

Ingest data from devices, applications, and transactions as it happens, supporting large volumes and high speed. Handle a variety of event types such as logs, telemetry, clicks, or inventory changes.



2. Store Streaming Data Reliably:

Use layered storage that supports quick writes and reads while preserving full history for auditing. Optionally, maintain a hot store for ultra-fast querying of recent data.



3. **Process and Refine Data on the Fly:**

Transform streaming data to clean, enrich, and aggregate it continuously. Maintain separate layers that progressively increase data quality and business value.

4. **Monitor and Manage Streaming Pipelines:**

Keep jobs running 24/7 with real-time alerts on failures or backlogs. Schedule maintenance activities such as file compactions without disrupting operations.

5. **Maintain Data Quality in Real Time:**

Validate data schemas and filter out bad records as they arrive. Handle schema changes and duplicates proactively.

6. **Keep Metadata and Catalogs Updated:**

Track schema changes and data freshness to inform users about data reliability. Make real-time datasets discoverable alongside batch data.

7. **Provide Complete Data Lineage:**

Trace the origin and transformations of streaming data in near real time. Support audit and troubleshooting even as data flows rapidly.

8. **Enforce Strong Security at All Layers:**

Control access to streaming sources and storage with strict authentication and network protections. Mask or restrict sensitive data within streams as needed.

9. **Deliver Live Dashboards and Alerts:**

Present continuously updating views of key metrics to business users. Use intuitive visualization tools designed for rapid refresh and interactivity.



CONCLUSION

A Medallion Architecture built on Azure offers a clear, scalable framework for managing data that aligns with a wide range of business priorities. By thoughtfully layering data storage and processing, embedding governance and quality at every step, and enabling appropriate access and reporting, organizations can build trust in their data assets and accelerate decision-making. Whether your focus is compliance, AI innovation, user empowerment, or real-time responsiveness, this architecture provides a roadmap to success.

