

From Spend Intelligence to Savings: Procurement Cost Management with the Cost360 Agentic AI Assistant

Utkarsh Tripathi, Principal Data Scientist

Smart Manufacturing & AI, Micron Technology

utripathi@micron.com

Sushobhan Nayak, Associate Engineer, Data Science

Smart Manufacturing & AI, Micron Technology

sushobhann@micron.com

Abstract

Procurement in the semiconductor industry faces unique challenges. These include complex supply chains, unpredictable markets, and high operating costs. Procurement teams often rely on various tools that were created separately. This situation leads to silos that limit visibility and slow down decision making. As a result, it becomes hard to gather data, track spending, and negotiate effectively with suppliers. Cost360 Assistant is important because it can bring these fragmented processes together. This allows procurement teams to move from reactive to proactive decision making. By centralizing insights and automating workflows, it turns cost management into a strategic advantage, rather than a manual and time-consuming task. It enables Micron to manage billions in spending more efficiently and transparently. Users can access data and clear summaries that aid in making informed choices. The solution features Agentic AI for independent reasoning and workflow management. It identifies savings opportunities and simplifies actions. Built on Snowflake Cortex, it uses three main services: Cortex Analyst for interpreting natural language queries and structured analysis, Cortex Agents for carrying out multi-step tasks, and Cortex Search Service for swiftly retrieving both structured and unstructured data. This integrated approach removes silos, speeds up analysis, and gives procurement teams valuable insight into better supplier negotiations. The paper examines how Cost360 Assistant changes procurement practices, increasing agility, improving resource use, and enabling sustainable savings in a competitive market.

Keywords: AI Agents, Cortex Analyst, Cortex Search, Cortex Agents, Procurement, Snowflake, Workflow Automation

I. INTRODUCTION

Procurement in the semiconductor industry has changed from a transactional role to a key factor in

competitiveness. However, this shift faces obstacles such as fragmented systems, unclear supply chains, and unstable market conditions that require quick, data driven actionable decisions [1]. Traditional procurement tools often work separately, creating visibility issues and slowing down negotiations. This ultimately affects cost efficiency and the ability of organizations to adapt quickly [2].

To address these challenges, companies are increasingly using AI powered solutions that streamline processes and provide usable insights. Cost360 Assistant, built on Snowflake Cortex and using Agentic AI, marks a significant change in managing procurement costs. Cortex Agents are vital to this transformation. They autonomously collect data from various procurement sources, simplify complex datasets into clear insights, and even allow users to update information easily through natural language interactions [3]. This feature ensures that procurement teams have real time visibility and control, reducing manual work and speeding up decision making. From a technical standpoint, Snowflake Cortex combines three main services to offer this functionality. Cortex Analyst interprets natural language queries and turns them into structured analytical tasks, allowing easy interaction with complex datasets. Cortex Agents carry out multi step workflows, such as pulling supplier data, summarizing spending patterns, and updating procurement records without needing manual involvement. Additionally, the Cortex Search Service provides quick access to both structured and unstructured data from multiple sources, ensuring timely and thorough insights. Together, these components form a cohesive system that breaks down silos, speeds up analysis, and gives procurement teams actionable intelligence.

This paper looks at how Cost360 Assistant changes procurement practices by improving transparency, speeding up analysis, and promoting long term savings, all while establishing a new standard for agility and resource use in semiconductor supply chains [4]

II. PRELIMINARIES

A. Problem Definition

To address procurement challenges such as fragmented tools, unclear supply chains, and reactive decision-making, the semiconductor industry needs an integrated solution. This platform would bring together data and automate workflows. It would allow for proactive strategies that improve agility and provide significant cost savings.

B. Text-to-SQL Using Snowflake Cortex

Cortex Analyst connects natural language and structured data by converting user queries into SQL commands. This removes the need for technical expertise, allowing procurement teams to get insights quickly and easily. The process uses schema aware semantic parsing, which makes sure that the generated SQL matches the underlying database schema. Formally, this can be expressed as:

$$Q_{SQL} = f(Q_{NL}, S) \quad (1)$$

Where:

- Q_{NL} = Natural Language Query
- S = Schema Context

This approach allows users to interact with data intuitively while maintaining accuracy and compliance with schema constraints [5].

C. Agentic Thinking

Cortex Agents apply reasoning to autonomously execute multi step tasks, such as gathering supplier data, summarizing insights, and updating records. This enables adaptive workflows and significantly reduces manual effort. The decision-making process of these agents can be modeled as:

$$U^{task} = \sum_{i=1}^n nw_i * p_i \quad (2)$$

Where:

- p_i = Success probability of step i
- w_i = Importance weight of step i

This formulation ensures that agents prioritize critical steps while considering their likelihood of success, leading to efficient and goal-oriented task execution [6].

D. Integrated Data Access

Cortex Search Service retrieves structured and unstructured data across procurement systems, ensuring comprehensive visibility. By combining Analyst, Agents, and Search, Cost360 eliminates silos, accelerates analysis, and provides real-time insights for better supplier negotiations.

E. Data Update (DML) using Custom Tool

Cost360 provides a custom tool that enables real-time data updates through natural language commands. This allows users to insert or modify procurement records without navigating complex interfaces, ensuring accuracy and seamless integration within existing workflows.

III. METHODOLOGY

To solve procurement challenges in the semiconductor industry, the Cost360 Assistant uses a strategy that combines Snowflake Cortex, Agentic AI, and cloud native deployment. This approach has a layered structure to simplify data intake, allow smart process automation, and make the system easy to scale for future needs. Figure 1 shows this structure, dividing it into four key areas: Data Ingestion and Processing, the Cortex Agent Workflow, Deployment on GKE, and Integration with External Data Sources.

A. Data Ingestion and Processing in Snowflake

Procurement data is often spread out and comes from various sources, such as supplier records, spending details by category, and operational logs. To combine these datasets, the ingestion pipeline in Snowflake uses a straightforward method:

1. Raw Data Capture

Procurement data is collected from different ERP systems and supplier portals into Snowflake through

secure connectors and scheduled triggers. This raw data includes supplier identifiers, purchase orders, category assignments, and spending details. At this stage, data integrity checks confirm that the data is complete and correct.

2. Enriched Layer Creation

Raw data is turned into Enriched Category Data Tables (Cost360) using dynamic tables and enrichment processes. Enrichment adds semantic tags, normalizes supplier names, and organizes category hierarchies for better processing later. Dynamic tables allow for updates, so procurement teams always have access to the latest data without needing manual refreshments.

3. Materialized Views for Curated Insights

Snowflake Materialized Views are created to provide focused insights at the category level. These views summarize spending patterns, supplier performance, and category trends, allowing for quick retrieval in analytical queries. By precomputing complex joins and aggregations, materialized views greatly reduce query delays and improve the user experience.

B. The Cortex Agent: The System's Core Intelligence

The Cortex Agent is the main intelligence and operational hub of the entire system. It manages complex, multi-step procurement workflows by integrating three key abilities: understanding user requests, finding relevant information, and carrying out tasks on its own. Its operation depends on several closely linked components that work together to provide clear and useful results.

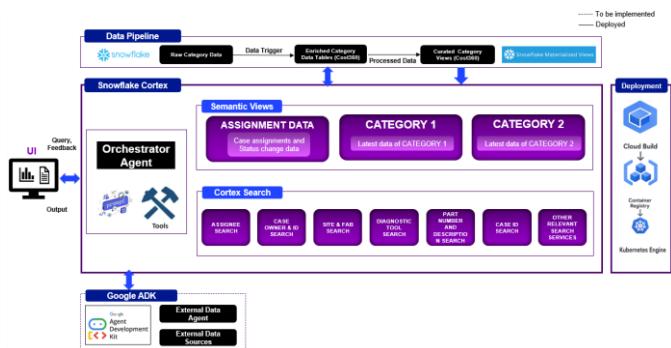


Figure 1 – Solution Architecture

1. Semantic Views

Semantic Views are an important translation layer that makes the system accessible to non-technical users. They provide structured, business friendly

representations of key procurement concepts, such as "Case Assignment Data," "Categories (Category 1)" and "Sub-Categories (Category 2)."

Rather than requiring users to understand complex database schemas, these views serve as a smart guide. They connect a user's natural language question to the specific dataset with the answer. For example, when a procurement manager asks, "Show me the latest spend for suppliers in the 'Logistics' category," the system does not see just a random string of words. It recognizes "latest spend" and "'Logistics' category" as concepts related to a particular Semantic View. This view has been set up to know which tables hold spending data and which contain supplier category information. The system can then automatically create the correct SQL query to get this information, fully shielding the user from the technical details.

2. Cortex Search Service

This component acts as a powerful, universal search engine for all procurement-related data. Its main strength is its ability to quickly scan and retrieve information from both structured and unstructured sources.

Structured data refers to organized information, like rows and columns in a database that contains purchase order numbers, invoice amounts, and delivery dates. Unstructured data includes other information, such as text within a supplier contract PDF, emails exchanged with a vendor, or comments in an operational log.

By searching both types of data at the same time, the service provides complete visibility. This saves time and reduces errors compared to the manual process where an employee searches for a purchase order in one system, then finds the related contract in a different folder, and looks for relevant emails in their inbox. This search capability ensures that no important information is missed. A very high-level diagram depicting how Cortex Analyst and Search service talk to each other is shown in figure 2 below [7].

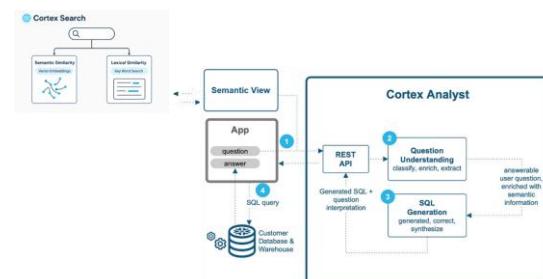


Figure 2 – Cortex Analyst with Search

3. Orchestrator Agent

The Orchestrator Agent is the main coordinator or "project manager" for the entire workflow. It interprets a user's intent and directs the other components to work together to fulfill the request. The process starts when it receives a query from the user through the Cortex Analyst interface. The Orchestrator breaks this query down into a logical sequence of smaller, manageable tasks. For instance, a request like "Summarize performance and total spend for our top three 'semiconductor' suppliers this year" would trigger a multi-step workflow:

Task 1 (Query): The Orchestrator instructs the Semantic Views to identify the top three semiconductor suppliers based on spending data.

Task 2 (Search): It then directs the Cortex Search Service to find all relevant performance reviews, quality reports, and email correspondence related to those specific suppliers.

Task 3 (Analyze & Synthesize): The Orchestrator collects the results, aggregates the spending data, and uses AI to summarize the key findings from the unstructured performance documents.

Task 4 (Deliver): Finally, it provides a single, clear, and actionable summary back to the user.

Additionally, the Orchestrator manages feedback loops. If a user corrects or refines an output, the agent learns from this feedback, improving the accuracy and relevance of its future responses. Figure 3 depicts how orchestrator agent utilizes semantic views, search service and custom tools^[8].

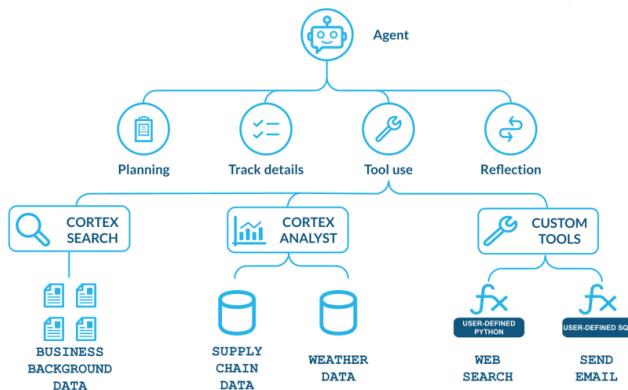


Figure 3 - Orchestrator agent

C. Custom Data Update Tool for Assignment Table

To keep case assignment records accurate, the Cost360 Assistant uses a custom data update tool with three main features:

1. Preserves a Complete Historical Record: Instead of overwriting or deleting old information, the tool always adds a new, time-stamped entry for every change. This method creates a full audit trail for each case, making it easy to see its entire history and analyze how it has changed over time.
2. Provides Flexible and Efficient Updates: The tool is built for both convenience and scale. It can handle quick, one-off changes to a single case just as easily as it can process large scale bulk updates affecting thousands of records at once. This versatility drastically reduces manual effort and saves time.
3. Integrates Seamlessly with AI and Modern Data Formats: The tool is activated automatically by the system's main AI agent whenever a user issues a command in simple language (like "*Update the owner for Case 12345*"). It uses a flexible JSON format to understand and apply these updates, which means it can adapt to future changes in the data structure without needing to be rewritten as shown in Figure 4.

```

CREATE OR REPLACE PROCEDURE DEV_BDM_EA_E.PROCUREMENT.COST360.INSERT_ASSIGNMENT_AGENT_ROW(:INPUT_ROW VARCHAR)
RETURNS VARIANT
LANGUAGE SQL
EXECUTE AS OWNER
AS
DECLARE
    input_variant VARIANT;
    input ARRAY;
    inserted_count NUMBER DEFAULT 0;
    case_ids ARRAY DEFAULT ARRAY_CONSTRUCT();
    fq_table STRING DEFAULT ''DEV_BDM_EA_E.PROCUREMENT"."COST360"."ASSIGNMENT_DATA"'';
    id_column STRING DEFAULT ''CASE_ID'';
    rec VARIANT;
    col_List STRING;
    val_List STRING;
    col STRING;
    val STRING;
    err STRING;
BEGIN
    -- FIX: Parse the input string into a VARIANT.
    input_variant := PARSE_JSON(:INPUT_ROW);

    -- Standardize the VARIANT to always be an array.
    arr := IF(TYPEOF(input_variant) = 'ARRAY', input_variant, ARRAY_CONSTRUCT(input_variant));

    LET rs RESULTSET := (SELECT VALUE FROM TABLE(FLATTEN(INPUT >> :arr)));
    LET ci CURSOR FOR rs;

    FOR r IN ci DO
        rec := r.VALUE;

        -- Build dynamic column and value lists from the JSON object
        SELECT
            LISTAGG(''||UPPER(f.key)|| '|| ''|| , ''|| )
            LISTAGG(|| f.value || , ''|| )
        INTO col_List, val_List
        FROM TABLE(rec);
    END FOR;
    -- Insert the row
    INSERT INTO :fq_table (:id_column, :col_List)
    VALUES (:case_ids, :val_List);
    inserted_count := inserted_count + 1;
    case_ids := ARRAY_CONSTRUCT(case_ids, :case_id);
END
  
```

Figure 4 – Custom Tool for Data Update

D. Agentic Reasoning and Decision-Making

To operate efficiently, Cortex Agents use a method of weighted decision making to prioritize tasks. This isn't just a "first-in, first-out" queue; it's a calculated judgment based on urgency and feasibility, following the formula in equation 2. The formula can be described as:

1. w_i (weight) shows how important a given task step is. For example, updating a master supplier record with

new banking information is critical and would receive a very high weight (w). Generating a weekly summary report, while useful, is less crucial and would have a lower weight.

2. p_i (probability) indicates the agent's confidence in completing that step successfully. A simple database lookup has a very high probability of success ($p \approx 1.0$), while extracting specific data from a low-quality scanned document might have a much lower probability.

By multiplying the importance by the probability of success for each possible step, the agent calculates a utility score. It then carries out the task with the highest score first. This smart prioritization ensures that the agent focuses its resources on the most critical and achievable tasks first, improving its efficiency, reliability, and overall impact on the procurement process.

E. Deployment on Google Kubernetes Engine (GKE)

To ensure scalability, resilience, and global accessibility, Cost360 Assistant uses containerized microservices on Google Kubernetes Engine (GKE). Cloud Build automates CI/CD pipelines. This enables rapid deployment of updates. Containerization packages Cortex components and orchestration logic into isolated units for portability and consistency. Kubernetes Engine provides dynamic scaling, fault tolerance, and load balancing to manage varying workloads.

IV. RESULTS

The deployment of Cost360 Assistant has brought significant improvements in procurement efficiency, especially with bulk case updates and automated case summaries. Previously, procurement professionals spent a lot of time manually updating individual cases and collecting data from various systems to find savings opportunities. These tasks took a lot of time and were also prone to mistakes and delays, which affected speed in negotiations. The bulk update feature, along with the Orchestrator Agent and custom Snowflake procedure, allows procurement teams to update many cases at once. This capability cuts update time from several hours to just a few minutes while keeping historical accuracy by adding new line items with timestamps instead of replacing existing records. This ensures a complete history for compliance and auditing. Being able to perform bulk updates during organizational

changes or priority shifts helps procurement managers respond quickly to changing business needs.

Similarly, the case summary function provides immediate visibility into potential savings opportunities. Instead of navigating through multiple ERP systems, spreadsheets, and supplier portals, procurement professionals can now access all insights in one place. This feature greatly cuts down the time needed to identify cost-saving opportunities from several hours to just a few minutes, allowing teams to act on multiple opportunities within the same day. By automating data collection and presenting actionable information, Cost360 enhances decision quality and confidence during supplier negotiations.

Overall, these improvements lead to measurable gains in productivity. Manual updates and data collection previously took up 30-40% of a procurement professional's time. With Cost360, this drops to less than 10%, freeing up resources for more strategic work. Quick access to insights and smoother workflows enables procurement teams to act proactively, boosting agility in a fluctuating semiconductor market. Quantitatively, the solution achieves a reduction of over 95% in time for retrieving case summaries and over 99% in time for bulk updates, resulting in a 30-40% increase in effective working hours for strategic tasks. A high-level comparison of before vs after cost360 implementation is shown in table 1 below.

Table 1 – Productivity Impact

Activity	Before Cost360	After Cost360	Improvement
Update 100 Case Assignments	~3 hours (manual updates)	< 2 minutes (bulk update)	>99% faster
Identify Savings Opportunities	~4–6 hours (manual data gathering)	< 10 minutes (case summary)	>95% faster
Time Spent on Manual Tasks	30–40% of working hours	<10% of working hours	30–40% productivity gain

V. CONCLUSION

The proposed Cost360 Assistant, built on Snowflake Cortex and powered by Agentic AI, offers a new way to manage procurement costs. It combines data ingestion, semantic search, autonomous workflows, and real time updates into one platform. Unlike traditional procurement systems that work in isolation and need a lot of manual effort, Cost360 automates everything from data collection to actionable insights and timely updates. A custom bulk update tool improves operational flexibility, allowing procurement teams to update hundreds of cases in seconds while keeping a complete lifecycle history for compliance and analytics.

This approach is unique because it uses natural language interaction, schema-aware Text-to-SQL conversion, and agentic reasoning. This helps the system not only respond to queries but also find savings opportunities before they are noticed. By using semantic views and coordinated workflows, procurement professionals can quickly see cost drivers and negotiation points. This reduces the time it takes to act on multiple opportunities from days to minutes. The result is a marked increase in productivity, decision-making quality, and organizational flexibility.

While this paper focuses on the semiconductor industry, the method and structure can easily scale and apply to other sectors. Large manufacturers in automotive, aerospace, electronics, and consumer goods face similar issues with fragmented systems, complex supply chains, and cost challenges. Cost360's integrated design can be tailored to these industries, helping procurement departments centralize insights, automate repetitive tasks, and unlock strategic savings on a large scale. By changing procurement from a transactional role to a data driven strategic function, this solution sets a new standard for efficiency and competitiveness in global manufacturing.

VI. FUTURE WORK

While Cost360 Assistant has shown significant improvements in procurement efficiency, there are several areas for future improvement that can elevate its capabilities further. The most important advancement will be integrating with external data sources and APIs using Google Agent Development Kit (ADK). This will allow the system to include real-time market intelligence, supplier risk scores, commodity pricing, and logistics data from third party platforms. By combining internal spend analytics with external signals, procurement teams can shift from reactive cost management to predictive and prescriptive strategies. Future work will also focus on:

1. Adaptive Workflows: Improving Cortex Agents to adjust procurement strategies based on external market conditions and risk signals.
2. API-Driven Data Fusion: Using Google ADK to connect easily with global procurement networks, financial indices, and sustainability metrics.
3. Advanced Analytics: Adding machine learning models for forecasting supplier performance and

spotting early warning signs for supply chain disruptions.

These improvements will change Cost360 into a complete procurement intelligence platform. This will help organizations achieve sustainable savings, reduce risks, and stay agile in competitive markets.

ACKNOWLEDGEMENT

We also extend our gratitude to Niladri Sekher Karmakar, Li Ye, Guo Jian Cheong, Joey Chong, Tong Jia from the Smart Manufacturing & AI team along with Procurement team members for their valuable ideas and suggestions during the development of the solution.

REFERENCES

- [1] Smith, A., & Lee, B. (2023). *Strategic Procurement in High-Tech Industries*. Journal of Supply Chain Management.
- [2] Johnson, R. (2024). *Breaking Silos: The Future of Procurement Tools*. Global Procurement Review.
- [3] Brown, T., et al. (2025). *Agentic AI in Enterprise Workflows*. AI Applications Journal.
- [4] Kumar, P., & Singh, R. (2022). *Cost Management, Procurement Management, Project Management*.
- [5] Yicun Yang, Zhaoguo Wang, Yu Xia, Zhuoran Wei, Haoran Ding, Ruzica Piskac, Haibo Chen, and Jinyang Li. 2025. Automated Validating and Fixing of Text-to-SQL Translation with Execution Consistency. Proc. ACM Manag. Data 3, 3, Article 134 (June 2025), 28 pages. <https://doi.org/10.1145/3725271>
- [6] Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
- [7] <https://docs.snowflake.com/en/user-guide/views-semantic/overview>
- [8] <https://docs.snowflake.com/en/user-guide/snowflake-cortex/cortex-agents-manage>