SQREAM

WHITEPAPER

MAKING THE MOST OF YOUR INVESTMENT IN HADOOP

Combining Hadoop with a modern approach to derive more insights from your data

sqream.com

WHO SHOULD READ THIS DOCUMENT?

This white paper is for any data professional – on the infrastructure, data engineering, data science or BI side - who is experiencing issues analyzing massive data due to data ingest challenges, lengthy and tedious data preparation cycles, or long-running queries that hinder comprehensive analytics. This paper will explain the root cause of these performance issues, and how a new approach to data analytics can alleviate them.

INTRODUCTION

Hadoop came to prominence when the web exploded with unstructured data. The use of unstructured data is common for web analytics, where flexibility is required for unknown or compound fields (arrays, nested objects, or just unknown). The popularity of Hadoop for these use-cases led to its adoption, also for structured use-cases. For these cases, SQL query engines have been bolted on Hadoop, and convert relational operations into map/reduce style operations.

The BI pipeline built on top of Hadoop - from HDFS to the multitude of SQL-on-Hadoop systems and down to the BI tool - has become strained and slow, resulting in three main problems:

- Tedious data preparation, requiring hours or days of coding
- Inflexible infrastructure that prohibits ad-hoc queries on large quantities of historical data
- Slow access to data, inaccurate results, and lengthy time-to-insight

These problems result in lost insight, troves of under-analyzed data, frustrated data teams, and ultimately, lost revenue.

This paper will explore the reasons behind these problems, and how companies can help alleviate data professionals' struggles at the source, with a new approach to storing, preparing, and analyzing big data. This new approach aims to help businesses reduce time-to-insight, increase productivity, empower data teams for better decision making, and increase revenue.

COMMON ISSUES WITH HADOOP

In order to get to the root of the problems, we must first understand what Hadoop is. Apache Hadoop is a collection of open-source software utilities that employ a large computer network to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

The core of Apache Hadoop consists of two main parts:

- Storage – known as **Hadoop Distributed File System (HDFS)**
- Processing – primarily a **MapReduce** programming model

Instead of copying data around, Hadoop copies code around. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel.

This approach takes advantage of data locality, where nodes manipulate the data they have access to. The downside of this approach is that a system dependent on data locality requires very careful planning. **If some data is on another node, it must be copied (for example, when performing a JOIN).**

Hadoop benefits from performing operations on unrelated entries, and combining the results in the end. For example, counting the number of times a certain parameter appears in the whole dataset – like a DISTINCT operation.

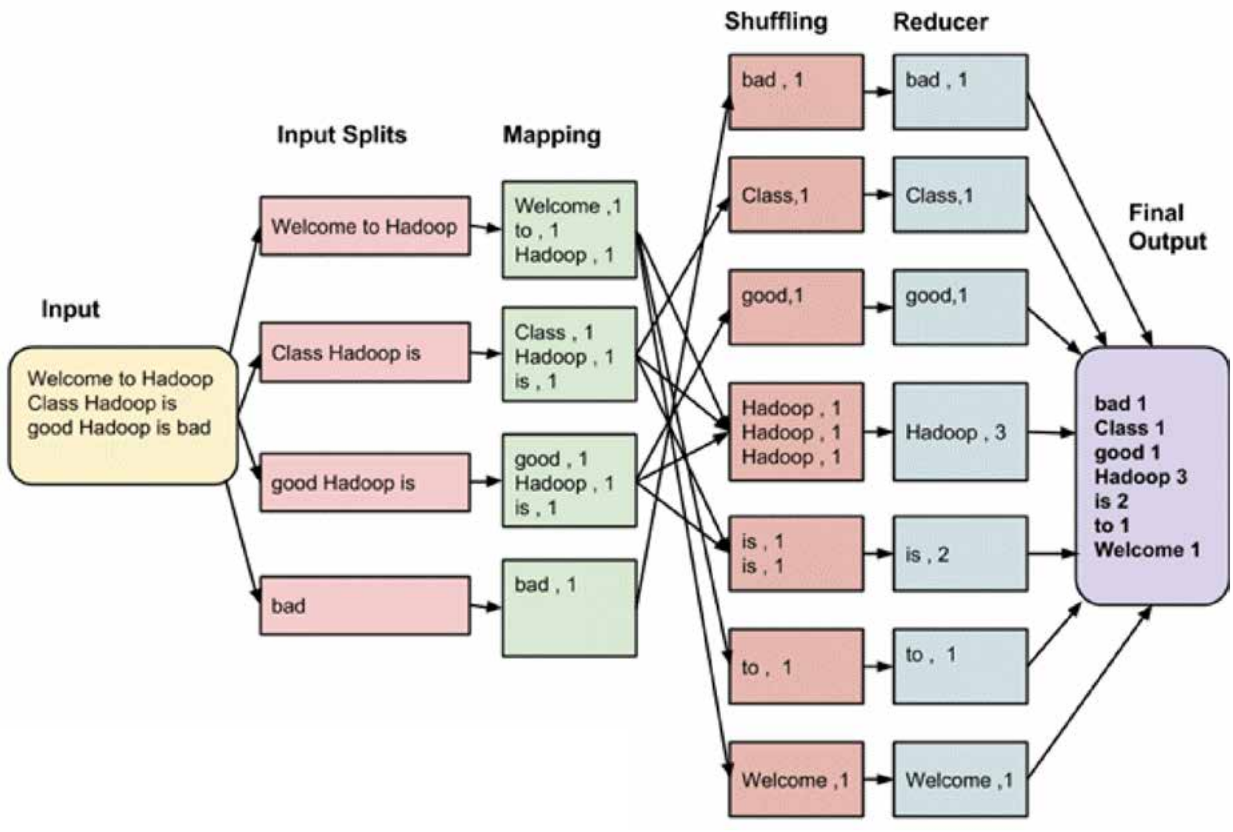
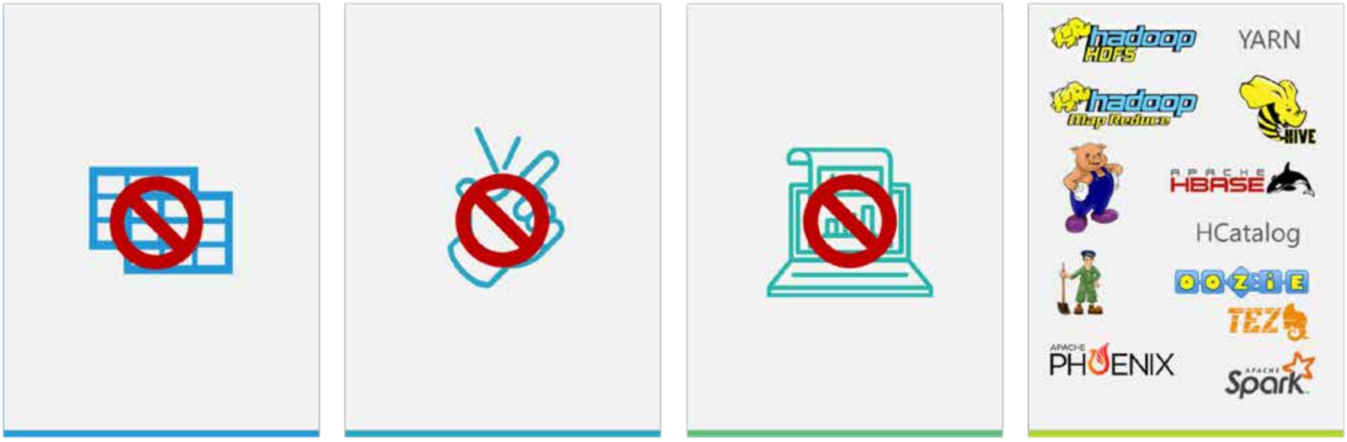


Figure 1 - MapReduce works by mapping out an operation on data that sits across many different nodes. The data is then sent to another node to be reduced (or aggregated). The result is collected again and sent to the client as a complete result set, or saved to HDFS as a result file.

(Source: Guru99 - https://www.guru99.com/introduction-to-mapreduce.html)

Because Hadoop was not designed as a database, in order to achieve the analytics goals, most companies deploy a huge patchwork of applications. These applications are difficult to maintain and deploy.



Not efficient for structured data

- Joins very difficult
- MapReduce doesn't translate well to relational operations

Difficult to use

- Relies on uncommon skills – Scala, MapReduce
- Finnicky SQL interfaces

Not designed for analytics

- Serial workloads
- Requires careful management of data

Patchwork of apps

- IT deployment is hard
- Difficult to secure

Figure 2 - Some of the issues Hadoop presents for analytics

To support the business requirements, large collections of Hadoop applications must be constantly tuned, with data duplicated in multiple distributions to closely match the expected business logic. Data must be loaded, and statistics must be collected. Collecting and loading data requires prior knowledge about which columns to collect statistics on, which further complicates the process. Furthermore, the tuning strategy must consider OS resources such as CPU, memory, and disk I/O calculations to prepare for generating indexes, writing metadata, materializing views, correctly distributing and partitioning data, and more.

This process requires significant time, which directly increases the latency of data availability for analysis. As the analysis becomes more complex and diverse, and the data increases in size, this trade-off becomes more difficult to manage. In addition, the constantly growing data stores will require additional costly investments in hardware, software, maintenance, and management.

DATA PREPARATION – TIME, COMPLEXITY, COMPUTE

The complexity of analytics required by data analysts and data scientists to get to results often causes great frustration. When queries are very complex and heavily compute-bound, even expertly tuned Hadoop clusters can struggle. Data must be carefully partitioned, distributed, and replicated, making joins and other complex queries nearly impossible. Such arduous data preparation severely limits the analysts' ability to perform ad-hoc data exploration. The dimensionality of the data is reduced (e.g. summarizing customer transactions, collapsing many transactions into one) to allow for faster querying later. This trade-off between coarser data granularity and faster querying comes at the high cost of making future fine-grained drilldown analysis impossible.

COMPLEX DATA PIPELINES

Poor performance, inflexibility, and difficult scaling ultimately lead to overly complex solutions. In addition to database administrators, many companies now also employ data engineers, data custodians, and data stewards. Data engineers work on the data-management side, maintaining the organization's data infrastructure, while database administrators focus on fine-tuning database performance. And while traditional data warehouse administrators have a large array of tweaks and optimizations that they can employ, Hadoop-based solutions usually involve adding more software and duplicating data further for the different use-cases.

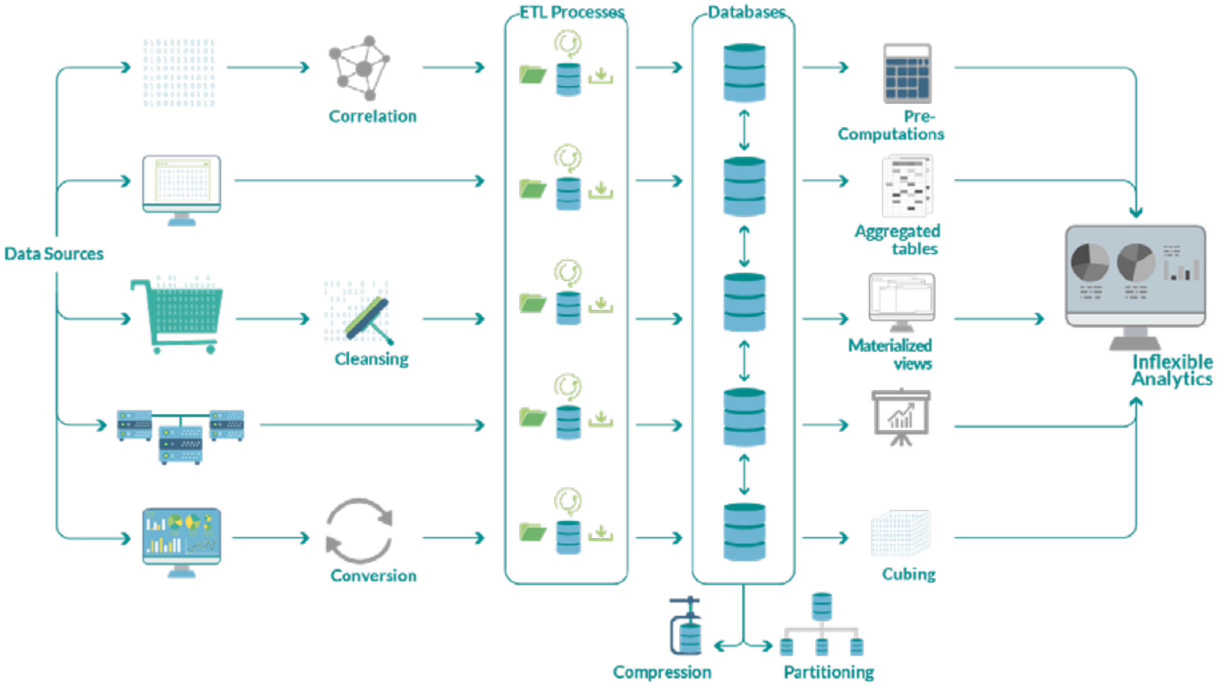


Figure 3 - Common data pipeline with Hadoop and traditional data warehousing methods are quite complex, and lead to inflexible analytics

COMPLICATIONS OF ADDING NEW DATA SOURCES

The problems Hadoop users experience trickle down to the BI tools and visualizers. One misguided attempt to create a new dashboard could generate complex queries that can bring the Hadoop cluster to its knees, causing outage for other lines of the business. These outages affect infrastructure teams and data consumers alike.

However, as BI analysts are the masters of the data in their domain, they are the ones most frustrated by the business impacts of data being inaccessible due to complexity, sub-sampling, and inflexible tools and infrastructure.

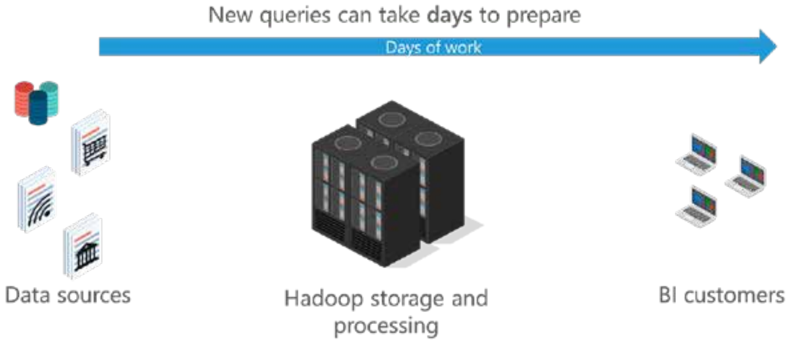

New queries can take days to prepare
Days of work
Data sources
Hadoop storage and processing
BI customers

Figure 4 - Integrating new data sources into Hadoop can take days or even weeks to prepare for analytics, causing frustration for data consumers like BI analysts and data scientists

ATTEMPTING TO BRIDGE THE GAP BETWEEN HADOOP AND DATA WAREHOUSES

As mentioned, Hadoop is not a database. It is a framework of applications and tools. Some of these libraries are incredibly solid and well-regarded, like HDFS. Others attempt to bridge the gap between Hadoop and what users want – SQL.

SQL is the de-facto standard for querying and BI. It offers immediate returns – most businesses are familiar with it and make use of it daily. For this reason, many SQL-on-Hadoop systems are now available that have enabled Hadoop to remain a viable data platform into the future. However, these tend to be fragile and incomplete.

DELIVERING RESULTS FOR AD-HOC ANALYTICS

When you have to deep dive into weeks, months, or years of data, even the most expertly maintained Hadoop clusters struggle to deliver. SQream DB was created to bridge the gap by bringing massive data analysis capabilities to Hadoop and other MPP ecosystems. SQream's GPU data warehouse enables rapid ad-hoc analysis of hundreds of terabytes to petabytes of raw data, minimizing the need for data preparation, while greatly reducing reporting time.

SQream complements Hadoop systems by accessing raw data directly, allowing analysts to focus on gaining new, business-powering insights.

SQream DB ingests data directly from raw files, or with any ETL-capable tool. With an ingest rate exceeding 3.5 TB per hour with a single GPU, SQream DB is an ideal complement to existing Hadoop installations.

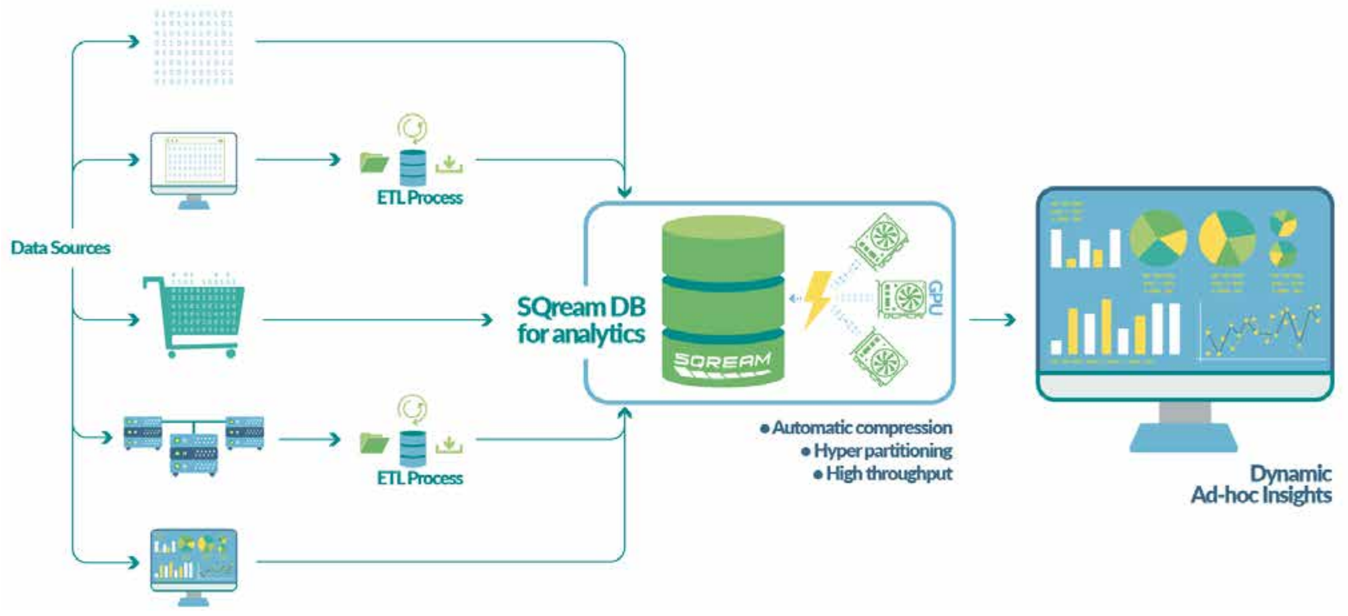


Figure 5 - A typical SQream DB implementation, with no intermediate data preparation needed

With SQream DB in their workflow, data professionals gain unrestricted access to any dataset with minimal preparation – enabling ad-hoc analysis. The combination of SQream's powerful GPU architecture and patented compression technology enables the ingest and analysis of significantly more data than before.

UNLIMITED SCALABILITY: AD-HOC ANALYSIS ON MORE DATA

Scaling to unlimited data sizes and any number of data consumers is key to SQream DB. By decoupling compute resources from storage, it is possible to scale only storage, only compute, or both - a key factor in providing excellent performance for organizations of any size, while reducing costs. Growing the system in either direction doesn't affect the data availability or integrity, which means that SQream DB can scale to virtually unlimited data sizes.

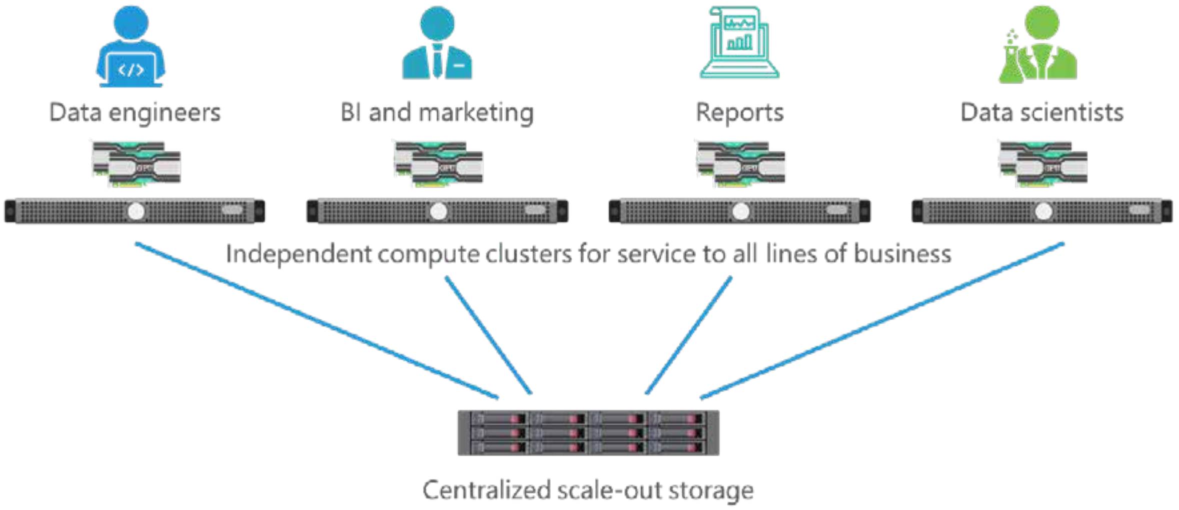

Data engineers
BI and marketing
Reports
Data scientists
Independent compute clusters for service to all lines of business
Centralized scale-out storage

Figure 6 - SQream DB has a shared-data architecture that enables independent scaling of storage and compute. The shared-data architecture lets you scale out to serve more lines of business and a wider range of data consumers as your needs change.

With effectively unlimited scalability, data consumers can analyze significantly more data without having to invest in complex, distributed solutions. SQream DB supports data consumers and infrastrcture teams alike by transparently and automatically optimizing, compressing, and partitioning data to ensure fast time-to-analysis.. This process, outlined in Figure 7 below is what we call "Load-and-Go".

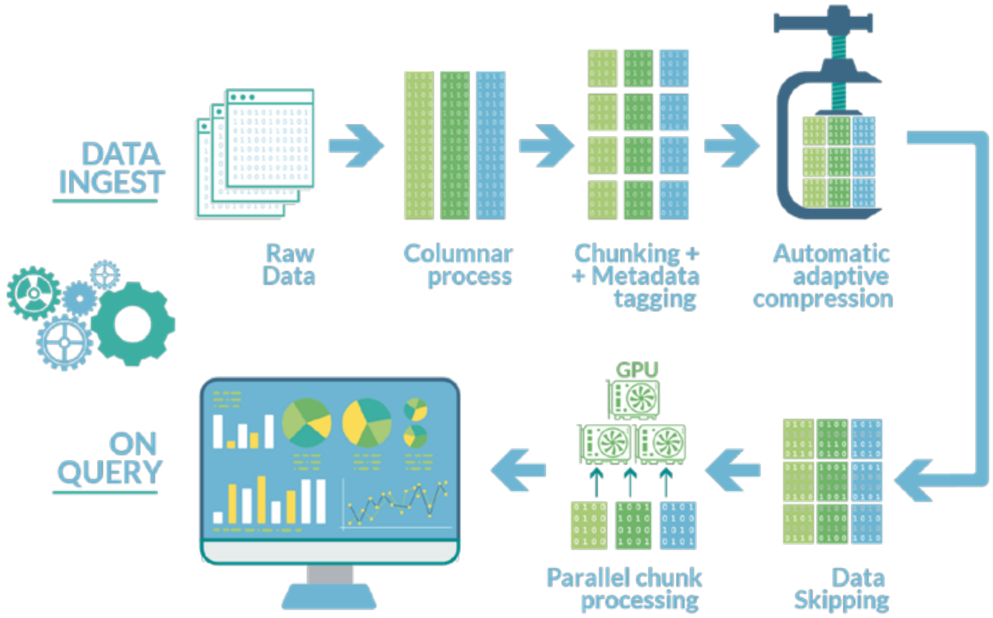

DATA INGEST
Raw Data
Columnar process
Chunking + + Metadata tagging
Automatic adaptive compression
ON QUERY
GPU
Parallel chunk processing
Data Skipping

Figure 7 – Load-and-Go architecture. Every step of the process is completely automated, resulting in faster time-to-analysis.

Unlike other Hadoop-based analytics solutions, SQream DB includes fully featured SQL and ODBC/JDBC connectivity, allowing users to connect and begin data exploration within seconds, without limiting the dimensionality or depth of the query. Joins are available on every column, with no indexing needed, making it easy to move from any database or SQL-on-Hadoop solution to SQream DB.

FASTER QUERIES, FASTER DASHBOARDS, MORE DATA ANALYZED

SQream DB can allocate additional resources to handle a varied workload by combining available CPU, GPU, RAM, and storage resources, enabling reports, interactive dashboards, and ad-hoc queries.

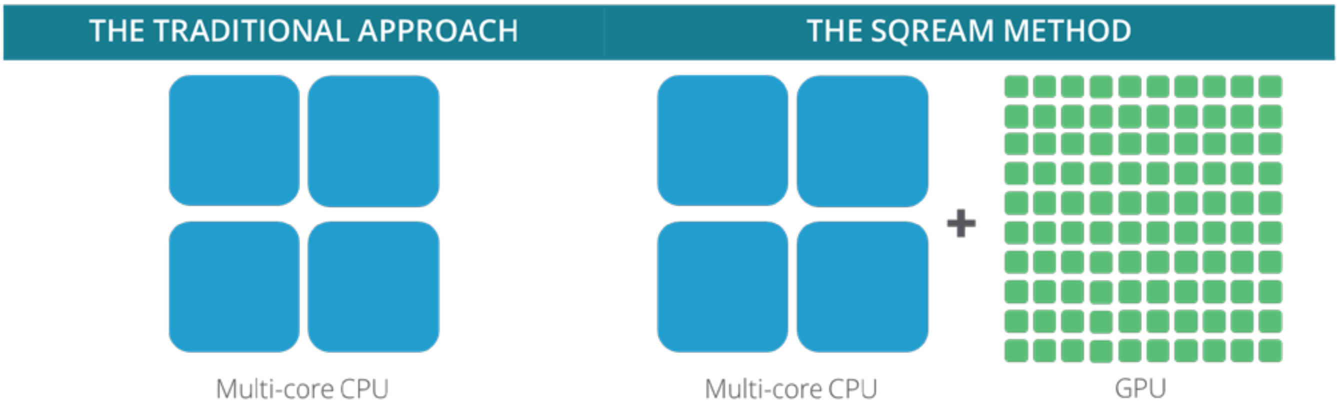


Figure 8 - CPU technology vs. GPU technology

This balance of CPU and GPU operations is key to ensuring optimal performance. GPUs excel at performing repetitive operations on large volumes of data in many streams. The result is faster response times, even on the most complex interactive dashboards.

Furthermore, SQream DB does not require the same careful replication and distribution needed with Hadoop systems, resulting in increased query flexibility, and a quick and straightforward data preparation process that virtually eliminates potential errors. Data is ready to be queried immediately upon loading.

EASY INTERGRATION WITH EXISTING TOOLS FOR AD-HOC ANALYSIS

SQream conforms to the ANSI SQL-92 standard, making interaction with existing systems no different than with any other RDBMS. With SQream DB, however, data professionals will benefit from many advantages under the hood. When a query is issued in a tool like Tableau via ODBC, the SQL command is instantly parsed and converted to relational algebra for further processing and optimizations inside the SQream DB query engine.

Furthermore, the Load-and-Go architecture outlined above, together with Automatic Adaptive Compression and Dynamic Workload Management (WLM) help SQream DB simplify data architectures and integrations even further. The system dynamically responds to changes in the analytic workload, automatically tuning queries and system resources on-the-fly, making it an ideal companion to your existing data warehouse.

Compression is automatic, as is the SQream DB hyper-partitioned table. All operations are performed via standard SQL interfaces and standard connectors for maximum flexibility.

SUMMARY

SQream DB is ideal for organizations seeking to get more insights out of Hadoop to take better business decisions. By combining Hadoop with SQream DB, companies are maximizing the economy of Hadoop where it shines, and the performance of SQream DB for analytics.

Customers using SQream DB load and query significantly more data, while retaining familiar tooling for data analysts and data scientists. Data from varied sources can be used immediately with minimal preparation, a key enabler for data teams' efficiency.

If your data teams are already spending way too much time preparing their data for analysis, if you are seeking to significantly reduce query execution time, or to enable queries that are currently just not running, we urge you to uncover the valuable business insights hidden in your data with a capable solution that complements your Hadoop system.

Contact SQream today to learn more about how SQream DB is helping enterprises worldwide maximize their Hadoop investments.

ABOUT SQREAM

SQream DB combines performance, flexibility, and ease-of-use, empowering and accelerating your data-discovery, so that you can focus on the core of your business, instead of on the infrastructure.

Bring the power of SQream DB to your business with a free trial in the cloud or on-premise at sqream.com/try-sqream-db.

✉ info@sqream.com
⌂ sqream.com
🐦 @SQreamtech