GIGAOM

MARKET LANDSCAPE REPORT

Revenge of the Data Warehouse

How a Classic Tech Category Has Evolved, and Triumphed

ANDREW J. BRUST

TOPIC: DATA WAREHOUSE

GIGAOM

Revenge of the Data Warehouse

How a Classic Tech Category Has Evolved, and Triumphed

TABLE OF CONTENTS

1. Summary

The pressure to leverage data as a business asset is stronger than ever. Enterprises everywhere are eager to devise sound data strategies that are realistic and achievable, based on available budget and sensitive to in-house technology skill sets.

For a while, it looked like open-source, specialized big data compute frameworks, including Hadoop and Spark, were the way to go. Enterprise organizations found them compelling for reasons of novelty, economics, and the apparent prudence of a future-looking technology. But those frameworks are at a bit of a crossroads: the hype around them has subsided and — while things are improving — the success rate of enterprise projects involving them has been modest.

Meanwhile, the data warehouse (DW) which, for decades has been a key technology platform for enterprise analytics, never went away. Yes, DWs struggled and incumbent DW platforms still do. Still, recent advances in storage costs and compute scalability issues, especially in the cloud, have addressed the most important challenges faced by DW platforms.

As a result, we are in a DW renaissance period. Problems with DWs have largely been solved, petabyte-scale data volumes no longer defeat them and the familiarity and ease of use that kept them viable all this time are now helping them face their open-source big data competition and, in many cases, emerge victorious.

Still, if DW platforms have changed, what should enterprises do to build an analytics strategy that integrates them? Even most DW-loyal shops will need to look at how DW platforms have evolved and adjusted strategies accordingly. Organizations that have committed to open-source analytics technologies will need to take a second look at DW platforms and consider an approach that combines them, essentially bringing the data warehouse together with the data lake.

The trick to adapting to the new world of data warehousing is understanding that it is not only the technology that has changed but the applications and use cases for DW technology as well. DW platforms do not need to be used exclusively for Enterprise DW implementations. The platforms are now more versatile and can be used for use-case specific workloads and even exploratory analytics. In a sense, the DW isn't just a DW anymore.

The juxtaposition of DW and Data Lake has even shifted – DW platforms today are increasingly able to ingest raw, semi-structured data, or query it in place. This means warehouse and lake technology can be used in combination and, sometimes, lake technology will not be necessary.

It is not just the Data Lake and its workloads that are becoming more integrated into the Warehouse. Streaming data, machine learning, and AI are onboarding as well. In addition, data governance and data protection are starting to enter the DW orbit.

While the familiarity of the relational model, dimensional design, and SQL are back; the application of

DW technology now covers territory that may be less familiar. Moreover, new vendors who have championed or been born in the cloud are emerging in leadership positions.

The new capabilities, new use cases, and new vendors make the space exciting but also difficult to navigate (for newbies and veterans alike). Enterprise customers will need to understand how DW platforms have morphed and shape-shifted; how best to use, deploy and implement them; combine them with other applications; understand the key differences between the vendors and their offerings.

Without this knowledge, enterprise buyers will freeze in indecision. With it, they will be armed to leverage today's DW platforms to their fullest and cherry-pick other technologies that can augment and optimize them. The end-result is organizations in the know will be ready to analyze all their data, in technologically familiar environments, for full competitive and operational advantage.

In this report, we map out today's DW landscape in the context of where the technology has been, where it is, and where it is going.

Key Findings:

- The Data Warehouse is alive and well, perhaps enjoying its biggest popularity wave to date
- open-source technology challengers addressed issues of storage costs and horizontal scalability but did not achieve parity in terms of enterprise skill-set abundance, interactive query performance, or acting as authoritative data repositories
- The advent of the cloud has helped DW platforms transcend their vulnerabilities, through the cloud's power of elasticity for compute and the use of economical, infinitely scalable cloud object storage
- Transcending their vulnerabilities and satisfying customers' need for familiar SQL-relational platform paradigms has turned out to be a one-two punch for DWs
- Customers need to bring themselves up-to-date on the latest DW innovations and product categories, understanding each in the context of the DW's historical evolution and market factors
- Customers must correlate DW product categories with corporate cloud (and multi-cloud) strategy

2. Market & Maturity Factors

Data warehouse methodology, from Ralph Kimball, optimized relational databases for analytics. Kimball prescribed a database schema that put all metrics data (e.g., sales data, customer count, or store visits) together in a single table (the "fact table"), and the categories by which those metrics would be analyzed (e.g. time, geography or salesperson) in their own, separate, tables (the "dimension tables").

This dimensional model and its associated database design, known as a star schema, are used to this day in data warehousing. These same concepts form the basis for OLAP (online analytical processing) business intelligence platforms (BI) as well, though OLAP systems are often on-relational.

Relational databases were designed to accommodate query workloads where only a small number of database rows (sometimes only one) are fetched. But in data warehousing, vast numbers of rows are typically queried instead, often causing a full table scan of the fact table. Because so much work is involved, data warehouse performance can be a challenge. As such, changes to both the underlying database software and hardware must occur.

MPP Architecture, Teradata and the Appliance Model

One approach to meeting this challenge is to involve multiple database servers, each one scanning just a part of the fact table, with each server doing their work simultaneously. In addition to these servers (the "worker nodes"), one additional server (the "master node") is present to accept the query, delegate a portion of it to each worker node, and then combine the result sets from each node to return to the client. The master node then creates an abstraction; wherein the client communicates with a single server, sends it a single query, and gets back a single result set. This architecture is called massively parallel processing (MPP).

A pioneer in MPP data warehouse technology was Teradata, a company formed in 1979 that hit its MPP stride in the mid-1980s when the dimensional model was born. Teradata added to its novel approach by selling its product not as shrink-wrapped software but as an appliance, with all the individual servers (master node and worker nodes) along with shared enterprise storage infrastructure, the operating system, and the software pre-installed. Many other vendors adopted both the MPP architecture and the appliance delivery model.

Columnar Storage and the Year of the Data Warehouse

Some MPP vendors, most notably Vertica, further optimized by implementing columnar storage, which arranges data in columns rather than rows. This improves table scanning efficiency since only the relevant column or columns are scanned. Placing values from the same column together also facilitates high rates of data compression, which allows much of the data to be loaded into memory, making table scans faster still.

In addition to Vertica, vendors focused on the MPP-columnar combination include Greenplum, Netezza,

Sybase (with its Sybase IQ product), ParAccel, and DATAllegro. Many of these companies were acquired between 2008 and 2011, by large enterprise software vendors: Vertica by HP, Greenplum by EMC/Pivotal, Netezza by IBM, Sybase by SAP, and DATAllegro by Microsoft. In addition, ParAccel was founded in 2007, and Teradata, which had been acquired in 1991 by NCR, was spun off as a newly independent, public company that same year.

There have been other milestones in the DW world during or after that watershed MPP acquisition era. In 2008, Oracle introduced its Exadata (a name undoubtedly chosen to tweak the folks at Teradata) data warehouse platform, a highly engineered system based on Oracle Database, and ParAccel was acquired by Actian in 2013.

DW Challenges and Big Data Obsession

Post-2007, several market factors have caused ebbs and flows in the popularity of DW platforms. The appliance model heavily monetized storage, engendering customer dissatisfaction, as even the need for a small, incremental amount of storage could force significant capital expenditure. The effect on the market was retrograde; the appliance model disincentivized customers from adding data to their warehouses.

This, combined with the top-down, centrally controlled nature of many enterprise data warehouse (EDW) implementations, and the risk and expense of their "big bang," waterfall-style implementation projects, caused a malaise in the data warehouse market. Such approaches increased the risk of project failure and, in any case, meant longer delivery cycles before the business could derive value. While smaller "data mart" applications could proceed on conventional RDBMS platforms, the phenomenon of big data reminded everyone that going smaller was not a forward-looking solution.

Big data, on the other hand, involved then-cutting edge new technology with economically attractive storage and gave some enterprise customers a sense that they would be taking a future-proof approach to analytics. That gave DW vendors incentive to provide "best-of-both-worlds" approaches.

Open-source big data platforms and SQL-on-Hadoop

As a reaction to ever-increasing data volumes, open-source big data platforms began to take shape. Most notable among these was Apache Hadoop, first released in 2006. Hadoop addressed the now common customer gripes around DW platforms, including licensing, hardware, and storage costs. Rather than expensive enterprise storage, open-source big data software uses commodity direct-attached disks, which are more cost-effective.

Hadoop architecture also offers numerous built-in replication and other fault-tolerance features that ensure the resiliency of its clusters in the event of disk failure. Unlike the appliance model, which has hard limits on the number of nodes in the cluster, Hadoop is expandable; nodes can be added at will, ensuring cost-efficiency and linear scalability. Plus, additional disks can be added to existing nodes, allowing the degree of compute and storage to scale independently.

To pivot toward new market trends, many DW vendors sought to integrate their platforms with Hadoop. This led to a number of acquisitions. Teradata acquired Aster Data, Hadapt, Revelytix, Rainstor, and consulting/service-oriented firms Big Data Partnership and Think Big Analytics. Aster Data was especially influential because it's SQL-MapReduce technology integrated the first generation MapReduce algorithm into a relational database platform. It also facilitated the development of the reverse: enabling SQL queries submitted to a relational database to trigger MapReduce operations on a Hadoop cluster. The latter technology, originally named SQL-H and later rebranded to QueryBridge, facilitated a SQL-on-Hadoop technology native to the Teradata DW.

Many database vendors have taken cues from Teradata's QueryBridge technology and implemented their SQL-to-Hadoop bridges. These technologies comprised the first major development in the fit-and-start process of integrating data warehouse and big data/data lake technologies. Examples include IBM's "Big SQL," Oracle "Big Data SQL," Microsoft's "PolyBase," Vertica "SQL on Hadoop," and ParAccel's "On-Demand Integration" (ODI). For enterprises working to make lake and warehouse work together, these technologies have matured, and now is an excellent time to implement. That said, further sophistication is coming — especially from Microsoft, in the 2019 wave of its SQL Server (and, likely, APS) technology — and things will get even better in the next year.

Why DWs Never Went Away

Even without the big data platform malaise, and despite customer satisfaction issues around the leading DW platforms, the data warehouse model never went away. In hindsight, there were several reasons why these platforms could not go away. Among them:

- The ubiquity of the SQL skillset
- Most business data is structured
- The popularity of the relational model
- Dimensional modeling works

Let's look at each of these, in turn.

SQL Rules

Structured query language has been around since the early 1970s; it is an intuitive, universal skill amongst developers and an increasingly common one amongst business users. All major DW platforms use SQL as the native query and administrative language, which is why the traditional data warehouse has always held court, even in the days of Hadoop romance.

Business Data Craves Structure

Next, the concept of rows and columns, foundational to most relational databases, is a natural structure

for business data. While Hadoop and some of its subcomponents (especially the HBase NoSQL database) may accommodate data that vary in structure from row to row, most business data does not require that.

Business data derives from business processes, which are largely deterministic and consistent. Even in the case where operational data may not follow a strict schema, analytics data typically does. So, while Hadoop might be more tolerant of semi-structured data, the fact remains that a requirement for such accommodation has been an enterprise edge case. Furthermore, many data warehouse platforms have semi-structured data features (e.g., Vertica's flex tables, Snowflake's variant data type) of their own.

Dimensional is Seminal

The dimensional model has been the foundation of DW and BI technology for decades, and the ecosystem still gravitates towards it. Even self-service tools like Tableau explicitly identify and embrace the concepts of measures and dimensions. The fact is, dimensional model and data warehousing define the enterprise analytics "template," and even when open-source technologies like Hadoop and Spark, are used, many enterprise practitioners prefer to make them look or work like DW platforms.

Beyond the Classics

As the technology landscape has changed, so too has the data warehouse. New architectures, products, and companies have emerged. New DW categories have emerged, as well. We discuss those in this section.

DW on Hadoop

Some vendors took SQL-on-Hadoop integration a step further by implementing MPP data warehouse platforms that leverage Hadoop's Distributed File System (HDFS) as their storage layer. These include two open-source projects: Apache Impala (originally driven by Cloudera and now one of the two technology options commercialized as Cloudera Data Warehouse) and Apache Hawq (driven by EMC/Pivotal/Greenplum and initially offered only as a commercial product).

For those who prefer to work with big data platforms directly, rather than bridging through DW platforms to run their queries, open-source SQL-on-Hadoop options like Apache Hive, Apache Drill, Presto, and Spark SQL fit the bill. While these components preserve the SQL and relational paradigms as an interface for data in Hadoop, they do not bridge the worlds of data warehouse and data lake. Instead, they effectively present Hadoop and Spark as data warehouse platforms in and of themselves (Hive explicitly calls itself a data warehouse), but without real MPP architecture or exclusively columnar storage.

While many SQL query options have broadened the accessibility of big data platforms, they have also confused and overwhelmed some customers, providing further enticement for them to rediscover their

DW platforms.

Cloud Data Warehouses

We have already mentioned Actian acquired ParAccel. The irony is that Actian retired the ParAccel MPP product, and the technology was perpetuated by Amazon Web Services (AWS) as the basis for Redshift, AWS's data warehouse as a service, and has proven very successful for the Seattle-based cloud computing leader.

But Amazon, which was one of ParAccel's pre-acquisition investors, took the technology in a different direction, making it more attuned to usage patterns in cloud computing. With Redshift, Amazon introduced the concept of data warehouse elasticity that can expand or contract at will. Instead of having to size the cluster for the maximum demand, sizing could instead be based on typical workloads. Beyond that, the cluster could be enlarged, temporarily, during spikes in demand.

Redshift also inherited from ParAccel the flexibility of being an appliance-free data warehouse product, so that an incremental expansion does not trigger an outsized expansion of infrastructure. With Redshift, the unit of scale is just a single node. Granular and elastic scaling made Redshift very popular. So did the fact that it was, at heart, an MPP relational data warehouse, with a query interface compatible with the open-source PostgreSQL relational database, to which most query and BI tools can connect.

The union of old and new, cloud innovation paired with familiar MPP DW paradigm, made Redshift very attractive to many customers. Additionally, Amazon advertised petabyte-scale service and rapid deployment, without incurring massive capital expense, certainly a value-add to any analytics-driven organization.

Decoupling Storage

Further innovation was forthcoming, this time around storage. Microsoft, with its Azure SQL Data Warehouse service, decided to leverage object storage. Redshift, meanwhile, used standard solid-state drives (SSDs) for storage, which provided compelling performance, relative to using standard hard drives (HDDs).

Microsoft's take was that persistent data in durable object storage meant the company could offer cloud customers the ability to pause the worker nodes in the cluster when the DW was not in use. Since each node is a separate cloud virtual machine and each node a resource billing by the hour, this approach allowed huge savings for customers whose data warehouse work tended to be intermittent. Redshift lost much, if not all, of its performance advantage recently, when Microsoft introduced its Azure SQL DW "Gen 2," which combines the use of cloud storage as a primary layer with an SSD-based caching scheme.

The decoupled compute-storage model also lets customers scale their computing power and storage

separately – effectively making the unit of scale even more exceptional than Redshift's node-level granularity, avoiding the situation where customers who need more storage must also buy more compute (or vice versa).

3. Considerations for using Enterprise Data Warehouses

Following is a guide to help buyers select the optimal data warehouse solution, presented in the format of need-determining questions and recommendation-based answers.

1. Will you be implementing an Enterprise Data Warehouse solution?
    1. Look to a vendor targeting EDW scenarios with robust SLAs (Service Level Agreements).
    2. Consider likely concurrent usage and confirm vendor can accommodate spikes and dips to keep user satisfaction high, across the organization. The ability to scale up and down is key.
2. Are data portability and the flexibility of a multi-cloud strategy critical to your organization?
    1. Consider cloud-based solutions from independent vendors or using an on-premises-capable DW, but deployed over cloud Infrastructure as a Service (IaaS) resource.
3. Will you be building domain-specific and/or departmental solutions?
    1. Select a cloud vendor offering a serverless architecture or economizing features like the ability to pause and resume compute clusters.
    2. Scrutinize data lake integration capabilities, as exploratory analytics needs may be significant.
    3. Consider products in the “Autonomous” DW category, below, if in-house database administration skills are scarce.
4. How important are data lake workloads and analysis of semi-structured data to you?
    1. Look for strong integration with data lake technologies, including Apache Spark.
    2. Confirm connectivity to standalone files in various formats (including CSV, JSON, and Parquet) in cloud object storage or the Hadoop Distributed File System (HDFS).
    3. Investigate DW platform’s ability to handle loosely-typed data and apply schema-on-read (i.e., impose schema at query time).
5. Do you have predictive analytics/AI needs?
    1. Investigate integration with data and machine learning platforms, especially Apache Spark, which serves both workloads. Products in the “Other” category, below, may be especially appealing.
    2. If you are using a specific cloud provider’s AI platform, consider using that same provider’s DW platform to maximize integration and supported workflows.
6. How important is Business Intelligence (BI)?
    1. Virtually all BI tools can connect to any DW platform. Still, some offer very tight integrations (e.g., Power BI with Azure SQL Data Warehouse and Looker, Tableau, and Power BI with Snowflake).
    2. If your DW will primarily serve self-service BI workloads, performance is of utmost importance. SSD technology, columnar storage, and vector processing capabilities will all be necessary. So will evaluation and testing under realistic data volume loads.

7. Are you looking past incumbent vendors toward adopting a DW solution from a newer player, but trying to map the risk/reward of such a strategy?
    1. If having a DW platform that is more cloud-savvy and versatile in storage technology compatibility is important, or if integration with open-source analytics technology is a priority, look at newer players.

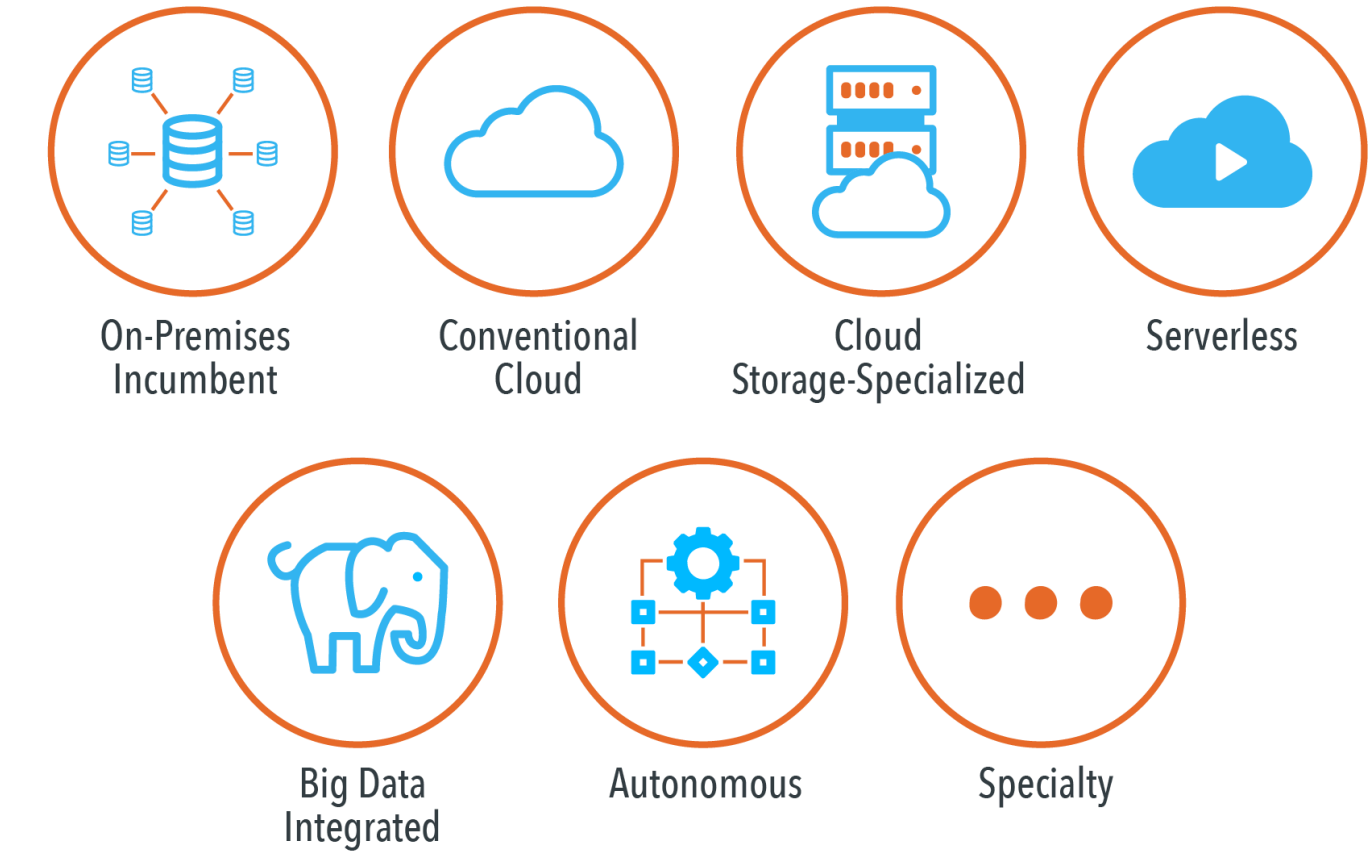
    2. If performance, enterprise manageability, skillset availability, or leveraging existing procurement arrangements outweigh new technology access, incumbents may be a safer bet.

    3. Consider that even incumbent players have modernized their platforms to take advantage of cloud and big data technologies, while still offering the assurances of a more prominent vendor with whom customers may already have relationships.

4. Key Players

Reviewing key players in the DW space is best facilitated by constructing a taxonomy of DW platforms. Figure 1, below, shows the overall DW product groupings.

FIGURE 1. **DW PRODUCT GROUPINGS**



On-Premises Incumbent (using MPP, column-store and/or other scale-out architectures)

These relational database platforms are purpose-built for data warehouse work. They may use MPP and columnar technology and sell in appliance form factors. This group includes products we have already discussed: Teradata, MicroFocus Vertica, Pivotal Greenplum, IBM Netezza, and Microsoft Analytics Platform System. We also mentioned ParAccel's DW platform, which became Actian Matrix and has been subsequently retired.

IBM Netezza came to Big Blue in one of the many 2010 DW pure-play acquisition deals. Although like several of its peers, Netezza is based on PostgreSQL at the node level, much of its technology is finding its way into IBM's legacy DB2 relational database product.

MicroFocus Vertica was one of the first data warehouse platforms to go beyond MPP and focus on columnar storage (hence the name). One industry juggernaut Michael Stonebreaker's progeny, original corporate entity Vertica Systems was acquired by HP in 2011, became an HP Enterprise product after the HP split and now owned by MicroFocus, which acquired HP Enterprise's software business.

Microsoft Analytics Platform System (formerly Parallel Data Warehouse, or PDW) is an MPP data warehouse built on SQL Server technology, engineered originally engineered by the DATAllegro team Microsoft acquired in 2008.

Oracle Exadata is a clustered, appliance-based version of the Oracle Database, engineered for data warehouse workloads.

Pivotal Greenplum, originally the product of the independent company named for the database, was acquired by EMC in 2010 and, along with fellow acquisition target Pivotal Labs, formed the Pivotal unit

captained by ex-Microsoft executive Paul Maritz, who is still Chairman of Pivotal's Board of Directors. This DW platform is based on PostgreSQL technology at the node level and, like its underlying relational technology, is now an open-source product. Greenplum technology is also the foundation for the Hadoop-based open-source database, Hawq, which was once a Pivotal product.

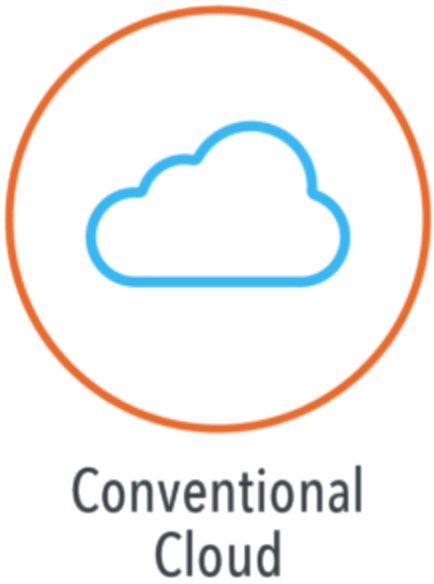
SAP BW4/HANA offers an in-memory database as a stand-alone product, which forms the platform for many of its software products. Although the original Sybase IQ product is still available as SAP IQ, the BW4/HANA product is the one on which the company's DW roadmap is now based.

There are a few niche sub-groupings. For example, some products are available as software-only purchases, allowing customers to use their hardware and avoid the appliance hardware lock-in. SAP's DW product, based on HANA, is in-memory. But, by and large, everything fits in a single category.

With the advent of the cloud, competition from big data platforms, and the pressure from self-service BI users to get data into the warehouse as quickly and with as much agility as possible, new categories have emerged. In the following sections, we investigate several such major categories, beyond the familiar on-premises MPP DW category.

Teradata is the granddaddy of MPP data warehouse products and the appliance form factor. Even in this age of cloud-native DWs, Teradata is the platform that challengers chase and seek to displace. Despite Teradata's significant investment in Hadoop-related companies and technologies, its classic DW pedigree makes it the standard by which others are judged.

Conventional MPP Cloud



This solution category encompasses data warehouse offered as cloud services based on conventional MPP architectures used by on-premises products. Products in this category offer managed services,

and are not just Infrastructure as a Service (IaaS) implementations.

Amazon Redshift. The key player in this category is Amazon Redshift. While several other cloud data warehouse services exist, virtually all of them leverage cloud object storage rather than more conventional storage infrastructure provisioned with the MPP cluster.

GoodData. GoodData offers a combination of services. GoodData embeds the Vertica data warehouse.

IBM Db2 Warehouse on Cloud. IBM is in this group as well, with its Db2 Warehouse on Cloud service. This service, though branded Db2, brings technology from the Netezza platform and, according to IBM, offers 90% or greater Netezza compatibility in terms of its data definition language, data manipulation language, and stored procedure capabilities. Data can reside on-premises (private cloud), in the public cloud, or a mix of the two (hybrid).

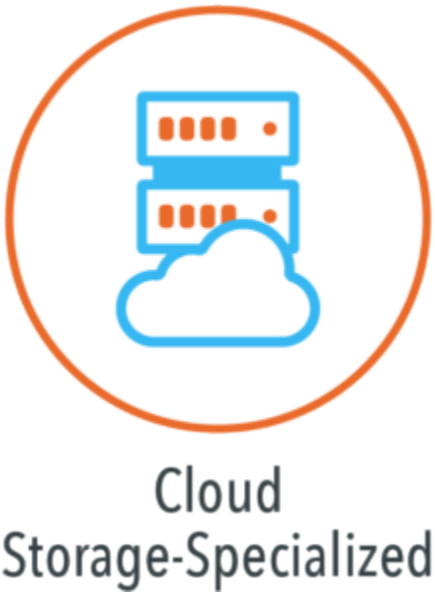
IBM provides a range of sizing options for the service. Its “MPP Small” service, available on AWS, best fits the conventional cloud DW grouping. For less demanding workloads, IBM also offers single-server symmetric multi-processing (SMP) options (in “small,” “medium,” and “large” sizes).

Db2 Warehouse allows data to reside on-premises (private cloud), in the public cloud, or a mix of the two (hybrid). The product also offers built-in analytics functions and integration with IBM Watson Studio, making it a compelling choice for IBM stack customers doing data science and AI.

Oracle Database Autonomous Data Warehouse. Autonomous Data Warehouse is Oracle’s lead Cloud DW service. It’s a fully-managed database, built on Oracle Exadata technology, running as a metered service on the Oracle Public Cloud, and optimized for DW workloads.

While Exadata leverages columnar data compression, it is not a full-fledged column store database and does not employ an MPP architecture.

Cloud Object Storage-Specialized



Several other cloud-based data warehouse services exist, both from public cloud vendors and independent providers. Each of these differ from the previous category in one key facet: they utilize cloud object storage instead of conventional cluster-based storage, allowing storage and compute to be scaled independently. Since cloud storage is durable beyond the life of any provisioned, compute infrastructure, many of these services will enable the warehouse service itself to be shutdown (or "paused").

IBM Db2 Warehouse Flex. IBM's Db2 Warehouse for Cloud, already covered in the Conventional Cloud product grouping, introduced "Flex" and "Flex Performance" sizing options recently (for dense storage and dense compute workloads, respectively) that, like Snowflake and Azure SQL DW, also leverage cloud object storage; offering independently scalable compute and storage as a byproduct. Db2 Warehouse Flex service is available on IBM's own cloud.

Microsoft Azure SQL Data Warehouse. Speaking of Azure, another key player in the Cloud Storage-Specialized product group is Microsoft, with its Azure SQL Data Warehouse ("SQL DW") service. Azure DW shares a technical heritage with Microsoft's Analytics Platform Service on-premises product and is ultimately based on SQL Server and its T-SQL query language.

Along with Snowflake, Microsoft pioneered the approach of using cloud object storage (Azure Blob Storage) and pause-able DW clusters. It recently improved on that with its "Gen2" product, which takes a hybrid approach to storage, bringing in conventional solid-state disk storage as a caching layer and continuing to use cloud object storage as the persistent store.

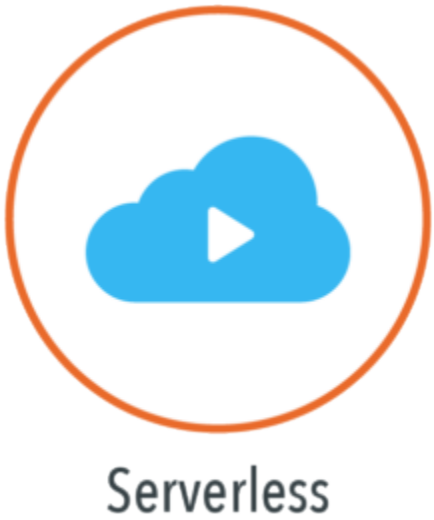
Snowflake. Foremost among vendors in this group is Snowflake, a pure-play/startup led by Bob Muglia, former President of Microsoft's erstwhile Server and Tools Business. Despite its startup status, Snowflake has had multiple rounds of funding, raising over $900M in aggregate. It is growing quickly

and has been an industry darling for some time.

Snowflake's product was originally launched on AWS, and an Azure-based offering was recently launched as well. As expected, Snowflake on AWS utilizes Amazon S3 storage. Snowflake on Azure uses Blob Storage.

Snowflake allows customers to spin up multiple clusters for greater concurrency (i.e., handling more simultaneous users and queries). The company has enjoyed notoriety and success with this approach and has arguably catalyzed innovation in its competitors.

Serverless



We have mentioned Amazon and Microsoft, but have not said much about the third major public cloud provider, Google.

Google BigQuery. Google's cloud data warehouse service is BigQuery. Unlike the DW services covered so far, BigQuery is not an MPP-based product, and it does not require customers to provision or size a cluster.

Instead, BigQuery is a serverless product. BigQuery decouples storage and compute; it utilizes its own managed, columnar storage and can connect to data in Google Cloud Storage as well as CSV, JSON, Avro, and Google Sheets data stored on Google Drive.

Users pay for the storage taken up by their data and also pay for compute power actually used. As such, there's no need to explicitly stop and start a cluster, as there would be with Azure SQL DW, and storage is, by definition, scaled separately from compute.

BigQuery also offers streaming data ingestion and supports operationalized machine learning via special SQL commands.

Big Data Integrated

As an outgrowth of the work done with open-source projects like Apache Impala and Apache Hive, open-source big data distributions have ushered in their own DW solutions, though some of them are more conventional in their use of the DW moniker than others.

Cloudera Data Warehouse. Cloudera approaches the analytics market with a "modern data warehouse" philosophy, combining elements of conventional MPP data warehousing and data lake implementations. Technologically, this translates into an endorsement of and a platforming approach based on a combination of the Apache Hive technology (especially Hive LLAP – described below) championed by Hortonworks and Impala technology, developed and advocated by Cloudera before the two companies merged.

This approach, in the context of the revamped Cloudera Data Platform (CDP) and its basis in Kubernetes, means that Cloudera customers can provision data warehouse instances that are workload-specific, persistent or ephemeral, supporting auto-scaling, based on Impala or Hive, in the cloud paired with Amazon S3, Azure Data Lake Store (ADLS) or Google Cloud Storage (GCS), or on-premises using HDFS or Apache Kudu (also developed by Cloudera, as a companion to Impala).

The company proffers the concept of data warehouse deployments dedicated to operations & events, research & discovery, as well as more traditional use cases. All Cloudera Data Warehouse infrastructure, regardless of underlying technology or location, can be managed in an integrated fashion using the company's Shared Data eXperience (SDX) framework and tooling, including unified security, governance, schema, and metadata.

Hawq. Pivotal, the unit of EMC, which acquired Greenplum's technology and team, developed its own big data ecosystem data warehouse by essentially re-engineering Greenplum to use HDFS for storage and releasing that technology as a new product called Hawq. Hawq is now also an open-source Apache Software Foundation project.

Hive, in its Many Versions. No discussion of big data ecosystem data warehouses would be complete without mentioning Apache Hive, the granddaddy of all SQL-on-Hadoop solutions. The problem is, despite billing itself as "data warehouse software" right on the project's home page, and despite Impala's SQL-level compatibility with it, Hive has historically not been a true data warehouse. Instead, it has been a standalone SQL query abstraction over Hadoop.

Originally, in the Hadoop 1.0 days, Hive generated MapReduce code, which was all Hadoop understood. The slow, batch nature of MapReduce meant anything close to DW-like performance on Hive was elusive at best. In the Hadoop 2.0 era, Hortonworks re-engineered Hive to work with Apache Tez, and Cloudera re-engineered it to work with Apache Spark. Both of these re-works made Hive faster, but neither really brought DW-level performance.

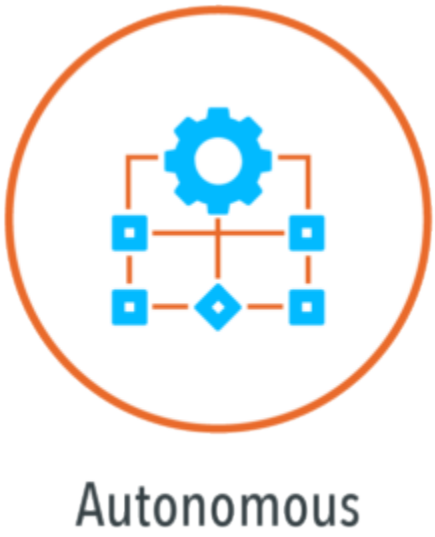
Later on, Hortonworks worked on a new version called Hive LLAP, the acronym portion of which can stand for "Live Long and Process" or "Low Latency Analytical Processing." Regardless of which acronym expansion is preferred, Hive LLAP was designed to make Hive a more responsive back-end for business intelligence tools, especially self-service products like Tableau and Power BI.

Now that we are in the Hadoop 3.0 age, Hortonworks had worked, prior to merging with Cloudera, to bring forward a 3.0 version of Hive. Hive 3.0 integrates Apache Druid in service of high-performance OLAP-style queries. Now that Hortoworks has merged into the new Cloudera and the latter has released its Cloudera Data Platform, Hive LLAP would appear to be the favored version.

Impala. Cloudera was the original developer of Impala before it contributed it to the Apache Software Foundation to become Apache Impala. The company has invested significant engineering resources into the product. While Impala supports most of the Apache Hive dialect of SQL (in addition to ANSI SQL) and can use HDFS (or cloud object stores) for its storage layer, it nonetheless uses an MPP architecture and was conceived as a data warehouse from its inception.

Cloudera's data warehouse approach leverages both Impala and Hive (described below). Specialty vendor Siren also embeds Impala in its platform. Impala was an optional component on Amazon's Elastic MapReduce (EMR) service for a while, but now must be installed manually to run on EMR clusters.

Autonomous



The industry has been at data warehousing long enough that patterns have emerged in the data modeling and ETL work that goes into transforming operational databases into data warehouses and dimensional models. Certainly, consulting organizations in the DW space have converted their experiences into formal methodologies that they apply as they embark on each project.

As an outgrowth of these methodologies, products that automate the modeling and design of a data warehouse and host the warehouse itself are now emerging. These products constitute a grouping of their own. Beyond methodology, these products may also use AI to help determine what is contained in various tables and how the data in those tables should be manifested as measures or dimensions.

Attunity. Attunity, though it does not offer a data warehouse product, deserves mention here. Through its Compose product, it provides offers an automated ETL solution that can support numerous data warehouse platforms, both cloud and on-premises. Attunity Compose and a DW platform, like Redshift or Snowflake, together provide an autonomous DW solution.

Infoworks. Another player in this space is Infoworks, which implements what it calls a “no-code transformation platform.” The Infoworks Autonomous Data Engine automates data ingestion (including ingestion of real-time streaming data) into Hadoop and automatically generates OLAP cubes. It provides monitoring, “collision detection,” auto abort/re-start, and other enterprise-grade manageability features to the ongoing operations of its automated pipelines.

Panoply. One such product is Panoply that automates the database maintenance

and data modeling/engineering that might ordinarily be done by a database administrator (DBA) or data architect. The system uses machine learning and natural language processing algorithms to

ingest data, reindex, and construct the DW's data model – a process Panoply refers to as "ETL-less data integration."

Panoply's DW materializes in Amazon Redshift, and BI tools need merely connect to that platform to consume the data and perform analysis. Panoply adds its own proprietary caching layer that "heats and cools" the data, aided by usage-based optimization (i.e., monitoring the most popular queries and optimizing accordingly) for maximum performance. In addition, Panoply automatically separates storage from compute.

Specialty

Not every DW product fits neatly into a taxonomy. This is especially true now that so many open-source analytics products can combine with relational database technologies in a mix-and-match fashion. In this section, we round-up a few more products that combine various technologies into platforms that can accommodate data warehouse workloads.

Actian Vector. Actian Vector has its roots in an academic project called MonetDB and a company called Vectorwise, which the former Ingres Corporation (now Actian) acquired in 2010. Vector and its precursors all utilize vector operations, supported by Intel's SIMD (single instruction, multiple data) CPU instructions, which process multiple values together rather than one at a time. This vector processing, pioneered in MonetDB, is now a staple a growing number of conventional DW platforms.

Siren. Like Splice Machine, Siren also mashes up Hadoop and Spark (as well as Impala) but adds other technologies to the mix as well. It uses PostgreSQL for relational database workloads and a combination of Solr and ElasticSearch for search-based analytics. Siren combines these technologies under a single analytics abstraction layer. It also sports its own dashboard/data visualization interface, combining search capabilities, BI, relational investigation, knowledge graph analysis, and alerting.

Splice Machine. Splice Machine uses a bit of a contrarian architecture: it is an ACID-compliant relational database that runs on top of HBase, a Hadoop-based NoSQL database, as well as Apache Spark. That combination yields a SQL-driven RDBMS on the Hadoop Distributed File System that can be used for operational or analytical workloads. Additionally, leveraging native Spark libraries, Splice Machine has machine learning native to the database.

5. Near Term Outlook

During the first several years of the big data era, many vendors and practitioners in the space proceeded on the premise that this new technology would supplant the then-current analytics platforms: data warehousing and business intelligence. Market factors and customer priorities have progressed to the point where big data platforms, which many now refer to as data lake technology, have found a peaceful and productive coexistence with data warehouse technology.

Data lakes have a unique ability to handle "raw" data. Raw data is sometimes only semi-structured and occasionally unstructured. But, even when fully structured, it tends to be in the form of individual files persisted in a storage layer, whether that is a distributed file system or a cloud object-store. Typically, each data set effectively combines fact and dimension data into a single file.

The curation of data in a data lake is, typically, minimal. Data is less vetted, redundant data may not be triaged away, and the very *goal* of a data lake is to err on the side of inclusiveness. All of this raises data governance and compliance risks with data protection regulations like GDPR and ePR (The European Union's enacted General Data Protection Regulation and draft ePrivacy Regulation) and CCPA (the California Consumer Protection Act).

Data warehouses are different. They tend to be repositories of consensus, with high degrees of curation and a very high bar for inclusion of data. As such, many practitioners consider the data warehouse the "single source of truth." As discussed, data warehouses employ a dimensional design, with dimension and fact data segregated into separate tables.

While both data lakes and warehouses store data ripe for analysis, they serve different goals. The data lake is more permissive, appropriate for experimental, one-off analyses. The data warehouse is more formal and "disciplined" fit for mission-critical, frequently-run analyses. It is also a great place to land data from the lake that turns out to be useful to a broad audience, where the ad hoc analyses of data conducted in the lake turn(s) out to be sufficiently valuable to run on a repeated, operational basis.

As such, the lake and the warehouse have complementary roles on a day-to-day basis, and they even have a workflow between them, where the lake can be a proving ground for data and analyses that may then "graduate" to the warehouse. The warehouse's use of schema-on-write makes it more suitable for systems of record. The data lake's use of schema-on-read makes it more flexible, less formal, and more appropriate for exploratory, ad hoc work.

One more thing to consider is that data warehouse platforms are becoming more viable for departmental or business domain-specific applications. Essentially, data warehouse platforms are now appropriate for more than Enterprise Data Warehouse implementations. The cloud makes this more feasible since it removes the significant capital expense of an on-premises appliance-based data warehouse, and allows data warehouses to be spun up for time-limited projects. Fundamentally, the cloud has made DW platforms more versatile, and this may cannibalize opportunities where data lake technology may otherwise apply.

Here is what to expect:

- The major RDBMS and DW platforms will embed Hadoop and Spark technologies into their platforms. This is already happening with Microsoft SQL Server, as the preview for the 2019 version of the product will integrate with Spark and HDFS. Expect other vendors to follow suit.
- Streaming data platforms will become more integrated into operational and analytics data systems too. And, while Kafka seems a big technology winner, expect the cloud providers to push their own streaming platforms, combined with Kafka API compatibility.
- Machine learning and AI will become more integrated, as well. The time horizon for this may be unclear for a while. Some point integrations have been created, and Databricks has integrated ML into the mainstream of its platform and workflow. But much more must be done. Auto ML and cloud cognitive services have to figure in more, and everything must become more embedded and seamless. ML and AI integration is an area that merits scrutiny to understand and leverage its benefits in the near future.
- Data governance and data protection must get easier so that companies can be both data-savvy and regulation-compliant.
- Finally, all the technologies discussed in this report – data warehouse, BI, and data lake – must become much more accessible to enterprise developers. The technologies must simplify, and the concepts must become more universally taught and conveyed. And, the more things that are available as a service, the better.

If the industry can do all of these things, and do them in the next year or two, then things will become much more manageable for many customers. And if machine learning is part of the platform, then digital transformation can finally begin in earnest.

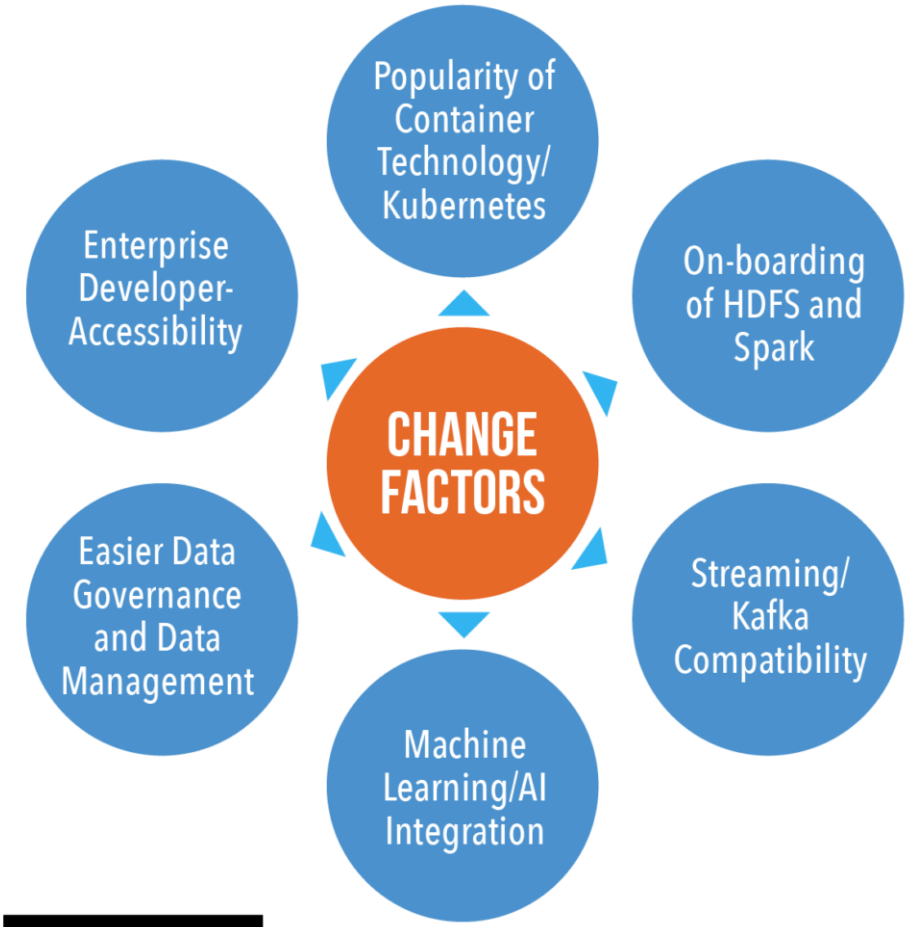
6. Key Takeaways

Taking an inventory of where things are and how they have progressed is valuable. The entire analytics market has shifted so much in the last few years that chronicling what has happened and summing up where we are is essential to planning and strategy. We hope this report has helped and will help in that regard.

But the $64,000 question is, where are things going? Will the current fabric of data warehouse, data lake, and streaming data technologies stick? Or will things change further? Will the incumbent mega-vendors like Microsoft, IBM, SAP, and Oracle win out over the startups and their disruptive technologies? Or will the cloud vendors like Amazon, Google, and – again – Microsoft win out as warehouse, lake, and streaming platforms continue their drift upwards to the cloud?

A number of change factors are present in the DW ecosystem today or will soon emerge. These are summarized in Figure 2, below.

FIGURE 2. **CHANGE FACTORS IN THE DW SPACE**



Summarizing each of these factors, the reality is that an increasing number of customers see cloud object storage as the repository of choice for all their data: operational, lake, warehouse, and streaming. They like running their warehouse in the cloud, and they like running Hadoop and Spark in the cloud. All of this points toward good things for public cloud providers. It also says that the real battle for the cloud is a battle for where customers store their data.

But customers also want portability. They want to use open-source platforms, deployed via container technology like Docker and Kubernetes. They want mobility for their workloads, so they can easily shift them between their own data centers and all three major public clouds. This points to independent companies (like Snowflake) that can run across public clouds, as major benefactors.

But, even in the latter case, the cloud providers win – as they will monetize the compute and storage, even when a third party DW platform is involved. Even if customers can keep their workloads mobile across clouds, the reality is that the lion's share of their data will be on just one of them. The cloud would seem to be the arena, no matter what; the only question is whether customers will buy their tickets at the arena box office or through an independent agency.

7. About Andrew Brust

Andrew has held developer, CTO, analyst, research director and market strategist positions at organizations ranging from the City of New York and Cap Gemini to Gigaom and Datameer. He has worked with small, medium and Fortune 1000 clients in numerous industries and with software companies ranging from small ISVs to large clients like Microsoft. Andrew's resulting understanding of technology, and the way customers use it, makes his market and product analyses relevant, credible and empathetic.

Andrew has tracked the Big Data and Analytics industry since its inception, as Gigaom's Research Director and ZDNet's lead blogger for Big Data and Analytics. Andrew co-chairs Visual Studio Live!, one of the nation's longest running developer conferences. As a longtime technical author and speaker in the database field, Andrew understands today's market in the context of its longtime Enterprise underpinnings.

8. About GigaOm

GigaOm provides technical, operational, and business advice for IT's strategic digital enterprise and business initiatives. Enterprise business leaders, CIOs, and technology organizations partner with GigaOm for practical, actionable, strategic, and visionary advice for modernizing and transforming their business. GigaOm's advice empowers enterprises to successfully compete in an increasingly complicated business atmosphere that requires a solid understanding of constantly changing customer demands.

GigaOm works directly with enterprises both inside and outside of the IT organization to apply proven research and methodologies designed to avoid pitfalls and roadblocks while balancing risk and innovation. Research methodologies include but are not limited to adoption and benchmarking surveys, use cases, interviews, ROI/TCO, market landscapes, strategic trends, and technical benchmarks. Our analysts possess 20+ years of experience advising a spectrum of clients from early adopters to mainstream enterprises.

GigaOm's perspective is that of the unbiased enterprise practitioner. Through this perspective, GigaOm connects with engaged and loyal subscribers on a deep and meaningful level.

9. Copyright