Introduction

In today's data-driven world, the rapid increase in data volume has introduced new challenges for traditional data analytics architectures. As conventional analytics infrastructure struggles to keep up with the demand for faster processing, the acceleration of AI and ML applications across industries places new demands on their ability to support the advanced analytics required by these workloads.

In response, organizations are consolidating their traditional Data Warehouses and Data Lakes into a Data Lakehouse architecture that combines the flexibility and raw data capacity of data lakes with the structured, query-optimized environment of data warehouses. In this evolving landscape, CPUs, which have been the workhorse of computing for decades, are not advancing fast enough to keep pace with improvements in network and storage performance. As a result, analytics are often bottlenecked on compute performance.

The NeuroBlade SQL Processing Unit (SPU) is purpose-built to bridge this performance gap. By offloading compute-intensive analytics tasks from the CPU, the SPU accelerates data analytics operations, resulting in enhanced performance, scalability, and cost efficiency. This specialized approach marks a pivotal advancement in addressing the challenges and the growing demands of data analytics workloads.

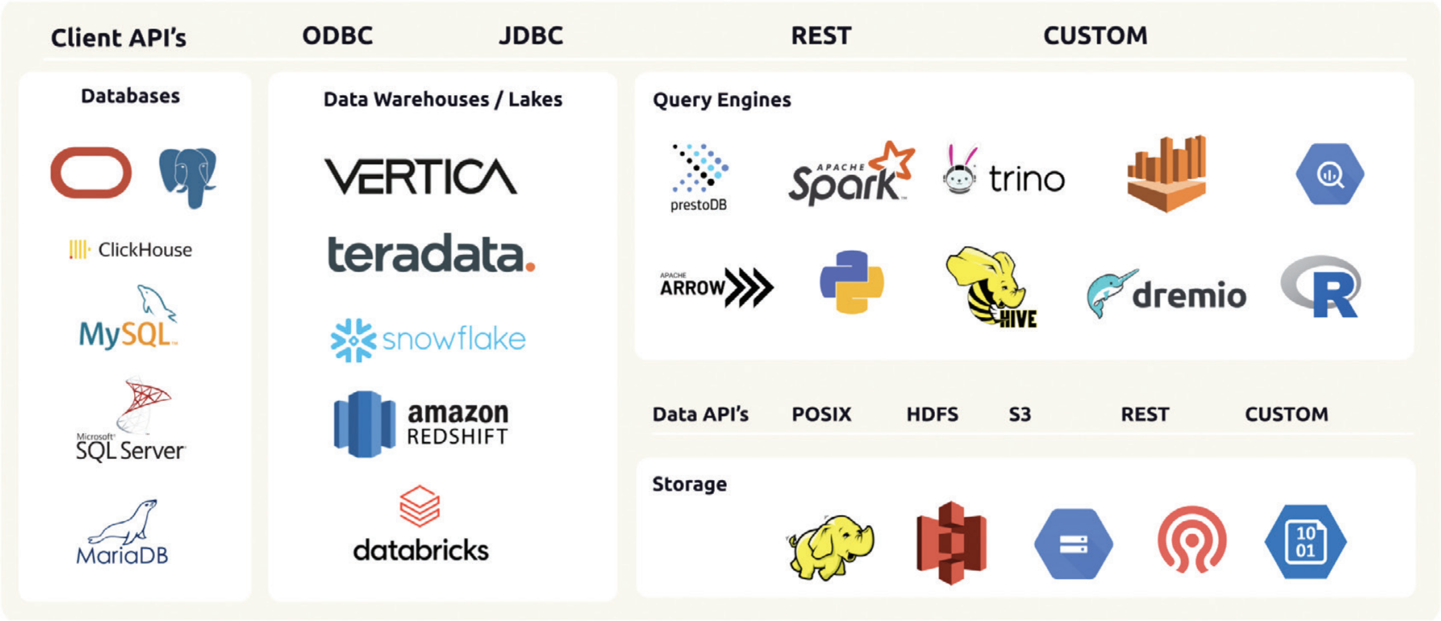


Figure 1 – Data Analytics Compute, Storage Metadata ecosystem

Modern Data Analytics – Overview

The Data Analytics architecture has undergone significant evolution in recent years, driven by the increasing data volumes as well as the growing demand for real-time analytics and insights. A key architectural evolution has been the separation of storage and compute. This has enabled storing vast amounts of data economically while dynamically scaling computing resources to meet the analytical demands, without being constrained by the physical limitations of either compute or storage.

But this new architecture results in performance bottlenecks caused by a mismatch between the performance of the compute, storage, and network components. While several architectural techniques have been developed to address these, none presents a comprehensive solution that addresses the root cause. The NeuroBlade SPU is specifically designed to provide a holistic solution that removes these bottlenecks and enhances the performance of a modern Data Lakehouse.

The Data Analytics Stack

Modern data analytics solutions encompass different tiers that contribute to efficient and scalable processing. These tiers include:

1. **Analytics Applications:** These include traditional applications such as Dashboards and Ad-Hoc queries, but is increasingly becoming dominated by AI-driven applications such as predictive analytics, recommendation systems, language processing etc.
2. **Compute Layer:** Processes data using frameworks like Apache Spark, Presto, and ClickHouse to enable distributed and parallel processing for scalability, and performance optimization.
3. **Storage Layer:** Stores and manages large volumes of data with distributed file systems and cloud object storage (e.g., Amazon S3, HDFS, etc.).

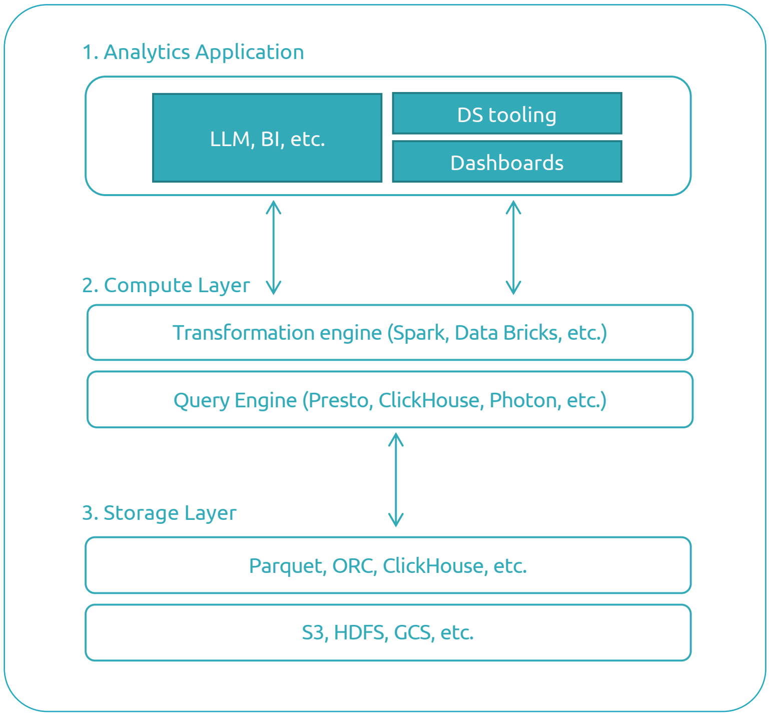

1. Analytics Application
LLM, BI, etc.
DS tooling
Dashboards
2. Compute Layer
Transformation engine (Spark, Data Bricks, etc.)
Query Engine (Presto, ClickHouse, Photon, etc.)
3. Storage Layer
Parquet, ORC, ClickHouse, etc.
S3, HDFS, GCS, etc.

Figure 2 – Modern Data Analytics Layers

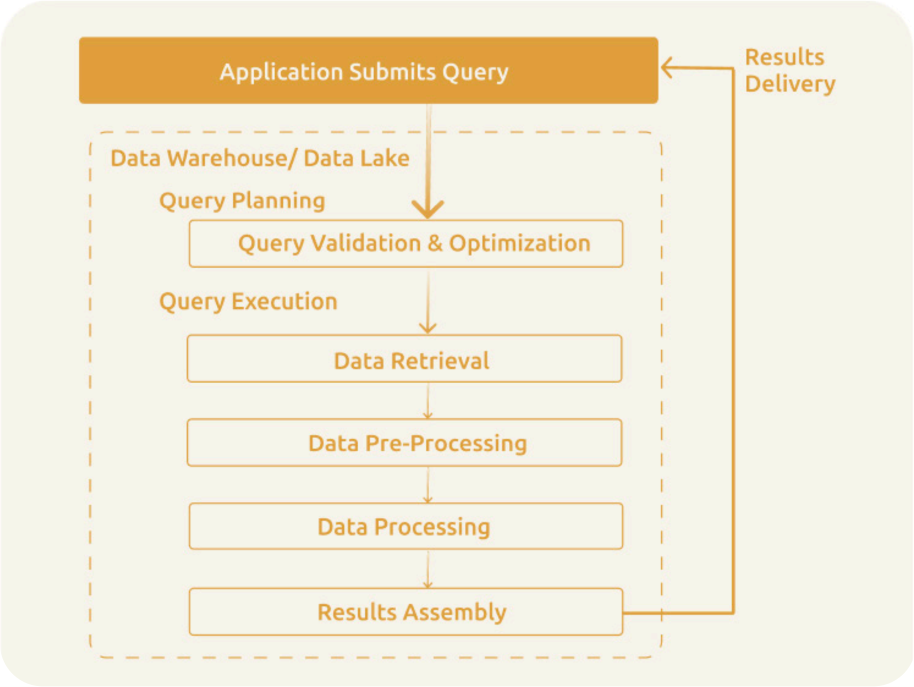


Figure 3 – Modern Data Analytics– Typical Query flow

Bottlenecks During Query Flow

The following describes the flow of events during query processing:

1. **Query Submission:** The compute layer receives a query request from a user or application. This query specifies the desired data manipulations or computations to be performed on the dataset.
2. **Query Optimization:** The compute layer optimizes the query execution plan by considering factors such as query complexity, available resources, and data distribution. It determines the most efficient way to process the query by considering techniques like query rewriting, join reordering, and predicate pushdown.

3. **Data Retrieval:** The compute layer interacts with the storage layer to retrieve the relevant data needed for the query execution. This may involve accessing data from distributed file systems, cloud object storage, or other storage systems.

4. **Parallel Processing:** The compute layer performs parallel processing on the retrieved data. It distributes the query workload across multiple compute nodes or cores to leverage parallelism and optimize query performance. This can involve splitting the query into smaller tasks and executing them concurrently.

5. **Data Aggregation:** The compute layer applies filters, aggregations, and other operations specified in the query. It processes the data according to the query logic, performing tasks such as data joins, groupings, sorting, and calculations.

6. **Result Assembly:** As the compute layer completes the query execution, it gathers the intermediate results generated by different compute nodes. It combines and merges these intermediate results to form the result set.

7. **Result Delivery:** Finally, the compute layer delivers the query result back to the user or application that initiated the query. This can involve sending the result directly or storing it in a designated location for subsequent retrieval.

Throughout this query flow, the compute layer interacts closely with the system memory and storage layer, leveraging parallel processing, and optimizations to efficiently execute the query. The result is obtained through the coordination of various computational and storage resources, providing users with the desired scale and performance. While there are many software-oriented approaches that the industry has taken to make the flow efficient, there are two major architectural bottlenecks arising from the underlying hardware that hinder efficient query execution that need to be addressed:

Bottleneck #1 – Computation and Memory Access

One significant bottleneck arises from the CPU-DRAM latency imbalance and throughput mismatch. CPUs operate at much higher frequencies and can exchange much more data than DRAM is capable of. This results in slow down in compute with every CPU-DRAM interaction. While hardware caching in the CPU and software optimization techniques have evolved to minimize the exchanges, this, remains one of the largest bottlenecks in computation of analytics. This imbalance results in slower query execution and inconsistent response times that impact the overall efficiency and responsiveness of the data analytics system.

Bottleneck #2 – Data Retrieval from Storage

The process of sourcing files becomes another bottleneck for query execution. Retrieving data from various file sources, such as distributed file systems or cloud storage, can be time-consuming, especially when dealing with large volumes of data. This file sourcing bottleneck slows down the overall query execution process, delaying the availability of data for analysis.

Addressing these bottlenecks is crucial to improving the query execution performance and overall data analytics experience. By implementing acceleration techniques and caching mechanisms, organizations can overcome these limitations, enabling faster computation and memory access, as well as more efficient file sourcing. These optimizations reduce execution times, improved consistency, and enhanced scalability in handling large datasets for data analytics workloads.

NeuroBlade Compute for Data Analytics

Addressing Bottlenecks

Analytical bottlenecks primarily emerge from two areas: the computation-memory access gap and the inefficiency of data retrieval from storage. The NeuroBlade SPU confronts these challenges head-on. It revolutionizes computation and memory access by directly reading and processing data in its original format, thereby mitigating CPU-DRAM latency issues. Simultaneously, by localizing data processing to where data resides, the SPU drastically reduces the overhead of data retrieval, ensuring swift and efficient analytics.

Addressing Bottleneck #1 – Computation and Memory Access

The SPU reads data in its original format, processes it and outputs directly in the memory format expected by the query engine. The entire query is processed on chip in the SPU eliminating the CPU-DRAM data exchanges required for processing operations, such as decompression, deserialization, processing, and re-serialization. This method minimizes data movement, bypassing the CPU for initial data handling, which enhances system efficiency. The CPU deals only with these processed, compact datasets, streamlining the data processing workflow and optimizing overall performance.

Addressing Bottleneck #2 – Data Retrieval from Storage

The SPU efficiently reads data from storage in its native format, processes data internally, and sends only the processed results to the CPU for final stages. This strategy significantly cuts down on data transfer overhead between storage and the query engine (CPU), leading to quicker query responses and enhanced efficiency in resource utilization. By streamlining the flow of data directly from storage to processing, the system minimizes delays and maximizes performance, ensuring resources are used more effectively.

NeuroBlade plugs into the query execution flow to update the flow as follows:

1. **Query Submission:** The compute layer receives a query request from a user or application. This query specifies the desired data manipulations or computations to be performed on the dataset.

2. **Query Optimization:** The compute layer optimizes the query execution plan for both accelerated and non-accelerated queries. For accelerated queries, the SQL accelerator within the NeuroBlade solution analyzes the query structure and modifies the plan to account for its specialized processing capabilities.

3. **Data Retrieval:** The compute layer interacts with the storage layer to retrieve the relevant data needed for the query execution. This may involve accessing data from distributed file systems, cloud object storage, or other storage systems.

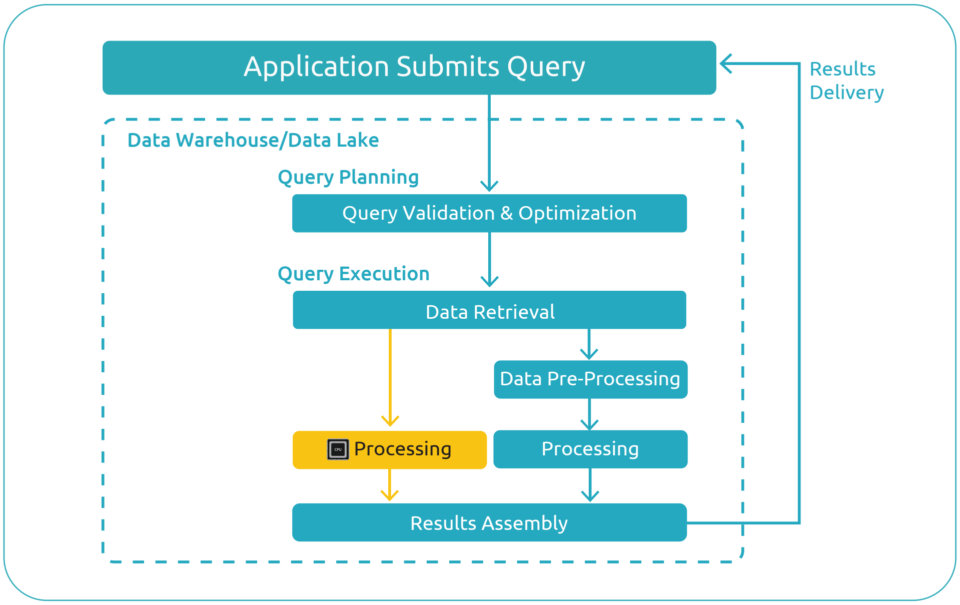


Figure 4 – Modern Data Analytics – with Acceleration path and Caching

4. **Accelerated Query Operations Execution:** Eligible queries are executed using the SPU. The SPU accesses the data stored within the analytics application database in its original format eliminating the need for Pre-Processing.
5. **Non-Accelerated Query Operation Execution:** Non-accelerated operations follow the traditional compute layer path, interacting with the analytics application database to retrieve the required data for processing.
6. **Result Assembly:** As the query execution progresses, both accelerated and non-accelerated queries generate intermediate results. The SQL accelerator and the compute layer efficiently collect and combine these intermediate results to form the result set.
7. **Result Delivery:** The compute layer, including the SQL accelerator, delivers the query results back to the user or application that initiated the query. This can involve sending the result directly or storing it in a designated location for subsequent retrieval.

Form Factor	Full Height Half Length, Single Width
SPU Processor	x1
PCIe Edge Connector	PCIe 4.0 x16 lanes
PCI Vendor ID	0x1E73
Power	150[W]
Extremal Power	8pin PCIe ATX
Thermal	Passive cooling
Environment	Storage: -40°C to 75°C Operating: 0°C to 55°C
Humidity	Storage: 5% to 95% Operating: 5% to 90%

SPU G200 PCIe Add-in Card

The NeuroBlade SPU-G200 PCIe card seamlessly integrates into data center infrastructures, offering unparalleled flexibility and ease of deployment across on-premises, shared services, or cloud environments. This adaptability makes it suitable for various infrastructure architectures, providing a broad range of deployment options.

By offloading analytics operations to a dedicated processor (SPU), the SPU-G200 enhances performance alongside CPUs in compute servers. This not only optimizes analytics processing but also reduces the need for additional servers, leading to cost savings and operational efficiency.

The SPU-G200 is designed for scale-out deployments, enabling you to handle even the most demanding analytics workloads. By providing both improved performance and increased capacity, the SPU-G200 is designed to meet your needs and seamlessly adapt to your evolving requirements.

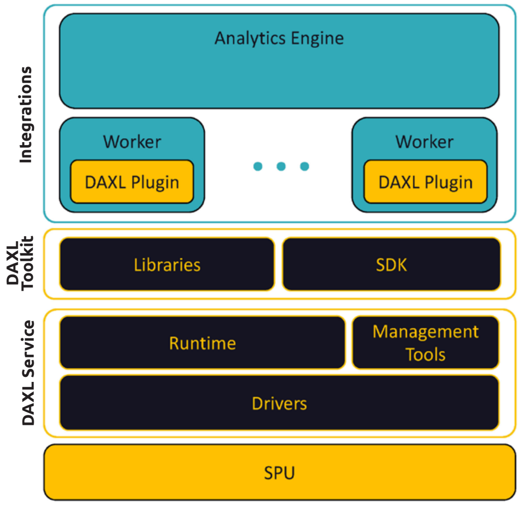
Moreover, the SPU can retrieve data from local cache or remote storage in the format stored by the data analytic system, eliminating the need for CPU-driven data transformation. This design contributes to a highly efficient and performance-oriented SQL processing solution.

DAXL Software Platform

The DAXL Software Platform must be installed on all SPU-enabled compute nodes to integrate it into your infrastructure. The DAXL Software Platform installs the following components:

- **DAXL Toolkit:** A development environment and runtime libraries for integrating the SPU with an analytics engine. DAXL Toolkit abstracts the capabilities of the SPU into abstractions that align with the workflows of a worker process in a modern distributed analytical engine. NeuroBlade has used this toolkit to build reference integrations with ClickHouse, Presto, and Apache Spark; these can be used as template for custom integrations.
- **DAXL Services:** A collection of drivers, runtime software and management utilities to manage the SPU hardware and execute operations on the SPU and stream results back to the CPU for further processing.

The DAXL Plugin integrates with the worker (even in the context of a table connector) using the DAXL Toolkit to enable pushdown of queries to the SPU.



Broad Support of Open Standards

Open-source projects and community-driven development have continued to be a significant force in shaping the modern data analytics stack. To simplify integration into existing software infrastructure, the SPU provides native support for open standard formats for files (Apache Parquet, Apache ORC, ClickHouse MergeTree) and tables (Apache Hive, Apache Iceberg, Apache Hudi). Data is read and processed in its native format, which dramatically improves efficiency and performance. This commitment to open standards underscores our dedication to interoperability, ensuring that organizations can leverage NeuroBlade's SPU without disrupting established workflows.

Reference Integrations

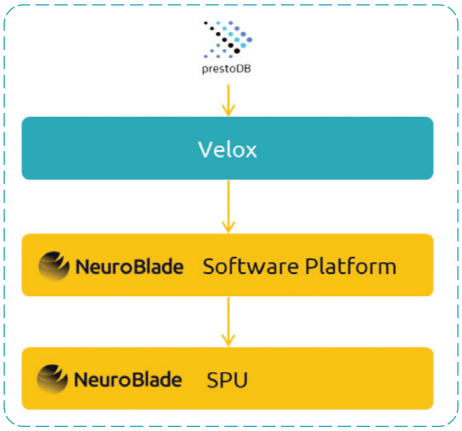
The ability to integrate with industry-leading analytical engines like PrestoDB, ClickHouse, and Apache Spark illustrates the SPU's versatility. For instance, the integration with PrestoDB leverages the Velox acceleration library to enhance query performance significantly. These practical applications highlight the SPU's capability to integrate effortlessly, offering a template for custom solutions that cater to unique organizational needs.

PrestoDB

PrestoDB offers powerful capabilities for processing large-scale data sets in a distributed computing environment. Velox is a native worker implementation that replaces the pre-existing Java based execution engine in PrestoDB.

NeuroBlade has developed a DAXL plugin for Velox that leverages the extensibility mechanisms built into Velox to integrate the SPU into Velox runtime. By enabling this plugin, PrestoDB workers that are running on a SPU-enabled compute server will transparently offload compute intensive analytics operations to the SPU.

This combined solution empowers organizations to unlock the full potential of their data analytics infrastructure, enabling faster insights and enhanced decision-making capabilities.

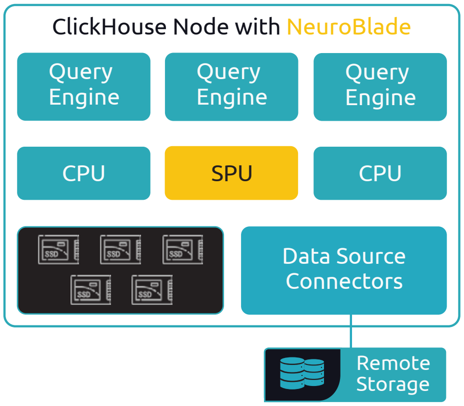


ClickHouse

The powerhouse behind ClickHouse's exceptional performance in data analytics is the MergeTree engine, which is optimized for analytics operations on very large datasets. To achieve this performance, MergeTree uses a proprietary Native ClickHouse file format and relies heavily on compute intensive operations such as decompression, fast indexing and data skipping to deliver high performance.

Key to the NeuroBlade integration with ClickHouse is the ability of the SPU to work with data in the Native ClickHouse file format and offload these compute intensive operations. This strategy not only minimizes CPU overhead but also reduces latency, ensuring the fastest and most efficient query execution speeds. Furthermore, the specialized processing capabilities of the NeuroBlade SPU, such as parallel processing and optimized algorithms, significantly speed up query execution.

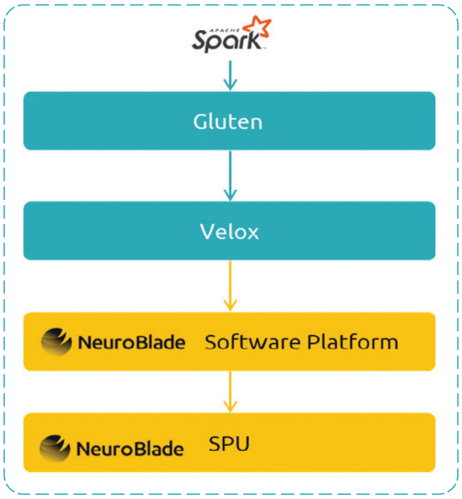
This seamless integration into ClickHouse enables organizations to experience marked improvements in query performance, leading to quicker data analysis and more informed decision-making. By leveraging the advanced capabilities of NeuroBlade's hardware acceleration within the ClickHouse ecosystem, organizations can extract valuable insights from their data more efficiently than ever before.

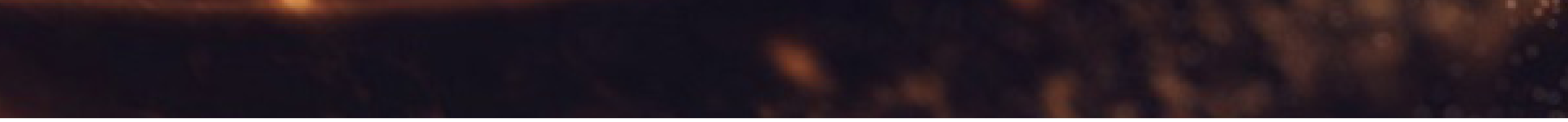


Apache Spark

Apache Spark is a widely used distributed processing framework that offers extensive capabilities for big data processing and analytics. Gluten enables Spark to pushdown operations to hardware accelerators.

The integration with Spark is simplified by leveraging the existing integration of Gluten with Velox. The same DAXL plug-in for Velox can be used to enable the SPU to read data directly from either disk cache or memory, bypassing the need to deserialize data before execution, resulting in improved performance. In addition to this, SPU's optimized architecture and parallel processing capabilities significantly accelerate query execution relative to running on a CPU.





Sample Use Cases

Social Media Interactive Queries

On a social media platform, where millions of users actively create and engage with content in real-time, the need for quick and interactive queries is crucial for maintaining a seamless user experience. Data analytics plays a pivotal role in these platforms, involving real-time interactions such as exploring trending topics, tracking user engagement on posts, and analyzing sentiments across various content types. However, traditional processing methods face challenges in providing instant responses to these dynamic queries, potentially leading to delays, and compromising the overall user experience.

Acceleration technologies, such as the SQL Processing Unit (SPU), play a significant role as specialized processing designed for data analytics, enhancing the speed and responsiveness of interactive queries within the social media landscape. For instance, when a user wishes to investigate trending hashtags or understand the impact of a recent event, the accelerated query processing facilitated by the SPU enables the platform to swiftly retrieve and present relevant insights, ensuring user engagement and satisfaction.

Consequently, the acceleration of interactive queries in social media analytics becomes a game-changer, ensuring that users can promptly access and interact with relevant information in a real-time context. By integrating cutting-edge technologies like the NeuroBlade SPU, social media platforms empower content creators with actionable insights, allowing them to stay at the forefront of the ever-evolving landscape of digital engagement.

AI/ML Pipelines

From personalized recommendations on streaming platforms to voice assistants that understand and respond to our commands, ML/AI technologies are transforming industries and revolutionizing the way we live. As organizations strive to leverage the power of AI/ML, there is a growing need to optimize the underlying infrastructure to achieve faster and more efficient outcomes. One critical aspect in AI/ML workflows is the latency involved in preparing the data before the actual modeling and training can begin, which typically accounts for a significant portion of the time spent, often gets to up to 80%. The NeuroBlade SPU enables organizations to achieve faster results, lower costs, and streamline data preparation.

For instance, in genomics research, the analysis of DNA sequencing data requires extensive preprocessing and variant calling, which can be computationally demanding. By utilizing the NeuroBlade SPU, researchers can significantly reduce the time required for data preparation, enabling faster identification of genetic variants, and accelerating the discovery of potential disease markers or treatment options.

A/B Testing

A/B testing is a critical practice in data analytics and experimentation, enabling organizations to make data-driven decisions and optimize operations. Whether it's refining website designs, fine-tuning marketing campaigns, or improving user experiences, A/B testing allows businesses to compare the results of different variations and identify the most effective strategies. While each A/B experiment may be small, running them at scale generates a significant amount of data. For example, LinkedIn reported that its A/B testing platform served up to 35,000 experiments concurrently, delivering 23 trillion evaluations daily. Handling such massive volumes of data and extracting valuable insights in a timely manner is crucial for organizations to gain a competitive edge.

The NeuroBlade SPU supports accelerated A/B testing capabilities that streamline the process and deliver faster insights. The accelerated data analytics capabilities provided by the SPU expedite the data processing, statistical analysis, and result interpretation, providing businesses with rapid insights into which variations are performing better. By reducing the time required for evaluation, NeuroBlade empowers businesses to iterate on their products, services, or marketing campaigns more quickly. This saves valuable time and resources, allowing organizations to conduct more experiments and get even more granular tests.

Summary

The NeuroBlade SPU (SQL Processing Unit) is a highly optimized data analytics compute accelerator designed to address the challenges of obtaining timely insights. Offering seamless integration with popular data analytics solutions such as Apache Spark, Presto, ClickHouse, and Velox, the SPU enhances performance and efficiency while tackling compute and memory bottlenecks and optimizing IO bandwidth to storage.

The DAXL Toolkit abstracts the capabilities of the SPU into a model that aligns with the workflow of a Data Analytics compute worker process in a modern query engine. This enables seamless integration with various data analytics tools, enabling organizations to work with diverse datasets and workloads. Reference implementations with Presto/Velox, Apache Spark and ClickHouse MergeTree highlight NeuroBlade's commitment to interoperability and collaboration within the broader open-source ecosystem. A standout feature of the SPU is its robust support for open standard formats for files (Parquet, ORC, MergeTree) and tables (Hive, Iceberg, Hudi), which greatly simplifies creating custom integrations using the reference implementations as a template.

The NeuroBlade SPU is available as a PCIe card (SPU G200) that seamlessly integrates into compute servers, providing organizations with a streamlined integration process into existing infrastructure. This approach maximizes benefits and enhances data analytics capabilities, empowering businesses to extract valuable insights from their data with greater performance and efficiency.

About NeuroBlade

NeuroBlade is unlocking data analytics by introducing its SQL Processing Unit (SPU) accelerator. This innovative technology significantly enhances query processing speed and scalability, offering up to a 100-fold improvement in performance-per-cost for Data Analytics workloads. Emphasizing a Compute Made For Analytics approach, NeuroBlade's advanced processor design optimizes throughput for petabyte-scale data, enabling queries to run exponentially faster.

Founded in 2018 by veterans of the systems, storage, and data analytics industries, NeuroBlade is headquartered in Tel Aviv, Israel, and has an office in Palo Alto, California, with operations in Taipei, Taiwan. **For more information, visit www.neuroblade.com.**