SQREAM

Accelerating SQL and BI Analytics

Extending Analytical BI with a GPU database

 www.sqream.com

This white paper explores the features that make GPU databases ideal for BI and incorporates real-world use-cases from actual customer implementations. It also explains how you can turn your existing BI pipeline into a more capable, next-generation big data analytics system using powerful GPU technology.

Some GPU databases offer extreme performance with fewer limits than traditional solutions. Ultimately, adopting GPUs and GPU databases to further your analytics capabilities and empowering your team will help you be more productive, and gain real insights into your biggest assets - your data.

Gaining an Edge in Analytics

Organizations worldwide are facing the challenge of effectively analyzing their exponentially growing data stores. The term 'Big Data' is becoming obsolete as we face data stores of massive new proportions.

Today's businesses increasingly rely on leveraging these growing data stores to extract actionable insights into customer behavior, security anomalies, risk analysis, stock and inventory predictions and more. This shift to a data-driven approach has caused many organizations to store their data in large data lakes, but the platforms that analyze the data struggle with the velocity and variety of data.

To get the most out of their data stores, enterprises must be able to access and analyze the raw data directly, so that data scientists and analysts can easily explore and receive fast, comprehensive answers to their queries. When evaluating data analytics platforms, it's important to consider their short-term and long-term capabilities – specifically, the ability to ingest, analyze, and scale queries for future data volumes and more end-users. In the long run, your data analytics platform will enable you to compete more effectively and make better-informed decisions.

THE NEED FOR FLEXIBLE ANALYTICS

Hadoop, NoSQL, and legacy platforms do not offer the level of analytics flexibility allowed by a full SQL analytics platform.

For example, for most SQL-on-Hadoop query engines like Hive and Phoenix systems, only a subset of features is available for understanding your data. Some NoSQL solutions force you to write custom code to achieve business goals. In both scenarios, developing your business logic with queries and moving data around becomes overly complex and taxing.

On the other end of the spectrum, OLAP cubes and pre-aggregated tables have become a common solution for slow SQL queries on products like Oracle Exadata and Teradata. This practice severely hinders BI, as each new question or drill-down into details forces a complete rethink of the entire BI process, which could drag on for many months.

SQream's customers already know how to solve these problems and have created an analytics-driven enterprise with the help of our GPU-powered analytics engine. Let's look at how SQream DB makes analytics work:

ANALYTICAL FUNCTIONALITY IN SQREAM DB

SQream DB conforms with most of the ANSI-92 SQL standard, but adds value-added SQL capabilities like window functions, regular expressions and more. These functions were designed by SQream to scale extremely well and make the most of the multi-thousand-core GPU technology.

Conforming to the ANSI SQL standard is important, because it makes interacting with SQream DB no different than interacting with any other RDBMS, even though SQream has many differences under the hood. A query is issued by the user either directly or through a connector like ODBC or JDBC, as well as native Python and others. The SQL command is parsed and converted to Relational Algebra for further processing and optimizations.

Many SQL engines use *Relational Algebra*, a powerful model based on mathematical theory. The internal operations described as filters and joins are such strong concepts that they are comparable to mathematical basics like addition and multiplication. Relational Algebra is therefore not only well studied, but comprehensively battle-tested in real world applications. By converting SQL queries into clever, highly parallelizable relational algebra operations, SQream DB can efficiently perform complex operations on the massively parallel GPU cores.

AGGREGATE FUNCTIONS

Aggregate functions summarize data over many rows and are commonly used with the GROUP BY clause in a SELECT statement.

SQream DB supports:

- Sum, Min, Max, Count, Average
- Statistical functions like Standard Deviation, Covariance, Variance, and Correlation

These functions were written to be extremely performant and scale well on the massively parallel GPU.

GEOSPATIAL FUNCTIONS

SQream DB includes built-in functions for geospatial analysis, allowing for the creation of geography-specific marketing or a variety of homeland-security use-cases, like geographical sensor fusion.

Target your customers and predict churn by combining location data with customer profiles, POIs, mobile application usage, and more.

ANALYTIC FUNCTIONS

Analytic functions like window functions allow for much faster and simpler queries. Instead of writing thousands of lines in Hadoop or other NoSQL solutions, just a few lines of SQL code are needed.

With a single query, SQream DB can rank the top earners or best-selling products across several time-ranges, calculate moving averages, per-group ranking, running totals, period-over-period reports, and more.

These functions make full use of the GPU's advanced mathematical capabilities and high throughput capabilities.

JOIN ON ANY KEY, EVEN WHEN THEY DON'T MATCH

On top of the standard data warehousing join operations, SQream DB also provides the ability to join on more complex conditions – even when the data types don't match up. This capability enables easily correlating different data sources, without having to resort to normalization of the data series. This is particularly useful in situations where data is typically siloed across different parts of a company's data architecture - often the case in telecoms and retail.

REGULAR EXPRESSIONS

SQream DB supports regular expression pattern matching. This extension lets you effectively filter and sift large amounts of data - particularly useful for clickstream and ad-tech, where you want to identify specific paths. Coupled with Window functions, regular expressions can be a powerful tool in funnel analysis.

PREDICTIVE ANALYTICS

Machine learning is an effective tool for predictive analytics – but it is most effective when applied to large amounts of clean data.

SQream DB interfaces with common machine learning frameworks like Spark MLib, R, and TensorFlow, and can feed them with fast data after it has been "sliced and diced" with the standard SQL preparation techniques. This allows data scientist to leverage the power of SQream DB's big data engine to accelerate business decisions with fewer compromises on data quantity and quality.

CONNECTIVITY

Depending on your current BI pipeline, you may wish to write SQL via your own applications, written in Python, C++, .Net, Node.Js, C++, Haskell, and others.

However, SQream DB supports common database connectivity standards out of the box.

- ODBC – Open Database Connectivity is a standard application programming interface for accessing database management systems. For example, ODBC is used by Tableau.
- JDBC - Java Database Connectivity is an API Java, and provides data access to many applications, like SQL Workbench.
- Ado.Net – ActiveX Data Objects is a data access technology from the Microsoft .NET Framework that provides data access through a common set of components.

QUERY LATENCY AT SCALE

Complex queries contain multiple filters, type conversions, complex *predicates*, exotic *join* semantics, and *subqueries*.

When running this kind of query on large data sets (>100 terabytes in billions of rows across several tables), the number of numerical computations performed is a product of the complexity of the query predicates and the number of rows to be processed.

Even when distributed, a conventional query engine using CPUs alone cannot deliver the result within an acceptable period. The query latency is huge, ranging from many minutes to hours. SQream DB can execute the same query on the same data set with a latency of seconds to minutes.

EXTREMELY FAST DATA INGEST WITH POWERFUL COLUMNAR ENGINE

Like some other analytics databases, SQream DB is a columnar database. This design aspect is well suited for GPU, because GPU operates optimally when the data types are consistent.

SQream DB tables support scalability by hyper-partitioning data in multiple dimensions. We call this chunking. Chunking is automatically and transparently performed during ingest. A user can query and interact with all of their data, just like a regular table. This allows SQream DB tables to grow to sizes that other databases can't support, while retaining familiar management functionality.

The SQream DB table is optimized for fast bulk loading, and can load tables at speeds of up to 3.3TB/h, giving you fast access to your data.

Figure 1 - The SQream DB hyper-partitioned table

SQream DB can ingest flat files like CSVs and Parquet, as well as via network sources like Spark or JDBC. It can be used by itself, or with 3rd party integration and ETL tools. This capability allows SQream DB to address your changing and growing data needs.

SEE THE RESULTS: LEADING TELECOM IMPLEMENTS SQREAM DB

The best way to understand what SQream DB can do for you is to see its impact in real-world customer scenarios.

USE CASE: SQREAM DB VS A LEADING DATA WAREHOUSE

With over 40 million subscribers, a leading network operator strives to differentiate itself through its high customer satisfaction. But in the highly competitive telecom market, keeping so many customers happy and loyal is a challenge. To offer customers a tailored experience, business intelligence users in the operator need to have fast, efficient, and cost-effective access to huge amounts of data.

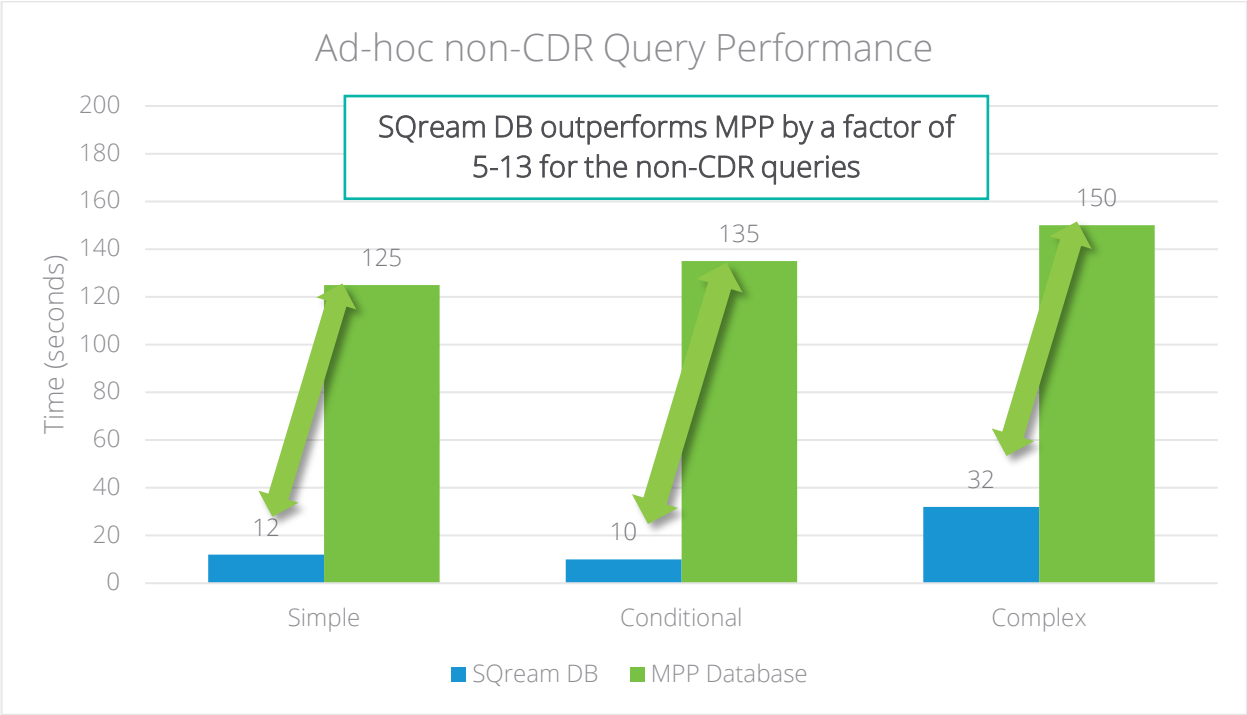
The operator decided to profile SQream DB in comparison with their existing MPP data warehouse, consisting of 40 compute nodes in 5 full racks. The operator wanted to start by analyzing a few months' worth of data, coming in at 1.6TB per week. The data represents CDR (call data records) and non-CDR data, such as customer profiles and customer-registered products.

The server used was a repurposed HP DL380g9 with a powerful NVIDIA Tesla card, further bringing down the project cost.

SQream DB outperformed the incumbent system in all tested scenarios by a factor of 5-18, including data ingest, data compression and query performance.

SCENARIO 1: AD-HOC QUERY PERFORMANCE

Query		Description	MPP	SQream DB	Ratio
1	Simple query	Number of transactions performed on specific products. 5-table join, GROUP BY on 8 columns, filter by day	2:05 m	**0:12 m**	**10.5x**
2	Conditional query	Count distinct mobile numbers with specific orders initiated by online service, that were completed with specific completion code	2:15 m	**0:10 m**	**13.2x**
3	Complex query	Find active or suspended accounts with service call opened on specific days and completed on the following day. Complex join on 6 tables	2:30 m	**0:32 m**	**4.7x**

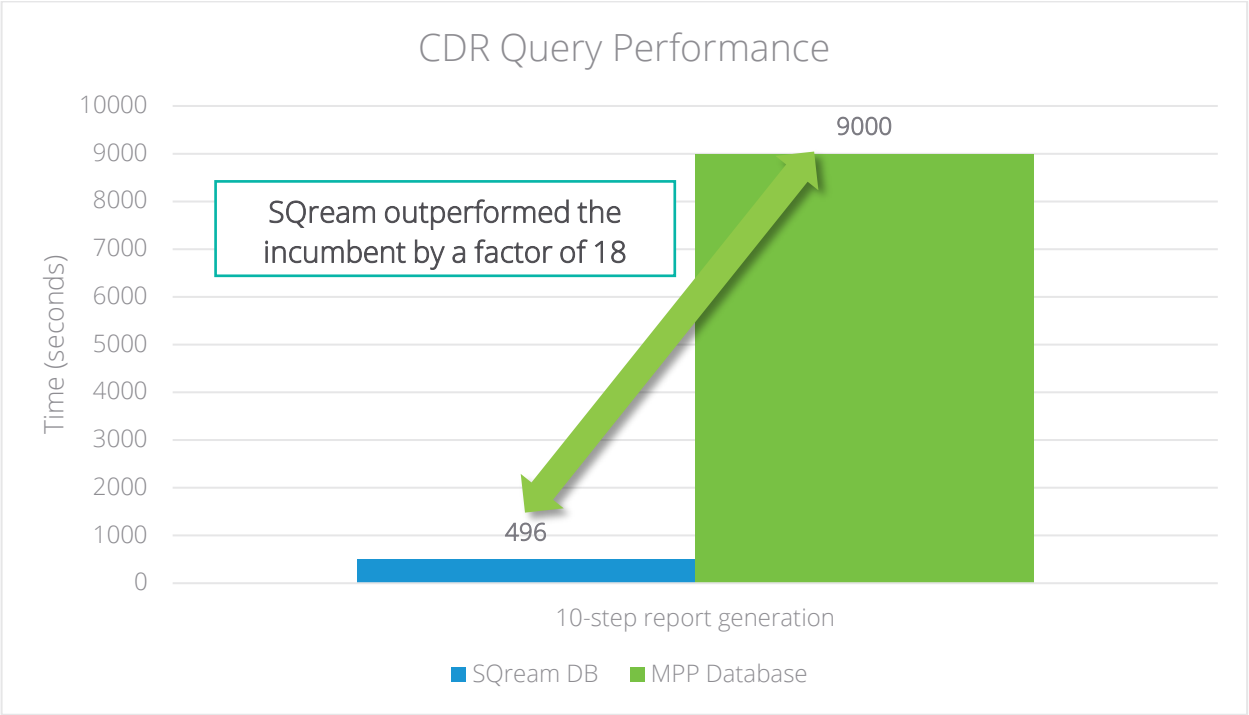


In these ad-hoc scenarios, SQream DB outperformed the MPP data warehouse by a factor of 5-13.

SCENARIO 2: COMPLEX DAILY GRINDER REPORT

In this test, a 10-step report was generated. The report queries data from 13 different tables and combines them into a single result-table used to identify top usage location, segmented by customer. The report uses advanced SQL features, including window functions.

Query		Description	MPP	SQream DB	Ratio
4	10-step report generation	Identify top 3 usage locations for each customer: Identify top 3 used cells by usage during weekends and weekdays, throughout several segments of a day	2-3 hours	**8:16 m**	**18x**



Using the existing MPP data warehouse, this report would typically take around 2-3 hours. SQream DB outperformed the incumbent by a factor of nearly 18, while providing accurate results from raw data, without pre-aggregations and limiting cubes.

SUMMARY

SQream DB is a next-generation data analytics system designed from the ground up to support large-scale analytics on massive datasets. Our platform provides a fast, cost-effective and energy efficient environment for advanced analytics, specifically in the multi-terabyte range where scaling with CPUs is not cost-effective. With standardized SQL, superior scaling and a robust architecture based on standard hardware, SQream DB is a future-proof big data solution.

SQream DB brings you the opportunity to do more with more of your data. Getting fast insights with hundreds of billions of data points is now within reach. SQream DB can be integrated as a standalone database solution or as a complementary analytics database to maximize your IT investments.

Integrating SQream DB is an easy transition from other SQL databases. There is little-to-no rewriting of SQL queries. SQream DB plugs in easily to your existing ecosystem. Because it uses standard SQL and common language bindings, deep learning technologies that also use GPUs, such as TensorFlow and Spark MLib, work “hand in glove” to reduce the time for modeling and learning experiments.

Ultimately, adopting GPUs and GPU databases will enable your data scientists to make better-informed decisions, be more productive, and gain real insight into your biggest asset – your data.

SQream DB combines performance, flexibility and ease-of-use, empowering your data science and making discovery insights in your data fast, allowing you to focus on the core of your business, not on the infrastructure – letting you develop your analytics-driven enterprise.

Learn more about SQream DB

Download the SQream DB white paper (info.sqream.com/download-sqream-db-white-paper) to learn more about how SQream DB can help you analyze more data than ever before, or contact us to speak with one of our database experts - visit sqream.com for more information.

