Market Guide for Analytics Query Accelerators

Published 14 March 2022 - ID G00741458 - 30 min read

By Merv Adrian, Adam Ronthal

Analytics query accelerators provide optimization for diverse data stores associated with data lake, data warehouse and lakehouse architectures. Data and analytics leaders should use these offerings to accelerate time to value of their data lake initiatives.

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 and performant 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 by enabling performance that meets requirements for both business and technical staff.

Recommendations

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

- Assess whether their performance line of “good enough” can be reached by running their most challenging workloads with high numbers of simultaneous users 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. “Coverage” is key.

- Evaluate security and governance capabilities to ensure that they meet enterprise standards and requirements by establishing clear governance and security “must haves.”
- Determine whether the product uses open standards for data like Apache Parquet, ORC or Apache Avro. The use of a proprietary format may have undesirable consequences around vendor lock-in or access via other APIs; however, it can provide more efficient access to the data.

Market Definition

This document was republished on 2 June 2022. The document you are viewing is the corrected version. For more information, see the Corrections page on gartner.com.

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 support BI dashboards, interactive query capabilities, data modeling and other analytics use cases. 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

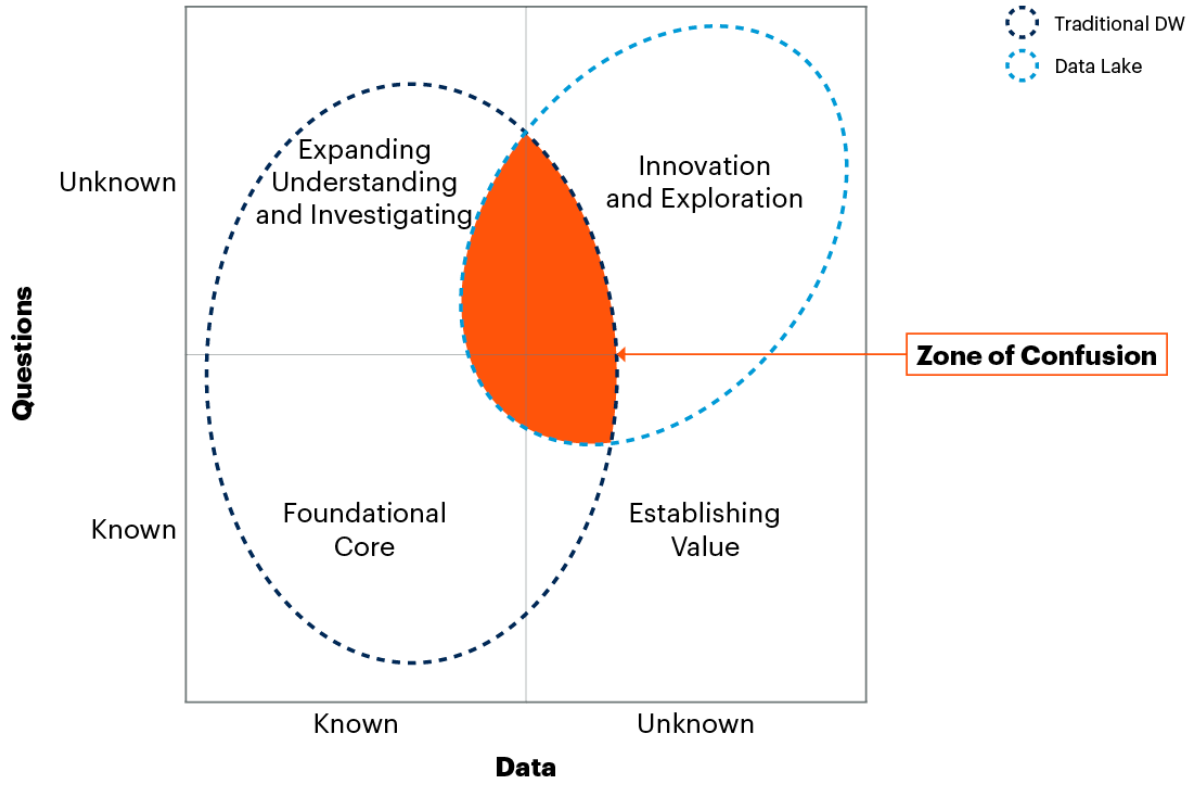
The optimization goals of the data warehouse and the data lake are different. The former is optimized for production delivery of semantically consistent, well-known data; the latter is optimized for semantic flexibility and rapid access to raw data. Data lake practitioners frequently try to deliver the 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 that 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, often begins to resemble rows and columns — similar to the structures inherent to the data warehouse based on a relational database management system (RDBMS).

Figure 1: The Zone of Confusion Within the Data and Analytics Infrastructure Model

The Zone of Confusion Within the Data and Analytics Infrastructure Model



Source: Gartner
DW = data warehouse
734032_C

Gartner.

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 (such as concurrency, latency and workload management) 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. Moreover, “schema on read” on a session-at-a-time basis can lead to conflicting definitions of the data in use, bypassing governance efforts and muddying the clarity of data definitions.

A more manageable 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). Depending on the complexity of a workload, data and analytics leaders may find that the products are sufficient to accommodate the service-level agreements (SLAs) for a significant percentage of production delivery workloads that originated in the data lake as well as to test and model them. The market is a logical extension of the SQL interfaces to Hadoop and the SQL interfaces to cloud object stores — both of which are featured in the Hype Cycle for Data Management, 2021.

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. Market development reflects the presence of the stand-alone vendors discussed here and the increasing addition of similar optimization features by DBMS vendors and analytics/BI tool vendors. The latter can compete with the specialists where they dominate internal use; independents will offer broader reach for both data and tools. As the offerings mature, they are increasingly adding governance capabilities as well.

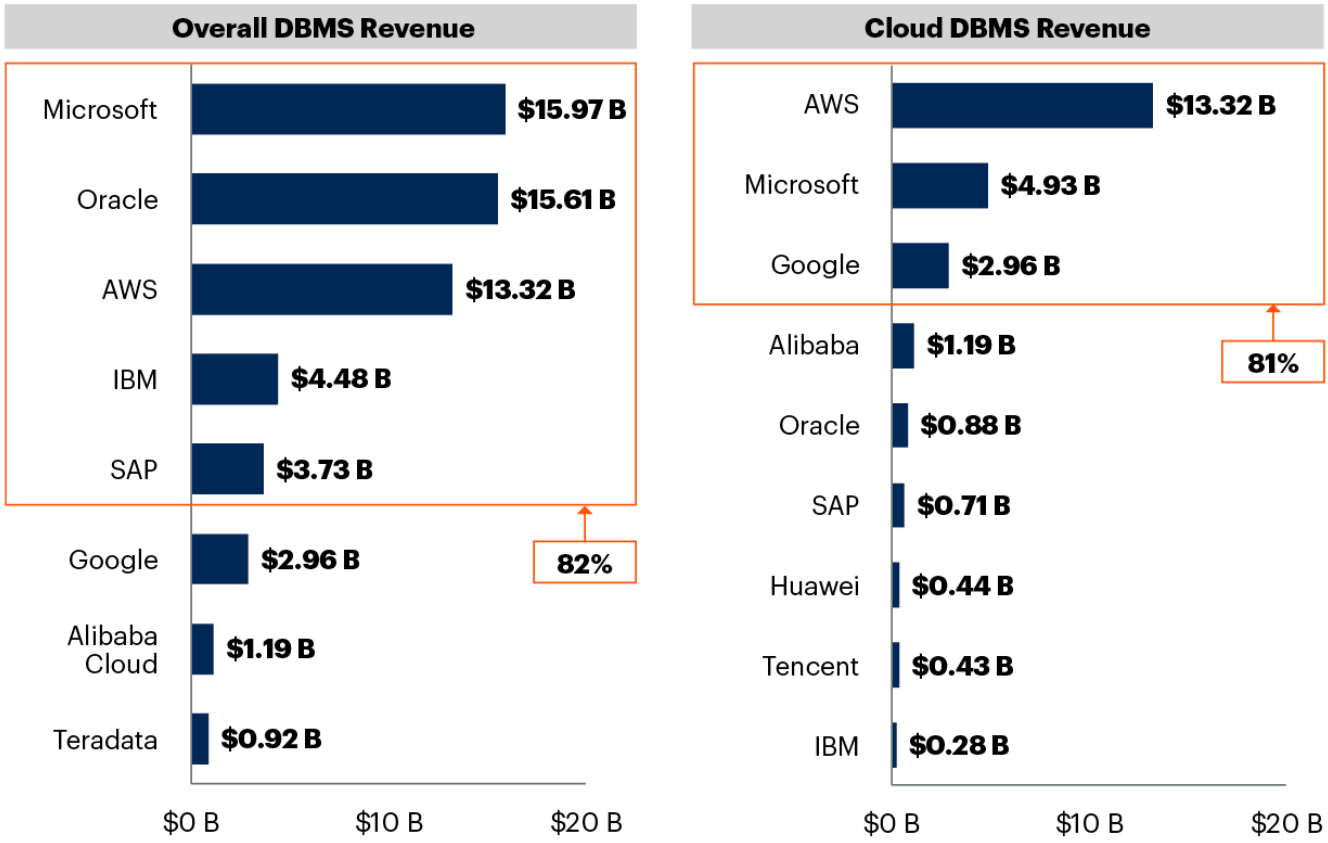
Early adopters should tread carefully to avoid the disappointment of “overpromise and underdeliver,” which plagued prior attempts at solving this problem. Care should be taken to ensure that the analytics query accelerator meets specific requirements for performance, integration and governance. Over 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 resembling Gartner’s Logical Data Warehouse.

Market Analysis

Each of the stand-alone 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% in 2022 (see Forecast: Public Cloud Services, Worldwide, 2019-2025, 4Q21 Update and Forecast: Enterprise Infrastructure Software, Worldwide, 2019-2025, 4Q21 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, over 80% of the revenue is attributable to a handful of dominant vendors (see Figure 2).

Figure 2: DBMS Market Dynamics: Overall and Cloud

DBMS Market Dynamics
Overall and Cloud



Source: Market Share Analysis: Database Management Systems, Worldwide, 2020

Source: Market Share: Enterprise Public Cloud Services, Worldwide, 2020

741458_C

Gartner

These market dynamics mean that these vendors will be competing against incumbent CSP native services that are far more likely to be well-integrated into a broader cloud data ecosystem (see The Impacts of Emerging Cloud Data Ecosystems: An Architectural Perspective).

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 the data stores of choice for many new use cases. Such file structures lack the performance optimization built into the DBMS typically used for data warehouses. Analytics query accelerator vendors will compete with the leading DBMS vendors for access to data in these other data stores, which are often referred to as “big data” repositories even though they may not be particularly large. The attractiveness of the market to investors is clearly shown by the sizable amounts of funding being raised. We have included information about funding raised since December 2020 in the vendor profiles, and numerous prior rounds were in place for several of them.

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, DBMSs, data integration (DI) and BI tools continue to develop their own capabilities.

Choosing to add products to your optimization toolkit should begin with an assessment of the capabilities of your existing portfolio of tools to provide this optimization, and whether you are using them. 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. 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 in-memory data grid (IMDG), it may also provide acceleration in addition to other features. Finally, some analytics/BI tools have added acceleration for non-DBMS-resident data to their capabilities. DBMSs and BI tools do not appear in this research. We highlight stand-alone offerings primarily focused on integrating with and supplementing the products already in place.

In this research, we look at specialists that may offer significantly more target data types, semantic mapping, security and other features. They may operate in deployment modes not available to your DBMS, such as multicloud, hybrid or container, or be tightly associated with other tools such as notebooks. We do not include vendors that primarily focus on other parts of the market like business intelligence visualization tools.

Representative Vendors

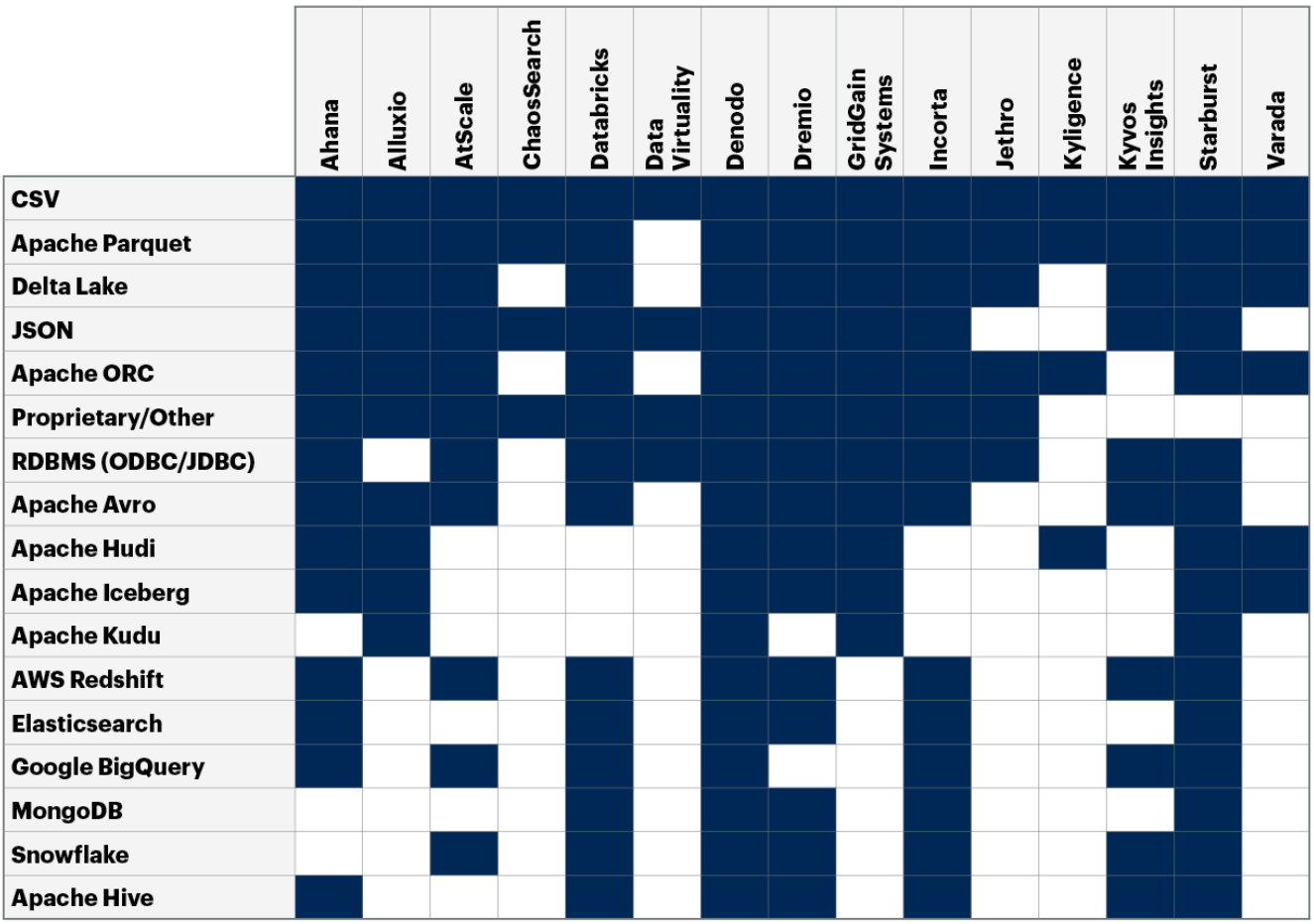
The vendors listed in this Market Guide do not imply an exhaustive list. This section is intended to provide more understanding of the market and its offerings.

Market Introduction

The analytics query accelerator market consists mostly of relatively new entrants. Some products are stand-alone, self-contained offerings, while others work in conjunction with broader product suites. Most are available in a variety of cloud settings, and a few are also available on-premises — potentially with the ability to query data in both locations at the same time in a hybrid architecture. The remarkable diversity of data formats common to data lakes is reflected in the 16 source formats at least three of these vendors support (see Figure 3). Comma-separated value format (CSV), a staple since first used in FORTRAN in 1977, is universally supported, followed by Apache Parquet, JSON and a relatively new entrant: the Delta Lake format pioneered by Databricks. There are many more sources — even in Table 1, we have not attempted to include all possible sources. Consult vendors directly for specific sources you are interested in.

Figure 3: Most Commonly Supported Data Sources by Vendor

Most Commonly Supported Data Sources by Vendor



	Ahana	Alluxio	AtScale	ChaosSearch	Databricks	Data Virtuality	Denodo	Dremio	GridGain Systems	Incorta	Jethro	Kyligence	Kyvos Insights	Starburst	Varada
CSV	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Apache Parquet	■	■	■	■	■		■	■	■	■	■	■	■	■	■
Delta Lake	■	■	■		■		■	■	■	■	■		■	■	■
JSON	■	■	■	■	■	■	■	■	■	■			■	■	
Apache ORC	■	■	■		■		■	■	■	■	■	■		■	■
Proprietary/Other	■	■	■	■	■	■	■	■	■	■	■				
RDBMS (ODBC/JDBC)	■		■		■	■	■	■	■	■	■		■	■	
Apache Avro	■	■	■		■		■	■	■	■			■	■	
Apache Hudi	■	■					■	■	■			■		■	■
Apache Iceberg	■	■					■	■	■					■	■
Apache Kudu		■					■		■					■	
AWS Redshift	■		■		■		■	■		■			■	■	
Elasticsearch	■				■		■	■		■				■	
Google BigQuery	■		■		■		■			■			■	■	
MongoDB					■		■	■		■				■	
Snowflake			■		■		■	■		■			■	■	
Apache Hive	■				■		■	■		■			■	■	

Source: Gartner
741458_C

Gartner

Some vendors create and 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. Some visualizations may be provided to show the structure, or even magnitude, of the data to help designers choose their best approach. 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 (such as Apache Avro, CSV, JSON, ORC or Apache Parquet) stored with the source data.

Numerous other software optimization technologies are used (see Note 2). Most vendors choose to highlight specific technologies to further their marketing differentiation, 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.

A variety of governance capabilities are a natural fit for these tools, beginning with authorization and authentication to ensure that the right users have access to the right data. As analytics query accelerator vendors mature, they are pursuing more sophisticated capabilities, such as integration

with existing data catalogs and security providers, data lineage and data modeling tools, and inbuilt data visualization capabilities.

Key use cases for an analytics query accelerator include:

- 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 analytics engines.

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

Table 1: Representative Vendors in Analytics Query Accelerators

Company Name, Product Name and Platforms	Data Sources Read	Is Data Persisted?	Is Visualization Included?	Governance Features

Ahana Ahana Cloud for Presto AWS	Apache Pinot, Apache Druid, Apache Kafka, Apache Hudi, Apache Avro, AWS Redshift, AWS S3, Azure Blob Storage, CSV, Delta Lake, Elasticsearch, GCS, Google BigQuery, Apache Iceberg, JSON, HDFS, MySQL, ORC, Apache Parquet, PostgreSQL, RDBMS (ODBC/JDBC), text delimited	No	Yes	Apache Ranger, AWS Lake Formation Integration, SQL standards-based authorization, file-based access control plugin at a global level, LDAP, Kerberos, Password file-based authentication, encryption, SSL Support for Presto coordinator, In-VPC deployment, private subnets and highly secure Kubernetes deployment
Alluxio Alluxio Data Orchestration Platform On-premises, Alibaba Cloud, AWS, Azure, GCP, IBM Cloud, Oracle Cloud, Tencent	Apache Avro, CSV, Delta Lake, Apache Hudi, Apache Iceberg, JSON, Kudu, ORC, Apache Parquet	Yes	No	Existing data catalogs, Apache Ranger, Styra OPA, Apache Hive Metastore, Kerberos, LDAP

AtScale AtScale semantic layer On-premises, AWS, Azure, GCP	Apache Avro, CSV, Delta Lake, JSON, ORC, Apache Parquet, RDBMS (ODBC/JDBC), Snowflake	Yes	Yes	Integrations with Alation, Collibra, enterprise directory services (AD, LDAP, Okta), row-level security, object (entity-level security), role-based access control, (RBAC)
ChaosSearch ChaosSearch Data Lake Platform AWS, GCP	Apache Parquet, CSV, JSON, text/log files	Yes	Yes	RBAC, integrates with common SSO providers, SAML 2.0

Databricks Databricks Lakehouse Platform Alibaba Cloud, AWS, Azure, GCP	Apache Avro, CSV, Delta Lake, JSON, ORC, Apache Parquet, Apache Hive, JSON, LZO, XML, Zip, Amazon Redshift, Amazon S3, Azure Blob, Azure Data Lake Storage, Azure Cosmos DB, Azure Synapse Analytics, Cassandra, Couchbase, Elasticsearch, Google BigQuery, Google Cloud Storage, MongoDB, Neo4j, Oracle, RDBMS (JDBC), Redis, Riak TS, Snowflake	Yes	Yes	Fine-grained access control, third-party catalogs — Immuta, Privacera for access control, Collibra, Alation for governance and discovery

Data Virtuality Data Virtuality Platform On-premises, AWS, Azure	CSV, JSON, Apache Parquet, RDBMS (ODBC/JDBC), >100 SaaS/Proprietary connectors, >50 DBMS connectors	Yes	No	Data lineage, column masking, metadata dependency viewer (data lineage and impact), metadata catalog with custom attributes and metadata search, versioning for custom metadata, open metadata interface, connectors to existing data catalogs

Denodo Denodo Platform On-premises, AWS, Azure, GCP	Apache Avro, CSV, Delta Lake, Apache Hudi, Apache Iceberg, JSON, Kudu, ORC, Parquet, RDBMS (ODBC/JDBC), enterprise package applications, flat and binary files. Refer to Denodo's datasheet for a comprehensive list of data sources.	Yes	No	Modeling tools (ERwin, ER/Studio), RDF/OWL ontologies, data security and authorization rules, masking, data lineage, integration with IBM Information Governance Catalog, Informatica Enterprise Data Catalog, Collibra, Alation, Talend Data Governance, ERwin Data Governance, Oracle Metadata Management, Ovaledge, metaintegration, identity providers including SSO and two-factor authentication, security key vaults, and integrates with CyberArk, HashiCorp, AWS Secrets Manager, Azure Key Vault

Dremio Dremio On-premises, AWS, Azure, GCP, own platform, Kubernetes	Apache Arrow, Apache Avro, CSV, Delta Lake, Apache Hudi, Apache Iceberg, JSON, ORC, Apache Parquet, RDBMS (ODBC/JDBC), XLSX, Elasticsearch, MongoDB, MinIO, Pure Storage, EMC, NAS, AWS S3, Azure Storage, GCP, GCS	No	No	RBAC, row and column masking, data catalog, data graph, security, integrations with ecosystems tools like Collibra, Okta
GridGain Systems GridGain In-Memory Computing Platform On-premises, private or public cloud infrastructure	Apache Arrow, Apache Avro, CSV, Delta Lake, Apache Hudi, Apache Iceberg, JSON, Kudu, ORC, Apache Parquet, RDBMS (ODBC/JDBC)	Yes	No	Authentication, authorization, encryption, integration with data lineage and data catalog tools

Incorta Incorta On-premises, AWS, Azure, GCP	Apache Avro, Apache Hbase, Apache Hive, Amazon Redshift, CSV, Delta Lake, Google BigQuery, Elastic, JSON, MongoDB, Snowflake, Apache Parquet, RDBMS (ODBC/JDBC), and numerous third-party drivers. For a complete list, click here.	Yes	Yes	Data lineage, fine-grained (row- and column-level) security, RBAC, SSO, views
Jethro Jethro On-premises, AWS, Azure, GCP	CSV, Delta Lake, ORC, Apache Parquet, RDBMS (ODBC/JDBC)	Yes	No	Integration with data virtualization tools
Kyligence Kyligence Cloud On-premises, AWS, Azure, GCP, Huawei Cloud	CSV, Apache Hudi, ORC, Apache Parquet, RDBMS (JDBC)	Yes	Yes	Active Directory, data cell-level ACL

Kyvos Insights Kyvos On-premises, Cloudera, AWS, Azure, GCP	Apache Avro, CSV, Delta Lake, JSON, Apache Parquet, RDBMS (ODBC/JDBC), ADLS, S3, Google Cloud Storage, Redshift, Snowflake, BigQuery, Delta Lake, AWS Glue, Apache Hive	Yes	Yes	APIs for integration with catalogs, AD/LDAP, RBAC, SSO, Okta, row- and column-level security

Starburst Starburst Enterprise, Starburst Galaxy On-premises, Alibaba Cloud, AWS, Azure, GCP, HPE Marketplace, Red Hat Openshift	Apache Avro, CSV, Delta Lake, Apache Hudi, Apache Iceberg, JSON, Kudu, ORC, Apache Parquet, RDBMS (ODBC/JDBC), BigQuery, IBM Db2, DynamoDB, Delta Lake, Greenplum, Apache Hive, IBM Cloud Object Storage, MinIO, Apache Iceberg, Kafka, MySQL, IBM Netezza, Oracle, PostgreSQL, Redshift, SAP HANA, Salesforce, SingleStore, Snowflake, Splunk, SQL Server, Microsoft Synapse, Teradata, Vertica, Trino, Accumulo, BlackHole, Cassandra, ClickHouse, Druid, Elasticsearch, Google Sheets, JMX, Kinesis, Kudu, MongoDB, Phoenix, Apache Pinot, Prometheus, Redis, Thrift	Optional	Yes	RBAC, data masking, audit, integration, Apache Ranger, Apache Atlas, Apache Sentry, Immuta, Privacera, Alation, Collibra, Amundsen

Varada Vrada On-premises. AWS, Azure, GCP	CSV, Delta Lake, Apache Hudi, Apache Iceberg, ORC, Apache Parquet	Yes	Yes	Apache Ranger

Source: Gartner (March 2022)

Vendor Profiles

Ahana

Ahana is headquartered in San Mateo, California. It offers Ahana Cloud for Presto, a managed service on Amazon Web Services (AWS). It is positioned for accelerated SQL performance on AWS S3 data and a large number of open-source data formats including Apache Druid and Apache Pinot as well as CSV, delimited text files and RDBMS. It integrates with AWS Lake Formation and features security based on Apache Ranger, including RBAC, adding encryption and private subnets and in-VPC support. Ahana is a key contributor to open-source Presto and a premier member of the Presto Foundation, which operates under the auspices of the Linux Foundation. Ahana provides visualization capabilities. It does not persist any data. Ahana raised $20 million in Series A funding in August 2021.

Alluxio

Alluxio is headquartered in San Mateo, California. It provides a free open-source Community Edition and a commercial Alluxio Enterprise Edition. Both editions can run on any suitable computing environment on-premises or in the cloud. Alluxio positions itself as a data orchestration tier that can help to unify multiple disparate data sources, as well as provide acceleration and performance caching capabilities for query access. It is integrated with both PrestoDB and Apache Spark for query acceleration, persists data, and supports Amazon EMR, cloud storage, HDFS, and multiple on-premises storage offerings with hybrid and multicloud support. No visualization capabilities are included. Alluxio also supports Intel Optane Persistent Memory. Alluxio has announced enhancements to its support for machine learning, as well as support for Kubernetes. The company recently raised $50 million in Series C funding.

AtScale

AtScale is headquartered in Boston, Massachusetts. It offers the AtScale semantic layer on AWS, Azure and GCP, as well as on-premises private clouds. It supports business views of data via its semantic layer and a number of open-data-source formats as well as RDBMS and CSV files. Integration with common BI tools is facilitated via MDX, DAX and SQL support as well as by integration with enterprise directory services. Alation and Collibra integration provide added catalog and governance support. AtScale provides its own visual modeling tool, automates

materialized view creation for persisting aggregates and cubes, and optimizes queries that are pushed to underlying data platforms. Role-based access control and object-level security are also included. AtScale recently introduced Python support, and created an executive-vice-president- and general-manager-level position for machine learning, signaling a key roadmap direction.

ChaosSearch

ChaosSearch is headquartered in Boston, Massachusetts. Its ChaosSearch Data Lake Platform is available on AWS and GCP. With a strong historic focus on log analytics, its supported sources are text/log files, Apache Parquet, JSON and CSV. ChaosSearch persists data inside a client's VPC. Its optimizations include those for data layout optimization, distributed shared memory, materialized views and the use of storage indexes. ChaosSearch provides visualizations via its own interface, Kibana, and Elasticsearch, and provides open APIs for expanded search and SQL access to BI and analytics tools. Role-based access control is provided, as well as integration with common single sign-on (SSO) providers. GDPR-compliant log analytics has been a key driver. The firm refers to client use of analytics tools like Kibana and the Elasticsearch API with ChaosSearch. ChaosSearch is extending its portfolio to support data lakes more broadly with availability to run on the Azure cloud platform and deliver direct API access for machine learning libraries. The company raised $40 million in Series B funding in December 2020.

Databricks

Databricks is headquartered in San Francisco, California. It offers data engineering, data science and machine learning as well as analytics. The Databricks Lakehouse Platform is available on Alibaba Cloud, AWS, Google Cloud Platform and Microsoft Azure (Azure Databricks), but not on-premises. The Lakehouse Platform consists of data stored in a data lake, including the open-source Delta Lake format it introduced, the high-performance query engine and the Unity Catalog, which provides fine-grained governance for data assets. The Delta Lake format adds metadata and structures to the underlying data to deliver the capabilities of a traditional data warehouse. Databricks also partners with leading vendors for catalog and governance support. In 2021, Databricks introduced Photon, a native vectorized engine that provides extremely fast query performance at low costs directly on the data lake. Optimizations include cost-based query optimization across SQL, Python and Scala, and query push-down. Databricks supports one of the largest numbers of data sources in this collection of vendors. The company raised $1.6 billion in a Series H round in August 2021.

Data Virtuality

Data Virtuality is headquartered in Leipzig, Germany. It offers the Data Virtuality Platform, primarily positioned as a data virtualization engine providing a common semantic access layer across multiple data sources. Data Virtuality aims to present the entire data space in relational form providing federation, connectivity, and the ability to transform, manage and move data. Materialization of data in an optional data persistence tier for enhanced performance for analytics relies on a choice of supported analytics storage platforms including Azure Synapse Analytics, Exasol, Google BigQuery, PostgreSQL, SAP HANA, SQL Server and Snowflake. Metadata management, security, data lineage and governance capabilities round out the solution.

Denodo

Denodo is headquartered in Palo Alto, California. It offers Denodo Platform (both on-premises and on AWS, Azure and Google Cloud Platform via Denodo Cloud), and Denodo Express, a free, community-supported version. Primarily a data virtualization platform, Denodo offers a sophisticated query optimizer. It provides access to a large set of data sources, including Apache Hudi, Apache Iceberg and enterprise packaged applications. Its Smart Query Acceleration for Analytics feature detects common query patterns and creates precomputed summaries, materialized in any data source to provide a fast response to frequently asked queries. Metadata management, data science and streaming integration round out the solution.

Dremio

Dremio is headquartered in Santa Clara, California. It is available on AWS, Azure and GCP, as well as on-premises. Positioning itself as an SQL Lakehouse platform provider without the need for additional copies of data or its own persisted data, Dremio leverages Apache Arrow, which it was instrumental in commercializing. It offers access to a large number of data sources, both open and proprietary, including RDBMS. Of particular note is its access to Elasticsearch, MongoDB and MinIO object stores. Dremio utilizes numerous optimization techniques including storage indexes, materialized views with rewrite and substitution, predictive pipelining and bytecode generation. It integrates with data catalog tools such as Collibra and identity providers like Okta. Data is encrypted at rest and in flight, and dynamic data masking and support for customer-managed encryption keys are provided. Dremio raised $160 million in a Series E round in January 2022.

GridGain Systems

GridGain Systems is headquartered in Foster City, California. The GridGain In-Memory Computing Platform is available on-premises and in private and public clouds, and supports numerous open and proprietary data sources including Apache Hudi, Apache Iceberg, Apache Kudu, MinIO object stores and RDBMS. It persists data but does not offer its own visualizations. Built on Apache Ignite, GridGain uses distributed shared memory, in-memory bitmap indexes, and offers user-defined functions. Security features include encrypted communications, role-based authentication and authorization, and performing audits based on events within the system. GridGain Nebula adds a cloud-native deployment option. The vendor’s roadmap includes a very strong focus on analytics with support for columnar storage, GPU-based processing and aggressive exploitation of Intel’s Optane Persistent Memory.

Incorta

Incorta is headquartered in San Mateo, California. The Incorta Direct Data Platform provides a unified analytics platform for data acquisition, storage, analysis, visualization and reporting. Its Direct Data Mapping technology acts as an acceleration tier on raw or normalized data to determine potential query paths and provide query performance without having to model or transform the data. It supports several open and proprietary data sources including Apache Parquet, CSV, Delta Lake and RDBMS. Incorta persists data in a columnar format on disk and in memory, dynamically moving compressed data into memory as needed. Data lineage and security capabilities round out the solution. In 2021, Incorta raised $120 million in series D funding.

Jethro

Jethro is headquartered in New York, New York. It offers Jethro on AWS, Azure and GCP, and on-premises. Jethro combines indexing technology and automated cubes to accelerate query performance against open-data formats such as Apache Parquet and Delta Lake as well as CSV and RDBMS. A relatively small vendor, Jethro stresses its partnerships with BI tool vendors, especially for their use with Hadoop. It uses in-memory bitmaps, materialized views, SIMD and storage indexes. Data is persisted. No visualization is included.

Kyligence

Kyligence is headquartered in San Jose, California, and it has several offices in the U.S. and in China. Kyligence Cloud is available on AWS, Azure, GCP and Huawei Cloud, as well as on-premises. It commercializes and extends Apache Kylin to provide high-performance and high-concurrency data services for analytics and applications. It supports several common open-data formats as well as RDBMS and CSV. Kyligence integrates with Active Directory and provides cell-level access control list-based security. User-defined functions, storage indexes, push-down queries and bytecode generation are among the supported optimizations. Kyligence raised $70 million in a Series D round in 2021.

Kyvos Insights

Kyvos Insights is headquartered in Los Gatos, California. The Kyvos platform leverages Smart OLAP technology to accelerate queries on data lakes on AWS, Azure and GCP, and with on-premises Cloudera deployments. Kyvos supports cloud object stores, Databricks Delta Lake, Hadoop, Snowflake, Amazon Redshift, Google BigQuery and other sources, providing a BI acceleration layer. It builds multidimensional cubes, persists data, and uses other data and query optimization techniques (including machine learning) to improve performance over very large datasets. Row- and column-level security and integration with security protocols like Kerberos, Apache Knox, Apache Ranger and Apache Sentry, along with a built-in visualization layer, complete the offering.

Starburst

Starburst is headquartered in Boston, Massachusetts. Starburst Enterprise is available on Alibaba Cloud, AWS, Azure, GCP, HPE Greenlake Marketplace, Red Hat OpenShift and on-premises as a query engine for data warehouse, data lake or data mesh. It offers a very large set of source data formats including Apache Hive, AWS Redshift, Delta Lake and Snowflake, as well as open-data formats like Apache Avro, Apache ORC, Apache Parquet and some less frequently supported sources such as ClickHouse, Salesforce and Splunk. Starburst Enterprise is based on open-source Trino (formerly PrestoSQL) and includes features such as a cost-based optimizer along with its massively parallel processing (MPP) engine, materialized views and user-defined functions. It includes native role-based access control as well as integration with Apache Ranger, Apache Atlas and Apache Sentry, provides data masking, and integrates with Immuta and Privacera. Its catalog and governance functions integrate with Alation, Collibra and open-source Amundsen, among others. Starburst raised $250 million in a Series D funding round in 2022.

Varada

Varada is headquartered in Tel Aviv, Israel. Varada, available on AWS, Azure and GCP and on-premises, uses dynamic indexing techniques combined with caching to enhance Trino (formerly PrestoSQL) and accelerate performance on cloud object stores, Apache Hive, and relational and nonrelational databases. Data tiering ensures that frequently accessed data is stored and indexed in SSD NVMe attached nodes, while cooler data remains in the data lake for retrieval if necessary. Varada uses a variety of indexes including bitmap, dictionary, trees and others, and automatically selects the most effective index based on data content and structure. Varada completed a $12 million Series A funding round in 2020.

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:

- Assess where their performance line of “good enough” is by running their 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, you will be able to make informed decisions about where to use it.
- Reassess the capabilities of their strategic DBMS vendor and analytics tool(s) of choice to optimize access to the external data they are storing in their data lake. If they perform 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 their enterprise standards and requirements by establishing clear governance and security “must haves.” Avoid conflicts with existing tools by setting clear coverage assignments for each and leveraging integration where available.
- Evaluate the degree to which an offering provides open-data access for persisted data by establishing whether the vendor uses open standards for data like Apache Parquet, ORC, Apache Avro or others. The use of a proprietary format may have undesirable consequences around vendor lock-in or impede 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 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; 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; and workload management.