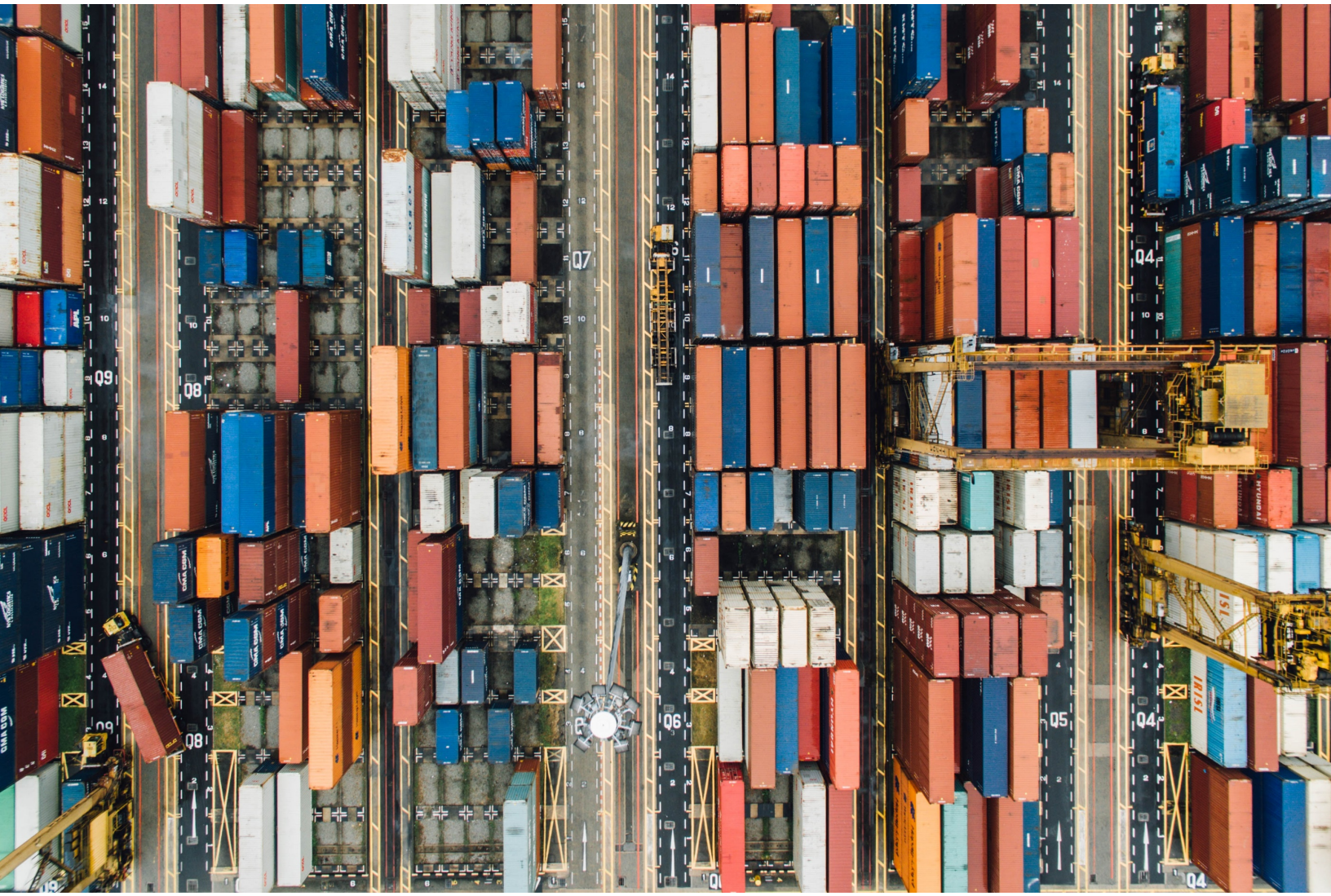
SQREAM DB MPP AUGMENTATION

THE MISSING LINK FOR UNCOVERING BUSINESS-CRITICAL INSIGHTS FROM MASSIVE DATA

SQREAM DB WHITE PAPER



v1.2

 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, or long-running queries that hinder comprehensive analytics. This paper will explain the root cause of these performance issues, and how a new GPU-based approach to data analytics can alleviate them.

INTRODUCTION

A recent Forbes article noted that for a typical Fortune 1000 company, just a 10% increase in data accessibility will result in more than $65 million additional net income. Using retail as an example, the article claimed that retailers who leverage the full power of big data could increase their operating margins by as much as 60%.

Given these numbers, we would expect that enterprises would be doing everything possible to maximize their ability to access and analyze their big data. Yet many organizations worldwide are still facing the challenge of how to effectively analyze their exponentially growing data.

The term 'Big Data' is becoming obsolete as we are faced with data stores of massive new proportions. Data is being ingested at unprecedented rates as existing MPP implementations struggle to keep pace. To get the most out of their growing data stores, enterprises must be able to access and analyze raw data directly and quickly, so that data scientists and analysts can easily explore and receive fast, comprehensive answers to their queries.

However, in a recent Market Guide for Data Preparation, the Gartner analyst firm estimates that organizations spend more than 60% of their time on data preparation, leaving little time for actual analysis. Additional studies reveal that most data scientists spend only 20% of their time on data analysis, while 80% is spent on finding, cleansing, and reorganizing huge amounts of data – resulting in a grossly ineffective data strategy.

The BI pipeline - from the database infrastructure down to the BI tool - has become strained and slow, resulting in three main problems:

- Slow access to data, resulting in an inability to analyze data effectively
- Out-of-date dashboards held-back by long-running queries that cannot be frequently refreshed
- Inflexible infrastructure that prohibits ad-hoc queries on large quantities of historical data

The culprit behind these issues is the lack of advancement in data processing and analytics technology. In the past 30 years, most advancements in this field were small innovations that focused on optimizing specific use-cases or workload sizes. These innovations, in both the infrastructure and the end-user BI visualizers, would become obsolete or outpaced within just a few years due to exponential data growth.

SQream DB introduces a new approach that eliminates data professionals' struggles at the source, resulting in fast, accurate, up-to-the-minute dashboards from which vast new insights can be extracted. SQream's GPU-accelerated data warehouse augments existing implementations to provide fast, unrestricted, ad-hoc access to an organization's full scope of data, even when data grows exponentially.

COMMON ISSUES WITH MPP DATABASES AND DATA WAREHOUSES

For most MPPs to achieve fast OLAP performance, the system must be constantly tuned, with data duplicated in multiple distributions so that the data model closely matches the report's 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 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, materialized views, projections, 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.

THE CPU BOTTLENECK RESULTS IN INCREASED DATA PREPARATION

When queries are very complex and heavily compute-bound, even distributed CPU-bound approaches have their limitations. 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 almost exclusively on fine-tuning database performance. MPP database and data warehouse administrators have a large array of tweaks and optimizations that have been built over 40 years of patching and addressing scalability issues with these solutions.

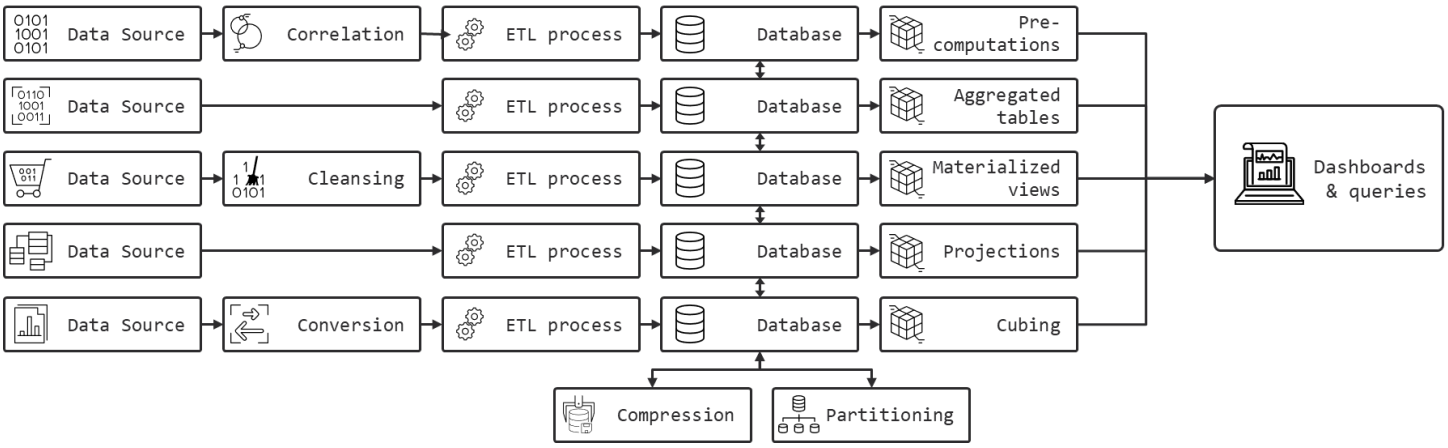


Figure 1 - Common data pipeline with traditional data warehousing methods are quite complex

The problems trickle down to the BI tools and visualizers. For example, one misguided attempt to create a new dashboard could generate complex queries that can bring the data warehouse 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.

INTRODUCING AN ACCELERATED DATA WAREHOUSE

When you have to deep dive into weeks, months, or years of data, even the most expensive and advanced CPU-based MPPs struggle to deliver. SQream DB was created to bridge the gap by bringing massive data analysis capabilities to traditional 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 MPP systems with a simple "lift and shift" of raw data into SQream DB, allowing analysts to focus on extracting new, business-powering insights.

SQream DB ingests data directly from raw files, or with any ETL-capable tool. With a fast ingest rate of 3 TB/hour/GPU, SQream DB is an ideal complement to existing MPP databases.

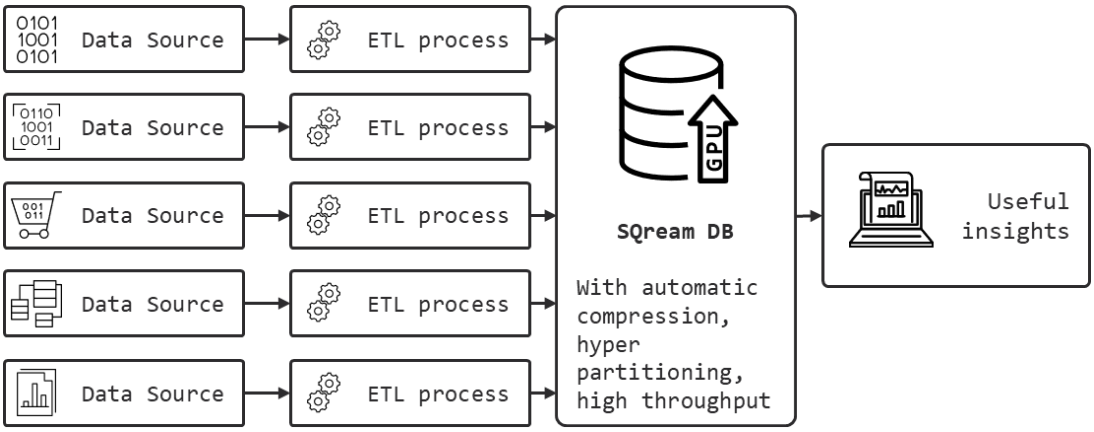


Figure 2 - A typical SQream DB implementation, with no intermediate data preparation needed

With SQream DB in their workflow, analysts can query massive volumes of raw data quickly, ad-hoc, and at scales previously infeasible. The combination of SQream's powerful GPU architecture and patented compression technology enables the ingest and analysis of 20 times more data, 60 times faster, and at 10% of the cost compared to expensive MPP hardware.

MINIMIZING TIME-TO-ANALYSIS: 'LOAD-AND-GO' ARCHITECTURE

SQream DB's GPU-accelerated architecture and automatic optimizations are a key enabler for analyzing data without intermediate steps. SQream DB was developed from the ground up to take advantage of the raw, brute power of the GPU, enabling data analysis immediately after loading. This capability is in stark contrast to most data warehouses, which require time-consuming and insight-limiting processes like indexing, cubing, and projecting. During the ingest process, SQream DB automatically and transparently prepares data for immediate, fast analysis with limited user intervention required.

UNLIMITED SCALABILITY: ANALYZE MORE DATA FOR BETTER INSIGHTS

SQream DB can scale to unlimited data sizes and to numerous data consumers. Because compute is decoupled 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.

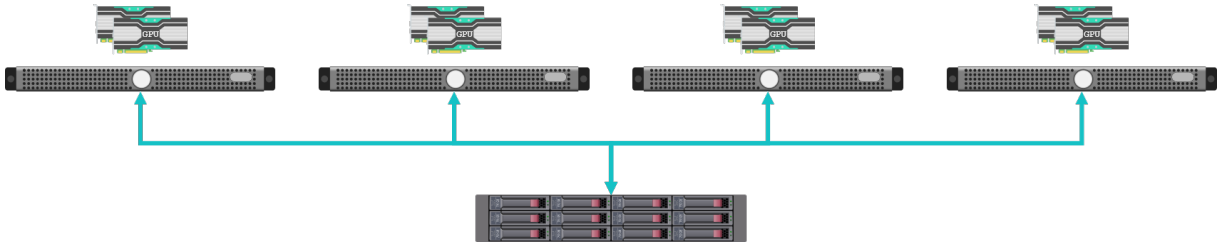


Figure 3 - SQream DB has a shared-data architecture that enables independent scaling of storage and compute

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 4 below is what we call "Load-and-Go".

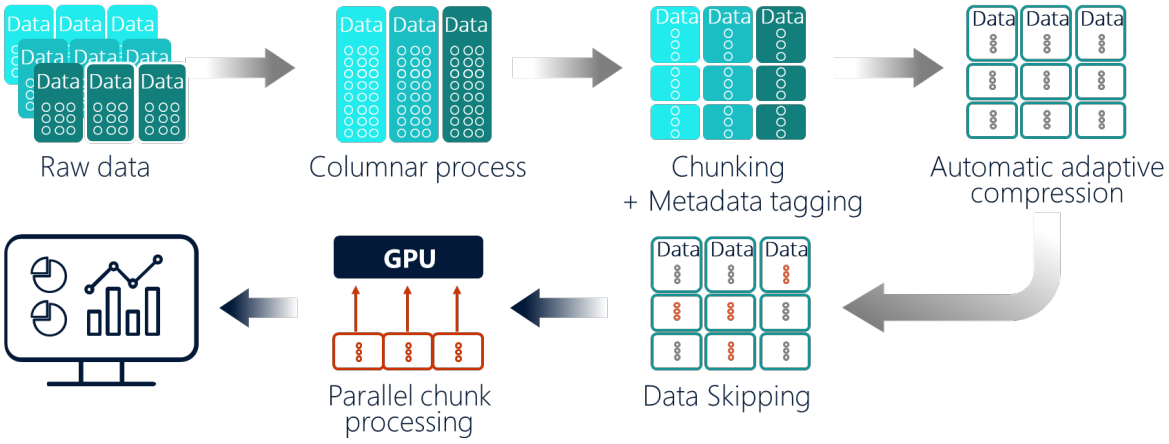


Figure 4 – Load-and-Go architecture. Every step of the process is completely automated, resulting in faster time-to-analysis.

SQream DB includes standard SQL and ODBC 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 an MPP to SQream DB.

FASTER QUERIES, FASTER DASHBOARDS

Traditional data warehouses rely on a fixed set of resources for running all scenarios. In contrast, 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.

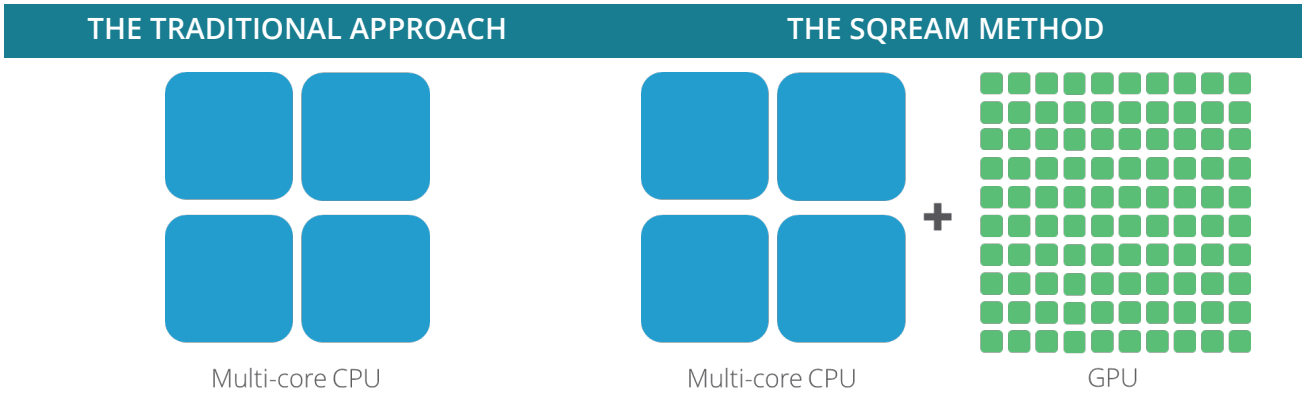


Figure 5 - CPU technology vs. GPU technology

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 CPU-bound MPP 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 MPP DATA WAREHOUSES

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 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 is ideal for organizations seeking to ingest much more data from varied sources in order to maximize the value of their existing data stores; data teams that are spending way too much time preparing their data for analysis and end up with inflexible queries; or enterprises seeking to significantly reduce query execution time, or to enable queries that they currently are unable to execute, and uncover the valuable business insights hidden deep in their data stores.

Contact SQream today to learn more about how SQream DB is helping enterprises worldwide to make the most of their growing data stores.

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.