Analytics query accelerators provide optimization on top of semantically flexible data stores, typically associated with data lake architectures. Data and analytics leaders should use these offerings to accelerate the time to value of their data lake initiatives as they move toward operational production delivery.

Overview

Key Findings

- Data and analytics leaders continue to struggle with getting value from data lake initiatives that have grown to be unwieldy or that cannot deliver adequate performance as they have evolved.

- Analytics query accelerators provide a means of making data in semantically flexible data stores more accessible for production and exploratory use. For those data lakes that store some of their data in semistructured or structured and understood form, the accelerators provide a means of accessing the data in situ.

- Analytics query accelerators are unlikely to replace the data warehouse, but they can make the data lake significantly more valuable for less skilled consumers.

Recommendations

For data and analytics leaders seeking data management solutions to improve the time to value of their data lake initiatives:

- Assess where your performance line of “good enough” is by running your most complex workloads on the evaluated target platform in a proof of concept (POC). Use these tests to determine how much improvement is possible.

- Test integration with surrounding cloud data management services and/or adjacent data management platforms and business intelligence (BI) tools by evaluating APIs and integration touchpoints.

- Evaluate security and governance capabilities to ensure that they meet your enterprise standards and requirements by establishing clear governance and security “must haves.”
Market Definition

Analytics query accelerators provide SQL or SQL-like query support on a broad range of data sources. They are most frequently used as a means of providing interactive and production-optimized delivery on semantically flexible data stores that do not inherently have the capabilities to provide sufficient performance or ease of use on their own. Commonly used in conjunction with data lakes, they aim to provide sufficient SQL or SQL-like support to deliver BI dashboards, interactive query capabilities and support for data modeling. Some also support relational databases as sources and may cross over into the data virtualization or BI markets, though this is not their primary function.

Market Description

The optimization goals of the data warehouse and the data lake are fundamentally different. The former is optimized for production delivery of semantically consistent, well-known data; the latter for semantic flexibility and rapid access to raw data. Data lake practitioners frequently try to deliver the optimization goals of the data warehouse on the architecture (or lack of any) of the data lake. Unsurprisingly, more often than not, they fail.

An easy way to visualize these optimization goals is to use the Data and Analytics Infrastructure Model (see Figure 1 and The Practical Logical Data Warehouse).

Within this model lies a zone of confusion, which arises because the type of work being done in the data lake starts with applying structure to data — sometimes referred to as “schema on read.” That structure, which is necessary for analysts to make sense of the data, more often than not begins to resemble rows and columns — similar to the structures inherent to the data warehouse based on a relational database management system (RDBMS).
The question then arises: “Why can’t we use the data lake exclusively, and retire the data warehouse?” The answer is that the data lake infrastructure, when based on a semantically flexible data store, is generally unable to optimize for the demands of production delivery (concurrency, latency, workload management, etc.) to the degree that the data warehouse can when built on a relational database. Some vendors are using the term “lakehouse” to assert that it is possible to build a structure within the zone of confusion that will surmount these performance challenges.

Another way to tackle the issue if the data lake structure has already been built is to add an analytics query accelerator.

---

The reason we build data warehouses is that we are providing sufficient optimization on known data to make it broadly consumable while meeting performance requirements. Everything else — governance, data quality, data integration, schema design and BI reporting — is a means to that end.

Analytics query accelerators seek to shrink the performance impact of the zone of confusion. Put another way, they are trying to move the “line of good enough” to the point where the data lake can
provide sufficient optimization on the data to make it suitable for an increasing percentage of workloads.

**Market Direction**

Interest in analytics query accelerators is increasing as data and analytics leaders continue to struggle with getting value from their data lake initiatives (see How to Avoid Data Lake Failures). The market is a logical extension of the SQL interfaces to Hadoop and SQL interfaces to cloud object stores — both of which feature in the Hype Cycle for Data Management, 2020. However, most vendors in this market are looking beyond providing a simple SQL query semantic access layer, and are taking an active role in performance optimization, scalability, security access and governance. Depending on the complexity of a workload, data and analytics leaders may find that the products are sufficient to accommodate the SLAs for a percentage of production delivery workloads that originated in the data lake.

Early adopters should tread carefully to avoid the disappointment of “over promise and under deliver,” which plagued prior attempts at solving this problem. Care should be taken to ensure that the analytics query accelerator meets specific requirements — for performance, integration and governance. Over the next three to five years, we expect this technology to play an active role in driving a unification of the data lake and data warehouse into a single, logically defined platform.

**Market Analysis**

Each of the vendors in this market will face challenges as the market continues to pivot toward cloud deployments. Gartner expects the percentage of revenue attributable to cloud in the overall DBMS market to exceed 50% by 2022 (see Forecast: Public Cloud Services, Worldwide, 2018-2024, 3Q20 Update and Forecast: Enterprise Infrastructure Software, Worldwide, 2018-2024, 3Q20 Update). Further, the same macro trends that have become inherent to the overall DBMS market are manifesting themselves in the cloud DBMS market as well: in both cases, nearly 85% of the revenue is attributable to a handful of dominant vendors (see Figure 2).
These market dynamics mean that most of these vendors will not only be competing for the approximately 15% of the market that is not already taken by very large dominant DBMS vendors; they will also be competing against incumbent CSP native services that are far more likely to be well-integrated into a broader cloud data ecosystem (see Cloud Data Ecosystems Emerge as the New Data and Analytics Battleground). The market will grow, however, for all players, as cloud object stores such as Amazon Simple Storage Service (Amazon S3), Microsoft Azure Data Lake Storage (ADLS) and Google Cloud Storage (GCS) are increasingly used as the datastores of choice for many new use cases. Such file structures lack the performance optimization built in to the DBMS typically used for data warehouses.

Proponents of analytics query accelerators will need to highlight their core differentiators, and clearly articulate why this technology should be a part of a broader data and analytics portfolio. Buyers run the risk of relying on a transient technology that is used to solve an immediate problem, but with less secure long-term prospects as competing native CSP offerings continue to develop their capabilities.

Choosing to add an additional product to your toolkit should begin with an assessment of the capabilities of your existing products. Your starting point may depend on where you choose to implement performance optimization for data not preoptimized by design, as is done by a data warehouse. We have identified this starting point as the “Primary Category” in the product table below (Table 1). This category includes the following tools:

- **Accelerator** = A stand-alone analytics query accelerator.
Your strategic DBMS may be the “center of gravity” for your planning, and it may have access to external data and provide optimization that meets your needs. If you are using a data virtualization offering or an IMDG, they may also provide acceleration in addition to their other features. Finally, some BI tools have added acceleration for non-DBMS-resident data to their capabilities.

In this research, we have mostly focused on specialists (called “accelerators” in the table below) that may offer significantly more target data types, semantic mapping, security and other features — and operate in deployment modes not available to your DBMS, such as multicloud, hybrid, container or tightly associated with other tools such as notebooks.

**Representative Vendors**

The vendors listed in this Market Guide do not imply an exhaustive list. This section is intended to provide more understanding of the market and its offerings.

**Market Introduction**

The analytics query accelerator market includes relatively new entrants as well as vendors with a strong foundation built on existing data platforms. Some products are stand-alone, self-contained offerings, while others work in conjunction with broader product suites. Some are available on-premises, some in the cloud and some in both — potentially with the ability to query both at the same time in a hybrid architecture. Numerous optimization technologies are used (see Note 3 for a list). Most vendors choose to highlight specific technologies and these appear in Table 1. Even if a technology is not listed in the table, it may well be included in the product’s capabilities. Some vendors maintain cubes, views and indexes in memory to accelerate performance and then update these when source data changes. Some persist those pieces to disk, meaning no “spin-up time” on the first few queries whenever the system is restarted for any reason. The stored acceleration files may be in a proprietary format available only to the product creating them, or they may be in an open format (Avro, CSV, JSON, ORC, Parquet, etc.) stored with the source data.

Key use cases for an analytics query accelerator include the following:

- Access, explore and combine diverse data types.
• Offload reporting to the data lake for structured data held there — with performance that is “good enough,” often at a cost less than the data warehouse.

• Make the data available for combining with the structured data in the data warehouse and/or data marts, either by providing data virtualization or being able to participate in access through a separate data virtualization software.

• Assist with understanding the underlying structure of the data, and the optimizations needed to be able to access it.

• Use in conjunction with more efficient open formats to make the data both more performant and more portable between analytic engines.

Not all vendors will support all of these use cases equally well.
### Market Recommendations

Data and analytics leaders considering analytics query accelerators to remediate data lake performance and governance concerns, or as a broader logical data warehouse play should do the following:

- Assess where your performance line of “good enough” is by running your most complex workloads on the evaluated target platform in a POC. If a workload fails due to complexity, workload management requirements, performance requirements or other reasons, it is not suitable for the platform, and the next most complex workload should be assessed. Once you have established what percentage of your...
workloads can be accommodated by an analytics query accelerator, they will be able to make informed decisions about where to use it.

- Assess the capabilities of your strategic DBMS vendor to optimize access to the external data you are storing in your data lake. If it performs well enough, an additional product and vendor relationship may not be needed.

- Test integration with surrounding cloud data management services and/or adjacent data management platforms by evaluating APIs and integration touchpoints.

- Evaluate security and governance capabilities to ensure that they meet your enterprise standards and requirements by establishing clear governance and security “must haves.”

- Evaluate the degree to which an offering provides open data access by establishing whether the vendor uses open standards for data like Parquet, ORC, Avro, or others. The use of a proprietary format may have undesirable consequences around vendor lock-in, or access via other APIs.

Evidence
The findings and vendors included in this research draw on:

- Gartner client inquiry data
- Data collected from interactive vendor briefings conducted for analysts
- Data collected by Gartner’s Secondary Research Services team

Note 1: Representative Vendor Selection
The vendors and their analytics query accelerator products listed in this Market Guide were selected because they offer the key capabilities listed in the Market Description section of this report. They are the vendors about which Gartner has received the most client interest (according to searches on gartner.com and our internal client inquiry service), as well as vendors identified as participating in this market by Gartner’s Secondary Research Services team.

Note 2: Gartner’s Initial Market Coverage
This Market Guide provides Gartner’s initial coverage of the market and focuses on the market definition, rationale for the market and market dynamics.

Note 3: Key Technologies Used
Vendors describe similar technologies under several names, some of which are branded. The following list has been curated to eliminate redundancies:
Aggregations, Apache Arrow, Apache Druid, Apache Hadoop, Apache Hive, Apache Iceberg, Apache Ignite, Apache Impala, Apache Kudu, Apache Kylin, Apache MapReduce, Apache Spark, artificial intelligence/machine learning (AI/ML), atomicity, consistency, isolation and durability (ACID) transactions, bytecode generation, caching, Ceph, column store, compression, cubes/multidimensional engine, data virtualization, data layout optimization, data sharpening, Delta Lake, distributed file system virtualization, distributed shared memory, in-memory database management system (IMDBMS), in-memory data grid (IMDG), indexing, Kyvos Engine (OLAP), LLAP (Apache Hive), massively parallel processing (MPP), materialized views, microqueries, parallel query execution, partitioning, Pilosa (in-memory bitmap index, feature-based data format), PrestoDB, PrestoSQL, pruning, push-down query optimization, query optimizer, single instruction, multiple data (SIMD), storage indexes, Tableau Hyper (IMDBMS), Tensorflow, tiered data storage, unified namespace, user-defined functions, workload management.

**Note 4: PrestoDB and PrestoSQL**

Presto is an open-source distributed SQL query engine originally developed at Facebook. Two different source code projects exist on GitHub: [Presto | Distributed SQL Query Engine for Big Data](https://github.com/prestosql/presto), and [Presto | Distributed SQL Query Engine for Big Data](https://github.com/prestosql/presto). This situation is commonly referred to as “forking.”

The Linux Foundation hosts the former PrestoDB Foundation, now called simply the PrestoFoundation (see [Facebook, Uber, Twitter and Alibaba form Presto Foundation to Tackle Distributed Data Processing at Scale](https://www.gartner.com/insight/newsroom/us/press-releases/2018/p4-facebook-uber-twitter-and-alibaba-form-presto-foundation-to-tackle-distributed-data-processing-at-scale)).

PrestoSQL has its own foundation website at [Presto Software Foundation](https://www.prestosql.org). Different commercial implementations base their offerings on one or the other, and we have represented this in the table by referring to PrestoDB or PrestoSQL.

**Recommended by the Authors**

- [Cloud Data Ecosystems Emerge as the New Data and Analytics Battleground](https://www.gartner.com/insight/search?term=Cloud%20Data%20Ecosystems)
- [How to Avoid Data Lake Failures](https://www.gartner.com/insight/search?term=How%20to%20Avoid%20Data%20Lake%20Failures)
- [Hype Cycle for Data Management, 2020](https://www.gartner.com/insight/search?term=Hype%20Cycle%20for%20Data%20Management)
- [The Practical Logical Data Warehouse](https://www.gartner.com/insight/search?term=The%20Practical%20Logical%20Data%20Warehouse)