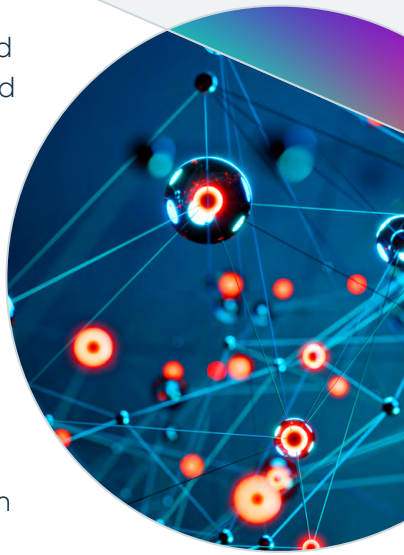
SQream

SQream's Unique Architecture:

Comparing and Contrasting to Leading Data Architectures

Introduction



Most businesses today generate and accumulate a massive amount of digital information. The data can come from sources that are spread out geographically and temporally, handled through multiple networks, systems and applications, and stored in a variety of databases and repositories, on mainframes or in cloud-based analytical systems.

Every organization understands the value of the collected data for business intelligence and more objective proactive decision-making. Many data managers SQream spoke with in a recent series of interviews made it clear that their goal is to help their companies become more data-driven, producing smarter, faster actionable business insights. They noted, however, that effective data usage raises several complex challenges.

Among the foremost issues are data quality and consistency, which requires distinguishing between data that can help meet their business goals and common digital clutter. Naturally, when data projects grow in scale, they become more complicated across the collection, analysis and implementation phases. In large enterprises, this is often due to a business infrastructure dependent on a patchwork of various vendors, licenses and solutions. The result is a continuous quest for an all-in-one platform, providing a single source of truth for data management and analysis.

Other challenges companies face are related to the on-time delivery of the data where it will do the most good, while maintaining business relevancy and regulatory compliance throughout.

Of course, any shift to a more data-driven business strategy requires increased expenditures on data collection, storage and analytics. The return-on-investment (ROI) companies seek from such an initiative is cost-effectiveness, as seen over time in increased revenue, operational optimization, and routine savings.

SQream steps into the breach with its enterprise-grade platform combining a centralized data warehouse, petabyte-scale data management, and GPU-accelerated SQL analytics in a single environment. This provides the consistent support needed for analytics of constantly-growing quantities of data and intensive data preparation pipelines, all while maintaining the simplicity of the familiar SQL interface.

The processing technology behind SQream's unique capabilities is differentiated both from the old-school Hadoop ecosystem and from the modern data lakehouses and warehouses.

Hadoop

Apache Hadoop is an open-source framework designed for distributed storage and processing of large volumes of data across clusters of commodity hardware.

Cloud data warehouse

A cloud data warehouse is a specialized type of data warehouse that leverages cloud computing infrastructure to store, manage, and analyze large volumes of structured and semi-structured data.

Cloud data lakehouse

A data architecture that blends the flexibility of a data lake (a single repository of raw data) and the management features of a data warehouse

Let's dive into a few comparisons.

Performance: Simultaneous, Not Sequential

For effective use of the complex information available to businesses, the data must be high quality, reliably collected, accurately and intelligently analyzed, and smoothly delivered to the right people. Of course, the processes involved should be able to scale for consistent benefit from all the data, all the time. However, every data analytics architecture has certain inherent performance bottlenecks that become increasingly challenging at scale. The time-consuming bottlenecks and data analysis silos that tend to creep into management processes can also pose an increased security risk due to inconsistent governance.

SQream was built ground-up to handle intensive data analytics workloads with patented GPU acceleration and latency-free architecture, which is the heart of its unique approach to common analytics bottlenecks.

I/O Bottleneck

The more complex a data query is, the more tasks the system is required to process in order to produce the desired output. An input/output (I/O) bottleneck happens when a system's throughput is not enough to handle the number of tasks it is being asked to process. This could be the result of data transaction processes lagging behind advancements in processor speed and memory capacity, as would be expected according to Moore's law regarding rapid exponential growth in speed and processing capabilities over time. acceleration and latency-free architecture, which is the heart of its unique approach to common analytics bottlenecks.

Hadoop

Hadoop-based massively parallel processing (MPP) frameworks address the I/O issue with a shared-nothing architecture - each node processes assigned tasks based on its local storage and communicates with other nodes during execution. MapReduce, a Hadoop module that helps programs carry out parallel computation, takes input data and converts it into a dataset that can be computed in key-value pairs.

Cloud data lakehouse

The cloud data lakehouse uses a decoupled, shared-storage architecture, but manages the data in an open-table format (such as Apache Iceberg, Apache Hudi or Delta Lake). Moreover, to further mitigate potential I/O bottleneck, lakehouse query processors were designed for the optimized reading and writing of open-source formats (both tables and files).

Cloud data warehouse

MPP is used in cloud-based data warehouses that support big data projects, as well. However, they address the I/O bottleneck problem with a hybrid shared-nothing and shared-storage architecture, with storage and compute completely layered in a decoupled manner. Multiple servers run in parallel to distribute processing and I/O loads, while a leader node manages the distribution of data among follower nodes to optimize performance. Also, multi-location redundancy is often available for faster network access to support large-scale global operations.

SQREAM

SQream is based upon a shared-storage architecture as well, with all the compute nodes linked to the same storage through a single mounting point, with separate memory and computing capabilities. There is no leader node for parallel operations, which eliminates the single-point-of-failure risk that big data analytical systems often face. Moreover, SQream's GPU-accelerated compute nodes are much more effective than CPUs in handling calculations in extreme throughput scenarios. The solution's enhanced resource utilization and simplified management of data provide the high scalability and flexibility needed to mitigate I/O bottlenecks.

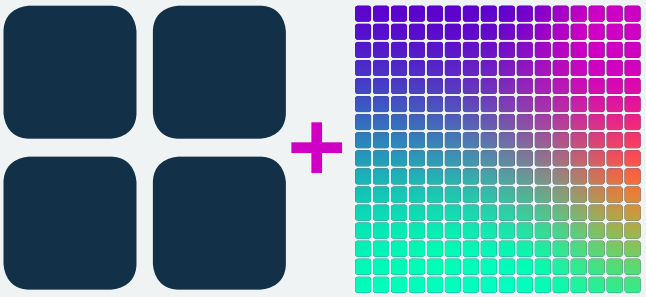
Memory Bottleneck

For data to be processed, it must be transferred from the storage, to the RAM, and from there, it can be read and processed by the CPU. This flow of data between memory layers often involves perceivable latency. As most data analysis use cases are not urgent, this time factor does not always pose a serious issue for organizations. However, there are certain mission-critical workflows where real-time or near-real-time data processing is absolutely necessary.

The traditional aproach

Multi-core CPU

The SQream method



Multi-core CPU GPU

Hadoop

Hadoop-based systems tried to break the limits of "in-memory" processing with caching, partitioning, and storing data in-memory in an efficient way. To that end, Hadoop breaks the data into essentially arbitrary blocks and processes it in batches, not in real-time.

Cloud data warehouse

Cloud data warehouses tried to resolve the memory bottleneck with columnar logic. That is, identifying the data actually needed to answer a specific query and only moving that data into the RAM and the CPU. For the aggregate queries most commonly used in business reporting, the columnar approach produces much faster results. However, if the amount of data to be processed is large, then the warehouse needs stronger computing capabilities or more compute machines with the same properties, which can then generate certain scaling challenges (see herein below).

Cloud data lakehouse

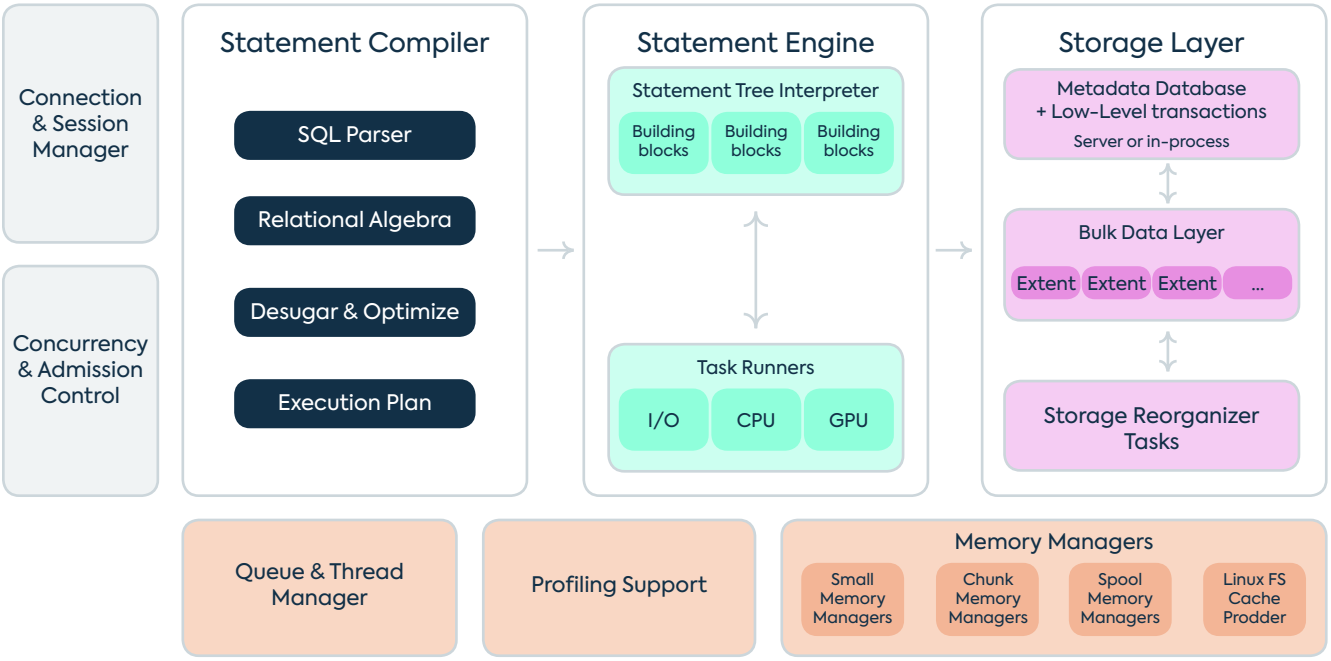
In the data lakehouse, open format tables act as a metadata layer between the lake storage and the execution engine. This structure limits the data with greater precision, helping prevent memory overload. Another strategy, also seen in cloud data warehouses, is auto-scaling. If the engine identifies a query that it assesses will overwhelm the memory, then it activates more resources to increase pipeline processing speed. Once they are no longer needed, the resources are freed or deactivated.

SQream has the advantages of both the data warehouse and the data lakehouse:

- Like cloud data warehouses, SQream's execution engine is columnar.
- Like cloud data lakehouses, SQream maintains metadata on the data it stores

SQream further addresses the memory bottleneck issue by ensuring more resources are available, with GPU-accelerated analytics leveraging the relatively large GPU-RAM. The flexible structure can allocate resources dynamically to handle varying workloads utilizing CPU, GPU, and RAM, maintaining a balance for optimal performance. The SQream interface layer's statement compiler is designed to analyze each query and determine which parts should run on the CPU or GPU. Some queries run solely on the CPU to avoid unnecessary overhead and the most complex tasks requiring parallelism are passed to the GPU.

SQream's Execution Engine Architecture



Scaling Bottleneck

There is an inherent limit to data volumes and workloads that a management and analysis system can handle. However, as these needs increase, the limitation is often overcome by scaling-out the number of nodes in use or upgrading them with much stronger hardware (scaling up). The downside to this strategy is enormous and growing infrastructure costs.

Hadoop

Hadoop-based systems distribute data processing among multiple nodes that work together to store and query data. However, with files randomly distributed across Hadoop clusters, JOIN operations require extensive and complex data shuffling, leading to potentially severe performance issues. In fact, the initial Hadoop promise of scaling out using commodity hardware lost its shine over the years, as the necessary hardware turned out to be not that basic and minimal node specifications have grown increasingly burdensome and expensive.

Cloud data warehouse

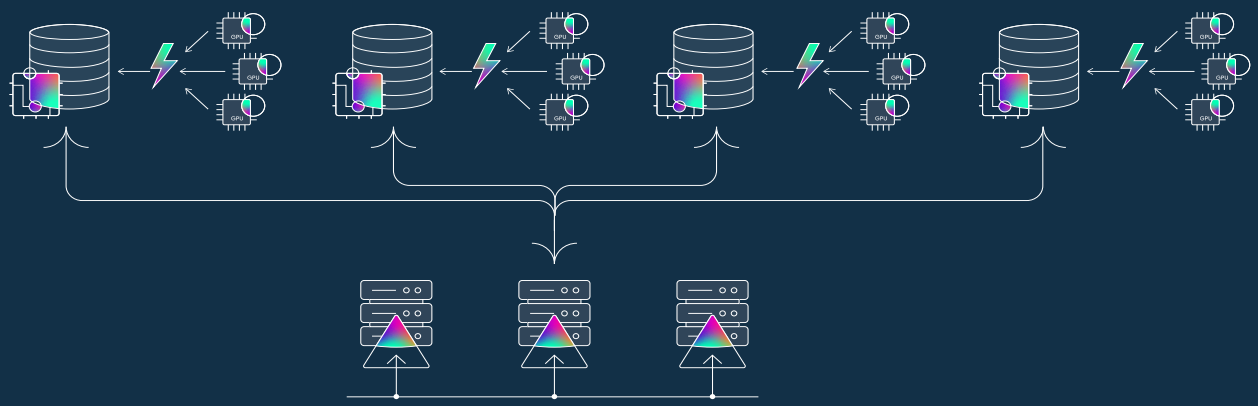
Cloud data warehouses split the growing volume of data into smaller parts, with each node processing its portion simultaneously and independently. Scaling up RAM to improve performance is accomplished by adding nodes, which can take a few minutes.

Cloud data lakehouse

Cloud data lakehouse storage can scale easily, as cloud-based object storage is conceptually unlimited, and new data sources need not be made to fit into an organization's existing data formats and schema. Computation capabilities scale separately, essentially through the same mechanism as a cloud data warehouse. Notably, resource managers may need to schedule large-scale projects for less busy hours or assign them specific resources designed to handle data-heavy tasks.

SQREAM

For meeting the needs of scaling up, SQream's unique GPU-acceleration combines two types of parallelism: between separate nodes (similar to any other MPP); and on a single GPU chip (MPP-on-chip). A single node can also host multiple GPUs to increase concurrency even further. In this way, the capabilities of every individual compute unit can be logically and rapidly multiplied without installing any additional nodes. It also means that each GPU can work simultaneously, rather than serially, on several different tasks.



Optimization Bottleneck

Ongoing (usually manual) maintenance of data to achieve better performance – such as ensuring consistent metadata - is usually time-consuming and delays the analytics workflow.

Hadoop

Hadoop-based systems, partitioned by the original design of numerous compute and storage resource nodes, require data to be partitioned and indexed manually. This is a time-consuming effort that has its own risks in terms of data integrity and consistency, as highlighted above.

Cloud data warehouse

Cloud data warehouses often offer built-in optimizations for partitioning and indexing, but the user still has to define the desired keys and schema. This manual configuration improves query performance, but it is not always valid for all types of queries. Moreover, if a performance boost or query acceleration is needed, then the data needs manual maintenance and optimizations (such as caching or materialized views).

Cloud data lakehouse

Cloud data lakehouses use the same optimizations as data warehouses (indexing, caching, materialized views, and more), but they usually address the optimization bottleneck by designing or optimizing the query engine for the open table format they support. For best performance, this often requires skilled data engineers familiar with specific characteristics and query patterns related to the data being managed.

SQream automatically saves the metadata and statistics of each data chunk it processes (equal to indexing), allowing efficient and flexible data skipping with no need for preliminary knowledge of the queries that will be run. This unique approach eliminates the need for manual optimizations for query acceleration, ensuring that queries are completed quickly and efficiently even in the case of high-intensity ETL/ELT processes. Ad-hoc queries are easily executed without the need for tuning, maintenance or preliminary adaptation.

SELECT * FROM
my_table WHERE year = 2018

Consistency And Integrity Across Data Analysis

As noted, a critical step in data management, tracking and analytics is ensuring the collected and collated data is accurate, reliable and relevant to an organization's business goals. The data must be normalized in some fashion for efficient organization and effective analysis, while data redundancy needs to be minimized or eliminated entirely. The quality of the data can also be compromised (data drift) as it is reused across growing datasets for various purposes.

Hadoop

Hadoop works with data lakes that are flexible and can store large amounts of different types of data, but the data can be very hard to govern and understand in its open file format. File directories are not very efficient in this case and intense data preparation is required due to the schema-on-read approach. Data analysts and scientists are forced to navigate difficult-to-use tools and jump through hoops to access the data they need. In addition, the data duplication necessary for redundancy and executing queries effectively creates endless copies that are a challenge to control over time.

Cloud data warehouse

One advantage of a cloud data warehouse is the support it offers for integration of new data sources. Another is built-in management features for saving space and streamlining performance, such as compression and deduplication. At the same time, the replication and synchronization required to import data from existing transactional databases, and the development effort needed for effective reporting and analysis, can impact the quality and timeliness of the data in the warehouse.

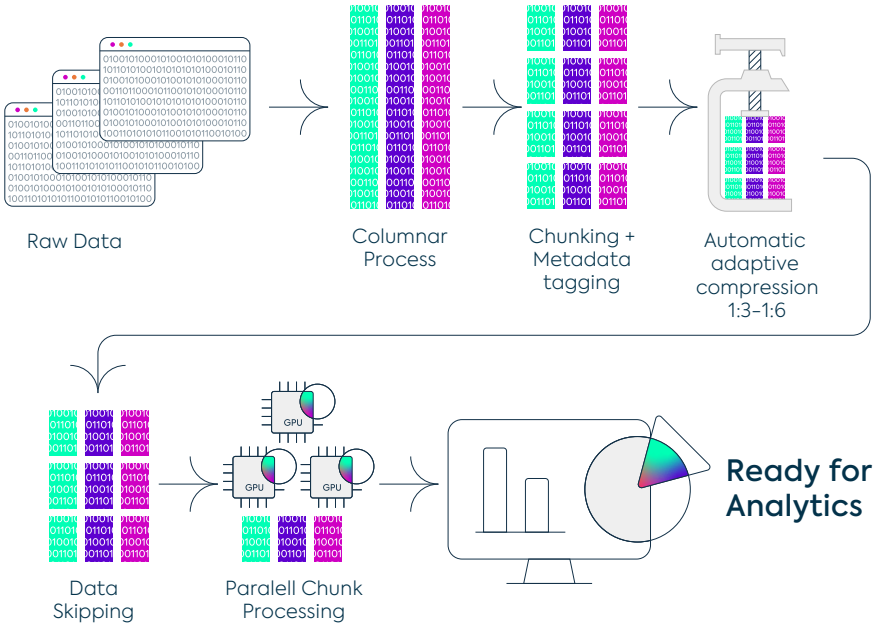
Cloud data lakehouse

A data warehouse stores historical data from various applications after the data has been filtered, organized, defined, and had metadata applied (i.e., schema-on-write). A data lake, on the other hand, is a repository that takes in the data as-is until needed for an analysis of some kind (i.e., schema-on-read). The data lakehouse aims to combine the best of both, ensuring new data imported into the lake conforms to a specific set of metadata conventions in open-table formats; however, the schema can automatically evolve over time. In a lakehouse, unlike in a typical data lake, the stored data can be queried directly on the lake, minimizing the need for copying data and making data governance easier. Data lakehouses also adopt reliability, consistency, access control and regulatory compliance best practices from data warehousing.

SQream automatically partitions the data in a columnar proprietary format column data as chunks, compressing them to disk using the CPU or GPU compression algorithm that best fits the input data (usually at a 1:3-1:6 compression rate). Unlike traditional systems that rely on data replication across nodes, SQream uses a shared storage approach that ensures data safety without the redundancy of replication.

For optimally efficient analytics, SQream's architecture leverages the power of GPUs for loading and analyzing data without the need for intermediate steps. SQream also incorporates features such as fine-grained, low-overhead zone maps and data skipping to facilitate rapid and efficient analysis of large datasets, especially use cases that require multiple JOIN clauses and aggregations. This data preprocessing and wrangling plays a critical role in transforming raw data into a "clean" format to train machine learning algorithms.

Data Loading Process into SQream



Calculating Cost And ROI

Data management and analysis technology has to be fit-for-need. That includes the overall objective of the adoption of new levels of analytics, the available budget, customization needs, and whether the tech is compliant with relevant regulatory demands.

Hadoop

Hadoop is an open-source framework, utilizing Java, multiple computers, and often a large number of relatively slow and inexpensive commodity servers. One of the most expensive resources in Hadoop is storage, as HDFS requires redundancy that consumes three times the size of the stored data itself.

As a complex software ecosystem, a Hadoop cluster can be quite expensive and time-consuming to set up, implement and maintain, requiring familiarity with an overwhelming number of different tools. Because of this, many enterprises use commercial and managed distributions of Hadoop, such as Cloudera, that are provided as a hardware-software bundle.

In order to support SQL queries for those organizations using data lakes, Hadoop solutions often use unsophisticated brute force methods. This is a highly inefficient use of machine resources, resulting in a high total cost of ownership (especially with a pay-as-you-go arrangement) and poor performance.

Cloud data warehouse

Cloud data warehouses are offered by cloud providers like AWS and Google Cloud or by companies like Snowflake at a fraction of the set-up costs of on-prem alternatives. In addition, MPP data warehouses typically use columnar logic, which is the most adaptable and cost-effective for the type of analytics most businesses need.

The long-term TCO advantages include provider commitments to set service levels for uptime and availability. Typically, the cloud data warehouse customer only pays for the storage and computing resources they actually consume, avoiding the costs and risks of under- and over-provisioning. However, with a pay-as-you-go pricing model, the monthly charge for the cloud data warehouse can be very unpredictable. Controlling costs requires constant attention, tracking and reconfiguring of the system if it is to remain affordable. Some companies even find it necessary to hire someone exclusively dedicated to those tasks.

Cloud data lakehouse

Cloud data lakehouses are cost-efficient due to the massive scale of data lakes: a single location often dependent on low-cost storage infrastructure. Data is maintained in generic, open formats and a single tool is used to process it, reducing data redundancy and costs. The connection to business intelligence and analytics tools, including AI and ML algorithms, is direct and requires no additional mediating software.

SQream achieves the same or better data management and analytics performance than Hadoop, cloud data warehouses or lakehouses, with the use of significantly fewer resources. Going beyond these traditional platforms, SQream accelerates data processing end to end, from data preparation to insights, with its unique multi-level GPU parallelizing technology, decoupled storage and compute, and rapid metadata standardization.

While a single GPU server is more expensive than a CPU server, the GPU has far better cost-performance. Hundreds of CPU servers would be necessary to handle the massive terabytes-to-petabytes of information enterprises typically need to process, while the same server power can be derived from just 10-20 GPUs in the SQream system.

SQream provides the best ROI compared to other available solutions - especially for large-scale enterprises - with its speed, flexibility and scalability.

Summary

Modern enterprises are outgrowing their existing data infrastructure (hardware and software), which is based on distributed data processing with CPUs - a technology that has been with us since the 1980s. As a result, enterprises with data-intensive workloads, complex queries and a need for large-scale storage are running into scaling challenges and notable performance limitations.

SQream was designed from scratch as a data processing and analytics platform to support heavy-lifting use cases and complex projects. It enables petabyte-scale data management in a single analytical environment, with innovative GPU-acceleration that brings unique multitasking and supercomputing power into CPU-based systems.

With SQream, data teams can comply with any new business-user request and even proactively suggest new use cases. They are much more productive with their time and don't have to constantly keep an eye on the analytics stack to prevent crashes. All stakeholders also have access to the same data at the same time, with the clean simplicity of a familiar SQL interface.

SQream puts into practice the undeniable paradigm that data needs to be ready when you need it (for analytics or machine learning) - not hours or days later - with faster data processing and more advanced analytics than any other alternative.

About SQream

SQream empowers companies to get value from their data that was unattainable before at an exeptional cost performance. Our data processing and analytics acceleration platform utilizes a GPU-parented SQL engine, that accelerates the querying of extremely large and complicated datasets. By leveraging SQream's advanced supercomputing capabilities for analytics and machine learning, enterprises can stay ahead of their competitors while reducing costs and improve productivity.

To learn more, visit sqream.com or follow us on