SQREAM

A Look Inside SQream

Whitepaper

It’s time to supercharge your data analytics

In today’s world, too many of the critical answers we need from our business data are still unreachable. Constraints in analytical firepower, time, and resources often prevent us from asking the questions of our data that shape the decisions driving the business forward.

Yet, if we don’t ask these questions – or think to ask them – we limit the growth of our business.

What if we could reduce time to insight, get deeper insights, and substantially reduce CAPEX and OPEX?

SQream empowers companies to get value from their data that was unattainable before, and at exceptional cost-performance. Our data processing and analytics acceleration platform utilizes a GPU-patented 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 improving productivity.

Modern GPU-acceleration technology presents amazing opportunities

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, organizations with data-intensive workloads, complex queries, and a need for large-scale data management are running into scaling challenges and notable performance limitations. Whether it’s Hadoop data lakes, modern data warehouses, or even data lakehouses - bottlenecks will arise at scale.

SQream was designed from scratch as a data processing and analytics platform to support heavy-lifting use cases and complex projects. It enables terabytes-to-petabytes-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 on GPU.

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 alternative. Take a look inside SQream, and see how we make this magic happen.

SQream: Accelerating SQL terabytes-to-petabyte scale analytics on GPUs

SQream is the first enterprise-grade GPU-accelerated platform for SQL analytics, enabling its users to analyze and process data with or without loading it into the database, and deploy it on-prem or private and public cloud.

The processing technology behind SQream's unique capabilities is differentiated both from the old-school Hadoop ecosystem and from the modern data warehouses and lakehouses. It was created to empower data consumers, harnessing the raw brute-force power and high throughput capabilities of the GPU, with MPP-on-chip capabilities at the simplicity of SQL.

SQream is not another in-memory processing platform or a GPU database. It is based on patented GPU-acceleration and latency-free architecture technologies, designed for larger-than-memory, constantly growing, data-intensive workloads. The platform accelerates data processing throughout the entire analytics and machine learning cycle - from data loading to preparation (pre-processing and transformations) and insights or predictions, SQream provides the best cost-performant solution.

The SQream platform offers a spectrum of deployment options designed to align with diverse customer needs, security protocols, privacy regulations, and infrastructural requirements:

1. On-premises: self-managed on the customer's data center.
2. Private cloud: self-managed on the customer's cloud tenant.
3. Public cloud: fully-managed SaaS on SQream's cloud tenant.

The SQream latency-free architecture

To effectively leverage 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 to consistently benefit from all the data, all the time. However, every data analytics architecture has certain inherent performance bottlenecks that become increasingly challenging at scale. SQream was built 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.

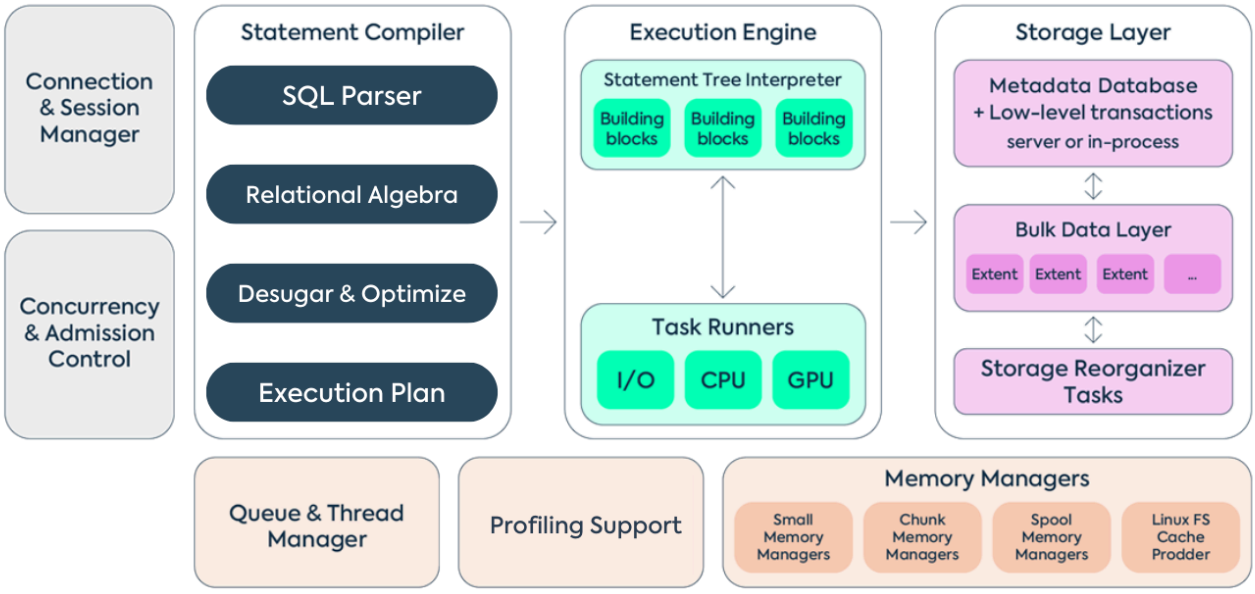

Figure 1 - SQream's architecture physically separates the compiler, runtime, and storage layer.

Independent scaling in any direction

SQream scales each layer independently, providing businesses with the flexibility of resources for their dynamic needs. SQream's persistent shared storage can run virtually on any file system – whether local, distributed, on-premise, or in the private cloud - ensuring reliability and performance. Moreover, SQream's engine can be used to query open-source file formats on any data lake directly.

As a result of the shared storage architecture, every SQream instance can be thought of as an MPP cluster by itself. Based on user role and permissions, it has full access to all of the data in the storage layer. For self-managed deployments, SQream's architecture further relies on storage caches to transparently and automatically cache data, which can be reused by other SQream instances if needed, boosting performance.

An additional benefit of SQream's architecture is that storage can easily scale up, and compute can easily scale-out (Fig. 2), as analytics requirements inevitably change. For fully-managed deployment of SQream, you have the flexibility to manually or automatically suspend and resume each of your compute nodes based on your specific requirements.

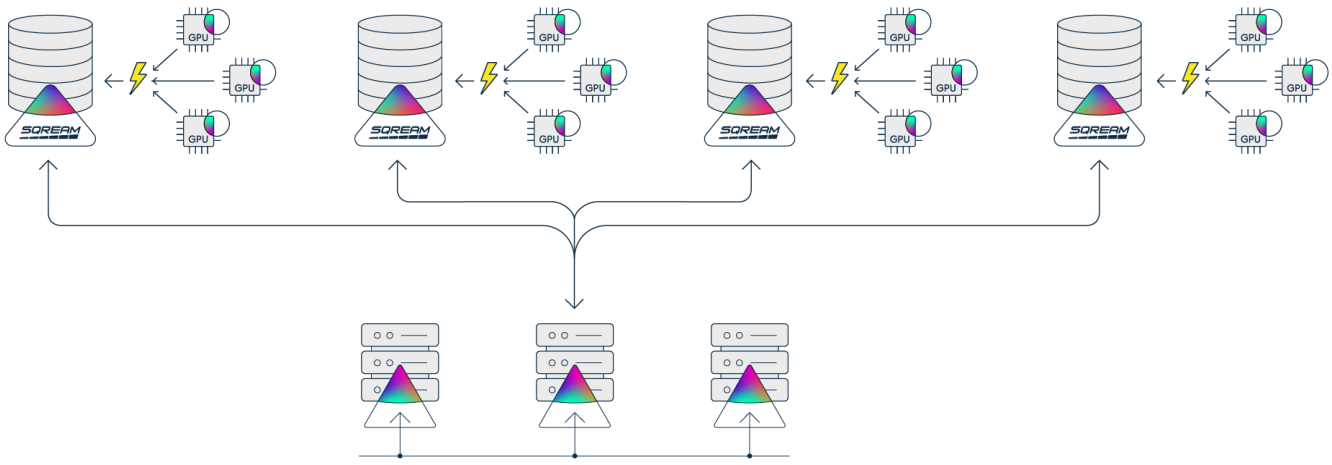

Figure 2 - SQream's shared storage architecture.

Combining CPU and GPU resources

Traditional data platforms all rely on a similar concept of CPU-based Massively Parallel Processing (MPP), where data has to be partitioned and processed in-memory. In contrast, SQream GPU-acceleration technology allocates more resources to handle a varied workload, by combining available CPU, GPU, RAM, and storage resources.

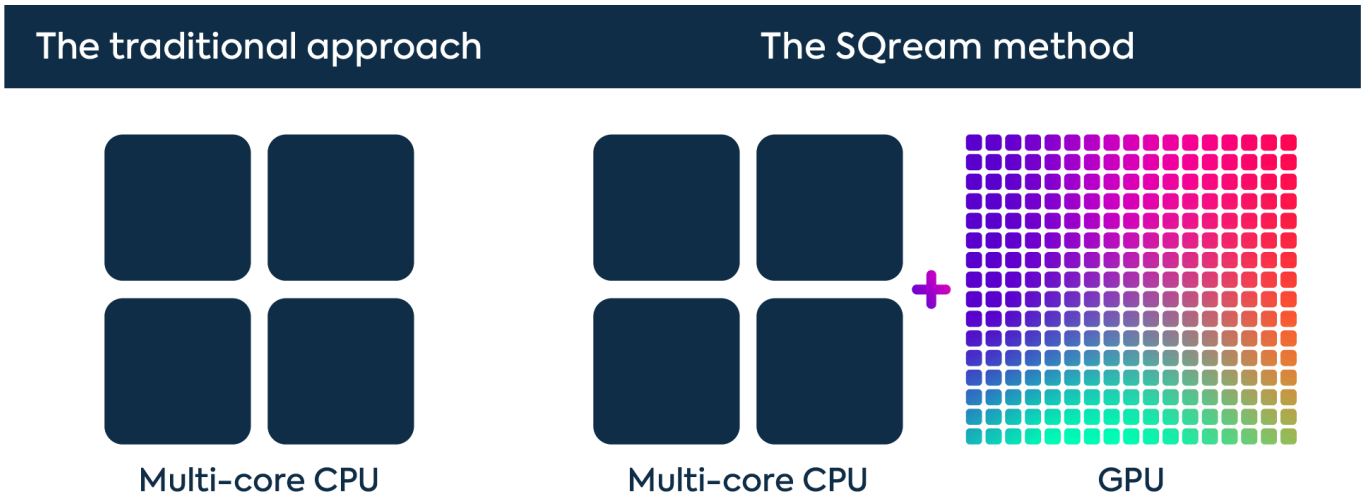

Figure 3 - CPU technology vs. GPU technology

This balance of CPU and GPU operations is key to ensuring optimal performance. While CPUs are optimal for data processing, GPUs excel at performing relatively simple, repetitive mathematical operations on large amounts of data in many streams. SQream's compiler was designed to accommodate and optimize the GPU throughput, so it may decide to run some parts of the query on the CPU if it calculates that the overhead of copying data to and from the GPU would slow down the query.

Parallelism

To accommodate 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.

Automatic hyper-partitioning designed for high throughput

If you choose to load data into SQream, it is automatically partitioned and tagged with metadata instantly without any user intervention or performance decrease. SQream's optimized columnar storage system is partitioned both horizontally and vertically for the best performance for heavy analytic operations like JOINs, aggregations, summarizations, and sorting.

Most data platforms require a team of administrators to finesse and manually tune processes, maintain indexing, update views and projections, etc. SQream was designed for frequently changing, modern workloads. It was built to handle worst-case scenarios and is optimized for huge datasets, where typical optimizations struggle.

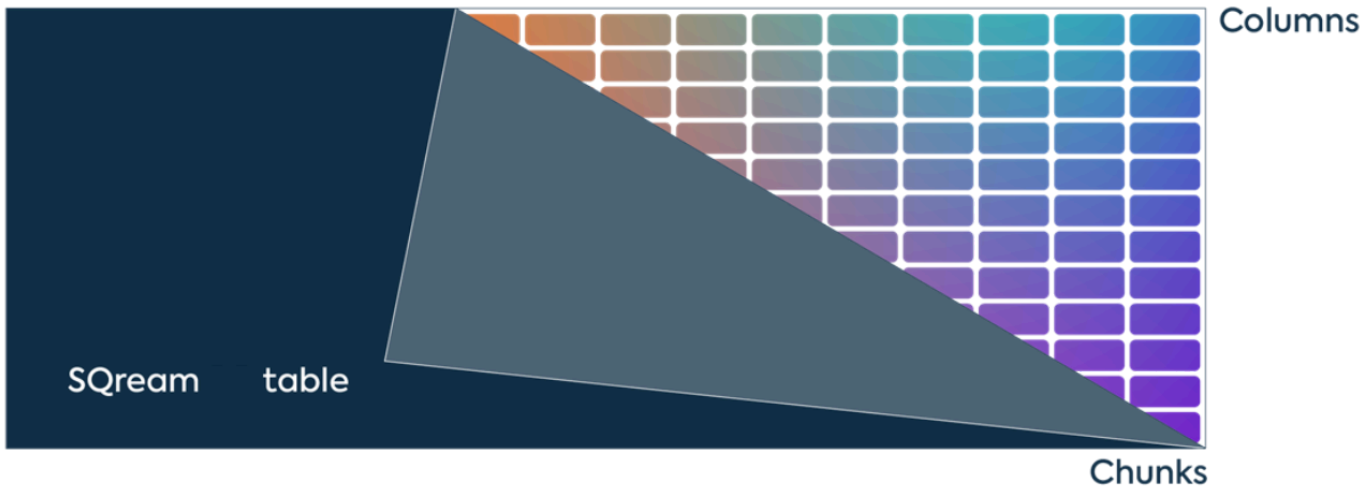

Figure 4 - The SQream table is partitioned vertically and horizontally

Vertical partitioning - columnar engine: This feature allows selective access to the required subset of columns (data skipping), reducing disk scan and memory I/O when compared with standard row storage. This concept is well-suited for parallelized compute, like the GPU.

Horizontal partitioning – chunks and extents: SQream automatically splits up the storage horizontally into manageable chunks (~1M rows each) enabling efficient usage of the hardware resources and relatively small G-RAM (GPU RAM) availability in GPUs. The intelligent use of spooling and caching helps make the most of the limited G-RAM.

SQream's GPU-accelerated architecture and automatic optimizations are key enablers for analyzing data ad-hoc without intermediate steps. SQream was developed to take advantage of the raw, brute power of the GPU, enabling data analysis immediately after loading.

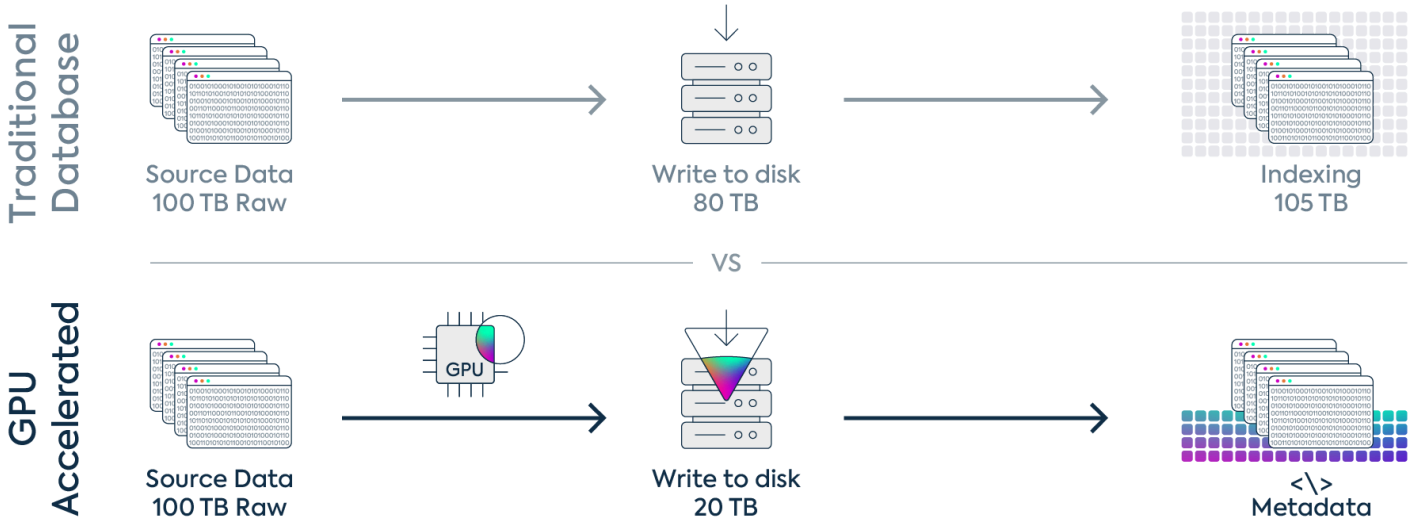

Figure 5 - SQream uses GPUs for compression and metadata collection during data loading, resulting in reduced I/O and higher throughput for load and queries

Upon data loading, SQream collects and stores metadata for all rows stored in a chunk. The most useful aspect of the collected metadata is the range of values and properties for the values being ingested. This metadata is stored separately from the compressed chunk.

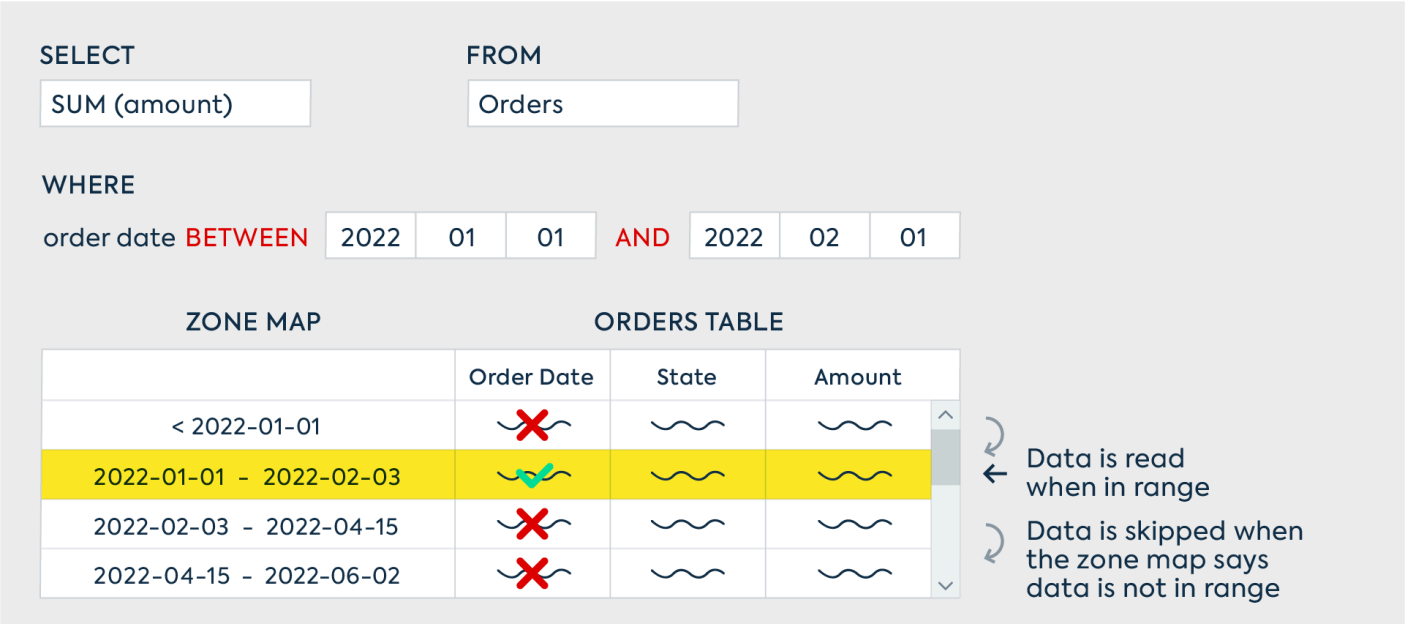

Figure 6 - SQream identifies and skips data that is unnecessary for queries

Unlike standard indexing, the metadata collected on these chunks is automatic and transparent across all data types and columns, requiring no intervention or maintenance. The metadata collection is space-efficient compared to columns, resulting in less than 1% overhead. Therefore, querying data becomes much faster, as the calculated zone-maps allow for efficient data pruning (also called skipping), eliminating the reading of irrelevant data.

Further notable SQream features

Fully featured SQL and industry-standard connectivity

SQream supports an ANSI-92 SQL-compliant syntax. It easily integrates into existing ecosystems, with support for industry-standard ODBC and JDBC connectors, as well as Python, .Net, Java, and others.

SQream's native SQL interface eases the transition from other databases. There's no need to maintain odd APIs and custom code. Full SQL support lets any existing ETL and applications connect and offload heavy workloads to SQream, minimizing the time needed to get up and running with the new platform. Moreover, SQream supports native batch and CDC data transfer from traditional databases using SQLoader.

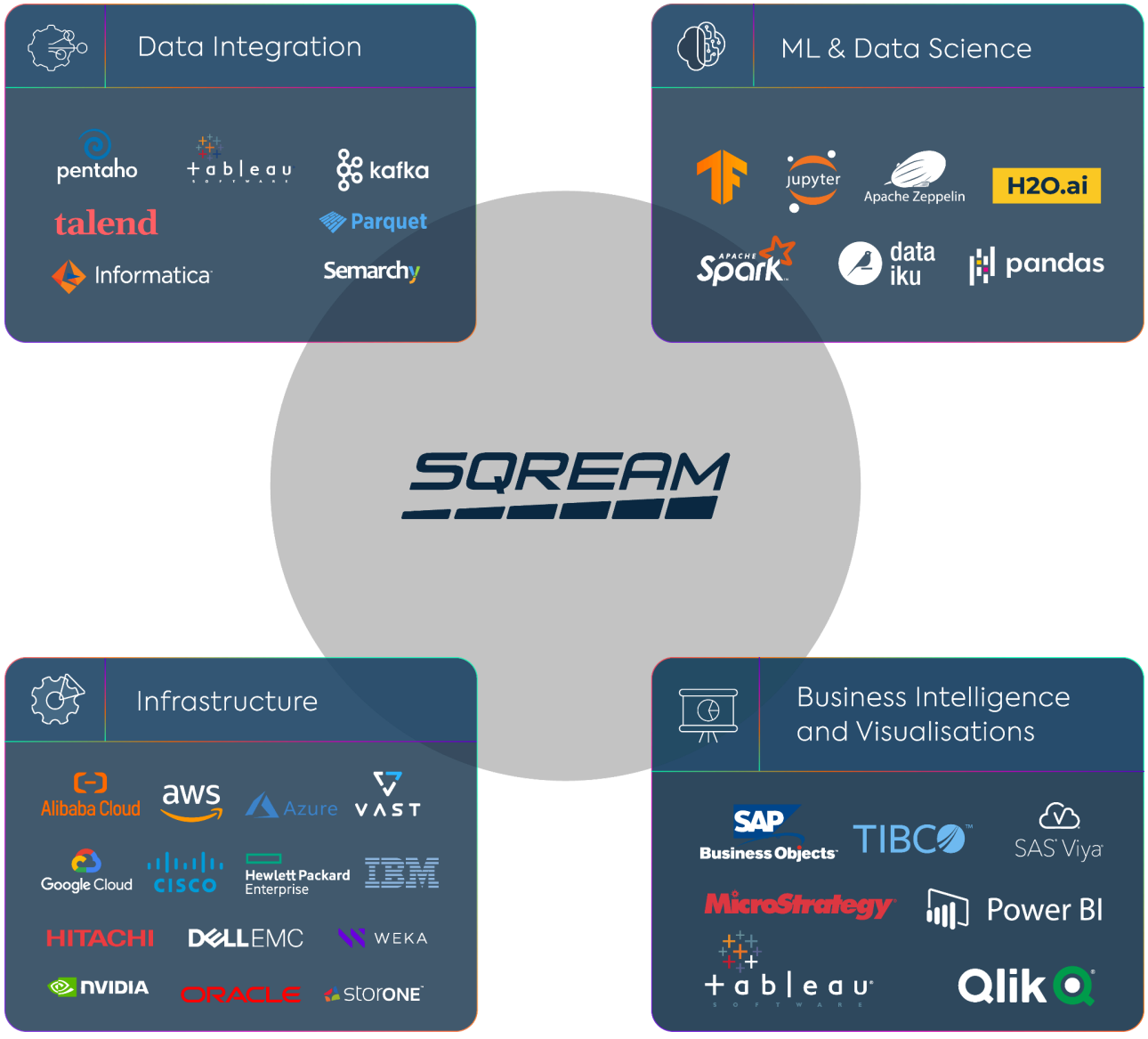

Figure 7 - SQream can be deployed anywhere through a widespread ecosystem of partnerships and integrations

Query from foreign tables

Foreign tables can be used to run queries directly on data without first inserting it into SQream. When working like a query engine, SQream supports read-only operations from external tables. Hence, you cannot use INSERT, DELETE, or UPDATE on them. Although querying foreign tables is slower than processing internal tables, one of their main benefits is expediting query time as no data loading is needed.

File format support

In addition to network-based connectivity, SQream also supports native reading and writing of the following popular file formats:

- CSV - a simple text-based format for storing tabular data
- Apache Parquet - a compressed columnar format, for which SQream is optimized
- Apache ORC - an alternative to Parquet, often used in Hadoop systems
- JSON - a well-used flexible format used to store dynamic and unstructured data
- Apache Avro - similar to JSON, but uses an efficient binary storage format

ETL/ELT and data pipelines

SQream excels at data loading and complex data preparation at scale. The core processing by GPU is both extremely fast and cost-effective. The separation of storage and compute, coupled with a unique approach of effective shared storage (as opposed to strictly ‘share-everything’ or ‘share-nothing’) enables each GPU to ingest and process without a need to coordinate with others, thus eliminating bottlenecks.

SQream scales across multiple machines and GPUs, and as such, has lots of extra bandwidth. This is used to perform instant processing on the fly, to accelerate other optimization operations without slowing down the process, i.e.: sorting, compressing/decompressing, and filtering.

For data pipeline fans, SQream’s fully-managed SaaS deployment has native data orchestration capabilities with visualized DAGs. Other deployments of SQream can integrate with the familiar Apache Airflow framework to externally schedule data preparation tasks.

Security and Permission System

SQream offers a secure and streamlined authentication process with LDAP or Auth0 credentials. After users have accessed SQream, their permissions will be managed on an object-level basis by their role (RBAC, Role-Based Access Control). Every user can be granted granular-level privileges for specific objects, such as Database, Schema, Table, Function, View, or Column.

SQream allows users to encrypt data in transit using TLS (Transport Layer Security) and data at rest with column-level encryption. For public cloud deployment, SQream leans on the cloud provider's inherent encryption capabilities. In addition, SQream meets the compliance of GDPR (General Data Protection Regulation) and SOC-2 Type II (for public cloud deployments).

Extensive Logging

SQream contains a built-in logger that monitors all critical information, enabling teams to gain insights into the platform’s operations from failed login attempts to GPU/CPU time spent per query and read-write cycles to memory.

Performance

To demonstrate SQream's cost-performance advantages, here are some benchmarks and POCs we performed in the past year.

TPCx-BB

On-prem (10TB)

- SQream (1xA100 GPU node) query time - 0:18:34
- Cloudera (11xCPU nodes) query time - 0:46:36

Cloud (30TB)

- SQream (4xA10 GPUs, OCI)
  - Loading time - 1:50:24
  - Query time - 0:48:36
- Snowflake (Medium Virtual Warehouse, AWS)
  - Loading time - 21:28:48
  - Query time - 3:26:02

Data Preparation Workloads

Cloud (1TB) External tables:

- SQream (8xA10 GPU nodes, AWS) - 2:27:06
- Snowflake (Large virtual warehouse, AWS) - 4:00:52

Cloud (1TB) Internal tables:

- SQream (8xA10 GPU nodes, AWS) - 2:31:50
- Snowflake (Large virtual warehouse, AWS) - 3:23:34

TPC-DS

On-prem (1TB)

- SQream (4xV100 GPUs)
  - Loading time - 0:14:52
  - Query time - 0:22:54

On-prem (10TB)

- SQream (4xV100 GPUs)
  - Loading time - 2:17:40
  - Query time - 3:38:03

Investment Bank POC

Cloud (5.6TB) External tables:

- SQream (1xA10 GPU node, AWS) query time - 0:18:18
- Snowflake (Small virtual warehouse, AWS) query time - 5:27:25

Cloud (5.6TB) Internal tables:

- SQream (1xA10 GPU nodes, AWS) load and query time - 4:08:30
  - Loading time - 3:54:44
  - Query time - 0:13:36
- Snowflake (Small virtual warehouse, AWS)
  - Loading time - 11:12:36
  - Query time - 0:41:12

Summary

In today's big data analytics market, SQream offers significantly better cost-performance than other players, specifically in the hundreds-of-terabytes to petabytes range where scaling with CPU-based MPP is not cost-effective. With the familiar SQL, superior linear scaling, and a robust architecture based on supercomputing hardware, SQream is a future-proof data platform.

SQream unlocks the opportunity to do more with more data. Fast insights with hundreds of billions of data points are now within reach. SQream can be integrated as a data warehouse, a query engine, or simply to expedite your data preparation pipelines, maximizing your analytics investments without replacing any part of your existing architecture.

SQream will fuel your data teams with exceptional productivity and the ability to innovate and become proactive. With data analytics and ML turbo-boosted by SQream's GPU-acceleration technology, you can finally start using valuable and qualified insights and predictions and enjoy superior cost-performance with a smaller hardware footprint on-prem or on the cloud.

About SQream

SQream empowers companies to get value from their data that was unattainable before at an exceptional cost performance. Our data processing and analytics acceleration platform utilizes a GPU-patented 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 improving productivity.