

Snowpark Unleashed: The Data Leader's Guide to Revolutionizing Collaboration & Efficiency in the Snowflake Data Cloud

October, 2023



Table of Contents



■ Legacy Technology vs. The Snowflake Data Cloud

Page 2

■ What is Snowpark?

Page 3

■ Delivering Dramatic Improvements

Page 4

■ Innovating Swiftly

Page 8

■ Collaboration in New Environments

Page 10

■ Bibliography

Page 15



Legacy Technology vs. The Snowflake Data Cloud



One of the biggest challenges facing organizations today is the shrinking availability of tech talent. As technology becomes more complex, the demand for skilled workers is outstripping the supply. As a result, many organizations struggle to find the specialized skills they need to maintain and develop their technology infrastructure. Those hit hardest by this shrinking supply and heightened demand environment are those organizations that rely on legacy solutions.

Legacy technology may feel like a safe-zone for leaders—after all, these solutions have been in place for years—but in the modern data landscape, the limitations of legacy technology outnumber the benefits. Legacy tech systems are expensive to maintain, require specialized expertise, and can be slow to adapt to changing business needs. In today's fast-paced technological landscape, relying on legacy technology can hinder an organization's ability to innovate and compete.

Legacy technology also dramatically increases the size of the team an organization needs to maintain their data, creating an unfillable gap in the availability of specialized expertise across a company. As technology becomes more complex, it will be increasingly difficult to find engineers with the necessary skills to manage and develop these legacy systems.

Consequently, many organizations facing large tech debt are reliant on a small group of experts, a situation that is both a risk to any project and a bottleneck for innovation and development.

As legacy systems age, they become increasingly difficult and expensive to maintain. In contrast, cloud data warehouses like Snowflake's Data Cloud technology offer a more cost-effective and flexible solution. With cloud data warehouses, organizations can leverage the power of large-scale computing resources without having to invest in expensive hardware or hire large teams of engineers to manage it. This enables organizations to scale their technology infrastructure up or down as needed, allowing them to adapt quickly to changing business needs or market climate.

Snowpark takes Snowflake's Data Cloud technology one step further. Snowpark is a cloud-native framework that enables organizations to develop data solutions using a process-first, enablement approach. In this ebook, we will address the needs of CIOs, CTOs, CDOs, CEOs, and other data practice leaders hoping to keep their organizations competitive in the fast-paced technological landscape. Discover a roadmap that focuses on the challenges of traditional infrastructure while offering modern data solutions to overcome them.

2

What Is Snowpark?

When Snowflake developed Snowpark, they set out to solve the challenges of an entire generation of business leaders. If the availability of tech talent is the problem, Snowpark is the solution. For leaders struggling to build a team that's versed across an entire host of coding languages and unique technologies, Snowpark makes it easy for teams to apply their skills to diverse business needs.

Snowpark is a new developer framework for Snowflake, composed of both a client-side library and a server-side sandbox. It brings deeply integrated DataFrame-style programming to the languages that developers like to use, such as Java, Python, and Scala.

Snowpark simplifies the development process by reducing the need for specialized expertise, enabling engineers to work in the language and tools they are most comfortable with. It also provides built-in features and capabilities that streamline the development process. Organizations can leverage cloud data warehouses without the need for Spark or DBX expertise.

Snowpark's reference architecture also streamlines the development process, making it easier for engineering teams to develop data pipelines and applications while reducing the time and resources needed to build and deploy solutions.



One of the key advantages of Snowpark is its ability to work seamlessly with other Snowflake products. The Snowflake Data Cloud provides a fully managed, cloud-native solution for data storage, processing, and analytics. Because Snowpark operates entirely within the Snowflake ecosystem, organizations can safely and securely develop end-to-end solutions that scale to meet their needs.

Snowpark also offers a variety of built-in features. For example, Snowpark includes support for common sources including JSON, CSV, and Parquet. Snowpark users can also quickly create User Defined Functions (UDFs) for their own custom logic that can be called from Snowpark. On the whole, Snowpark represents a dramatic intervention into the development lifecycle for cloud leaders across industries. By leveraging Snowpark's process-first, enablement approach, organizations can reduce their reliance on specialized expertise, streamline the development process, and accelerate innovation.





3 Delivering Dramatic Improvements

When it comes to performance, Snowpark is capable of delivering high quality outputs with increased efficiency. In fact, Hakkoda clients that migrated to Snowpark off of alternative platforms saw a **96% performance improvement**.¹ So how is Snowpark able to deliver these dramatic improvements while also saving you money? Through centralization, speed enhancements, and simple solutions paired with sophisticated tooling.

Centralization

Snowpark centralizes all of your data processing in Snowflake, removing the need to transfer large volumes of data between Snowflake and external processing engines. Reducing the amount of data transferred between your client and the Snowflake database leads to improved performance.

How is centralization possible? Through the use of scalable, resizable warehouses that allow for cost control per second (credits are billed per second).

This reduces the burden of finding and retaining specific engineers in a competitive labor market, and enables your users to learn and operate in Snowflake directly. By giving users access to data without the need to work with analysts or IT teams, you reduce friction, increase productivity, and facilitate faster results.

Speed and Performance

In the case of Snowpark, improvements to speed are a direct result of improved processing performance. For example, the Snowflake platform is exceptionally effective at **running SQL, ML models, and ETL transformations**, eliminating the need to cycle through thousands of lines of code. Such effectiveness allows teams to spend less time provisioning. By removing the need to wait for servers and pipelines to be created, Snowpark allows users to instantly connect to a Snowflake warehouse and run their code.

The coding language flexibility offered by Snowpark allows your team to work in the language of their choice (for example, Java, SQL, Python, Scala) without leaving the platform. Snowpark operations are also executed lazily on the server, meaning can retrieve the data when you are ready to perform an action versus data being retrieved the moment you construct dataframes.

¹ Donlin, J. A Faster, More Efficient Snowflake Takes the Stage at Snowday.



The coding language flexibility offered by Snowpark allows your team to work in the language of their choice (for example, Java, SQL, Python, Scala) without leaving the platform. Snowpark operations are also executed lazily on the server, meaning can retrieve the data when you are ready to perform an action versus data being retrieved the moment you construct dataframes.

Snowflake also offers a host of cost-effective features, including the `Explain()`² function and vectorized UDFs³. With `Explain()`, your team can view which operations Snowflake would perform in order to execute a given query. Such a function can help teams evaluate performance and speed beforehand without incurring high costs. Vectorized UDFs, on the other hand, help process batches of rows rather than row by row, improving speed and performance.

In other words, your team has full control over query execution—be it a request for data results from your database or for action on the data, or both. A query can give you an answer to a simple question, perform calculations, combine data from different tables, and add, change, or delete data from a database.

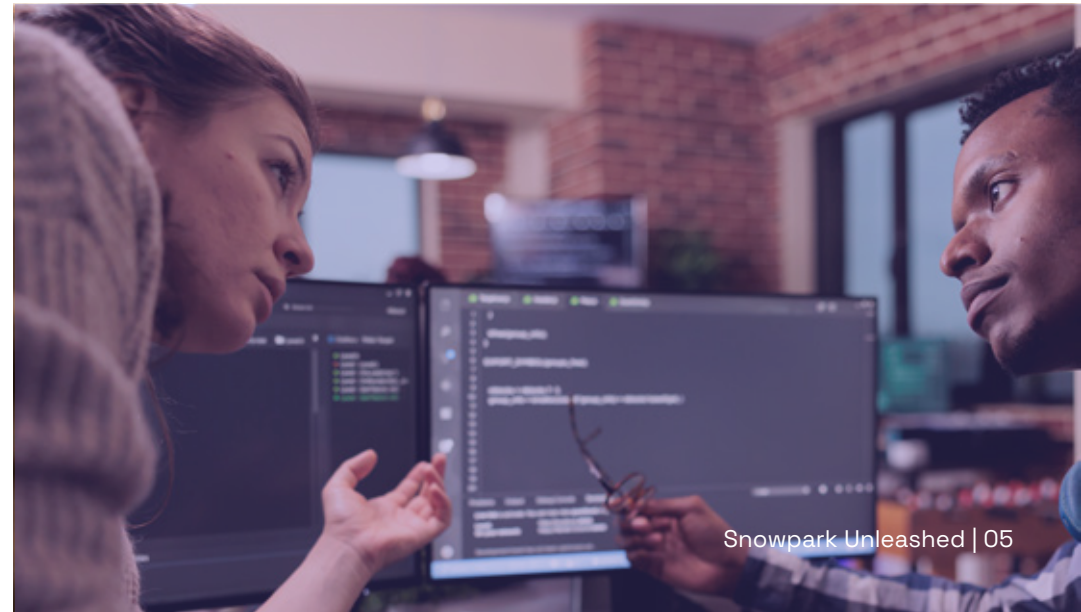
² Snowflake. `Explain()` function.

³ Snowflake. Vectorized Python UDFs.

Cost-effectiveness

We've talked about all the ways that using Snowpark can save time and money, but how about the cost of the solution itself? With some technologies, there are hidden costs lurking under the surface. But because Snowflake keeps its compute and storage price separate from the monthly fee, users have transparency and control over their total cost.

Aspects such as efficient query execution, meanwhile, help reduce cost by eliminating silos and resources dedicated to this specific task and by reducing the time your data scientist has to wait for the result of the computation.





Elevating Data Governance and Security Practices

Data governance procedures control how information is gathered, stored, processed, and disposed of. These procedures ensure that information remains trustworthy, setting standards to keep data secure, private, accurate, available, and usable across the enterprise. Because of the sensitivity of the data being gathered and managed, it is imperative that any technology being used is best-of-class and that it enables best practices.

A report done by Mordor Intelligence⁴ shows that the data governance market is poised to grow at a compound annual growth rate of 20.83% by 2028. They further state that “the growing adoption of data science in various sectors, such as IT & telecom, retail, and healthcare, is further expected to boost the demand for the data governance market.” The report’s conclusions are further supported by Cisco⁵ data, which states that around five quintillion bytes of data are produced every day. So, it is hardly surprising that one of the biggest concerns organizations face when adopting new data technology is security and governance.

⁴ Mordor Intelligence. Data Governance Market Size.

⁵ Stack, T. Cisco Data Center.

Security Practices

With Snowpark, you get to utilize Snowflake’s fully-managed services, including enterprise governance and security. Your data does not move for transformation, data science, and business intelligence reporting. It lives inside a single platform versus living on several systems, reducing data silos and eliminating risk. With everything in one place, you can have a unified data governance model that follows the same security protocols and best practices.

“A survey of technology leaders found that the data governance market is poised to grow at a compound annual rate of 21% by 2028”

– Mordor Intelligence



Through a secure environment, Snowpark allows you to develop data pipelines and easily process and integrate these pipelines into your Snowflake platform. Although pulling from third-party Java and Python libraries to build user-defined functions (UDFs) poses an additional risk, Snowpark engineers have created multiple independent layers of defense. Snowflake implements a multi-faceted security architecture that protects the kernel, the network, and the workload orchestration process.⁶

Data Governance Practices

An additional level of security provided by Snowpark is treating all network traffic from compute clusters as untrusted. In case of an attack scenario, Snowpark's combination of security-first design, on-host access controls, isolation mechanisms, and egress proxy offers a high degree of protection from malicious code. By using the power of machine learning with curated data libraries, you can build scalable, secure data pipelines and machine learning workflows right where your data lives.

You can also expand access while increasing control, easily giving the right people access to the right data at the right time. Snowpark also proactively protects against compliance issues, and allows you to gain clear visibility into your data while enforcing security policies and granular access controls.

Take Snowpark's approach to access control as an example. Snowpark combines aspects from Discretionary Access Control (DAC) and Role-based Access Control (RBAC). In other words, each object has an owner with access privileges, and access privileges are assigned to roles and users. At the same time, Snowpark can generate data access policies for query runtimes, specifically for row-level access. Snowpark also offers dynamic data masking, where sensitive data is loaded and encrypted at the time of query for unauthorized users.

Whenever a column with a masking policy is attached, Snowflake rewrites the query at runtime. This means that authorized users can view sensitive data in plain text, while unauthorized users will see it tokenized or masked. Dynamic data masking can work on more than one column by using several values to mask sensitive data.

⁶ Halcrow, M. Snowpark Security.

4 Innovating Swiftly

Snowpark helps drive results and innovation with machine learning for data science and AI-driven applications. Snowpark's performance speed, a critical factor in supporting robust ML models, married with its tight integrations to Spark, Python, Qubole, etc., makes it an indispensable machine learning technology. Moreover, Snowflake's partnership with Anaconda gives users curated data libraries where datasets are FAIR (Findable, Accessible, Interoperable, and Reusable), adding value by enhancing data sets for current use and future reuse.

The impact of Snowpark ML on machine learning is evident through a series of prominent features that reshape the way data scientists approach model development, testing, training, and storage within Snowflake pipelines. Snowpark ML offers a Python software development kit (SDK) with a range of APIs to support the entire machine learning lifecycle.

- The first component, **Snowpark ML Development**, empowers data scientists with Python APIs to efficiently develop models directly within Snowflake. It incorporates features like data preprocessing, feature engineering, and model training, utilizing Snowpark Optimized High Memory Warehouses for scalable data transformations on large datasets.
- The second component, **Snowpark ML Ops**, complements the development API and enables model management and integrated deployment into Snowflake. Notably, the FileSet API materializes data into a Snowflake internal stage, while the Model Registry facilitates ML model deployment into Snowflake Warehouses as vectorized UDFs, enhancing their usability and efficiency within the Snowflake environment.
- One standout feature is the introduction of **Snowpark Container Services**, a runtime option allowing secure deployment and execution of sophisticated AI models and full-stack applications within Snowflake. This simplifies compute and cluster management, granting developers and data scientists the ability to concentrate on solving business problems without the complexity of managing separate tools.





At Hakkoda, we have been able to harness the power of Snowflake's near-unlimited secure access to data and its elastic processing engine paired with Anaconda's data science workflows to develop cutting-edge applications. These applications are opening up a new array of data capabilities, and we are seeing incredible results.

Innovation in the Financial Services Industry

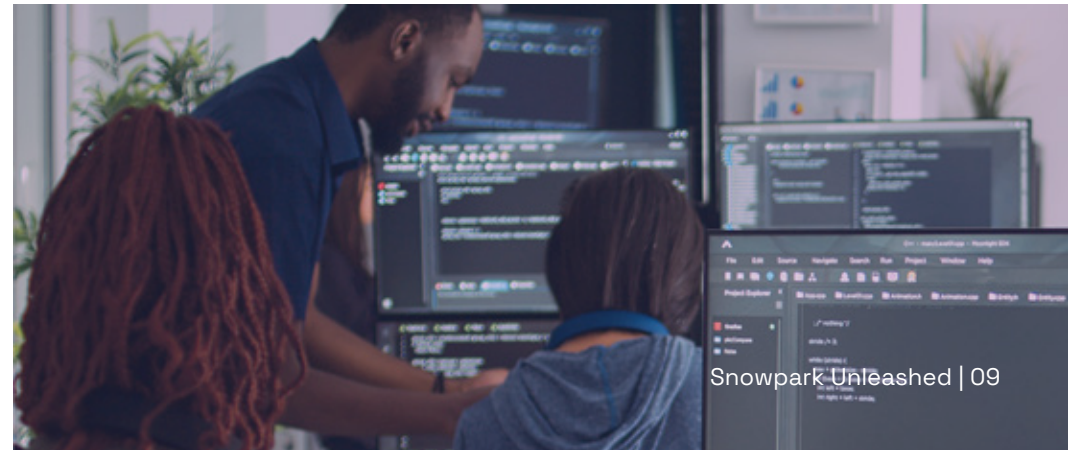
Many financial services organizations are evolving past Excel spreadsheets and siloed ML code to centralized pipelines, delivering powerful analytics and ML capabilities. These ML models are now forecasting, providing powerful early fraud detection, and running analytics on a secured, data-governed central location in the user's language of choice. The financial services industry is revolutionizing in real time and Hakkoda, with our expertise across Snowflake and modern data stack, is able to help our clients lead the innovation.

Developing a Data Tokenization Native Application

CoreLogic, a data analytics and technology company in the real estate industry, partnered with Hakkoda to revolutionize the industry's approach to data analytics with Snowflake. The collaboration between CoreLogic and Hakkoda resulted in the development of the CoreLogic Integrated Property (CLIP) application.

CLIP assigns unique identifiers to properties, enabling users to access comprehensive historical data, valuations, ownership changes, tax information, and environmental data. This user-friendly application empowers clients to gain valuable insights from their property portfolios, saving time and effort on data management. CoreLogic and Hakkoda plan to expand their capabilities on the Snowflake platform, including property climate risk analytics and precision marketing tools based on property data and behavioral patterns.

Overall, the successful partnership between CoreLogic and Hakkoda demonstrates the transformative potential of specialized systems integrators and multi-cloud strategies in the real estate industry. By leveraging Snowflake's platform and Hakkoda's expertise, CoreLogic delivers exceptional value and empowers clients with seamless access to comprehensive property data and innovative analytics tools.





Innovation in Healthcare

Data has the ability to help solve some of the world's toughest challenges, and cancer prevention is definitely on that list. Although it is rated as the number one employer health concern, there are very few disease-specific solutions. As a matter of fact, cancer accounts for 10-12% of employer healthcare spending, despite impacting only 1-2% of their health plan members.

The average annual cost per cancer patient exceeds \$150,000, which is six times more than other high-cost conditions. Addressing a very small number of individual cases can have a significant impact on your bottom line. For example, the top 5% of cancer cases typically represent 50% of cancer spend and the top 20% account for over 95% of cancer spend.⁷

⁷ Puuri, C. and Savacool, M. Fighting Cancer with Snowflake.

⁸ Ibid.

Armed with the information above, Hakkoda saw a need for a way to predict cancers in a given population, which in turn would help self-insured companies and their employees better manage the corresponding financial and healthcare burdens.

Improving Cancer Detection on Snowpark

A healthcare startup providing concierge medical services and specializing in highly acute cancer cases partnered with Hakkoda to use machine learning to project cancer case numbers. Hakkoda helped them streamline their provider database and improve the delivery of high-quality care by collaborating with top doctors and healthcare professionals.

Not only did Hakkoda collaborate in cleaning up provider data and creating a data roadmap and analytics architecture aligned with the organization's strategic goals; they also improved the accuracy and performance of the previous codebase, thus democratizing the data.

In other words, members within the healthcare startup could execute the model easily. Finally, our data experts provided guidance on data governance, data quality, and training the organization's data team to ensure best practices in data management.

As stated by the organization's Chief Clinical Officer, "From a health care economics perspective, the tool enables an understanding of the specific potential cancer burden that allows for a proactive approach to prevention through screening and also allows for a proactive approach in the design or selection of a network that best meets the needs of their employees." By leveraging the power of Snowpark as well as other tools in the modern data stack, the healthcare startup was able to approach cancer prevention under an innovative scope.

5 Collaboration in New Environments

Data Clean Rooms

Snowflake's overall architecture and capabilities can be leveraged to create data clean room environments. The following capabilities can be combined with Snowpark to establish secure and isolated environments:

- **Virtual Warehouses:** Snowflake's virtual warehouses enable the isolation of compute resources for specific workloads. This means you can allocate dedicated resources for data clean room activities, ensuring performance and isolation from other workloads.
- **Role-Based Access Control (RBAC):** Snowflake's RBAC model allows fine-grained access control to databases, schemas, tables, and even individual columns. You can define roles and privileges to restrict access and enforce data governance policies within the data clean room.
- **Data Masking and Encryption:** Snowflake provides features for data masking and encryption to ensure sensitive data remains protected. You can configure policies to mask or encrypt certain columns or rows of data, further enhancing security and privacy within the clean room.

"Capitalizing on the power of Snowpark as well as other tools in the modern data stack, a healthcare startup was able to leverage machine learning at scale to deliver cancer prevention data to employers."

With Snowpark, you can create data clean room environments where custom code can be executed to perform advanced analytics and transformations on data, while ensuring data security, privacy, and isolation.

Snowpark also allows you to leverage familiar programming languages and libraries to work with the data directly within Snowflake, making it easier to implement data clean room workflows.

Moving Across Data Environments

The combined use of Snowpark and Snowflake's features can help minimize the need for data movement across environments. By leveraging Snowflake's capabilities for data sharing, secure data exchange, and in-database processing, organizations can enable collaborative data analytics and processing without the need to move large volumes of data between environments, reducing costs, breaking down silos, minimizing complexity, and improving security.

Streamlit and Snowpark

Streamlit is a powerful solution for building interactive data applications and performing advanced data processing within Snowflake.

Streamlit is an open-source Python library that allows you to create web-based user interfaces for data applications, while Snowpark enables custom code execution and data transformations within Snowflake. Here's how Streamlit can be used with Snowpark:

- **Data Exploration and Visualization:** Streamlit provides an intuitive and interactive way to create visualizations and explore data. You can leverage Snowpark to query and process data within Snowflake, and then use Streamlit to present the results in interactive charts, tables, or custom visualizations. Streamlit's simplicity allows you to quickly iterate and experiment with different visualizations, enhancing data exploration capabilities.
- **User Input and Dynamic Interactions:** Streamlit allows you to create user input widgets such as sliders, dropdowns, and text boxes. These widgets can be used to dynamically interact with Snowpark code running in Snowflake. For example, you can create a Streamlit interface that allows users to input parameters for a Snowpark data transformation or analysis script, providing an easily accessible way to interact with the underlying Snowflake data processing.

- **Real-Time Dashboards:** Streamlit can be used to build real-time dashboards that display data processed by Snowpark in Snowflake. By continuously querying and updating data from Snowflake, Streamlit can present dynamic visualizations and metrics to users, giving them up-to-date insights and analysis.
- **Collaboration and Sharing:** Streamlit applications can be easily shared and deployed, making it convenient for collaborating with team members or sharing results with stakeholders. You can build Streamlit apps that incorporate Snowpark code to perform complex data transformations, enabling others to interact with and explore the data using a user-friendly interface.

With Streamlit and Snowpark, you can create interactive and easy-to-use data applications that leverage Snowflake's data processing capabilities. Streamlit handles the user interface and visualization aspects, while Snowpark enables the execution of custom code within Snowflake, allowing for advanced data transformations and analytics. This combination provides a powerful solution for building data-driven applications and exploring data residing in Snowflake.

Native Apps

The use of Snowflake data sharing as well as UDFs and stored procs has helped data experts develop the Snowflake Native Application Framework. This allows developers to create, distribute, and utilize applications easily by leveraging the core features of Snowflake. This framework provides convenience for customers by consolidating their data and simplifying their search for effective applications.

Hakkoda helped Very Good Security (VGS) create a Snowflake Streamlit application that performs tokenization and de-tokenization in Snowflake by connecting to the VGS vault.

Hakkōda empowers customers to harness the power of Snowflake Native Apps and create secure and innovative applications.

This application expands VGS's customer base and total addressable market by offering a Snowflake Native Application through the Snowflake App Marketplace. By leveraging Snowflake's application building capabilities, VGS streamlines its data processing and enhances security.

The VGS Vault Tokenizer, available as a Snowflake Native App, allows Snowflake customers to tokenize sensitive data within Snowflake's Data Cloud, minimizing security risks and compliance concerns. It replaces raw data with tokens, preserving privacy and enhancing data analysis capabilities. With the Snowflake Native App Framework, developers can build and test apps natively within the Data Cloud, providing high availability and auto-scalability while reducing security obstacles. Hakkoda empowers customers to harness the power of Snowflake Native Apps and create secure and innovative applications.

Collaboration of Code

Since Snowflake's release of Snowpark in late November 2020, the solution has simplified organizations' architecture by including more pipelines into Snowflake's data platform. Doing so has allowed data professionals to improve scalability, performance, security, and collaboration. By leveraging existing skill sets, Snowpark users can improve team productivity while reducing cost. The following are some examples of how Snowpark improves code collaboration between teams:

- **Multi-Language Support:** Snowpark supports multiple programming languages such as Java, Scala, and Python. This flexibility allows developers to choose their preferred language for coding, ensuring that diverse teams can collaborate using their language of choice.
- **Integrated Development Environment (IDE):** Snowpark integrates with popular IDEs like IntelliJ IDEA and VS Code, providing a familiar and feature-rich coding environment for developers. This allows team members to work on Snowpark code within their preferred IDEs, enhancing collaboration and productivity.

- **Version Control Integration:** Snowpark code files can be managed using standard version control systems such as Git. Developers can collaborate on code changes, track revisions, and resolve conflicts using their preferred version control workflows. This ensures that the codebase remains organized and facilitates collaborative development practices.
- **Code Sharing and Reusability:** Snowpark allows developers to create and share code snippets or libraries that can be reused across projects or shared with other team members. This promotes code sharing and reuse, accelerating development and fostering collaboration by leveraging existing code components.
- **Testing and Debugging:** Snowpark supports testing and debugging capabilities, enabling developers to validate their code and identify and fix issues during development. This allows team members to collaborate effectively by reviewing and troubleshooting code together, ensuring code quality and reliability.
- **Execution within Snowflake:** Snowpark code is executed directly within the Snowflake platform, leveraging Snowflake's scalable and elastic compute resources. This enables multiple developers to work simultaneously on code execution within their separate Snowflake sessions, facilitating collaboration while utilizing the platform's power and performance.



6 Bibliography

1. Donlin, J. (2022) "A Faster, More Efficient Snowflake Takes the Stage at Snowday". Snowflake. Retrieved from: <https://www.snowflake.com/blog/performance-enhancements-making-snowflake-faster-more-efficient/>
2. Halcrow, M. (2022) "Snowpark Protection Through Java/ Scala and Python Isolation". Snowflake Medium. Retrieved from: <https://medium.com/snowflake/snowpark-protection-through-java-scala-and-python-isolation-f8d10be61d56>
3. Mordor Intelligence. (2022) "Data Governance Market Size and Share Analysis". Growth Trends and Forecasts (2023 - 2028). Retrieved from: <https://www.mordorintelligence.com/industry-reports/data-governance-market>
4. Puuri, C. and Savacool, M. (2023) "Fighting Cancer with Snowflake: A Data Apps Success Story". Snowflake Summit 2023. Las Vegas, Nevada.
5. Snowflake. (ND) EXPLAIN(). Retrieved from: <https://docs.snowflake.com/en/sql-reference/sql/explain>
6. Snowflake. (ND) Vectorized Python UDFs. Retrieved from: <https://docs.snowflake.com/en/developer-guide/udf/python/udf-python-batch>
7. Stack, T. (2022) "Internet of Things (IoT) Data Continues to Explode Exponentially. Who Is Using That Data and How?". Cisco Data Center. Retrieved from: <https://blogs.cisco.com/datacenter/internet-of-things-iot-data-continues-to-explode-exponentially-who-is-using-that-data-and-how>