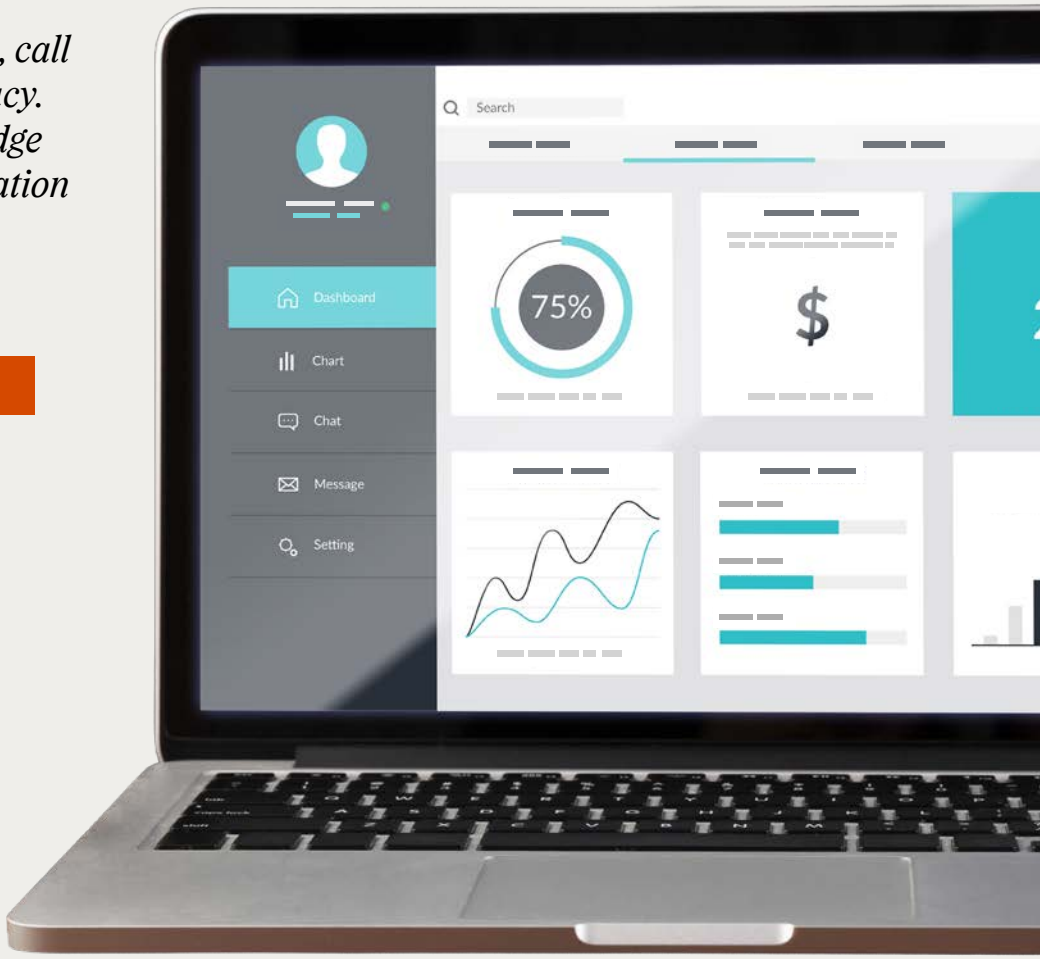
Altimetrik

WHITEPAPER

AI-Powered Service Intelligence for WEALTH MANAGEMENT

Connecting knowledge access, call evaluation and factual accuracy. AI-Powered real-time knowledge and automated claims verification on Snowflake



AUTHOR

Shivkumar Krishnan

Contributing Team Members

Mahesh Varavooru, Amol Brahmankar, JB Johnson, Shravan Gopalakrishnan, Shashwat Jha, Vidhi Tralsawala, Aditya Chobe

Client Relationship

Michael Woodall, Gautam Makhani

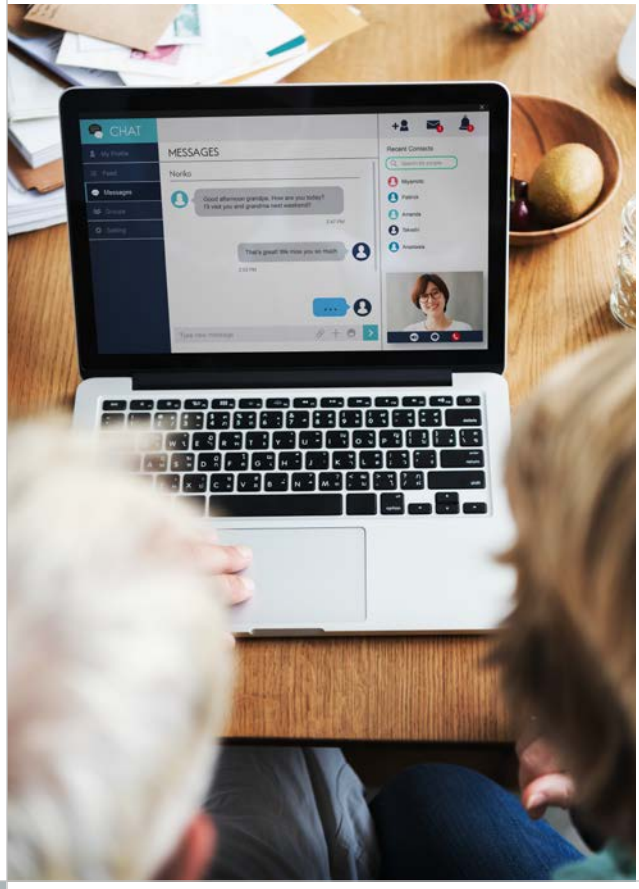
TABLE OF CONTENTS

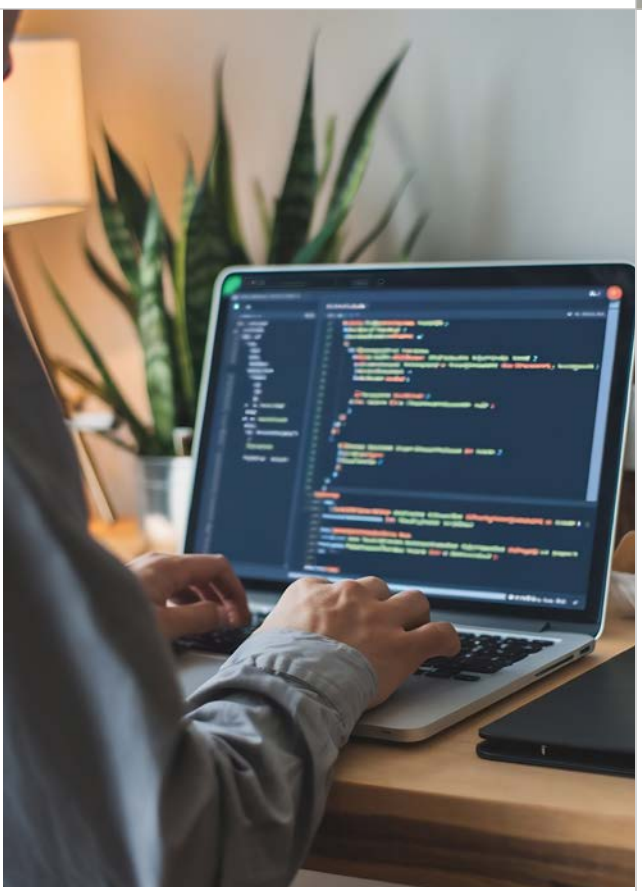
	Abstract	3
	Executive Summary	3
01	Business Context and Challenge	4
	1.1 Business Context and Challenge	4
	1.2 Inadequate Call Quality Review	4
	1.3 PII Exposure in Call Transcripts	4
02	Why Existing Solutions Fall Short	4 - 5
	2.1 Evaluation Isolated from Organizational Knowledge	4
	2.2 The Ground Truth Differentiator	5
	2.3 Scale and Recency	5
03	Solution Overview	6
04	Knowledge Ingestion Platform	7 - 8
	4.1 Salesforce Knowledge Articles	7
	4.2 SharePoint Documents	7
	4.3 PolicyBase Policy Documents	8
	4.4 ContactIQ Transcripts and 3-Level PII Redaction	8
	4.5 Unified Vector Store	8
05	Advisor Knowledge Agent	9 - 10
	5.1 Architecture and Request Path	9
	5.2 Performance	10
06	Post-Call Training Agent	10 - 12
	6.1 Pipeline Architecture	11
	6.2 Claims Funnel and Groundedness	11
	6.3 Quality Scorecard	11
	6.4 Dashboard and Conversational Analytics	11
	6.5 Business Impact	12
07	Snowflake Capabilities Leveraged	12 - 13
08	Evaluation and Quality Assurance	13
	8.1 RAG Accuracy with RAGAS	13
	8.2 Adversarial Security with Black Ice	13
09	Engineering Challenges and Design Decisions	14
	9.1 Heterogeneous Source Ingestion	14
	9.2 PII Redaction at Scale	14
	9.3 Speaker Diarization Without Native Labels	14
	9.4 Claims Verification Without Ground Truth Labels	14
	9.5 Vectorization at Source Scale	14
10	Conclusion	15
	Glossary	16

Abstract

This paper describes a dual-agent AI platform built natively on Snowflake for a wealth management firm. It addresses two recurring problems in large contact centers. During live calls, service representatives lack unified, real-time access to knowledge that can be easily searched without combing through various information sources. After calls, quality review covers fewer than 1% of interactions with results arriving weeks after. The Advisor Knowledge Agent supports representatives during live calls by by retrieving grounded, cited answers from a unified knowledge base through a Model Context Protocol (MCP) plugin. The Post-Call Training Agent automatically scores every transcript against a quality rubric and checks representatives' claims against the organization's knowledge base. This enables fact-level accuracy validation that existing contact center tools cannot provide. In validation by Subject Matter Experts (SMEs), the solution achieved a 100% SME approval score across a 649-question golden evaluation dataset, and deflected all 610 adversarial security probes. It processes the full call volume within two hours of call completion.

Note: Certain third-party tool and platform names referenced in this document have been replaced with neutral identifiers to protect the confidentiality of client vendor relationships.





Executive Summary

Altimetrik built a dual-agent AI platform natively on Snowflake. The Advisor Knowledge Agent tackles fragmented knowledge by giving representatives, grounded answers while a call is still in progress. The Post-Call Training Agent transforms quality assurance from a monthly sample of a few tens of calls per month into a continuous process that covers every call, with claim-level verification that no existing contact center tool provides.

01 Business Context and Challenge

The organization handled thousands of inbound calls a day across account inquiries, reporting, and how-to guidance. The representatives were expected to give answers that were accurate and empathet , drawing on a knowledge base distributed across three separate platforms.

1.1 Siloed Knowledge

Representatives rely on three separate knowledge sources: Salesforce Knowledge (~2,500 articles), SharePoint (~400 documents), and PolicyBase (~60,000 policy documents). With no, unified search representatives couldn't check all three during a call, and this fragmented setup resulted in poor response quality, slower call resolution, lower customer satisfaction rates and escalations.

1.2 Inadequate Call Quality Review

Quality assurance was manual and retrospective. At the end of each month, supervisors would select 10 calls at random, listen to recordings, and score them against a 17-item rubric. This amounted to a fewer than 1% of calls being reviewed; weeks later, the call was handled. Most importantly, there was no mechanism to detect when a representative gave a caller information that contradicted the official knowledge base.

1.3 PII Exposure in Call Transcripts

ContactIQ, the firm's contact center platform () provided raw call transcripts containing personally identifiable information: customer names, account numbers, and sensitive financial data. Any team wanted to use the transcripts for analytics, training or reporting would be working with the exposed customer data as there was no automated redaction creating compliance risk.

02 Why Existing Solutions Fall Short

Assessing call quality in wealth management requires an understanding of both the conversation and the information shared. Established platforms, such as NICE CXone and Nuance, offer call scoring, sentiment analysis, and rubric-based evaluation. The gap emerges when evaluation needs to go beyond call behavior and verify a representative's guidance against the organization's knowledge base.

2.1 Evaluation Isolated from Organizational Knowledge

Existing quality tools evaluate calls against in-call behavioral signals such as, whether they greet the caller appropriately, verify identity, use positive language, and resolve the query within the expected time. These are observable process metrics, but they answer only one question: "Did the representative follow the script?" They cannot answer the more important question: "Was the information the representative gave the caller actually, correct?"

This gap exists mainly because existing tools have no connection to the organization's knowledge base. A representative could pass every behavioral rubric item and still give a caller policy information that was updated six months ago, quote a product feature that no longer exists, or describe a process that was discontinued and the existing tools have no way to detect this.



2.2 The Ground Truth Differentiator

Our Training Agent is built on a fundamentally different premise: every factual claim a representative makes during a call is extracted and verified against the real organizational knowledge base. This knowledge base, which contains over 62,000 documents across Salesforce Knowledge, SharePoint, and PolicyBase, is the same source of truth used to train and certify representatives.

The Training Agent compares the representative's call log with the organization's knowledge base to assess the accuracy of the information shared, going beyond whether the representative followed the script. Each claim is matched against the knowledge base by cosine similarity and then judged by a language model for semantic alignment, producing one of four verdicts: Supported, Partially Supported, Contradicted, or Not Found. Contradicted claims, where a representative's statement directly conflicts with the source, are flagged automatically for coaching without any human listening to the call.



2.3 Scale and Recency

Existing tools require manual review of workflows and provide metrics at the call or agent level. The Training Agent processes the entire call volume automatically, every two hours, and produces claim-level evidence. When the knowledge base is updated, the impact propagates calls where representatives referenced outdated information can be retrospectively identified. Feedback also flows in the other direction, as repeated contradictions may signal that a knowledge base article needs review. This creates a continuous quality loop, helping teams improve both representative guidance and the information they rely on.

Capability	Existing Call Center Tools	This Solution
Quality scoring	Rubric-based behavioral metrics	Behavioral scoring plus claim-level knowledge verification
Ground truth integration	None; evaluation isolated from knowledge base	Every claim verified against 62,000+ source documents
Error detection	Process compliance only	Factual inaccuracies and contradictions automatically flagged
Review coverage	Random sample, typically under 1%	100% of calls, results within 2 hours
Knowledge feedback loop	Not available	Contradictions surface knowledge base gaps for content teams



03 Solution Overview

Altimetrik delivered two complementary AI agents sharing a common knowledge platform built natively on Snowflake. Every capability, from ingestion and vectorization through retrieval, evaluation, PII redaction, and analytics, lives within or is governed by Snowflake, maximizing data security, role-based access control, and operational simplicity.

Advisor Knowledge Agent	Real-time RAG agent for service reps via a custom ChatGPT MCP plugin. Returns grounded Detailed Answer and Customer-Ready Summary with inline citations.
Post-Call Training Agent	Automated transcript evaluation. Scores every call against a 17-item quality rubric, extracts representative claims, and verifies each against the knowledge base.
Shared Knowledge Platform	Unified vector store (KB_UNIFIED_RAG) fed by Salesforce, SharePoint, and PolicyBase. All ingestion, chunking, embedding, and delta-load pipelines managed in Snowflake via dbt Cloud.
Evaluation Framework	RAGAS regression testing for retrieval accuracy; Black Ice (Databricks open-source) for adversarial red teaming and security validation.

AI-Powered Contact Center Platform - End-to-End Architecture

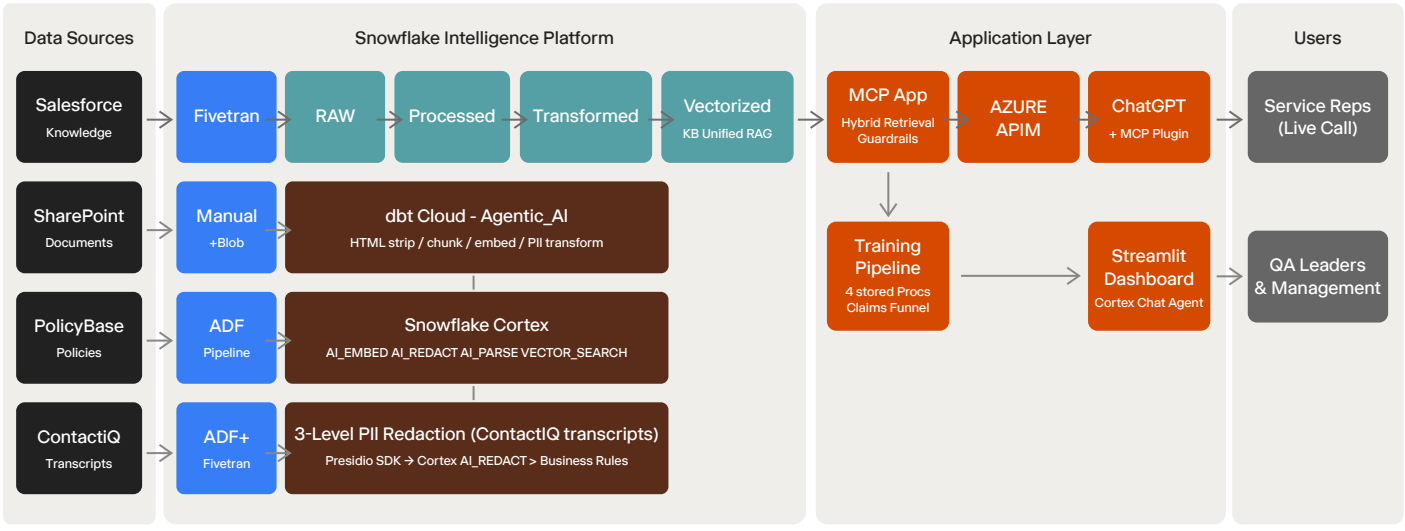


Figure 1: End-to-end system architecture from data sources through Snowflake to the serving layer



04 Knowledge Ingestion Platform

The foundation of both agents is a unified vector store holding chunked, embedded documents from four source systems. The ingestion architecture follows a Snowflake medallion pattern (Raw, Processed, Transformed, Vectorized), with per-source deviations documented and justified. dbt Cloud orchestrates Python models for HTML stripping, chunking, and embedding.

Multi-Source Data Ingestion: Snowflake Medallion Architecture

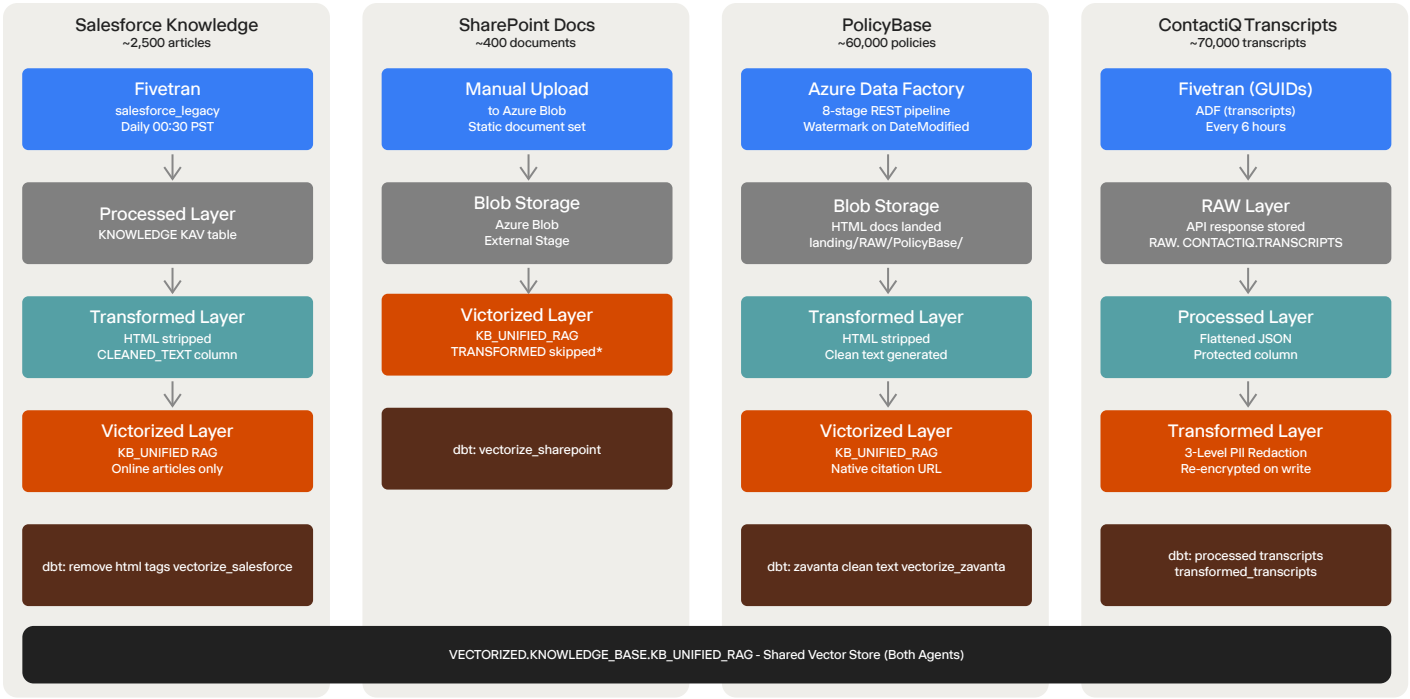


Figure 2: Multi-source ingestion lanes with Snowflake medallion architecture per data source

4.1 Salesforce Knowledge Articles

Ingestion uses the existing Fivetran salesforce_legacy connector (daily 00:30 PST). A dbt Python model strips HTML from the detail_c column; only articles with PUBLISH_STATUS "Online" proceed to chunking and vectorization. Delta handling tracks ID and FIVETRAN_SYNCED; deletes propagate via the IS_DELETED flag.

4.2 SharePoint Documents

Approximately 400 static reference documents are manually uploaded to Azure Blob Storage. A dbt Python model (vectorize_sharepoint) reads directly from blob, skipping the Transformed layer to avoid redundant intermediate copies of static files. This is a deliberate and documented deviation from the standard medallion pattern.

4.3 PolicyBase Policy Documents

PolicyBase contains approximately 60,000 HTML policy documents accessible via a paginated REST API. The team built an 8-stage Azure Data Factory pipeline: document discovery, change detection via DateModified watermark, batch creation (100-document API limit), bulk export with job-ID capture, job status polling, HTML extraction, blob landing, and dbt Python processing for chunking and embedding. All pipeline configurations are held in PIPELINE_CONFIG and read dynamically by ADF; no parameters are hardcoded.

4.4 ContactIQ Transcripts and 3-Level PII Redaction

ContactIQ transcripts are ingested incrementally via Fivetran (interaction GUIDs) and ADF (transcript retrieval), with complete raw API responses retained in RAW.CONTACTIQ. TRANSCRIPTS for audit and reprocessing. A three-level redaction pipeline runs before any downstream use:

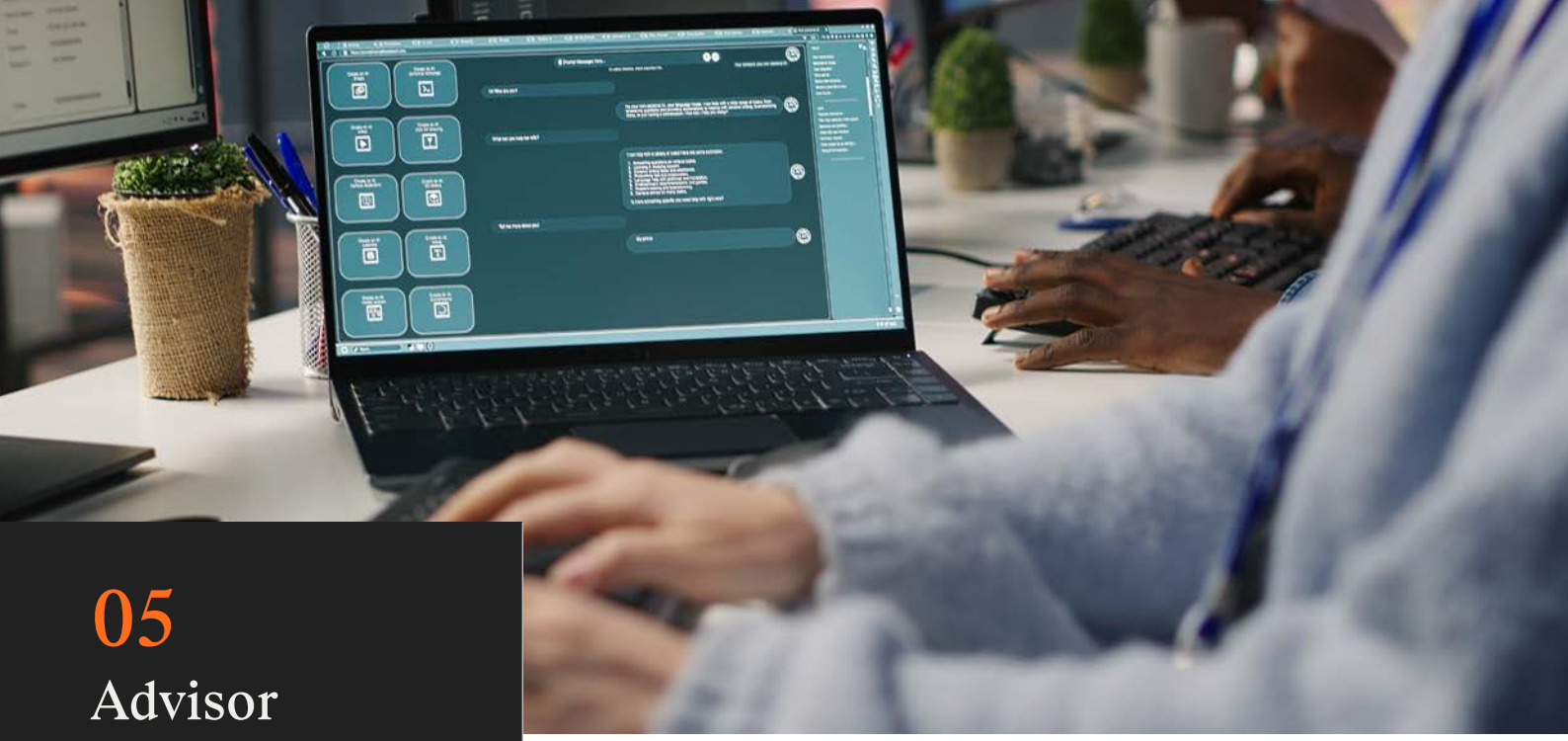
Level	Method	Purpose
01	Microsoft Presidio SDK with custom pattern rules	Structured entities and internal agent identifiers
02	Snowflake Cortex AI_REDACT (contextual model-based)	Entities that rule-based systems miss
03	Business-specific domain rules	Wealth-management financial identifiers and account formats

Redacted transcripts are re-encrypted on write. This redaction framework, delivered as a bonus capability beyond the original statement of work, is now being evaluated by the client for broader organizational adoption.

4.5 Unified Vector Store

All knowledge sources feed VECTORIZED.KNOWLEDGE_BASE.KB_UNIFIED_RAG. Each record holds chunked text, its vector embedding using the Snowflake snowflake-arctic-embed-l-v2.0-8k model (integrated immediately following Snowflake Summit), metadata, and a canonical source URL that surfaces as a native citation in ChatGPT.





05 Advisor Knowledge Agent

The Advisor Knowledge Agent gives service representatives instant, grounded answers during live customer calls via a custom ChatGPT plugin. The agent returns a Detailed Answer for the representative and a Customer-Ready Summary, both grounded in retrieved knowledge with inline citations linking back to the source article.

Troubleshooting Agent: Real-Time RAG Pipeline

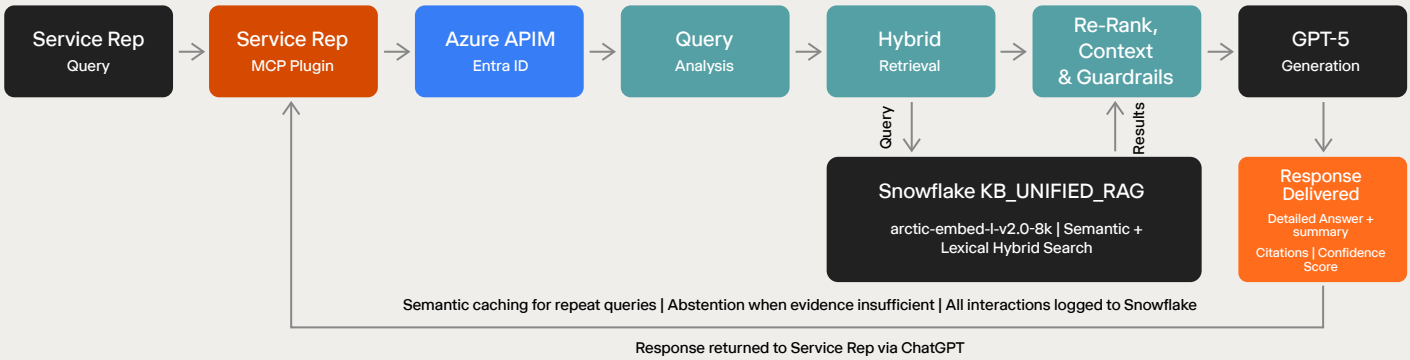


Figure 3: Advisor Knowledge Agent RAG pipeline from user query to grounded, cited response

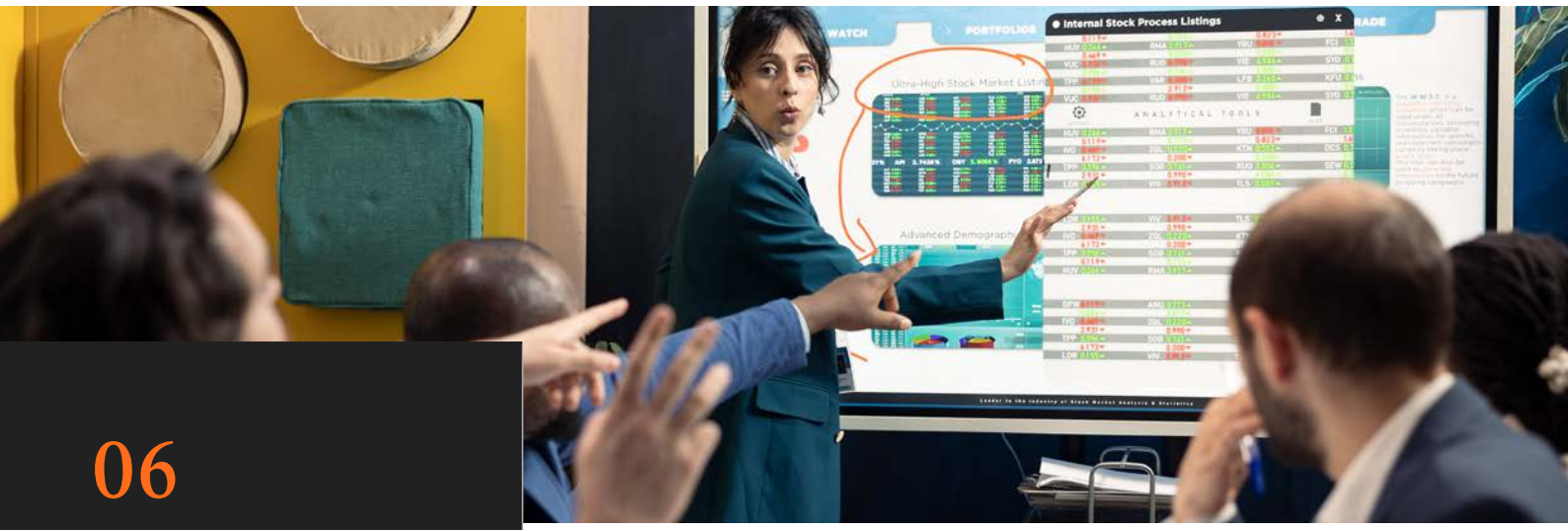
5.1 Architecture and Request Path

The application is a Python service on Azure App Service exposing capabilities via the Model Context Protocol (MCP), behind Azure API Management with Microsoft Entra ID authentication. The in-application retrieval pipeline runs six sequential stages per query:

- Query Analysis - derives retrieval intent, keywords, phrases, and named entities
- Hybrid Retrieval - combines semantic vector search with lexical and keyword signals against KB_UNIFIED_RAG
- Re-ranking and Selection - ranks candidate chunks and selects the strongest supporting evidence
- Context Construction - assembles ranked chunks into the answer context
- Guardrails and Confidence Handling - validates evidence; abstains when insufficient
- GPT-5 Generation - produces Detailed Answer and Customer-Ready Summary grounded in retrieved context

5.2 Performance

SME approval score	100% on the 649-question golden dataset (1 excluded: source article absent from sandbox)
Response time	10 to 15 seconds uncached; under 5 seconds from semantic cache
Grounding	All answers sourced from KB_UNIFIED_RAG; abstention when evidence is insufficient
Citations	Canonical URLs render natively in ChatGPT, linking to Salesforce, SharePoint, or PolicyBase source
Security	All 610 Garak probes and all custom PromptFoo attacks rejected



06
Post-Call Training Agent

The Training Agent transforms call quality review from a manual retrospective spot-check into a continuous automated evaluation pipeline. It ingests every ContactIQ transcript, scores it against a 17-item rubric, extracts factual claims made by the representative, and verifies each against the unified knowledge base, flagging knowledge gaps, coaching opportunities, and compliance issues at scale.

Post-Call Training Agent: Claims Verification Pipeline

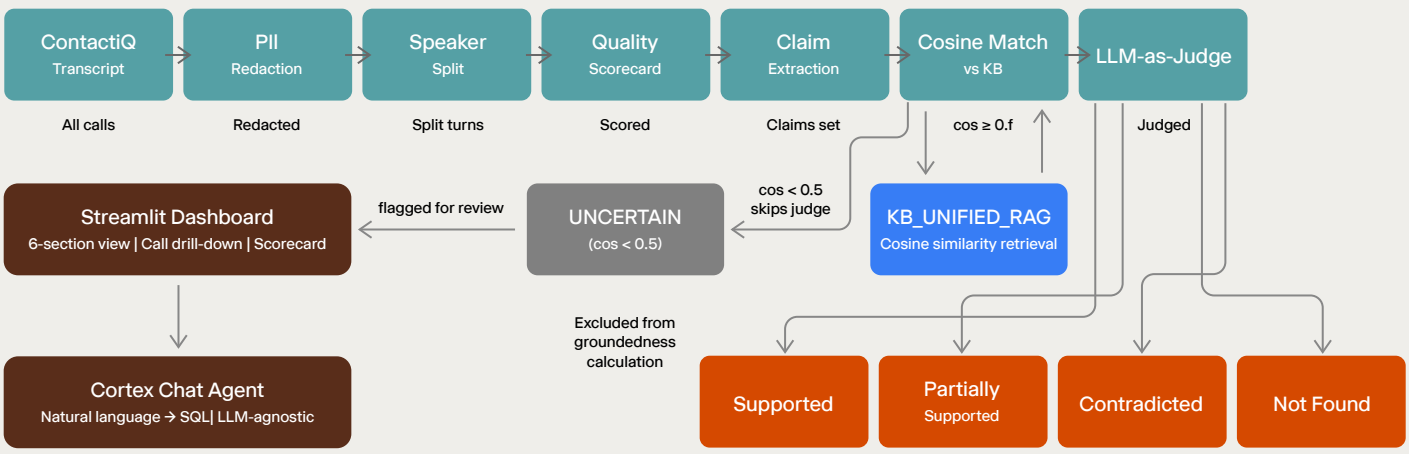
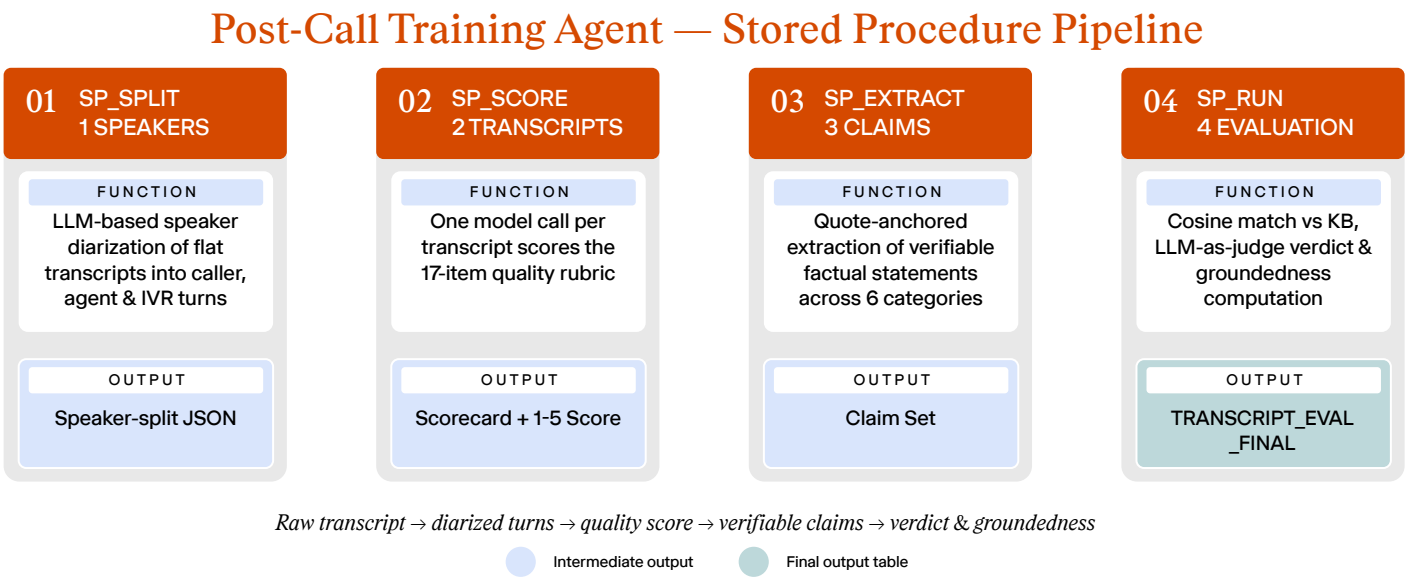


Figure 4: Training Agent claims verification funnel from raw transcript to actionable quality verdict

6.1 Pipeline Architecture

The entire pipeline runs natively within Snowflake as four stored procedures in VECTORIZED.CALL_EVAL, eliminating external orchestration dependencies:



6.2 Claims Funnel and Groundedness

Each extracted claim is matched against KB_UNIFIED_RAG by cosine similarity. Claims below a 0.5 threshold are marked Uncertain and excluded from LLM judging. Claims above the threshold receive one of four verdicts:

Verdict	Condition	Meaning
Supported	Cosine > 0.5, judge confirms alignment	Representative statement matches the knowledge base
Partially supported	Cosine > 0.5, information incomplete	Correct but incomplete response given
Contradicted	Cosine > 0.5, context contradicts	Representative gave information conflicting with the knowledge base
Not found	No relevant chunk retrieved	Claim cannot be verified against available knowledge

Groundedness Rate = (Supported + Partial) / (Supported + Partial + Contradicted). This metric is computed per call and aggregated portfolio-wide, giving leadership a precise measure of knowledge accuracy at scale.

6.3 Quality Scorecard

A 17-item rubric mirrors the firm's existing call evaluation scorecard, covering opening greeting, caller verification, purpose identification, positive phrasing, active listening, empathy, jargon avoidance, resolution delivery, and closing. Every affirmative answer carries verbatim evidence from the agent's transcript turns. Scorecard definitions load at runtime from SCORECARD_ITEMS; rubric changes require no code deployment.

6.4 Dashboard and Conversational Analytics

A Snowflake-native Streamlit dashboard presents output across six sections: portfolio overview, call deep-dive, agent-level scorecard, claims and KB verification metrics, statistics, and a split viewer with judge output and contradiction detail. A Snowflake Cortex chat agent enables natural-language querying of evaluation data, translated to SQL and governed by Immuta data masking policy.

6.5 Business Impact

Previously fewer than 10 calls per month were reviewed, covering under 1% of volume, weeks after the fact. The automated pipeline processes every call with results within two hours:



- Contradiction detection: calls where a representative's statement conflicts with the knowledge base are automatically flagged for coaching
- Knowledge base maintenance: systematic contradictions signal articles that need updating, closing the content feedback loop
- Onboarding acceleration: high-scoring calls surface as training examples with claim-level evidence for coaching
- Compliance gap detection: calls missing mandatory caller verification are flagged without manual review
- Leadership reporting: portfolio-level sentiment, topic, scorecard trends, and groundedness replace manual report compilation



07 Snowflake Capabilities Leveraged

The solution maximizes the latest Snowflake AI capabilities, consolidating the entire intelligence stack within a single governed platform.

Capability	Snowflake Feature	Application
Vector embeddings	Cortex AI_EMBED (snowflake-arctic-embed-l-v2.0-8k)	All chunk and query embeddings; adopted immediately on Summit release
Hybrid retrieval	VECTOR_SEARCH and cosine similarity	Semantic and lexical combined retrieval with re-ranking in a single pass
PII redaction	Cortex AI_REDACT	Level 2 contextual redaction in the 3-level transcript pipeline
Document parsing	Cortex AI_PARSE	Structured extraction during PolicyBase and ContactIQ ingestion
LLM inference	Cortex (Mistral-2 and GPT-5 via Azure)	All 4 training pipeline stages: speaker split, scoring, extraction, judging

Capability	Snowflake Feature	Application
Pipeline orchestration	Snowflake Stored Procedures	4-stage training pipeline; incremental, idempotent; SP_EVAL_ONE_CALL for single reruns
Evaluation scoring	CORTEX.EMBED_TEXT_1024	RAGAS semantic similarity scoring for golden-dataset regression
Analytics UI	Streamlit on Snowflake	Six-section evaluation dashboard hosted natively in Snowflake
Conversational analytics	Snowflake Cortex chat agent	Natural-language-to-SQL over evaluation data; LLM-agnostic across Claude, GPT, and Snowflake-native
Data transformation	dbt Cloud (Agentic_AI project)	Python dbt models for HTML stripping, chunking, vectorization, PII transformation



08 Evaluation and Quality Assurance

8.1 RAG Accuracy with RAGAS

A golden evaluation dataset of verified question-answer pairs was compiled from real knowledge articles. The RAGQA harness submits each question to the deployed MCP endpoint, extracts the Detailed Answer, and scores it using RAGAS AnswerSimilarity, a semantic metric using Snowflake Cortex embeddings that compares meaning rather than exact wording. Result: 100% SME approval score across the golden dataset.

8.2 Adversarial Security with Black Ice

Security testing used the Black Ice toolkit (Databricks open-source) with two tools:

- Garak: 610 built-in adversarial probes across broad attack categories. All 610 rejected by the agent.
- PromptFoo: custom adversarial prompts covering prompt injection, data leakage, hallucination induction, and jailbreak attempts. All rejected.



09 Engineering Challenges and Design Decisions

9.1 Heterogeneous Source Ingestion

Each of the four sources required a completely different ingestion mechanism. Rather than forcing a uniform pattern, the team accepted per-source medallion deviations where the standard pattern added cost without value, documenting each deviation with production remediation steps. This pragmatic approach kept the POC on schedule while preserving a clear production path.

9.2 PII Redaction at Scale

Redacting 70,000 transcripts required three complementary approaches: Presidio for structured entities and internal identifiers; Cortex AI_REDACT for contextual entities that rule-based systems miss; and domain-specific business rules for wealth-management-specific patterns. Processing approximately 200 transcripts takes around 1.5 hours; parallelization of the full backfill is a production prerequisite.

9.3 Speaker Diarization Without Native Labels

ContactIQ transcripts arrived as flat text with no speaker labels. LLM-based speaker splitting (SP_SPLIT_SPEAKERS) was implemented to output structured JSON with ordered speaker turns. Long transcripts hit the Snowflake 600-second statement timeout; the production path uses native ContactIQ API pre-split transcripts, which became available in mid-engagement.

9.4 Claims Verification Without Ground Truth Labels

Evaluating call quality against a knowledge base poses a fundamental challenge: there is no pre-defined correct answer for open-ended conversation. The two-tier approach, deterministic cosine similarity for retrieval-based filtering followed by LLM-as-judge for verdict classification, grounds subjective quality assessment in mathematical similarity against authoritative knowledge content.

9.5 Vectorization at Source Scale

The PolicyBase corpus of 60,000 HTML documents required a multi-platform pipeline spanning the PolicyBase REST API, Azure Blob, and Snowflake dbt. The team standardized chunking logic across all sources into per-source Python dbt models that union into KB_UNIFIED_RAG, while documenting the linear maintenance cost scaling as a production backlog item with a clear consolidation path.



10 Conclusion

Altimetrik delivered a dual-agent AI platform that fundamentally transforms how a wealth management firm's service operation accesses knowledge and measures quality. The Advisor Knowledge Agent gives every representative instant, grounded answers during live calls. The Training Agent converts a process covering fewer than 1% of calls retroactively into a continuous, automated quality function that scales to 100% of call volume, with claim-level verification against real organizational knowledge that no existing contact center tool provides.

The decision to build natively on Snowflake, using Cortex AI_EMBED, AI_REDACT, stored procedure orchestration, Streamlit, and the Cortex chat agent, produced a solution that is secure by design, manageable within the client's existing data governance framework, and positioned as a reusable platform for AI capabilities well beyond the original scope.

This engagement demonstrates that Snowflake's latest AI capabilities, combined with the Model Context Protocol and rigorous evaluation practices, can deliver production-quality enterprise RAG and automated quality analytics at the pace of a proof of concept.

Glossary

Terms and abbreviations used in this document.

Term	Definition
ADF	Azure Data Factory. Microsoft cloud-based ETL and data integration service.
APIM	Azure API Management. Gateway service managing API security, rate limiting, and routing.
Cortex	Snowflake Cortex. Suite of ML and AI functions built into the Snowflake platform.
dbt	Data Build Tool. SQL and Python transformation framework; dbt Cloud is the managed hosted version.
Entra ID	Microsoft Entra ID (formerly Azure Active Directory). Identity and access management service.
Fivetran	Managed data connector platform used to ingest Salesforce and ContactIQ data into Snowflake.
Garak	Open-source LLM vulnerability scanner with 610+ adversarial probe types.
GPT-5	OpenAI large language model used for answer generation in the Advisor Knowledge Agent.
Groundedness	Metric measuring the fraction of representative claims that are Supported or Partially Supported by the knowledge base. Formula: (Supported + Partial) / (Supported + Partial + Contradicted).
KB_UNIFIED_RAG	VECTORIZED.KNOWLEDGE_BASE.KB_UNIFIED_RAG. Snowflake table holding all chunked, embedded knowledge documents from all sources.
LLM	Large Language Model. Transformer-based AI model trained on text.
MCP	Model Context Protocol. Open standard for exposing tools and data sources to LLM applications.
Medallion	Layered data architecture pattern: Raw (landing), Processed (cleaned), Transformed (business-ready), Vectorized (embedded).
PII	Personally Identifiable Information. Data that can identify an individual.
PolicyBase	Third-party policy management system containing the firm's approximately 60,000 policy documents.
Presidio	Microsoft open-source SDK for PII detection and anonymization.
PromptFoo	Open-source LLM testing and adversarial red teaming framework.
RAG	Retrieval-Augmented Generation. Pattern of grounding LLM responses in retrieved documents.
RAGAS	RAG Assessment framework. Evaluation library for measuring RAG system accuracy.
ContactIQ	Contact center telephony and transcript platform used as the call recording source.
Snowflake	Cloud data platform. All data storage, transformation, vectorization, and inference runs within Snowflake.
Streamlit	Python-based UI framework; Streamlit on Snowflake hosts apps natively within the Snowflake environment.
VECTOR_SEARCH	Snowflake Cortex function performing approximate nearest-neighbor search over embedding columns.

About Altimetrik

Altimetrik is an AI-native engineering company helping some of the most revered and iconic enterprises modernize systems, data, and processes at the heart of their business, so they can move faster, operate more efficiently, and innovate continuously. Through ALTi AIOS™, its AI-native operating system, Altimetrik combines the latest AI capabilities with deep engineering expertise to help clients solve complex challenges, accelerate modernization, and deliver measurable outcomes at scale.

Altimetrik's clients get access to the latest AI innovations while maintaining flexibility to choose the right technologies for their business, through trusted relationships with OpenAI, Google Gemini, Anthropic, Databricks, and major hyperscalers.

A member of the World Economic Forum's Centre for AI Excellence, the Forum's global hub for shaping responsible AI, Altimetrik is also recognized in the 2025 Constellation Research ShortList™ for Global AI Services and named a Major Contender in multiple Everest Group PEAK Matrix® assessments, including Software Product Engineering Services (2026), Enterprise Quality Engineering Services (2025), and Digital Engineering Services for BFSI and Life Sciences.

Learn more at
altimetrik.com.

