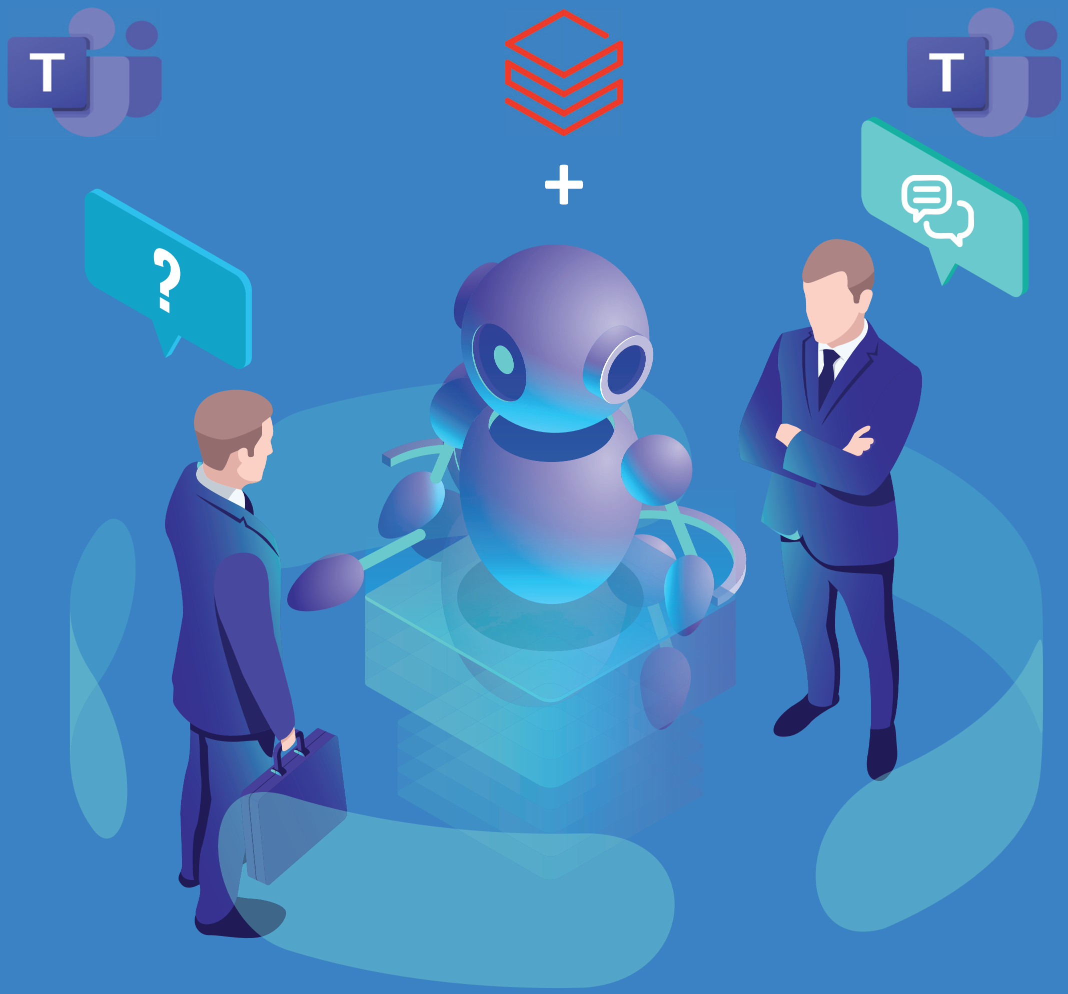
Computomic

Accelerating Product Intelligence with Generative AI



Abstract

This white paper presents a successful implementation of a Retrieval-Augmented Generation (RAG) application, developed by Computomic in collaboration with a leading Manufacturing enterprise. The project demonstrates how the Databricks Data Intelligence Platform was used to transform unstructured product knowledge into a real-time, AI-powered assistant integrated within Microsoft Teams. The goal was to solve a critical sales enablement issue, enabling frontline employees to access accurate product information across thousands of permutations and formats instantly.

Introduction

In today's fast-paced enterprise landscape, having real-time access to accurate and contextual product information is essential for business success. Our client faced challenges in product recommendation due to the fragmented nature of their product documentation scattered across hundreds of manuals, specification sheets, and tribal knowledge.

To address this, Computomic built an enterprise ready GenAI solution powered by Databricks, Azure, and Microsoft Teams, leveraging Retrieval Augmented Generation (RAG) architecture to enable a contextual, searchable interface for complex product data.

Business Challenges

The customer sought to:

- **Evaluate Databricks** as the core data platform for their LLM and AI readiness
- Handle **heterogeneous data** sources like SAP, Salesforce, IoT, and document
- Identify **high impact use cases** to validate their commitment to Databricks as their enterprise-wide Data & AI platform

The **selected use case** focused on improving sales productivity:

- The client has 160,000+ product SKUs with thousands of configuration permutations, making it hard to search and find the right product
- Salespeople were unable to confidently recommend the right products
- No centralized repository: knowledge was scattered and inaccessible
- Sales managers were manually fielding queries, slowing response times

Solution Overview

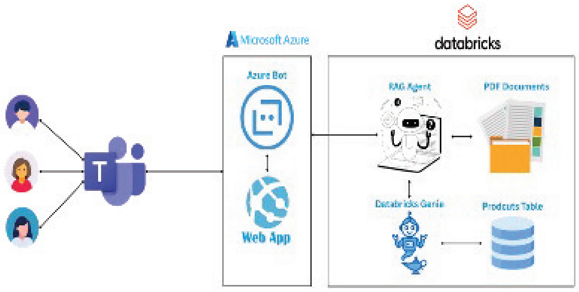
Computomic designed and implemented an AI-powered Teams chatbot with RAG capabilities. The solution integrated Databricks with Azure services to create a secure, scalable, and intelligent assistant capable of:

- Summarizing documentation
- Performing vector similarity search
- Generating contextual answers from structured and unstructured data

Architecture & Technical Components

High-Level Architecture

Data Sources → Processing → RAG System → Genie Space → Teams Bot



Data Sources:

- Structured: Product tables
- Unstructured: PDF manuals, specification documents

Key Technologies:

- Databricks: RAG pipeline, Genie Space, vector search
- Azure Bot Service: Teams integration
- Azure Web App: Bot hosting
- OAuth: Secure authentication & RBAC
- Custom Serving Endpoints: Real-time API response handling

The Technical Challenge

Our enterprise client had 10,000+ PDFs and structured SAP HANA data, but their existing search could only return products by name and specifications. Sales teams needed contextual intelligence: "What product that we have will be well suited for a place like Florida (Humid Conditions)? instead of generic product lists.

The Databricks-Powered Solution

Core Architecture Components

Databricks GenAI Serving Endpoint

- Custom RAG pipeline deployed as a **managed serving endpoint**
- Real-time inference with **sub-30-second** response times
- Automatic scaling based on query volume

Databricks App Connector Integration

- **OAuth2** authentication with user-level access control
- Seamless token management for enterprise security
- Fine-grained permissions tied to organizational roles using **Unity Catalog**

Hybrid Data Processing

- **Structured Path:** SAP HANA,SQL Server →Databricks Delta Table → Databricks **Genie Space**
- **Unstructured Path:** PDF ingestion → Databricks Volume → Vector embeddings → Vector Search
- Real-time data fusion for comprehensive recommendations

Technical Implementation Deep-Dive

RAG Pipeline Architecture

```
Query → Intent Classification → Dual Retrieval:
   ├──── Structured: Genie Space SQL queries
   └──── Unstructured: Vector similarity search
   ↓
Context Synthesis → LLM Generation → Response
```

- **Data Preparation:** PDFs ingested and summarized using NLP
- **Vector Embedding:** Semantic search setup for product descriptions
- **Dynamic Prompt Engineering:** Custom instructions based on user input
- **Context Management:** Enables deep drill-down and multi-turn conversations

Genie Space Customization

- Added metadata: Column-level descriptions, SQL examples
- Optimized cluster configuration
- Enabled user-specific instruction sets

Authentication & Security

- OAuth 2.0 token-based authentication
- Role-Based Access Control (RBAC) for data governance
- API token refresh and session management

Microsoft Teams Integration

- Built as a native Teams App via Azure Bot Framework
- Supports interactive cards, dynamic responses
- Hosted on Azure Web App with telemetry for observability

Key Databricks Features Leveraged

- **Vector Search:** Embedded PDF content with semantic similarity matching
- **Genie Space:** SQL interface for structured product data queries
- **Model Serving:** Custom endpoint with enterprise-grade SLA
- **Unity Catalog:** Centralized governance for data and model access
- **App Connector:** Secure authentication without credential management

Data Flow Example

1. User asks: "What product is best for a humid climate?"
2. Databricks processes query through dual retrieval:
   - Genie Space: Queries environmental ratings and specifications
   - Vector Search: Finds relevant PDF sections about humidity protection.

Each query is issued once; results are merged and polished by an LLM so the user gets a single, clear answer in Teams. This directly replaces the “email the tech team and wait” flow with an always-on, self-serve assistant that speaks the same language as clients' product experts and cites the same authoritative sources.

It is designed to be quick, consistent, and defensible: structured data provides precision, while PDF retrieval brings nuance from official documentation. Identity comes from Teams; access is enforced by Unity Catalog, ensuring every answer is generated within the user’s permissions.

Performance Metrics

Databricks-Specific Wins

- **Query Latency:** 30 seconds average (vs. 45 minutes manual search)
- **Serving Endpoint Uptime:** 99.9% availability
- **Vector Search Accuracy:** 92% relevance score on contextual queries
- **Cost Optimization:** Auto-scaling reduced compute costs by 40%

Technical Challenges & Solutions

Challenge: Combining structured and unstructured data without latency penalties
Solution: Parallel retrieval with Databricks Vector Search + Genie Space, results merged before LLM generation

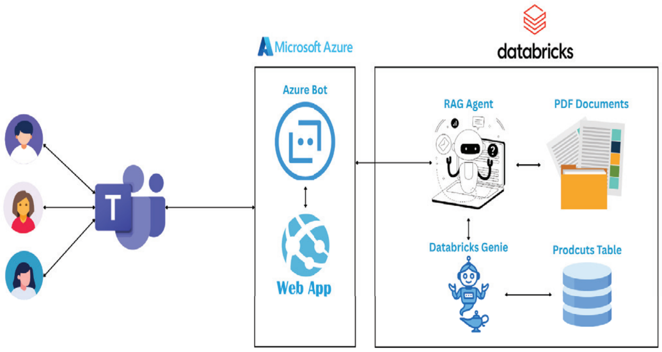
Challenge: Enterprise security with external Teams integration
Solution: App Connector handles OAuth2 flows, eliminating credential storage

Challenge: Model performance tuning for domain-specific recommendations
Solution: Fine-tuned embeddings on product documentation, custom prompt engineering for contextual reasoning

Deployment Architecture

Teams Integration Flow

Microsoft Teams → Azure Bot Framework → Azure Web App → Databricks Serving Endpoint → [Genie Space + Vector Search] → Synthesized Response



Security Stack

- Microsoft Entra ID authentication
- Databricks App Connector for secure API access
- Unity Catalog for data governance
- User-level permissions enforced at query time

Production Results

- **User Adoption:** 92% positive satisfaction during pressure testing
- **Performance:** Sub-30-second contextual recommendations
- **Security:** Zero security incidents, full audit trail
- **Business Impact:** Sales teams focus on selling, not searching

Key Learnings for Databricks Implementation

- **Start with Vector Search early:** Embedding strategy impacts retrieval quality more than model choice
- **Leverage Genie Space:** SQL interface dramatically simplifies structured data integration
- **App Connector is a game-changer:** Eliminates authentication complexity for external apps
- **Model Serving auto-scaling:** Essential for unpredictable enterprise query patterns
- **Unity Catalog governance:** Critical for enterprise compliance and audit requirements

What's Next

Currently preparing for full production rollout with enhanced features:

- Multi-language support via Databricks Model Serving
- Real-time feedback learning through Delta Lake
- Advanced analytics on query patterns via Databricks SQL

The future of enterprise knowledge isn't better search—it's conversational AI that understands context, powered by platforms like Databricks that make complex AI architectures enterprise-ready.

Results & Business Value

- **Faster response times** for complex product queries
- **Increased confidence** among sales staff in product recommendations
- **Eliminated bottlenecks** by reducing dependency on sales managers
- Demonstrated **AI-readiness** of Databricks to leadership
- Enabled **secure, scalable, real-time** product intelligence

Model Evaluation: RAG Query Accuracy

Metric	Manual	Model 1	Model 2	Model 3
Total Questions	546	546	546	546
Correct	402	166	188	412
Partially Correct	-	19	18	119
Incorrect	-	21	21	13

Model 3 (Databricks RAG setup) outperformed others in both correctness and contextual depth.

Conclusion

The RAG-based chatbot solution proved to be an effective starting point for the client's AI adoption journey. By empowering users with intelligent, real-time access to product information, Computomic was able to demonstrate the tangible value of the Databricks platform.

This success story showcases how Retrieval-Augmented Generation, combined with enterprise-grade security and integration, can unlock value from unstructured data and elevate business productivity.

Computomic

Authors & Contributors

- **Swagatika Rath** - Senior Consultant
- **Genevive Mendonca** - Senior Consultant
- **Benjamin Rafatian** - Data Scientist
- **Mick Szabo** - Project Manager

www.computomic.ai