Critical Capabilities for Data Warehouse and Data Management Solutions for Analytics

Published: 21 April 2015

Analyst(s): Roxane Edjlali, Mark A. Beyer

This Critical Capabilities research combines experience of existing client implementations with an evaluation of vendor capabilities. It will help information leaders to select the right technology to evolve their information infrastructure in support of diverse and evolving analytical needs.

Key Findings

- Data management solutions for analytics have evolved into four critical use cases: traditional, logical, operational and context-independent. The "traditional" use case continues to be the most common, and trending indicates it will remain so in the majority of the market through 2018.

- Incumbent vendors have evolved their product offerings to address emerging, modern use cases, but their installed base is still mainly using traditional data warehouse scenarios.

- Emerging vendors are entering the market addressing wide access to varied types of data and reduced dependency on a predefined data model.

- The adoption rate for modern use cases is increasing year over year by more than 50% — but the net percentage for the "context-independent" and "logical" data combined remains below 15% of the total market.

Recommendations

Information leaders should:

- Evaluate their current data warehouse (DW) and data management solution for analytics (DMSA) products for their ability to evolve traditional and "operational" DWs or otherwise address the newest use cases of context-independent and logical DWs, and plan on skills training for leveraging additional features and functions.

- Use this research to identify solutions to augment current information infrastructure in order to support new use cases but understand that it does not replace either an RFP or a proof of concept (POC).

Use this research to evaluate adoption of your current vendor technology in support of new use cases alongside our "Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics," which has insights into the timing for introducing new use cases and workloads.

- Create a plan for reaching a level of proficiency in your implementations within two to three years to meet the requirements of your new upcoming use cases and increased data variety.

What You Need to Know

This research is aimed at information management leaders, business intelligence (BI) strategists and advanced analytics implementation teams and will help them to evaluate the capabilities of current vendors. In this Critical Capabilities document, we have focused on the 10 most important functional capabilities that are required to support the major use cases identified. Importantly, this research combines both the customer experience and prevalence of capabilities among vendor customers as an indication of use cases in which clients are using the vendor technology, as well as specific vendor product capabilities as described in the capability definition. User experience is evaluated based on companion Magic Quadrant reference survey, Gartner inquiries, in-depth reference calls and interactions with vendors. In addition to customer experience, capability ratings include Gartner analysis of differentiating product capabilities as described in the capability definitions. Finally, vendors that have multiple products are evaluated across their combination of products and not for each product separately. Last year's Critical Capabilities research note was almost exclusively based on customer experience, while in this year's revision, we are extending the capability definitions to take into consideration both customer feedback and specific vendor capabilities. This new approach has created some major shifts in the capability ratings.
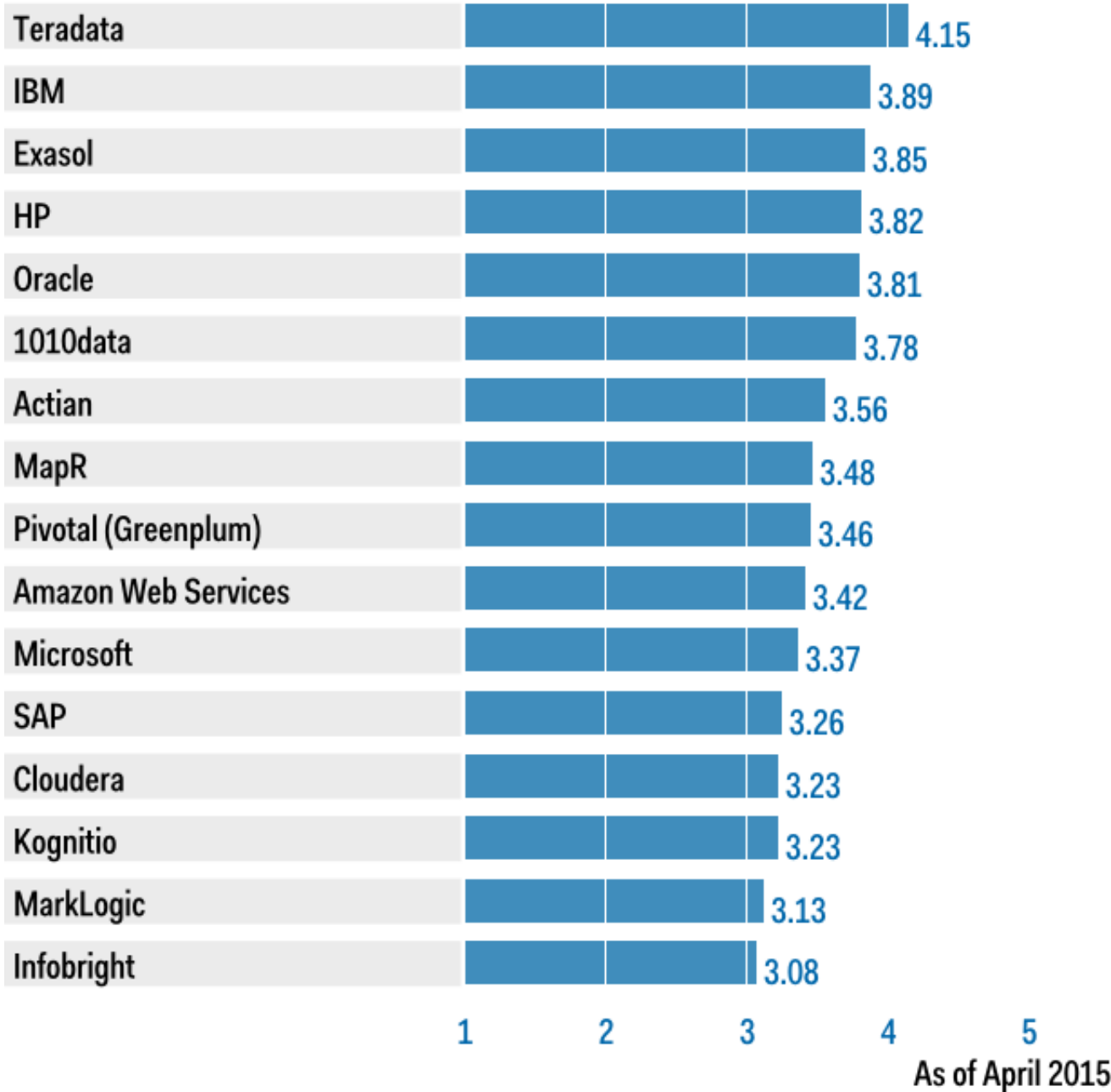
This research does not include all criteria that should lead organizations to selecting a data warehouse (see Note 1) or DMSA vendor. There are many other criteria that will come into play in selecting a vendor — whether it's a stand-alone DBMS software package, appliance or cloud solution — that are not included in this analysis. Other requirements — such as pricing, vertical industry offerings, the availability of services and many other criteria — are not included but would need to be included in a formal RFP process.

Analysis

Critical Capabilities Use-Case Graphics

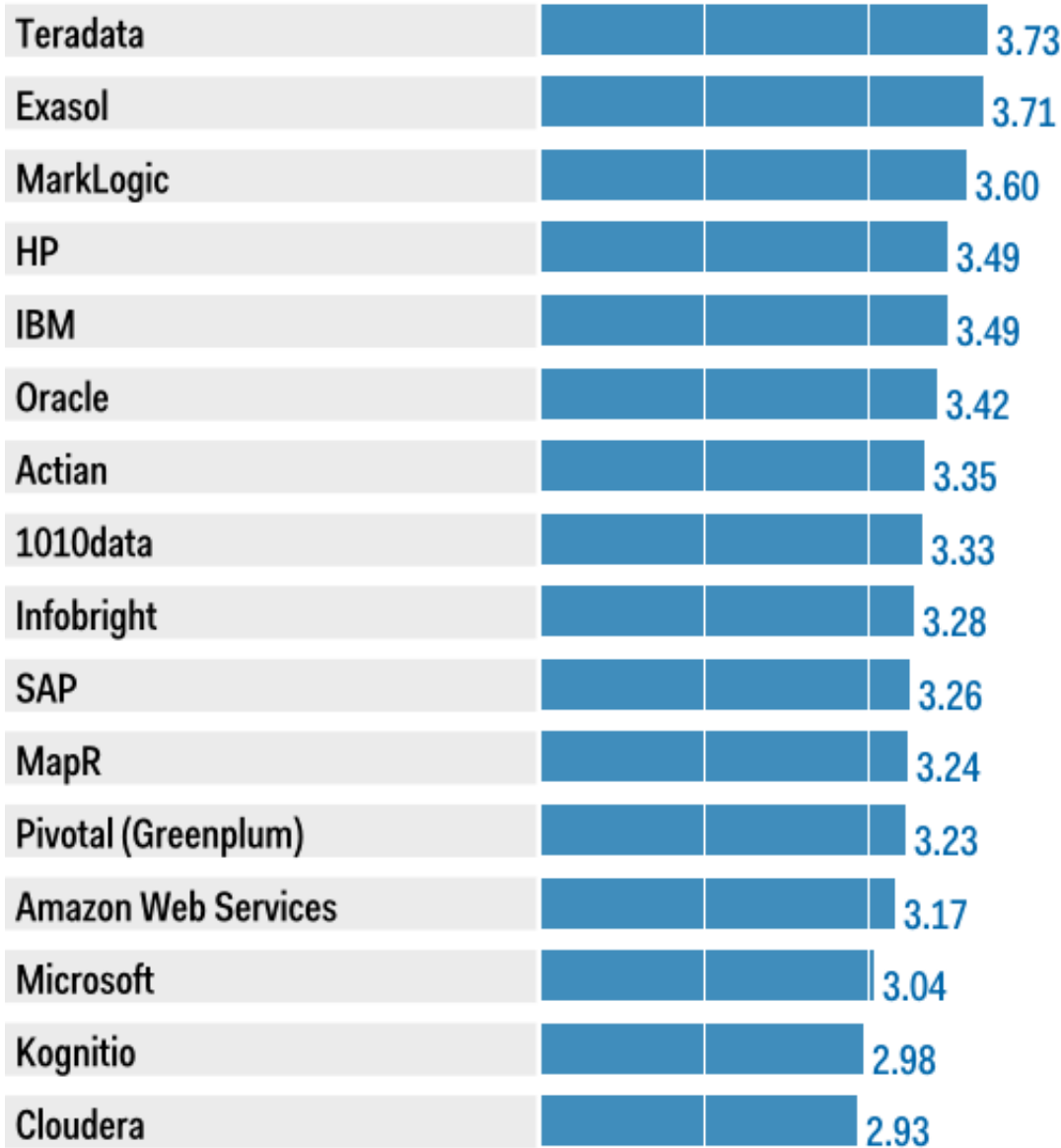Figure 1. Vendors' Product Scores for Traditional Data Warehouse Use Case

Product or Service Scores for Traditional Data Warehouse

Vendor	Score
Teradata	4.15
IBM	3.89
Exasol	3.85
HP	3.82
Oracle	3.81
1010data	3.78
Actian	3.56
MapR	3.48
Pivotal (Greenplum)	3.46
Amazon Web Services	3.42
Microsoft	3.37
SAP	3.26
Cloudera	3.23
Kognitio	3.23
MarkLogic	3.13
Infobright	3.08

As of April 2015

Source: Gartner (April 2015)

Figure 2. Vendors' Product Scores for Operational Data Warehouse Use Case

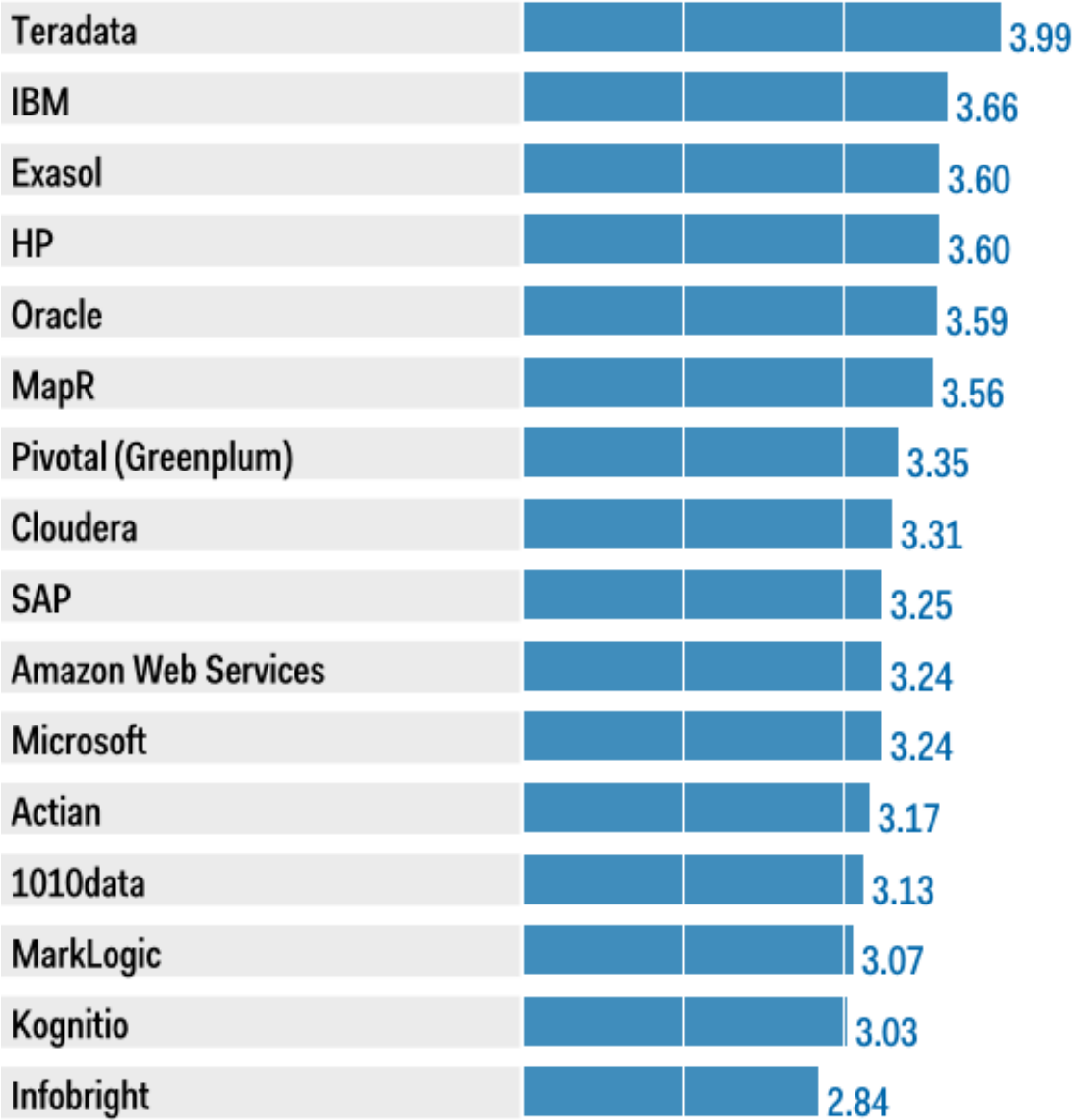## Product or Service Scores for Operational Data Warehouse

Vendor	Score
Teradata	3.73
Exasol	3.71
MarkLogic	3.60
HP	3.49
IBM	3.49
Oracle	3.42
Actian	3.35
1010data	3.33
Infobright	3.28
SAP	3.26
MapR	3.24
Pivotal (Greenplum)	3.23
Amazon Web Services	3.17
Microsoft	3.04
Kognitio	2.98
Cloudera	2.93

As of April 2015

Source: Gartner (April 2015)

Figure 3. Vendors' Product Scores for Logical Data Warehouse Use Case

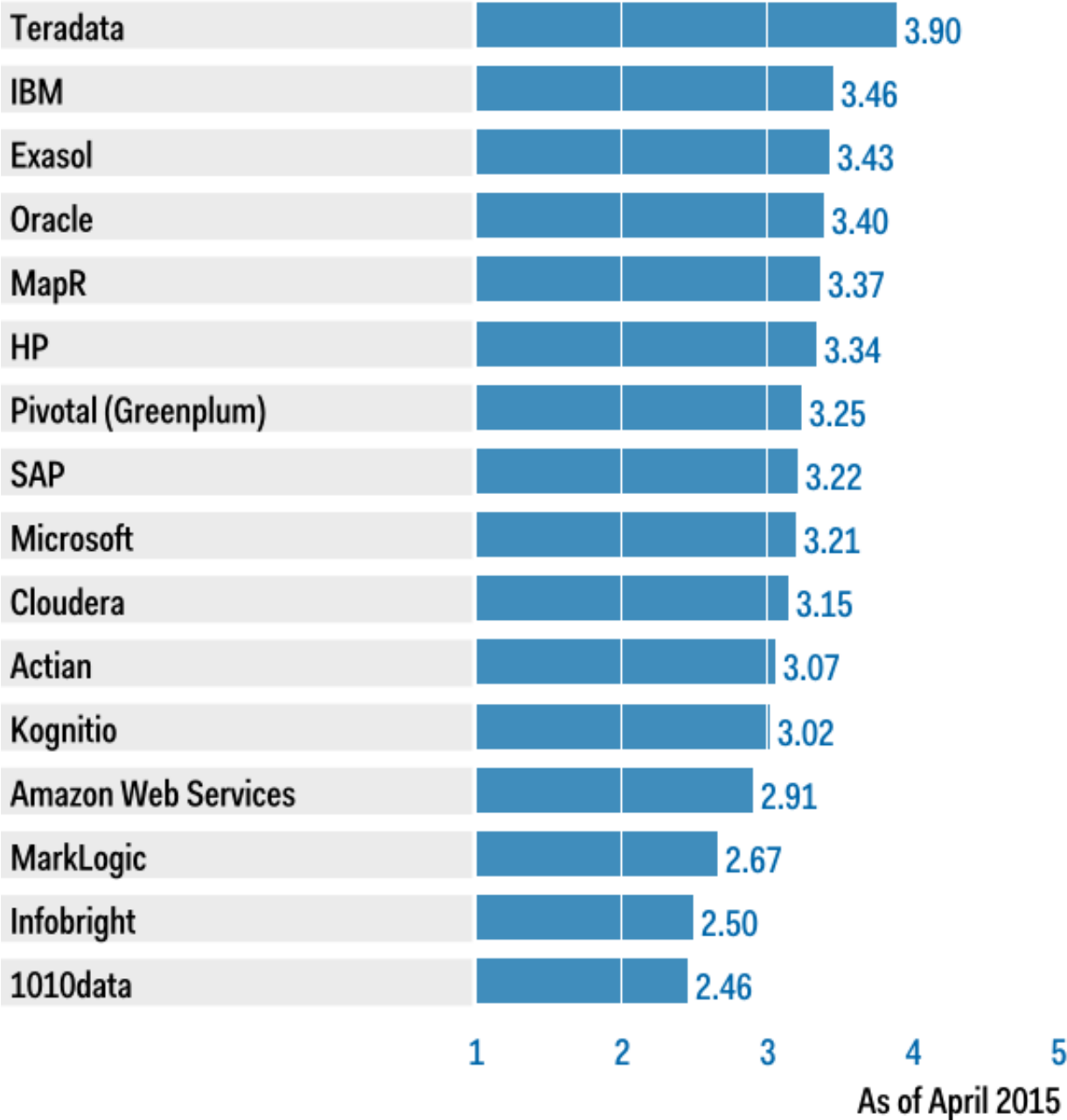## Product or Service Scores for Logical Data Warehouse

Vendor	Score
Teradata	3.99
IBM	3.66
Exasol	3.60
HP	3.60
Oracle	3.59
MapR	3.56
Pivotal (Greenplum)	3.35
Cloudera	3.31
SAP	3.25
Amazon Web Services	3.24
Microsoft	3.24
Actian	3.17
1010data	3.13
MarkLogic	3.07
Kognitio	3.03
Infobright	2.84

As of April 2015

Source: Gartner (April 2015)

Figure 4. Vendors' Product Scores for Context-Independent Data Warehouse Use Cases

Product or Service Scores for Context-Independent Data Warehouse

Vendor	Score
Teradata	3.90
IBM	3.46
Exasol	3.43
Oracle	3.40
MapR	3.37
HP	3.34
Pivotal (Greenplum)	3.25
SAP	3.22
Microsoft	3.21
Cloudera	3.15
Actian	3.07
Kognitio	3.02
Amazon Web Services	2.91
MarkLogic	2.67
Infobright	2.50
1010data	2.46

As of April 2015

Source: Gartner (April 2015)

Vendors

1010data

1010data (www.1010data.com) was established in 2000 as a managed service DW provider with an integrated DBMS and BI solution primarily for the financial sector, but also for the retail/consumer packaged goods, telecom, government and healthcare sectors.

The company's main focus is on traditional DW, managing historical data, and is best-suited to addressing the needs of casual business users. Its ability to easily load data from multiple sources and merge them directly in the platform makes it attractive to data scientists and data analysts. 1010data also offers the ability to run advanced analytics queries and has added integration with R. As a result, 1010data can be used as a sandbox environment supporting the context-independent use case, bringing raw data from multiple sources, but this type of usage remains a small portion of overall use. Finally, since much of the usage is driven by bringing data into 1010data, it has limited support for the logical use case. Continuous data loading — an important capability for the operational DW — is rarely used by 1010data clients, making it a solution only partially used for this use case. Managing large data volumes is something that 1010data customers do. The company's customers reported a wide range of environments including some that manage of over 100 terabytes (TB) of data involving tables with billions of rows.

Actian

Actian (www.actian.com) offers Matrix for advanced analytics support and Vector for operationally embedded analytics as well as the general-purpose, open-source Ingres DBMS. Actian presents integrated and complementary components for analytics data management, data integration and embedded analytics. Actian products are sold as part of the Actian Analytics Platform.

Actian offers the ability to manage all use cases, but has specific capabilities for the logical and context-independent use cases. With an overall growing number of data analysts and data scientists reported in our customer survey, the adoption of the logical and context-independent use cases are progressing. Surveyed customers said that they also used Actian for traditional DWs, although these remain mainly small in size with few more than 100TB. In traditional and operational data warehouse scenarios, the completeness of the solution (such as high availability and disaster recovery) has affected the extent to which the products are used in these use cases. However, the vendor is further investing in these capabilities for its Matrix and Vector products.

Amazon Web Services

Amazon Web Services (AWS) (aws.amazon.com) offers Amazon Redshift (a DW service in the cloud), AWS Data Pipeline (designed for orchestration with existing AWS data services) and Elastic MapReduce.

With a growing number of large deployments cited in our customer survey, Redshift adoption is maturing but continues to be mainly centered on the traditional use case (with bulk and batch loading). Adoption of the other three use cases has been affected by the completeness of the

product offering when it comes to workload management with concurrent ad hoc queries, or richness of SQL capabilities. The survey also showed limited use of continuous loading, which can be related to the greater complexity of managing cloud and on-premises environments. The user skills distribution is showing an above-average number of data scientists and data analysts. This is further demonstrated by the above-average score in the variety of data used in addition to structured data, which has been made possible by the addition of new functionality such as JSON functions, allowing easy integration of JSON data. However, limited support for advanced analytics such as user-defined functions has affected Redshift's ability to address the context-independent use case.

Redshift is delivered as a cloud-based solution and is still a new offering in the market. As a result, it is a platform that continues to evolve quickly, both in terms of new features and functions being delivered, as well as in terms of customer maturity and the variety of use cases being supported. For example Redshift added the ability to visualize the time spent in various parts of the query to help optimize complex queries more easily.

Cloudera

Cloudera (www.cloudera.com) provides a data storage and processing platform based upon an Apache Hadoop Project, as well as proprietary system and data management tools for design, deployment, operation and production management.

Surveyed customers use Cloudera to manage large volumes of data of various types using the Hadoop file system, with loading patterns including both bulk and batch loading as well as continuous loading of streams of data for use cases such as machine-generated data. In 2014, surveyed customers cited a user skills bias toward data scientist and casual users, but the proportion of data analyst and business analyst has been growing, demonstrating a wider usage of the technology. This is further visible by the spread across the various query types. With usually highly skilled clients, administration and management continues to be reported as "difficult" when compared to other solutions, although system availability is high.

Addressing the needs of enterprise customers is at the center of Cloudera's roadmap. The company has demonstrated progress by delivering data security, data governance and administration, and management capabilities as part of Cloudera Manager and Navigator. This interesting combination of capabilities makes Cloudera best suited for the logical data warehouse and context-independent use cases. Logical DW implementations based on Cloudera are founded on the ability to colocate and join data of varied types and formats without using federation and virtualization capabilities.

Exasol

Exasol (www.exasol.com) is an in-memory column-store, massively parallel processing DBMS. EXASolution is used primarily as a data mart for advanced analytical scenarios enabled by its in-memory capabilities.

Survey customers report using Exasol for advanced use cases, in particular context-independent DW. This is made possible by Exasol's use of Hadoop integration, support for user-defined

functions, and database analytics execution of R, Python, Lua and Java. At the same time, the solution is used for operational and traditional DW use cases, thanks to good system availability, administration and management capabilities. Exasol's versatility can be explained by the changed positioning from traditional and operational data warehousing to more advanced use cases like logical and context-independent. This can also mean that the technology could support moving directly from data mining/science functions into casual use cases via embedded analytics and reporting. The ability to manage a variety of use cases is a specific strength of in-memory/column capabilities, but not necessarily one that is applied by customers.

HP

HP (www.hp.com) has a portfolio that addresses DMSA and is based on HAVEn, a concept that combines many analytics acquisitions under one banner. HP's offering is anchored by Vertica (a column-store analytic DBMS), Autonomy and Hadoop. Vertica is delivered as software for standard platforms (excluding Windows); as a Community Edition (free for up to 1TB of data and three nodes); and as a cloud-based service on HP Helion as Vertica OnDemand. HP Factory Express (a predefined certified configuration) is also available for Vertica.

Vertica provides an interesting combination of capabilities for managing large volumes of data, supporting various types of queries and good system availability, administration and management. Vertica clients have started to include content data types and the ability to access data outside of the relational DBMS. Surveyed customers and Gartner analysis rate Vertica as particularly suitable for the logical and context-independent use cases. Surveyed customers indicated that data loaded continuously still represents a small proportion of the overall data, although the product can support continuous loading. HP Vertica's capabilities, as well as customer experience, places Vertica in the top five position across all use cases.

IBM

IBM (www.ibm.com) offers stand-alone DBMS solutions with DB2 and Informix, as well as DW appliances and a z/OS solution.

Its various appliances include:

- IBM zEnterprise Analytics System

- IBM PureData System for Analytics

- IBM PureData System for Operational Analytics

- IBM DB2 Analytics Accelerator

IBM offers a very complete but disjointed set of capabilities through various tools and platforms that support all four use cases demanded by modern analysts. As a longtime player in the DW space, IBM demonstrates functional delivery with new styles of data management and integration that take advantage of the use of new technology advancements, such as fully exploiting in-memory. For example, in 2014, IBM moved to the cloud for fast deployment and instant-on POC or Pilot deployment with DashDB and DataWorks. Bluemix has introduced community developers and

marketplace contributions to leverage execution experience for a better deployment experience among end users.

Although his research focuses on technical capabilities evaluated and reported by user organizations, Gartner inquiry clients and technical capabilities assessments have revealed a set of functionality that supports all user types — a change from February 2014 when IBM demonstrated a gap in use by business analysts.

IBM customers report that most loads take place daily or intraday (bursts throughout the day); however, this is based on "practice" rather than "capability" as IBM can perform continuous load with varying technology solutions.

Infobright

Infobright (www.infobright.com) is a global company offering a column-vectored, highly compressed DBMS. With open-source — Infobright Community Edition (ICE) and commercial Infobright Enterprise Edition (IEE) — versions, the company also offers an Infopliance database appliance.

Surveyed customers mainly use Infobright for small traditional DWs, and our survey data shows that Infobright is occasionally used for large deployments (over 100TB). Surveyed customers mainly load data in batches, with a small subset being loaded continuously. Infobright also offers continuous loading capabilities that would allow for a wider set of data being loaded continuously.

Some gaps in product functionality — such as completeness of SQL support and user-defined functions — make it challenging to support the logical data or context-independent use cases as well as more sophisticated users such as data scientists or data miners. As a result, Infobright is mainly suited for traditional and operational use cases. However, even with the traditional DW use case, some missing capabilities such as scale out make it more complex for customers to grow their deployments.

Kognitio

Kognitio (www.kognitio.com) offers the Kognitio Analytical Platform, an in-memory row-store DBMS that supports the analytical use case.

It is offered as:

- An appliance

- A DW DBMS engine

- DW as a managed service (hosted on hardware located at Kognitio's sites or those of its partners)

- A DW platform as a service using AWS

Kognitio's small customer base makes it possible to determine that, over several years, very few of its customers are exceeding daily load frequencies (verified by surveys in November of 2011, 2012,

2013 and 2014). This reinforces that the scenarios in which the platform is primarily used are data-mart support to tactical analysis, and as a basis for operational analytics. This role is further reinforced by issues reported when utilizing parallel or distributed load, some missing functionality, and a particular note of missing connectivity/third-party access.

Investment in integration with Hadoop distribution, combined with in-memory capabilities, have also allowed Kognitio to be used for the context-independent use case — although this is done by very few customers.

Kognitio has been in the DW DBMS business for over 20 years with customers of various levels of maturity and size of deployments. While surveyed customers generally have deployments that are mainly small (below 5TB), some are larger (around 20TB).

MapR

MapR Technologies (www.mapr.com) offers a Hadoop distribution with performance and storage optimizations, high availability improvements, and administrative and management tools. It offers training and education services and quick-start solutions for common Hadoop scenarios.

The technology is based on the open-source Apache Hadoop project to which MapR also contributes. MapR offers multitenancy that supports the isolation of distinct datasets, user groups, and jobs in the same cluster. The Apache Drill project is also supported, which provides federated querying across multiple data sources, including MongoDB.

The MapR file system (commonly known as MapR-FS) is an NFS- and Posix-compliant file system that enables several functions. Hadoop begins with an append-only file system; MapR-FS has full, random read/write capability and eliminates many of the layers required to deploy Hadoop run in a Java Virtual Machine, for example. MapR-FS effectively eliminates the requirement to write a temporary file before a subsequent write to HDFS — providing a real-time, NFS interface. Datasets are protected by cross-data-center mirroring and point-in-time consistent snapshots. The distribution includes an integrated in-Hadoop NoSQL database (MapR-DB) to support an alternative to HBase (which can be used to support very low-latency writes) that is compatible with the HBase API. MapR has no "NameNode," This function is effectively distributed as a service across the entire cluster in MapR. This provides additional assurance in node failure situations as well as removing any upper limits on the number of files under management in their cluster — enhancing its scalability and thereby reducing overall hardware resources and associated costs. The company's JobTracker, in combination with YARN, provides high-availability support. MapR offers their optimization techniques without special configuration as part of their distribution. Importantly, the NameNode limit is reached in Hadoop in the tens of millions of files, and it is possible to optimize the performance characteristics in Hadoop using other approaches.

MapR is beginning to appear in situations where organizations are seeking to integrate Hadoop deployments with existing production data warehouses or other analytics data use cases. These capabilities contribute to the high rating from surveyed customers for administration, management and system availability. MapR ranks high in the logical data warehouse but, the support for this use case is driven by the variety of data stored in MapR.

MarkLogic

MarkLogic (www.marklogic.com) offers an ACID NoSQL database that utilizes JSON and XML storage, and offers a strong metadata-driven, semantic-access management layer.

MarkLogic customer implementations generally range up to approximately the 15TB range, but a small number of references report very large deployments (one at over 500TB). The use cases are diverse by vertical and applied solution, but are primarily based on converting search into an active function of analytics. It supports search and what it calls "data unification" across both structured and unstructured information assets — this is basically advanced business-analyst-level work beyond casual users (just short of data mining tasks). Customers report a very high proportion of casual users to other categories of users. This shows that, while the solution is suited to more advanced users, the requirement for the company to support a large number of casual users is a concern as it will affect their productivity.

MarkLogic is best-suited to both context-independent and operational use cases given its combination of capabilities, but reference customers favor the latter. This can be explained by the high proportion of casual users further demonstrated by a high proportion of repetitive queries.

Microsoft

Microsoft (www.microsoft.com) markets SQL Server (a reference architecture) and Microsoft Analytics Platform System (which combines SQL Server Parallel Data Warehouse and HDInsight), as well as Azure HDInsight for Hadoop. Surveyed customers mainly used SQL Server 2012 as the 2014 release was still new during the time frame of this research.

Microsoft customers have added new use cases beyond traditional and are reporting clear use cases across all data user types (casual, analyst, miner and scientist). For all the improvements in the last three years, Microsoft references and Gartner inquiry clients continue to report gaps in management and administration, and difficulties with deploying parallel sessions with Parallel Data Warehouse. The important thing is that Microsoft is innovating and challenging the market with new deployment models and capabilities.

Since February 2014, Microsoft customers have been reporting a significant increase in the number of deployments involving machine data (operational technology), which supports digital business and the Internet of Things (IoT). Similarly, there is an increase in intraday load frequency (more than once per day), but customers do not yet report continuous loading. As a result, surveyed customers using machine data are mainly focusing on post-event analysis. This is both the result of the use cases selected by customers, as well as the functional limitations of the Microsoft technology for continuous loading. However, Microsoft has positioned itself to move into continuous capability sometime in the next three years, or sooner.

Oracle

Oracle (www.oracle.com) provides the Oracle DBMS in multiple offerings as well as a Hadoop appliance called Oracle Big Data Appliance. Customers can choose to build a custom warehouse, a certified configuration of Oracle products on Oracle-recommended hardware, or on appliance. Oracle is expanding its delivery model to cloud, and its capabilities and practice are focused on a

product-delivery approach (see "Oracle's Approach to Information Infrastructure and the Gartner Information Capabilities Framework").

Oracle user experience indicates a thriving practice of performing traditional analytics on structured data, and a willingness to add new data types such social profile data, document and email parsing, and even operational technology data (IoT and digital business). Additionally, a mixed set of use cases for continuous data loading crosses higher education, banking and telecom. Because of its large customer base, Oracle has a wide range of data users, including casual, analysts, miners and data scientists. Usage does tend to follow a pattern of isolating on *either* advanced users (miners/scientists) *or* those in production jobs (analysts/casual).

Oracle capabilities now include an in-memory columnar option, which enhances analytical query speeds. This expands the Oracle relational database management system (RDBMS) and, as a result, benefits from all of the standard Oracle reliability (fault tolerance and DataGuard, for example). Oracle has combined its infrastructure as a service and platform as a service (see "Hype Cycle for Information Infrastructure, 2014") through capabilities such as virtual machines and operating system isolation in virtual images, as well as full provisioning, and provides database schemas as a service.

Oracle often provides for higher priority for supporting functionality specific to its customer base first (first with multitenant) and then adds functionality, sometimes later than other vendors (in-memory columnar, for example). Oracle provides continuing enhancements with numerous SQL extensions (adding support for JSON in 2014), integrations with R and others. Oracle deployment of emerging technologies and practices is robust, but is often a variant of emerging market practices. It is sometimes first to market and other times introduces new capabilities after other competitors in the market. Oracle focuses on production requirements, which often require product and system stability as the highest priority (and their customer base agrees) as opposed to immediate rollout of "the next new thing."

Pivotal (Greenplum)

Pivotal (www.pivotal.io) is a wholly owned subsidiary of EMC financed by EMC, VMware and General Electric (GE owns 10%). It carries the following products:

- Pivotal Labs

- Greenplum DB

- Pivotal HD

- GemFire

- GemFire XD

- Cloud Foundry

Pivotal also combines and delivers these various products through its Big Data Suite, which is also available as open source.

Pivotal reference implementations tend to be large at over 40TB, with a few over 100TB. In November, 2014, references report a rebalancing of all user types (scientists, miners, analysts and casual) that is more aligned with the market than reported in 2013. This means that Pivotal is now facilitating movement between analytic use-case models. In February 2014, Gartner anticipated that the growing adoption of Pivotal HD would mean that the use of other data types beyond relational would also expand. Approximately 50% of Pivotal references report they have (or will) complete loading of documents, email or other text-based unstructured data within the next 12 months. Pivotal is well-suited to the context-independent and logical scenarios as well as traditional DW. While possible, continuous loading is not commonly used by Pivotal references with most reporting intraday (frequently throughout the day but not yet continuous).

SAP

SAP (www.sap.com) offers both SAP IQ and SAP Hana. SAP IQ was the first column-store DBMS. It is available as a stand-alone DBMS and on an OEM basis via system integrators. SAP Hana is an in-memory column store that supports the operational and analytical use cases, and is also offered as an appliance and reference architecture (tailored date center integration [TDI]). Customers either use one or a combination of both as their DW solution.

SAP reference customer deployments are very diverse in size. References and Gartner inquiry clients report that specific industries are pursuing intraday or continuous loading of data. However, these types of usage are often driven by specific and isolated scenarios. Overall, most organizations are not evolving to (or combining) the different analytics use cases. Infrequently, SAP solutions are used for all types of big data and all four use cases. However, this is usually limited to organizations that have solved the issues of mixed use and mixed workload themselves. While a challenging environment, this is evidence that SAP covers a breadth of technology in 2015 that matches its competition. Additionally, SAP customers exhibit a preference for utilizing the platform in production reporting and embedded analytics, which are predominantly casual or analyst user scenarios.

In late 2014, Hana 9 introduced critical improvements to security, hardware virtualization, text analytics/mining, machine learning/graph, streaming data, smart data integration (adapters to bulk load/replicate data from remote sources), smart data quality (algorithms to cleanse addresses and more) improvements and administrative/management enhancements. The cooperative effect of adding Hadoop to Hana (with high-value data going to the faster Hana platform and early adoption of Spark through a partnership with Databricks) presents the opportunity to expand the customer experience. At the same time, multitenancy enables Hana for the cloud. These new SPS 09 capabilities are expected to enhance SAP's ability to address data science needs and improve transferability of data science/mining work to analysts. Finally, the addition of data tiering — allowing the combination of in-memory columnar storage and disk-based storage — should mitigate some of the issues for Hana adoption with large volumes of data and enable greater adoption in traditional DW deployments, for example.

Teradata

Teradata (www.teradata.com) has more than 30 years of history in the data warehousing market. Teradata offerings include: DBMS licenses, appliances and cloud solutions. It offers products aimed at traditional and logical DW use cases, which Teradata calls the Unified Data Architecture. It offers

a combination of tuned hardware and analytics-specific database software, which includes the Teradata database (on various appliances), the Aster Database and Hadoop.

Surveyed clients mainly use the Teradata database in traditional DW use cases that have now grown to large environments (over 50TB), which is in line with the overall market situation, and is representative of successful and mature enterprise DW deployments. The diverse spread of query and user types — with a much higher-than-average proportion of business analysts and data analysts among survey respondents — demonstrates the company's ability to address mixed workloads.

Although reference survey clients report a below-average proportion of data loaded continuously, this is also related to overall large DW environments where it is neither appropriate, nor meaningful, to load a large percentage of the data continuously. Throughout 2014, Gartner has observed a growing number of logical and context-independent use cases being implemented within the Teradata installed base. Customer experience combined with rich functional capabilities puts Teradata in first or second position across all four use cases.

Context

Over the past few years, data warehousing technology and practices have evolved to meet various enterprise needs. Well over 80% of the market adheres to traditional approaches. We are, however, seeing a growing number of leading organizations evolving their DW toward the logical use case to address volume, variety and velocity needs driven by big data. At the same time, mature organizations are making their DWs active contributors to their operational needs (for example by embedding analytics in their applications), hence adding to their mission-critical nature.

Overall, administration capabilities remain a concern across all use cases and, while vendor capabilities play an important role, availability of skills remains an important factor, specifically for smaller vendors or for those that are new to the market.

As organizations start using their DW and DMSA beyond descriptive requirements, and evolve it toward predictive and prescriptive analytics, additional user profiles are emerging (data scientist) or growing (data analysts). These new types of users are being added to business analysts and casual users, and are creating new types of demands and interactions with the DMSA. These include demanding native support for the new data types and in-database programming capabilities such as R, Lua or Python, affecting its workloads and type of queries.

Operational, logical and context-independent DW use cases still constitute just a small proportion of deployments. The operational use case has only seen a modest adoption as the number of organizations requiring real-time analysis and reporting remains limited (less than 1% of the total market). Logical DW (less than 11% of the market) and context-independent DW (less than 3%) are only emerging to support the volume, variety and velocity requirements of big data as well as to perform new styles of analysis on these new data types.

Product/Service Class Definition

The various capabilities identified below address the major needs identified above.

Critical Capabilities Definitions

Managing Large Volumes of Data

This capability reflects if the volume of data managed by customers is large. This applies to data of multiple data structures and formats.

It plays a role in all use cases, but to various degrees, as it may not be equally important for all use cases. In this context, we have defined "small" as being below 10TB and "large" as being over 40TB.

In addition to customer experience, this capability takes into consideration the ability for the vendor to address management of query workloads, the availability of price performance optimization options as well as strategies for query optimization in isolation.

Loading Data Continuously

This capability represents the prevalence of data being loaded continuously by customers. Some use cases more than others require data to be loaded from the operational sources in real time.

This capability does not assume all data is loaded continuously. Bulk and batch loading remains the most common loading process.

In addition to customer experience, this capability takes into consideration the vendor's ability to support ingestion of streaming data as well as the ability to perform continuous updates for read optimization.

Other Data Types Beyond Structured

This capability represents the prevalence of customers in using other data types in addition to structured data (such as machine data, text, documents, images or videos). This capability plays primarily in the logical and the context-independent use cases.

In addition to customer experience, this capability takes into consideration the vendor ability to store data of varied data types (polyglot persistence), to access multiple repositories of varied data types, and to process and link diverse data types.

Repetitive Queries

This capability reflects the prevalence of the use of repetitive queries by customers across all types of queries. Repetitive queries are used in support of reporting requirements addressing a wide set of casual users.

The prevalence of repetitive queries in workloads is indicative of DWs supporting a wide set of casual users that only consume data that has been prepared for them. Repetitive queries can also extend to supporting dashboards that offer limited and prebuilt interactions.

In addition to customer experience, this capability takes into consideration the ability to perform automation of optimization and workload management for repetitive queries.

Queries Support Advanced Analytics

This capability reflects the prevalence of advance analytics queries by customers across all types of queries.

Advanced analytics queries support forecasting, predictive modeling, in-database execution of R (an open-source statistical programming application), user-defined functions or other mining styles, as well as free-form analytical queries based upon data marts, views, cubes or a semantic-enabled modeling interface.

In addition to customer experience, this capability takes into consideration the ability to support complex ad hoc queries, perform mixed workload and data temperature.

Queries on Many Data Types/Sources

This capability reflects the prevalence of queries across multiple data types and sources by customers across all types of queries.

Queries across multiple data types and sources access data in other sources beyond the database management system such as other relational DBMSs, Hadoop distributions or others.

In addition, this capability takes into consideration the ability to process data of various types managed on the same repositories or over multiple repositories. This includes features such as push down of predicates or JSON support.

Operational BI Queries

This capability reflects the prevalence of queries for operational reporting or embedded analysis in applications as part of the overall set of queries among customers.

While operational BI queries are primarily used in support of the operational use case, other types of queries will also play a role. All of our use cases attempt to support operational BI in some way, but do not have the same level of success.

In addition to customer experience, this capability takes into consideration the ability to perform operational queries with high SLA levels as well as to support running in database prescriptive analysis.

System Availability

This capability outlines the ability for the vendor to support general production availability, high availability and disaster recovery as expressed by customers.

In addition to customer experience, this capability takes into consideration the vendor's ability to provide high availability and disaster recovery as well as the ease of deployment and management of these features.

User Skill Level Support

This capability represents the distribution of user types among customers across data scientists, data miners, business analysts or casual users.

The score takes into consideration the relative proportion of these user types across surveyed clients. Low scores reflect a below-average proportion of user types among one or more of the categories (data scientists, data miners and business analysts) as compared to the overall user type.

In this context we have defined the following skills for each user type.

- **Data Scientist:** Expert in statistics, abstract mathematics, programming, business processes, communications and leadership.

- **Data Miner:** Expert in data, statistical software, statistical models, fully aware of computer processing "traps" or errors.

- **Business Analyst:** Utilizes online analytical processing and dimensional tools to create new objects, some faculty with computer languages and computer processing techniques.

- **Casual User:** Regularly uses portals and prebuilt interfaces, minimally capable of designing dimensional analytics (if at all).

In addition to the distribution of user types, this capability takes into consideration the ability to address the specific needs of the various user types such as in-database programming of Python, Lua and R, ability to perform any type of join, run ad hoc queries, ability to build own queries and reports leveraging data marts or data specifically prepared, performing root-cause analysis, leveraging guided analysis and drill downs.

Administration and Management

This capability demonstrates the vendor or product's ease of implementation, upgrade and ease of use as expressed by customers.

This capability covers not only the overall ease of administration and management during implementation, but also during ongoing use as well as during upgrade phases. Scoring is also affected by the complexity of deployment and by vendor history. Some vendors have recent offerings for which upgrades may not have been released.

In addition to customer experience, this capability takes into consideration the completeness of vendor administration capabilities such as role-based activities, advisors, utilization and capacity planning, resource allocation features and user interface as well as complexity of deployment and management.

Use Cases

Traditional Data Warehouse

This use case involves managing historical data coming from various structured sources. Data is mainly loaded through bulk and batch loading.

The traditional DW use case can manage large volumes of data and is primarily used for standard reporting and dashboarding. To a lesser extent, it is used for free-form querying and mining, or operational queries. It requires high capabilities for system availability and administration and management, given the mixed workload capabilities for queries and user skills breakdown.

Operational Data Warehouse

This use case manages structured data that is loaded continuously in support of embedded analytics in applications, real-time DW, and operational data stores.

This use case primarily supports reporting and automated queries to support operational needs and will require high availability and disaster recovery capabilities to meet operational needs. Managing different types of users or workloads, such as ad hoc querying and mining, will be of less importance as the major driver is to meet operational excellence.

Logical Data Warehouse

This use case requires managing data variety and volume of data for both structured and other content data types.

Besides structured data coming from transactional applications, this use case includes other content data types such as machine data, text documents, images and videos. Because additional content types can drive large data volumes, managing large volumes is an important criterion. Logical DW is also required to meet diverse query capabilities and support diverse user skills. This use case supports queries reaching into other sources than the DW DBMS alone.

Context-Independent Data Warehouse

This declares new data values, variants of data form and new relationships. It supports search, graph and other advanced capabilities for discovering new information models.

This use case is primarily used for free-form queries to support forecasting, predictive modeling or other mining styles as well as queries supporting multiple data types and sources. It has no

operational requirements and favors advanced users such as data scientists or business analysts resulting in free-form queries across potentially multiple data types.

Vendors Added and Dropped

Added

MapR

Dropped

InfiniDB

Inclusion Criteria

- Vendors must have DW or DMSA software that has been generally available for licensing or supported download for approximately one year (since 10 December 2013).

- We use the most recent full release of the software to evaluate each vendor's current technical capabilities. We do not consider beta releases. For existing data warehouses, and direct vendor customer references and reference survey responses, all versions currently used in production are considered. For older versions, we consider whether later releases may have addressed reported issues, but also the rate at which customers refuse to move to newer versions.

- Product evaluations include technical capabilities, features and functionality present in the product or supported for download through 8:00 p.m. U.S. Eastern Daylight Time on 1 December 2014. Capabilities, product features or functionality released after this date can be included at Gartner's discretion and in a manner Gartner deems appropriate to ensure the quality of our research product on behalf of our nonvendor clients. We also consider how such later releases can reasonably impact the end-user experience.

- Vendors must have generated revenue from at least 10 verifiable and distinct organizations with a DW or DMSA in production. Revenue can be from licenses, support and/or maintenance. Gartner may include additional vendors based on undisclosed references in cases of known use for classified but unspecified use cases. For this year's Magic Quadrant, the approved questionnaire was produced in English. Gartner exercises its option to provide for other languages as deemed appropriate only in the case of an extreme exception.

- Gartner should have evidence of a minimum of 10 distinct DW or DMSA in production either through reference survey, inquiries or reference calls conducted between 10 Dec 2013 and 10 Dec 2014.

- Customers in production must have deployed DWs that integrate data from at least two operational source systems for more than one end-user community (such as separate business lines or differing levels of analytics).

- To be included, any acquired vendor product must have been acquired and offered by the acquiring vendor as of 30 June 2014. Acquisitions after 30 June 2014 are considered a legacy offering and will be represented as a separate vendor until publication of the following year's Critical capabilities.

- Support for the included products must be available from the vendor. We also consider products from vendors that control or participate in the engineering of open-source solutions and their support.

- We also include the capability of vendors to coordinate data management and processing from additional sources beyond the DBMS, but continue to require that a DBMS meets Gartner's definition in particular regarding support of at least one of the four major use cases (traditional data warehouse, operational data warehouse, logical data warehouse or context-independent data warehouse).

- Vendors participating in the DW and DMSA market must demonstrate their ability to deliver the necessary services to support a data warehouse via the establishment and delivery of support processes, professional services and/or committed resources and budget.

- Products that exclusively support an integrated front-end tool that reads only from the paired data management system do not qualify for this research.

- Gartner analysts remain the sole arbiter of which suppliers/vendors and/or products will be included in this research.

Vendors and products were also determined for inclusion or exclusion based upon the following:

- Relational data management.

- Nonrelational data management.

- No specific rating advantage is given regarding the type of data store used (for example, RDBMS, Hadoop Distributed File System [HDFS], key-value, document; row, column and so on).

- Multiple solutions in combination to form a DMSA are considered valid (although one approach is adequate for inclusion), but each solution must demonstrate maturity and customer adoption.

- Cloud solutions (such as platform as a service) are considered viable alternatives to on-premises warehouses. An ability to manage hybrids between on-premises and cloud are considered advantageous.

- Data management solutions for analytics are expected to coordinate data virtualization strategies for accessing data outside of the DBMS, as well as distributed file and/or processing approaches.

Table 1. Weighting for Critical Capabilities by Use Case

Critical Capabilities	Traditional Data Warehouse	Operational Data Warehouse	Logical Data Warehouse	Context-Independent Data Warehouse
Managing Large Volumes of Data	15.0%	5.0%	15.0%	10.0%
Loading Data Continuously	5.0%	25.0%	5.0%	5.0%
Structured Data and Other Types	0.0%	0.0%	15.0%	25.0%
Repetitive Queries	15.0%	15.0%	10.0%	0.0%
Queries Support Advanced Analytics	10.0%	5.0%	10.0%	25.0%
Queries on Many Data Types/Sources	5.0%	5.0%	10.0%	15.0%
Operational BI Queries	10.0%	15.0%	10.0%	0.0%
System Availability	20.0%	15.0%	10.0%	5.0%
User Skill Level Support	10.0%	5.0%	5.0%	10.0%
Administration and Management	10.0%	10.0%	10.0%	5.0%
Total	100.0%	100.0%	100.0%	100.0%

This methodology requires analysts to identify the critical capabilities for a class of products/services. Each capability is then weighed in terms of its relative importance for specific product/service use cases.

Source: Gartner (April 2015)

Critical Capabilities Rating

Each of the products/services has been evaluated on the critical capabilities on a scale of 1 to 5; a score of 1 = Poor (most or all defined requirements are not achieved), while 5 = Outstanding (significantly exceeds requirements).

Table 2. Product/Service Rating on Critical Capabilities

Product or Service Ratings	1010data	Actian	Amazon Web Services	Cloudera	Exasol	HP	IBM	Infobright	Kognitio	MapR	MarkLogic	Microsoft	Oracle	Pivotal (Greenplum)	SAP	Teradata
Managing Large Volumes of Data	3.9	2.7	2.7	3.7	2.8	3.9	3.9	1.7	2.2	3.9	2.2	2.4	3.9	3.1	2.8	4.1
Loading Data Continuously	1.6	2.8	1.8	1.8	2.8	2.2	2.2	3.2	1.8	2.2	4.5	1.8	2.2	2.2	3.0	2.4
Other Data Types Beyond Structured	1.1	2.7	3.3	4.1	2.9	3.0	3.2	2.3	2.5	4.3	3.0	3.9	3.1	3.2	3.2	4.0
Repetitive Queries	4.7	4.5	4.7	4.0	4.2	4.0	4.1	4.8	3.4	4.1	4.7	4.8	4.2	4.1	4.4	4.8
Queries Support Advanced Analytics	2.5	3.5	1.9	1.7	3.4	2.9	2.8	2.3	3.4	2.2	1.8	2.8	3.0	3.3	3.0	3.5
Queries on Many Data Types/Sources	1.9	1.9	3.5	4.0	3.8	3.5	3.9	2.9	3.4	3.9	3.7	2.8	3.5	3.7	4.0	3.9
Operational BI Queries	3.6	2.4	2.8	2.9	3.9	4.0	3.0	3.1	3.2	3.6	2.8	1.9	2.6	3.3	3.0	3.0
System Availability	4.8	4.4	4.5	3.7	4.5	4.3	4.8	3.6	3.9	3.9	4.0	4.2	4.5	4.2	3.0	4.9
User Skill Level Support	3.8	4.4	3.2	3.2	4.4	4.4	4.3	2.6	3.7	3.2	1.2	4.0	4.0	3.0	3.5	4.5
Administration and Management	3.6	3.3	3.5	2.6	4.0	3.6	4.1	2.8	3.2	2.9	3.0	3.5	4.5	2.8	2.8	4.2
																As of April 2015

Source: Gartner (April 2015)

Table 3 shows the product/service scores for each use case. The scores, which are generated by multiplying the use-case weightings by the product/service ratings, summarize how well the critical capabilities are met for each use case.

Table 3. Product Scores by Use Case

Use Cases	1010data	Actian	Amazon Web Services	Cloudera	Exasol	HP	IBM	Infobright	Kognitio	MapR	MarkLogic	Microsoft	Oracle	Pivotal (Greenplum)	SAP	Teradata
Traditional Data Warehouse	3.78	3.56	3.42	3.23	3.85	3.82	3.89	3.08	3.23	3.48	3.13	3.37	3.81	3.46	3.26	4.15
Operational Data Warehouse	3.33	3.35	3.17	2.93	3.71	3.49	3.49	3.28	2.98	3.24	3.60	3.04	3.42	3.23	3.26	3.73
Logical Data Warehouse	3.13	3.17	3.24	3.31	3.60	3.60	3.66	2.84	3.03	3.56	3.07	3.24	3.59	3.35	3.25	3.99
Context-Independent Data Warehouse	2.46	3.07	2.91	3.15	3.43	3.34	3.46	2.50	3.02	3.37	2.67	3.21	3.40	3.25	3.22	3.90
																As of April 2015

Source: Gartner (April 2015)

To determine an overall score for each product/service in the use cases, multiply the ratings in Table 2 by the weightings shown in Table 1.

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

"Magic Quadrant for Data Warehouse Database Management Systems"

"The State of Data Warehousing in 2014"

"How Products and Services Are Evaluated in Gartner Critical Capabilities"

Evidence

This research is based on:

- Gartner inquiry data on the topics of data warehousing, data integration and metadata management.

- In-depth interviews with reference customers provided by vendors.

- Gartner strategic advisory service, full-day sessions with end-user organizations in which Gartner is asked to review client implementation plans and designs.

- Gartner's respondent survey (265 respondents) for "Magic Quadrant for Data Warehouse Database Management Systems" for year 2015 (survey completed in 4Q14).

Note 1 Definition of a Data Warehouse (DW)

A "data warehouse" is a solution architecture that may consist of many different technologies in combination. At the core, however, any vendor offering (or combination of offerings) must exhibit the capability of providing access to the files or tables under management by open-access tools.

A data warehouse is simply a warehouse of data, not a specific class or type of technology. In particular, we include in the data warehouse DBMS market relational and nonrelational data management systems such as Hadoop Distributed File System (HDFS), relational, key-value, document; row and column.

Critical Capabilities Methodology

This methodology requires analysts to identify the critical capabilities for a class of products or services. Each capability is then weighted in terms of its relative importance for specific product or service use cases. Next, products/services are rated in terms of how well they achieve each of the critical capabilities. A score that summarizes how

well they meet the critical capabilities for each use case is then calculated for each product/service.

"Critical capabilities" are attributes that differentiate products/services in a class in terms of their quality and performance. Gartner recommends that users consider the set of critical capabilities as some of the most important criteria for acquisition decisions.

In defining the product/service category for evaluation, the analyst first identifies the leading uses for the products/services in this market. What needs are end-users looking to fulfill, when considering products/services in this market? Use cases should match common client deployment scenarios. These distinct client scenarios define the Use Cases.

The analyst then identifies the critical capabilities. These capabilities are generalized groups of features commonly required by this class of products/services. Each capability is assigned a level of importance in fulfilling that particular need; some sets of features are more important than others, depending on the use case being evaluated.

Each vendor's product or service is evaluated in terms of how well it delivers each capability, on a five-point scale. These ratings are displayed side-by-side for all vendors, allowing easy comparisons between the different sets of features.

Ratings and summary scores range from 1.0 to 5.0:

1 = Poor: most or all defined requirements not achieved

2 = Fair: some requirements not achieved

3 = Good: meets requirements

4 = Excellent: meets or exceeds some requirements

5 = Outstanding: significantly exceeds requirements

To determine an overall score for each product in the use cases, the product ratings are multiplied by the weightings to come up with the product score in use cases.

The critical capabilities Gartner has selected do not represent all capabilities for any product; therefore, may not represent those most important for a specific use situation or business objective. Clients should use a critical capabilities analysis as one of several sources of input about a product before making a product/service decision.

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp