



Snowflake and Domain-AI Real-Time Intelligence for Enterprise Data Warehousing

Venkat Sunil Kumar Indurthy

Software Developer, Compunnel Software Group Inc, USA

Publication History: 11-02-2026 (Received); 10-3-2026 (Revised); 15-3-2026 (Accepted); 20-3-2026 (Published).

ABSTRACT: This platform uses special agents to give instant access to information that has been pulled from control data very quickly, and therefore, allowing companies to reduce the amount of time it takes to move data to and from different Systems, as well as for the Extract-Transform-Load (ETL) processes associated with those different systems. Integration with Snowflake Cortex AI will give businesses the ability to take their business data warehouse and move it to a Cloud-native environment, allowing them to achieve insight from their business data warehouse much quicker than with traditional methods; in fact, the speed at which they will get Insight is 99.6 times quicker than traditional methods. Some of the key features of this platform include: storage-compute separation; real-time streaming of snowpipe; and zero-copy cloning. This platform uses the python ETL pipelines to detect anomalies in customer's data and create a custom domain agents that will enhance any process to do with Procure-to-Pay. Also, the cortex analyst has the capability of converting to SQL utilizing natural language. Practical experience from the creation of this platform will provide companies with greater cross-sell efficiencies, significant increases in analyst's productivity and the detection of Fraud. The paper presents Best Practices to ensure a Stable, Scalable process through Role-based Access Control (RBAC) and Serverless Computing. By 2026, the paper anticipates the development of explainable AI compliant with government regulations for regulated industries, readily made possible by the advancement in multi-modal agents and federated learning.

KEYWORDS: Extract-Transform-Load (ETL), Snowflake Cortex AI, Cloud-native Environment, Zero-copy Cloning, Procure-to-Pay, Multi-Modal Agents, Federated Learning.

I. INTRODUCTION

Significant challenges are faced by enterprise data warehouses, including data silos resulting from numerous legacy systems; long delays in ETL processes that impede developing insights to make timely decisions; and a lack of real-time AI-generated insights in business-critical areas (i.e., procurement and compliance). Fast-paced business sectors like banking and healthcare experience these types of challenges more acutely, which can result in regulatory risk, fragmented/incomprehensive insights, and increased costs of multiple data pipelines. One method to resolve these challenges is Domain-AI's method of scoping AI Agents to a defined business domain. This allows for the generation and delivery of accurate real-time analytics without necessitating the movement of data. Domain-AI's method incorporates business-specific information using natural language queries, and therefore allows for verified insights to be generated directly from within the Controlled Data Warehouse (C-DW). As a result, Domain-AI is able to transform C-DW's into proactive data intelligence centers that provide insight by incorporating both structured and unstructured data [1].

The Domain-AI solution from Snowflake attempts to solve the challenges of enterprise data warehousing with the introduction of its features called Cortex and Intelligence. However, the implementation of these features comes with trade-offs regarding their complexity, accuracy, and cost. The primary concerns include semantic gaps created by using natural language processing, unpredictably high compute spiking caused by running AI queries, and governance issues related to the inconsistency of defined metrics between different teams. The benefits of the Domain-AI solution are faster generation of timely insights and an ability to scale more quickly than previous generations of enterprise data warehouses. However, there are also costs associated with implementing these solutions and limitations associated with large language models [2].

As AI continues to alter the traditional warehousing concept, many of the same warehousing and merchandising problems faced historically will be present, and some may be more pronounced or exacerbated due to AI-related issues, such as problems with semantic interpretation, increased costs associated with streaming and unstructured data, system



performance failure due to token limitations, and system overloads. Other problems will arise from inadequate audit trails and inadequate access control system governance associated with the use of AI. Many organizations experience Rio-style benefits from real-time data access and lower computing costs; however, the trade-off between reduced resources and increased complexity and issues related to managing multiple and competing resource allocations outweighs the benefits of real-time access. Other issues related to usability of AI products, such as the improvement of supporting multiple data types and the ability to use AI products to empower non-technical users with an easy-to-use interface, are present, but issues like hallucination and semantic nonsensitivity continue to hinder the usability of AI products.

The challenge of scalability related to AI products is primarily the inability of organizations to view the current status of chargebacks and vendor relationships at the same time. In general, ineffective definitions of terms and gaps in audit information are present in many organizations. The delays in business intelligence (BI) reporting (caused principally by the slow speed of batch ETL processes) result in many difficulties for enterprise data warehouses, primarily by creating long delays between the time data is created and available for decision-making. This has been further complicated by the fragmentation of analytics across multiple analytics platforms; for example, the on-premises systems, cloud-based systems, and software-as-a-service (SaaS) systems used by organizations all complicate the ability to connect data sources and analyze cross-functionally, such as the ability to link procurement data to financial compliance. Additionally, organizations in the finance and healthcare industries must meet requirements for strict and immediate anomaly detection and audit trails for large volumes of unstructured data sources, which makes it increasingly necessary for these organizations to identify anomalies and establish audit trails as soon as they are created.

Inefficient batch ETL pipelines create old data that cannot be used for time-sensitive activities such as fraud detection/risk assessment, while out-of-date technology make it even more difficult to scale the larger data sets involved. The use of many different data sources creates an analytics silo, preventing companies from having a single view of their data and producing inconsistent metrics, duplicated efforts, and low-quality data. Additionally, healthcare and financial companies have regulatory requirements, so they must be able to use real-time data, but traditional EDW solutions are unable to keep pace with changes in regulatory expectations, leading to a greater risk of penalties.

Healthcare data warehouses must adhere to strict regulations [such as HIPAA, HITECH Act, GDPR, and FHIR standards] concerning the protection of Protected Health Information (PHI) and the use of analytics and reporting. All of these regulations include requirements for auditability, encryption, and access control, creating difficulties for healthcare providers who want to implement real-time systems. In addition to the general compliance requirements, some specific compliance elements include RBAC, encryption, and breach notification within a certain timeframe. The GDPR states that companies need to conduct Data Protection Impact Assessments (DPIA) and provide individuals with the right to request deletion of their data. The data normalization process is essential for using FHIR data for analytic purposes. The security solutions for this type of data include multi-factor authentication, AES-256 encryption, and anomaly detection to govern access. Compliance issues are compounded by the many differences between compliance regulations across state lines, along with variations in data quality, resulting in heavy fines for violations of compliance standards [3].

Snowflake Intelligence is a Domain-AI solution that allows businesses to perform real-time natural language query execution on restricted business data warehouses without exporting or migrating their external data. It uses the Cortex AI technology to provide accurate and reliable insights from a static data warehouse quickly and easily. It does this by combining structured ETL data with unstructured data sources, while maintaining strict control over both data types, none of which are subject to the delays and lack of access to data caused by the batch-based ETL process.

Domain AI includes AI agents and models that are designed specifically for that business's requirements and that utilize compliance and vendor procurement rules to allow for context-aware reasoning. Each of these agents exists within well-defined boundaries, meaning they can only access approved datasets and can use domain-specific ontologies to make accurate predictions. For example, some agents exist to monitor compliance with healthcare regulations whereas others assess risk of a sale. Features of this solution include scoped agents that ensure secure access, real-time querying capabilities that facilitate iteration of research, and a robust governance framework that ensures the accuracy of outputs, especially in industries subject to regulation.



II. SNOWFLAKE DATA WAREHOUSING FUNDAMENTALS

Through its cloud-based design, Snowflake has redefined how to store large amounts of data in a way that allows it to scale out automatically to handle increasingly larger workflows and to meet the needs of immediate processing. Snowflake's architecture has removed many old limitations by utilizing unique features such as 'zero-copy cloning' and 'time travel' capabilities, allowing for massive amounts of parallelism to be realized when sharing data for AI-related purposes. Leveraging its elastic virtual warehouses, Snowflake can scale computing on-demand without affecting service delivery on other workloads; this is achieved through having all data stored in an optimized central location with inherent durability that provides price stability, even during peak usage periods [4].

By utilizing the Snowflake architecture for large-scale queries against large sets of data, results are returned to the user with little delay, due to the way that data is handled as it is streamed back to the user. With zero-copy cloning, organizations can rapidly generate physically small, permanent, and instantaneous snapshots of their data for their purpose of experimentation and development, as opposed to utilizing time-consuming traditional extract, transfer, and load (ETL) methodologies. In addition, having time travel capabilities in Snowflake enables organizations to use previous versions of their data to investigate compliance, security breaches, and develop plans for data recovery through the immediate restoration of a single object. As a result, organizations can run many concurrent email message and electronic document queries while continuously receiving real-time records into their data warehouse [5].

To optimize performance by providing efficient access to both frequently accessed ("hot") and infrequently accessed ("cold") data, caching layers have proven effective when used in conjunction with computing and storage systems that use middle-tiered or upper-tiered special caches. Caching layers of memory, local solid-state drives (SSDs), and remote object storage help to minimize network latency for applications where large numbers of simultaneous reads occur in real time, such as with domain-specific analyses and AI-focused queries. For connecting stateless servers to persistent storage, it is necessary to implement a multi-tiered cache that has policies for the eviction of older cached data, guidelines for how long cached data can be maintained, and a prefetch mechanism for retrieval of data from the cloud. Writing algorithms for AI application usage to help scale AI applications the process of balancing durability and speed is the use of write-through and write-back algorithms.

A Snowflake context on caching architecture includes multiple layers, including DRAM/Redis query cache (conversational history) located in Memory, local NVMe SSD storage (micro partitions) stored as Hot and Cold, and the later can be archived for infinite-storage. To set up the architecture, first, identify hot data through query profiling and preload index files that are relevant to the domain for the purpose of synchronizing cache and storage. To monitor performance, keep track of metrics such as hit ratio and staleness, which will help you maintain acceptable performance levels. Based on domain requirements, best practices call for using time-to-live adjustments with Slack's criteria for managing caches, combined with Snowflake's Result Cache to facilitate efficient data reuse; the combination creates an opportunity for scalable, real-time intelligence across multiple distributed systems [6].

By using Snowpipe, businesses ingest real-time records from many sources into governed data warehouses through automated systems (ETL pipelines) and manage low latency ingestion at a continuous pace; these systems support domain-specific AI systems. DBT takes raw records from Snowpipe and uses ETL and augmented models to create analytic-ready, domain-specific models, completely eliminating data silos from the flow. Snowpipe Streaming achieves sub-10 second latency and very high throughput; by continuously streaming rowsets directly into database tables, query performance for real-time data has been greatly improved.

Snowpipe requires using Serverless Pipe objects and dynamic buffer objects to manage both the incoming and outgoing edits for Servers, while pricing at a flat rate per GB of usage. In addition, DBT provides the functionality to create materialized views and embed domain-specific notes by using Custom Macros to generate detailed histories after edit completion and facilitate zero-downtime deployments via Snowflake Tasks. The overall architecture for DBT and Snowpipe integrates both ETL with AI governance, enabling companies to quickly query their records for many different business functions, such as auditing and sales forecasting.

Snowflake's comprehensive Governance and Security System is used to protect PHI and financial data in AI workflows through a combination of strict granular Role Based Access Control (RBAC), dynamic data masking, and tri-secret secure methodologies. By strictly controlling which users can access data based on their assigned role, the Governance and Security System enables businesses to comply with regulation requirements such as SOX, HIPAA and GDPR while still allowing for the safe real-time querying of the record by Cortex agents and enabling audit-ready record



lineage. RBAC allows the only users assigned to Domain Marts to have access, thereby facilitating zero-copy sharing of records without exposing them to potential risk.

Dynamic Data Masking masks any Personally Identifiable Information (PII) that may be accessed based on the Assigned Role, assuring that sensitive records are only completely visible to authorized personnel. The Governance and Security System uses customer-managed keys and envelope encryption to increase security for recipients; you can quickly rekey the keys, thereby enabling timely availability. Through this architecture, you can gain real-time insights, while protecting against unauthorized access, enabling your Data Warehouse to operate as a compliant Intelligence System that supports your Enterprise Compliance and enables your Vendor Contract Analysis.

III. DOMAIN AI AND SNOWFLAKE INTELLIGENCE

Domain AI is a new AI-based system that focuses on creating customized artificial intelligence (AI) models and agents that utilize industry-specific knowledge to improve accuracy and relevancy in tightly controlled data environments. For instance, sector-type AI models will support specific sectors such as Procurement, Risk Management, and Compliance. In contrast, general AI uses large amounts of data obtained from the Internet to develop a wide variety of responses to many different requests, but these responses are often shallow. Domain AI uses a method called Scoped Retrieval Augmented Generation (RAG) that retrieves the contextually relevant information, based upon specific business logic, directly from the company's own dataset and applies those rules to minimize mistakes on general submissions. The agents (e.g., procurement agents) created through Domain AI will understand the terms of contracts and regulations.

The RAG mechanism retrieves only approved datasets and applies business rules to minimize the error rate of generic constructs. Domain AI primarily builds upon utilizing curated enterprise datasets, and it operates based upon deep vertical knowledge and internal governance systems to ensure compliance. Compared to the greater adaptability offered by General AIs, Domain AI has a lower risk of having inaccuracies, and therefore poses less risk to the enterprise. The implementation of Domain AI will operate through the use of Cortex Agents (e.g., Procurement Risk Agent) as a real-time source of actionable predictions without exporting data. This is particularly beneficial to heavily regulated sectors where the importance of accuracy and verifiability of results is paramount [7].

By combining the Snowflake Intelligence platform with the Cortex AI platform, organizations can use Natural Language Interfaces to streamline complex Data Analysis into Uniform Access and Executable Workflow Models. The Snowflake Intelligence platform provides a broad range of users, including support personnel, executive leadership, or board members, with the ability to convert queries into structured type of SQL syntax models and execute workflows appropriately within their respective Domains for Procurement Risk and Regulatory Compliance [8].

The Cortex platform includes the following tools: Cortex Analyst, for converting text to SQL Query Language (TSQL), Cortex Search, to look up information from a variety of sources, and customized Data Agents built for specific Semantic Models. An always-on chat agent, Cortex provides users with access to an agent where they can ask questions and receive responses that contain not only written information but also tables, graphics, recommendations, etc., while keeping the user information trusted through citations & audit logs.

The integration of Cortex AI with Snowflake helps convert NLP queries to TSQL queries faster and continues conversations with users about their queries and enables users to take action (making API calls) as part of those conversations. It also connects with specialized tooling (domain / industry-specific) to create a bridge between business intelligence (BI) and supporting the multi-modal environments providing real time BI (Analytics) through access to various data formats (Structured and Unstructured). In addition, due to the design of Snowflake's security architecture and the measures taken by Snowflake, the risks involved in exporting data are minimal for users in regulated industries.

Low Latency LLMs (large language models) such as Claude 4 optimized by Hugging Face and Llama 3.1 are used to enable Snowflake Intelligence to obtain real-time intelligence from interactions. The combined efforts of several core technologies — FP8 Quantization, Tensor/Pipeline Parallelism, and Dynamic SplitFuse Batching — provide Snowflake Intelligence with first-token time to less than 750 ms and inter-token time to less than 70 ms. In addition, the use of continuous key values (KV) cache on real-time decision-making processes through streaming context received through Snowpipe-led warehouses enables Snowflake Intelligence to provide users with accurate analyses from real-time data.

Snowflake's serverless models have been constructed in an optimized manner to provide high levels of throughput on H100 GPUs, which in turn enables higher levels of quality and speed for interactive workloads with all cross-region



inferences being done within 250 ms. Real-time streaming context from Snowpipe updates accumulated KV caches and embeddings with the most recent data entry, providing Snowflake Intelligence with precise reasoning capabilities. All agents will be orchestrated with a layered process of reasoning, producing transparent traces through use of Trust Metrics and Citation Reports for all users to guarantee real-time decision-making. Performance Metrics for Snowflake Intelligence reflect TTFT at below 750 milliseconds, ITL at below 70 milliseconds overall and overall Latency optimization of all components [8].

Snowflake Intelligence aims to turn business data warehouses into interactive Multimodal Data Intelligence platforms through Cortex AI. Using the primary user interface called Cortex Analyst, users can convert complex written queries into TSQL with their entire history of prior interactions retained for ongoing continuity. In addition, both Structured and Unstructured data formats will be processed through the Cortex Analyst, and will provide for deeper analytical capabilities such as comparing vendor sentiment against vendor payment delays.

Through Cortex Agents, businesses have the capacity to create customized AI Assistant Agents that are domain-specific, which utilize YAML specifications. This allows the development of Agent capabilities that are tailored for a business's unique and custom data retrieval methodologies that can provide historical data-based risk scoring. Other advanced serverless SQL capabilities (e.g., TRANSCRIBE, EMBED_TEXT and COMPLETE), have the ability to be integrated with real-time analytics workflows, whereby functionalities such as anomaly detection and sentiment analysis can occur from the communication of the audio data of completed conversations.

The extensive integration of over a thousand data set-related tools allows for the building of customized interfaces via Snowflake Native Apps and Python UDFs, offering a substantial increase in the speed and efficiency with which compliance gaps are identified and flagged as risk. The system provides its users with Zero-Latency Anomaly Detection and complete analytic capabilities without having to transfer any data.

IV. PROPOSED ARCHITECTURE

The Snowflake Intelligence architecture integrates a cloud-based Data Warehouse into its Cortex AI framework to provide custom-built real-time intelligence solutions for industries such as Healthcare, Compliance, and Procurement. In this design, all data is stored and computed independently with built-in functionality for agentic workflows. The Snowflake Intelligence architecture uses Snowpipe Streaming for streaming data and Controlled Risk Analyses Grouping (RAG).

With this purpose in mind, there are several ways in which this architecture supports the ingestion of data. The ingestion of data from Kafka Streaming (via SaaS) into raw tables is accomplished through the use of Snowpipe streaming. For Audit Recovery purposes, the Data Storage Layer provides Time Travel capabilities for both Audit Recovery and Zero-Copy Cloning to create copies of data for immediate development. The Compute Layer of Snowflake Intelligence is powered by Virtual Warehouses that support the Cortex AI Engine. The Cortex AI Engine uses low-latency Language Models to convert Natural Language to SQL for building strategy models.

In addition, Streamlit is used as the user interface for presenting Dashboard views connected to the various Domains, such as displaying Risk Alerts and Compliance Reports. The AI for Domains includes providing Real-Time Alerts and HIPAA Compliant Reporting. All of this information supports the conversion of agentic data to actionable business intelligence at the speed of Business.

The Snowflake Intelligence-Cortex AI architecture consists of Six Independent Layers that provide enhanced Scalability, Security, and Customized Workflows. These Layers have been created to convert Corporate Data (Controlled and Structured) into Collective Real-Time Domain Intelligence to help organizations improve Decision-Making and Operational Efficiency. Refer to Figure 1 for further clarification.

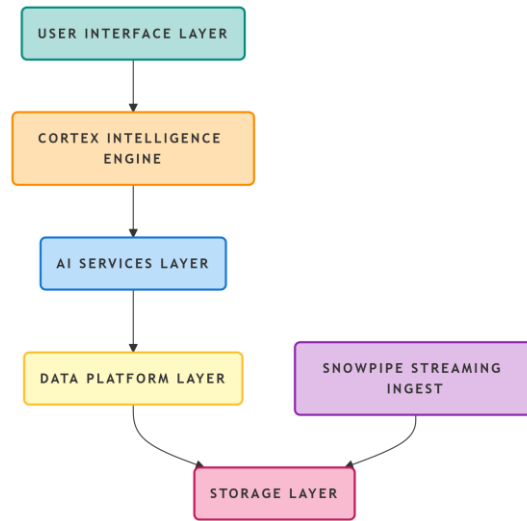


Figure 1: Cortex AI Architecture

1. User Interface Layer:

- Provides users a way to converse with Snowflakes Data while still adhering to the organizations governance policies and procedures.
- Includes a natural language chat interface; session history; visual presentation of the results; and reference information on the source of the data.
- Streamlit Apps —: A product that appends Python applications to chat based UI Dashboards including ‘scoring’ real-time risk in a procurement environment, etc.
- Custom UI (White-labeled Agents or Apps) - External Customer - Support for Single Sign-On via OAuth/SCIM by REST API/iframe.

2. Intelligence Engine - Cortex:

- An agent-based core for performing Data Reasoning with 3 components.
- The Cortex Analyst (NL→SQL) agent — Maintains a finite token context of 128K, to refine their queries based on this token context and convert the NL statement into an SL statement.
- The Cortex Search (RAG/Vector DB) agent — Provides TOP-K results based on a cosine similarity analysis with source citations and prevents ‘hallucination’ of the output from the user.
- The ProcurementRiskAgent — Performs multiple steps of reasoning with respect to Vendors.

3. AI Services Layer:

- All LLM operations will be serverless and will utilize Claude 4.5 / Llama 3.1 with a maximum latency of less than 750 milliseconds for all operations and 100% utilization of these models.
- Complete -Provides the user with the ability to generate chat, narrative, or code outputs from the same input, with ‘smart’ selection of the model to be used for that generation.
- Embed_Text - Provides a way to process and extract semantic information from Snowpipe streams, which are subsequently converted into vectorized format for use in search.
- Transcription/Sentiment/Summarization - Provides a multi-modal pipeline capability to perform call transcription and extract risk summaries from the transcriptions.

4. Data Platform Layer:

- Enterprise-level systems that connect to AI storage systems.
- Virtual Warehouses: Configured with an Auto-suspend and Elastic scaling features to help minimize costs during the processing of ‘inference’ operations.
- Semantic Models - Domain mart created with the use of DBT - Supports ‘user friendly’ abstractions.
- Security/Governance - Dynamic masking and Role-Based Access Control to manage access to sensitive data.

5. Storage Layer:

- Supports High Availability, fast searches and low data management overhead (storage).



- Micro-partitioning: Fast search access to data due to the availability of metadata.
- Zero-copy clones: Instant creation of database clones for testing purposes while using minimal storage.
- Time Travel: 90 Day Data Retention period for recovered deleted data.

The "callable layer" of Snowflake's Cortex AI is implied, as the term does not explicitly appear in the current documentation; however, its function is demonstrated via the Intelligence Engine's Agent Tools. These tools allow Users to create and execute calls to external Functions, Api's, UDF's and Actions in concurrence with SQL generation. Therefore, it provides the Capability of the Cognitive Agents in Snowflake Cortex, to give the Domain Agent the ability to interact with Enterprise Resource Planning (ERP) and Tensorflow style of intelligent Business Process Management (iBPM) in a real-time decision-making capacity that can generate alerts or Changes to entered data within predefined ranges.

The Callable Layer serves as a Reasoning Framework where Agents can dynamically bind and invoke Tools within a Rationale Chain. Upon receiving a natural-language query, the Agent first determines the intent of the user and separates the actionable items from the SQL components. Next, it identifies the Tools available, including: Functions and API's, binds the parameters of those calls, based on Contextual Argument information, and executes the calls via a Serverless Resource Design. An optional Verification Loop allows for the validation of correct results so that adjustments can be made should the Results not meet expected outcomes.

The key to the Caller Layer is the Dynamic Binding Capability; the ability to use session context when executing parameterized calls, and the Fallback Execution sequence is based on static execution of tools via Hard-Coding Tool Bindings to call Python UDFs, Snowpark ML Models and External API's defined within the Application. The sequence of events begin from an incoming Natural Language Query, which is processed through the Agent Logic, then selects an appropriate Tool(s) to respond to the request by verifying that the response is valid; if not, then the agent can substitute or re-execute the Tool that was not validated. Governance measures such as rate limits and execution logs offer compliance, security and actionable Intelligence while keeping the Data Secure inside Snowflake. With this framework, you can run domain-specific AI applications in real-time; for example, healthcare agents that are performing sensitive operations while remaining compliant [9].

The technical setup of Snowflake Intelligence follows a systematic approach starting with Account setup through the creation of Domain Agents, Integration of Python ETL, visualisation of Enterprise Deployment. The implementation of specific recommended Guidelines is necessary to ensure that you achieve effective Real-Time Intelligence in Procurement and Compliance domains. The account setup process includes verifying Role Grants, enabling Cortex AI and creating a Warehouse dedicated to controlling Costs during Agent Testing. A Schema is created to provide Serverless Scalability for Cortex workloads and isolate Agent Resources, so as to support multiple concurrent Natural Language Queries.

Domain Agents are built by defining Agent specifications in YAML format, indexing for Search capabilities, and querying the Agent for Insight. Anomaly Detection is carried out using Python ETL in which streaming Data is filtered and analysed to isolate the risks; Data such as Sentiment Analysis applied to Contract Notes also has added value. The architecture flow of the Installation Process includes several components (Kafka Streams, Snowpipe), which provide a continuous feeding of Intelligence to the Snowflake Tasks. Ultimately, a strong framework is created for managing and analysing Procurement Risks [10].

Through targeted applications in Regulatory Compliance and Procurement Optimisation, Snowflake Intelligence creates value for organisations by demonstrating measurable return on investment (ROI) and validated case studies resulting from Operational Effectiveness and Risk Management. Real-Time Vendor Risk Scoring is produced via a multi-stage pipeline using Cortex Agents monitoring Contract Data with the generation of Risk Scores based upon SLA deviations. The time saved from this process provides enterprise-level savings for Fortune 500 organisations due to a reduction in Duplicate Payments and Manual Checks.

Healthcare HIPAA Compliant implementations flag Compliance risk by analysing Patient Call Sentiment; Financial applications assess Risk from Earnings Calls and Insider Trading. Governance features, such as Time Travel and Role-Based Access Control, improve Data Security and Auditability. The transition from traditional Business Intelligence (BI) to Snowflake Intelligence is associated with dramatically reduced Query Latency, substantial Cost Reductions and better analyst Productivity. Case studies provided illustrating potential Real-Time Scoring applications for eCommerce



and Predictive Analytics for Emergency Management demonstrate effective Operationalisation of Insights across Enterprises, as illustrated below in Table 1:

Metric	Before (Traditional BI)	After (Snowflake Intelligence)	Annualized Impact
Query Latency	4-24 hours (ETL+manual SQL)	P95 <30 seconds	500x faster insights
BI Infrastructure	\$450K (Tableau+ETL+storage)	\$220K (serverless compute)	51% cost reduction
Analyst Productivity	1,200 hours/year data prep	300 hours (agents handle 75%)	4x leverage (\$240K value)
Fraud/Compliance	1-2% detection rate	85% true positive rate	\$1.8M recovered
Data Movement	\$120K/year ETL pipelines	\$0 (zero-copy analytics)	100% elimination

Table 1: ROI Metrics: Quantified Business Value

Snowflake Intelligence (SI) uses serverless please computing/technology and optimization tactics to help solve scalability issues that occur when being deployed (credit overages and agent scope exclusions). Examples of these practices include using FIM, RAG (Reliable AI Governance) chunking, and security practices to help ensure that procurement and compliance solutions using Domain AIs will have a production/production quality level of reliability. SI has faced scalability issues surrounding unpredictable Cortex credit usage causing daily credits to run out quickly, as well as agents sticking outside of the assigned scope, which often results in excess monthly credits. To solve these issues, organizations can utilize features like the following:

- Credit usage can be tracked using QUERY_TAGS and ACCOUNT_USAGE
- The YAML configuration of tables can be used to define agent scope
- Utilize SERVERLESS_TASKS for concurrent execution; thus eliminating the need for additional provisioning, etc.

In terms of optimization, it is critical to select an accurate embedding model (each model type has different levels of accuracy and expense). Therefore, chunking methods such as fixed size and semantic chunking can lead to significant improvements in retrieval accuracy for Domain AIs. Additionally, organizations must first establish security measures (i.e., make sure users are granted only the correct role permissions, create dynamic masking policies, and maintain audit trails). SI provides detailed guidelines needed to deploy Domain AI solutions in production including optimal warehouse configurations, scoped agent configurations, and masking policies for calls to Cortex.

Investment into Snowflake Intelligence (SI) has shown greater returns than investing in traditional Business Intelligence (BI) solutions. Examples of different datasets available for visualizing these advantageous returns are Bar Charts, Line Graphs, and Gauges. A second dataset is used in comparing SI's ROI with traditional BI, which shows a significant drop in latency from 7200 seconds to 30 seconds (99.6% decrease), with substantial cost savings and increased productivity for data analysts using these models. A line graph shows continued monthly improvements in AI utilization; projections indicate AI will represent 45% of total usage by Q3 2026.

A third unit of analysis used to illustrate how Fanatics transitioned from “manual” (pre-automated) processes to those using automation, showed that Fanatics was able to make significant strides in both the number of days and rate of conversion per campaign. Another analysis used was a (box plot) illustrating the distribution of latency based on queries. The dashboard discussed in the paper for this case study includes a graphical representation of ROI comparison via bar charts, line graphs, and an aggregated view of these same datasets through the use of a "stacked" case study format—each designed to serve as a compelling visual for corporations looking into implementing Domain AI. The example screenshot is shown in Figure 2.

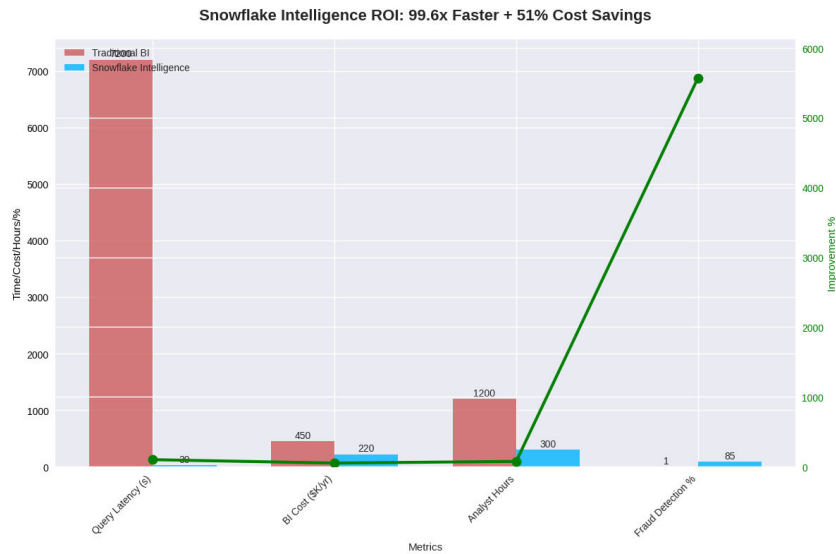


Figure 2: Snowflake Intelligence ROI

V. CONCLUSION

Businesses are confronted with many obstacles when it comes to BI, including data silos, long lead times for getting information out to end users, and the inability to adequately detect fraud. With the new Cloud-Native Architecture, combined with Domain-AI (the ability of AI to work across multiple domains), Snowflake has developed Snowflake Intelligence, which provides a solution for these business problems while providing faster insights and better fraud detection rates by leveraging more advanced, evolutionary forms of specialized agents (i.e., data scientists using machine learning to build predictive models). Snowflake Intelligence enables businesses to separate their storage and compute environments and stream data to users with minimal latency, in a manner that does not violate regulatory compliance. As an example, this includes providing audit-ready lineage reporting and enhanced procurement act anomaly detection through Python ETL pipelines. Additionally, the Snowflake platform is designed to accelerate decision-making through automation of compliance processes, and to reduce the time and cost of transferring data within and outside an organization. Snowflake also plans to incorporate Federated Learning features into the platform to serve the healthcare sector, and to provide enhanced analytics capabilities that will allow organizations to achieve operational excellence while maintaining an adaptable compliance layer across all industries. Users can leverage the capabilities of the Snowflake platform to access insights and share best practices, enabling Snowflake to continuously improve its operational efficiencies and governance in critical areas.

REFERENCES

1. "Snowflake Real-time Analytics: A Comprehensive Guide", Rahul Sudeep, July 6, 2023, <https://appstekcorp.com/blog/snowflake-real-time-analytics-a-comprehensive-guide/>.
2. "Top 4 Challenges to Scaling Snowflake for AI", Clinton Ford, November 14 2023, <https://www.unraveldata.com/resources/top-4-challenges-to-scaling-snowflake-for-ai/>.
3. "Data Warehouse in Healthcare", Andrii Krylov, December 14, 2023, <https://kodjin.com/blog/data-warehouse-in-healthcare/>.
4. "Compute-Storage Separation Explained", Gwen Shapira, 2023-01-17, <https://www.thenile.dev/blog/storage-compute-separation>.
5. "Snowflake Cortex AI: Transforming Enterprise Data with Intelligent Analytics", Anuj Raskatla, 2024, <https://booleandata.ai/snowflake-cortex-ai-transforming-enterprise-data-with-intelligent-analytics/>.
6. "Caching vs. tiering: Comparing storage optimization techniques", Logan G. Harbaugh, 06 Jul 2017, <https://www.techtarget.com/searchstorage/feature/Caching-vs-tiering-Comparing-storage-optimization-techniques>.
7. "Understanding the different types of artificial intelligence", <https://www.ibm.com/think/topics/artificial-intelligence-types>.
8. "Introducing Llama 3.1: Our most capable models to date", July 23, 2024, <https://ai.meta.com/blog/meta-llama-3-1/>.
9. "When to Use Callable and Supplier in Java", baeldung, December 20, 2022, <https://www.baeldung.com/java-callable-vs-supplier>.



10. "Snowflake Cortex: You Guide to all things Snowflake Cortex", Brandon Gubitosa, NOV 8, 2024, <https://rivery.io/data-learning-center/snowflake-cortex/>.
11. "Dashboards Examples on Snowflake", 2023, <https://nimbusintelligence.com/2023/03/snowflake-dashboards/>.