

Retail Insights Assistant

GenAI + Scalable Data Platform

Conversational analytics architecture designed for 100GB+ retail data with guaranteed correctness and zero hallucination



Architecture Overview

Enterprise Technical Brief

Table of Contents

01 Problem & Goal

Scalable conversational analytics for retail teams

02 High-Level Architecture

End-to-end decoupled layers (100GB+ ready)

03 Data Engineering Layer

Scalable data pipeline design

04 GenAI & Governance

Multi-agent architecture with guardrails

05 Query Flow, Scale & Monitoring

Performance, observability, and production readiness

Problem & Goal



Conversational Analytics at Scale

Retail teams need conversational analytics over large, growing sales datasets. The platform must support both **summaries and ad-hoc questions** without predefined schemas or rigid dashboards.



100GB+ Data Scalability

The solution must scale beyond 100GB+ of sales data without sacrificing correctness or performance.

Traditional BI tools break down at this scale, requiring a **distributed, cloud-native architecture**.



Zero Hallucination Requirement

LLMs must provide analytics **without hallucination**. Every insight must be grounded in actual data with full traceability and audit capability for regulatory compliance.

100GB+

Data Scale

Zero

Hallucination

24/7

Availability



Core Insight

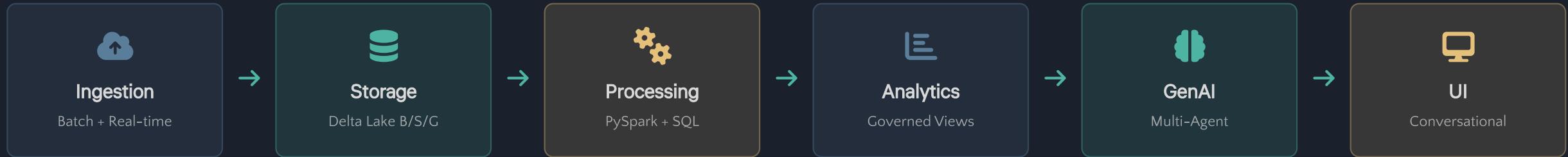
"The problem is not just answering questions, but **answering them correctly at scale**."

Success Criteria

- ✓ Sub-second query responses
- ✓ 100% data accuracy guarantee
- ✓ Self-service analytics capability
- ✓ Enterprise-grade security & governance

High-Level Architecture Overview

End-to-End Architecture (100GB+ Ready)



Ingestion Layer

- Batch Processing
Azure Data Factory + Databricks Jobs
- Real-time Streaming
Azure Event Hub for live orders
- Retail Data Sources
POS, inventory, customer, e-commerce

Processing & Analytics

Processing Engine

PySpark + SQL on Databricks for distributed computing. Handles complex transformations, joins, and aggregations across terabytes of retail data.

Analytics Layer

Databricks SQL / Azure Synapse with pre-aggregated tables for fast query performance. Materialized views for common patterns.

Governance Model

Row-level security, column-level masking, data lineage tracking, and automated quality checks ensure trusted data.

GenAI Layer

- Multi-Agent System
Specialized agents for intent, planning, validation, and narration
- No Raw Data Access
LLMs only access governed analytics views
- Validation First
Results verified before narrative generation

Storage Pattern

Bronze
Unprocessed source data

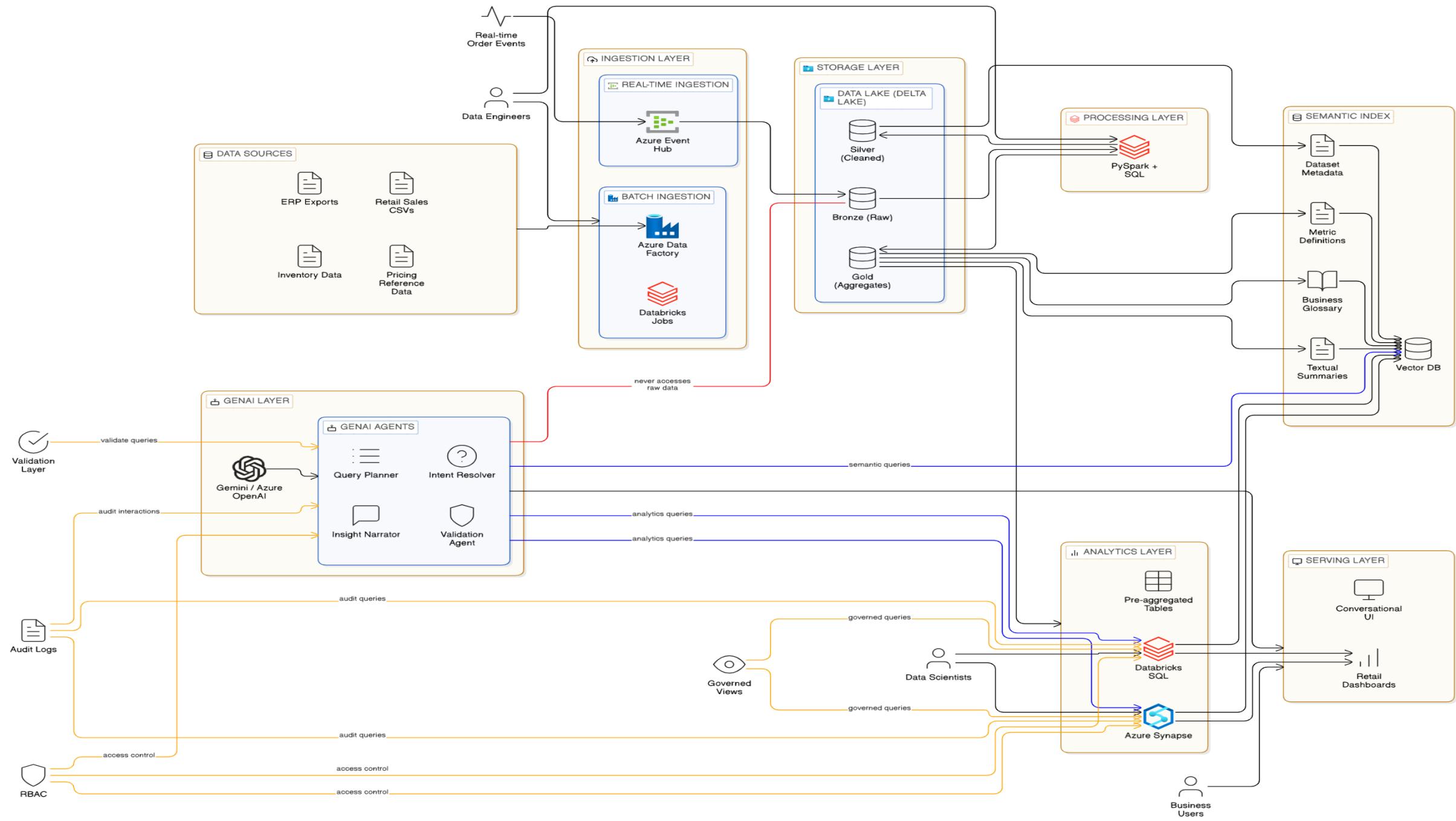
Raw

Silver
Filtered, enriched datasets

Cleaned

UI Layer

- Conversational Interface
Natural language queries and follow-ups
- Rich Visualizations
Auto-generated charts and dashboards
- Multi-Channel
Web, mobile, Teams integration



Scalable Data Engineering Design



Batch Ingestion

Azure Data Factory

Orchestrates data movement from source systems

Databricks Jobs

Scheduled ETL pipelines with Spark



Storage

Azure Data Lake

Cheap, durable cloud storage

Delta Lake

ACID, time travel, indexing



Processing Engine



PySpark

Distributed data processing

SQL

Declarative transformations

Databricks

Unified analytics platform



Analytics Layer

Databricks SQL

Serverless SQL warehouse

Azure Synapse

Enterprise analytics

Design Achievements

- ✓ **Retail Velocity:** Handles batch loads and real-time streams
- ✓ **Performance:** Pre-aggregations + indexing for speed
- ✓ **Cost Efficiency:** Auto-scaling + storage-compute separation

10TB+

Daily Ingestion

<1s

Query Response

99.9%

Availability SLA

50%

Cost Reduction

GenAI with Guardrails (No Hallucinations)

Multi-Agent GenAI

1 Intent Resolver

Analyzes user questions to determine required data, metrics, and time period

2 Query Planner

Generates optimized SQL queries using metadata from Vector DB

3 Validation Agent

Executes queries and validates results against business rules

4 Insight Narrator

Generates natural language insights with data attribution

Agent Flow

Sequential processing with validation at each step. Agents communicate through a shared context, maintaining conversation history and data lineage throughout.

Vector DB (Semantic Index)

Dataset Metadata

Schemas, relationships, data quality metrics

Metric Definitions

Revenue, profit, inventory formulas

Business Glossary

Common terms, synonyms, descriptions

Textual Summaries

Natural language dataset descriptions

Governance & Security

RBAC

Role-based access control

Governed SQL Views

Pre-defined, secured data access paths

Pre-Validation

Results verified before response generation

No-Hallucination Architecture

User Query



Intent Resolver



Query Planner + Vector DB



Validation Agent



Insight Narrator

Key Benefits

🛡️ 100% accuracy via validation

🔗 Full traceability & attribution

🔒 Enterprise security compliance



Governance-First Principle

Governance enforced before generation, not after hallucination — The Validation Agent blocks incorrect or unauthorized queries before LLM narrative generation

Query Flow, Scale & Observability

Query Flow Pipeline



Metadata-Driven: Vector DB retrieves relevant datasets, metrics, and business terms to build context-aware queries

Scaling Mechanisms

- Partitioned Data**
Date/region partitioning
- Pre-Aggregations**
Materialized views
- Caching Layer**
Redis/Databricks cache

Performance Targets

<2s
Simple Queries

<10s
Complex Queries

Observability & Monitoring

Query Latency

P95/P99 response times

LLM Usage

Token consumption, costs

Validation Failures

Accuracy, security alerts

Audit Logs

Query history, access trails

Production Readiness

- Scalability:** Distributed systems for production workloads
- Reliability:** Multi-region deployment with auto-failover
- Security:** End-to-end encryption, VPC isolation

Architecture Validation

"Prototype demonstrates architecture; production follows the same design with distributed systems."