Options for analytic databases and warehouses

Introduction

It is our view that if you are an Oracle or Microsoft shop, are dedicated to big blue or, God help you, committed to SAP HANA, then you are not going to look outside of these vendors for a data warehousing or analytic database solution. However, if that is not the case, or you are tired of the adverse price/performance you get from these vendors, then this Market Update focuses on the alternatives. For this reason the mainstream products that are offered by these vendors (IBM Db2 Event Store is an exception for reasons discussed below) are not discussed in this report. If you want to imagine them on the Bullseye chart then, with the exception of SAP (challenger) you can envisage each of them as champions but (with the possible exception of Azure Synapse Analytics) on the challenger rather than the innovative side of the sector. For example, Oracle has just announced (February 2020) the availability of Python support in the Oracle Autonomous Database. You wouldn't exactly say that this was ground-breaking.

The products that are included here are heterogeneous. Some might be deployed as either data warehouses or marts, while others are more oriented towards data lakes. Some may even be deployed in combination. These are colour-coded appropriately on our Bullseye diagram. In the case of Actian, which has multiple products that might qualify for this report, we have opted to focus specifically on Actian Avalanche. While there are several other products that we might have included in this paper the one whose absence is most obvious is Amazon Redshift. We would have liked to include it though we suspect – for reasons discussed later – that it would not have fared too well in our comparisons.

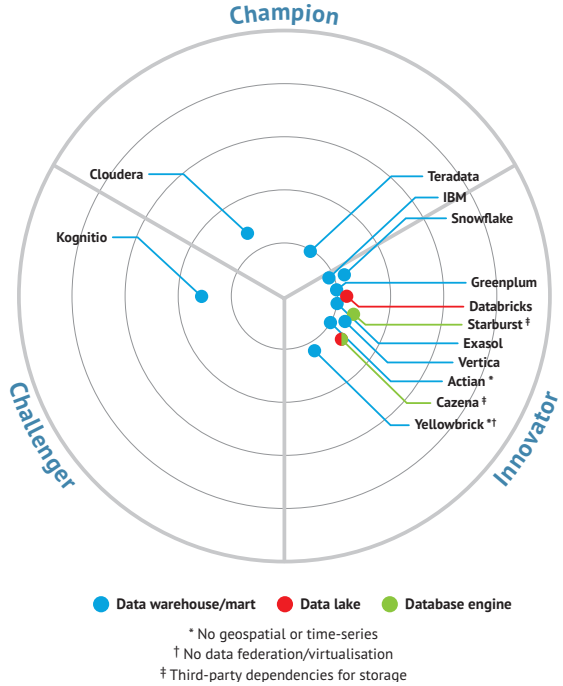
As far as the actual vendors included are concerned, most can be implemented as either data warehouses or marts. The exceptions are Databricks' Delta Lake and Cazena, which are both targeted at data lakes; and Starburst Presto which is a database engine rather than a database per se. That is, it relies on third party data storage. The same is true for Cazena. It should be noted that the Cloudera Data Warehouse, as evaluated here, is just one example of a use case for the Cloudera Data Platform.

Market trends

The biggest trend in data warehousing, if trend is the right word, is in deployment flexibility. With the exception of some cloud-only providers almost everyone else wants to offer a choice between traditional on-premises, in-cloud (private or public), hybrid and managed service deployments. However, vendors are taking different routes to this goal with some offering a single product with multiple implementation options and others offering different products, even though they may (not always) have the same underlying database engine, depending on the environment.

A trend that is effectively over but which has been an issue for the last several years is in the separation of storage and compute. This was once a differentiator but that is no longer the case as it is now commonplace, though there is at least one vendor that offers this separation for on-premises

Figure 1: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator or Challenger segments, depending on their innovation score. The exact position in each segment is calculated based on their combined innovation and overall score. It is important to note that colour coded products have been scored relative to other products with the same colour coding.



as well as in-house deployments, which might be considered a differentiator. It is also worth noting that this separation should be optional as not all use cases lend themselves to this principle.

A third trend, which is not vendor specific, but related to user demands is that we see a growing demand for object storage, mostly Amazon S3 and Azure, as opposed to the use of Hadoop (which is diminishing). Together with the growth in use of data lakes this has led to warehousing and analytic database vendors increasingly supporting data virtualisation and federation across multiple sources (not just object storage). We make a distinction here between virtualisation and federation where the latter simply allows you to access data in third-party environments, but the former lets you push-down query processing into those data sources which is, needless to say, significantly more efficient. Some vendors are much more advanced and support a far wider range of these sources than others, and this represents a major competitive advantage for those suppliers. In the longer term we expect this capability to become ubiquitous and it will put at threat pure-play data virtualisation vendors such as Denodo and TIBCO (Composite Software).

Fourthly, with the growth in interest in Internet of Things analytics, there is an increasing push to support both geospatial and time-series capabilities within the database. Some suppliers are much further ahead than others in this area.

Finally, of course, there is the neverending pursuit of improved performance. Everyone now has in-memory capabilities, supports columnar processing, has in-database analytic capabilities and supports machine learning. These are all de rigueur, as is ease of use and administration (though some are better than others), high availability, security, disaster recovery and so on. This is, after all, a mature market.

The vendors

We started on this research with the intention of ignoring the 800lb gorillas of the warehousing world and concentrating on other options. However, we have made an exception for Db2 Event Store, partly because it is probably less well known outside the IBM customer base and partly because it is essentially addressing a different issue, namely how to support analytics, including complex analytics, against both streaming and historic data, within a single environment and without requiring either a Lambda or Kappa architecture. Actian is another company with multiple database offerings but we have chosen to focus on its newly launched Avalanche cloud-based product. To some extent this is unfortunate for Actian because Avalanche does not yet have some of the features – especially geospatial and time-series support – that are incorporated into the company's more traditional products and which will no doubt be ported to Avalanche at some point in the future.

The other notable exclusions from this report are Splice Machine, Esgyn and Amazon RedShift. We approached Splice Machine for inclusion, but they did not respond. Esgyn told us that it is focusing on the Chinese market. As far as Amazon is concerned, at the time of writing we do not regard its omission as any great loss. Indications suggest that Redshift has been losing traction and is being outperformed by competitors. The latter also appears to be true for Google BigQuery. As far as Redshift is concerned this may change, as Amazon has announced some significant enhancements going forward, but it is always difficult to regain lost momentum.

With respect to other suppliers, we considered including the graph database vendors Cambridge Semantics and TigerGraph but have opted not to, both because we will shortly be publishing the 4th Edition of our Graph Database Market Update and because it would further complicate this report with yet another technology.

Finally, we mentioned previously that Hadoop seems to be losing market share. Also, IBM is no longer supporting Netezza, though it does have a replacement. Inevitably, other vendors in the market are targeting these areas as migration opportunities. Amongst others, Teradata is targeting Hadoop while both Actian and Yellowbrick are targeting Netezza. Cazena is one of several others targeting refugees from the warehouse market behemoths in general. In this context, tools to support migrations from third-party environments will be useful capabilities.

Metrics

The metrics against which we have compared the various vendor products include:

- Performance: database optimiser, workload management, in-database analytics, ingest capabilities via Kafka and other streaming technologies, and other features that pertain to performance.
- Architecture: including scalability (auto and elastic scaling), compression, high availability, disaster recovery and failover, support for Kubernetes, and so on.
- Ease of use and administration: including autonomics, the provision of a managed service, command modules, security and so forth.
- Datatype support: notably geospatial, time-series, temporal, text and image support.
- Machine-learning support: for libraries and languages (R, Python and Scala) as well as Jupyter and other notebooks, PMML, and any other specific capabilities like support for MLFlow.
- Data federation and virtualisation: the former is useful the latter is better. The extent of support for third-party environments is relevant.

Features not used in our comparisons include tools for migrating to the relevant database environment, as well as data stewardship tools – available from some vendors – that enable the discovery of sensitive data.

Conclusion

Data warehousing is a mature market and even for companies such as Cazena, Databricks and Starburst, that are in allied spaces, it is not as if these are new technologies. It is often, therefore, difficult to choose between products unless there are specific features that are required, such as support for time-series or Scala or PMML. The advantages of managed services and cloud-first development are well known but these do not come without cost, while most other capabilities you might want, are ubiquitous. Most typically, however, performance is a key determinant. Bloor Research has consistently recommended that users perform their own proofs of concept as performance claims and counterclaims are common currency in this market. Unfortunately, this is becoming more difficult as users increasingly want to access data stored in a variety of different places. Where putative suppliers have push-down data virtualisation capability, or even simple data federation capabilities, this will make the establishment of a realistic proof-of-concept more difficult. With the growth in use of Amazon S3 and similar storage mechanisms it will certainly make sense to include this at least in any proof of concept.

Choosing between the various offerings in the market is non-trivial. The majority of products discussed in this report have been available for more than a decade and they have the capabilities that you would expect from such maturity. New entrants into the market may have had some initial advantages when they first appeared, but those have now largely disappeared. The truth is that all the products discussed in this Market Update are worthy of serious consideration.

About the authors

PHILIP HOWARD
Research Director / Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), and dining out.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

tel: **+44 (0)1494 291 992**
web: **www.BloorResearch.com**
email: **info@BloorResearch.com**

Bloor

InBrief

Philip Howard – Research Director, Information Management

CLOUDERA

www.cloudera.com

395 Page Mill Rd
Palo Alto, CA 94306 USA
Phone: +1 888 789 1488

Cloudera Data Warehouse

The company

Cloudera was founded in 2008. Initially it was backed by venture capital but in 2017 it floated on the New York Stock Exchange. The following year it announced a merger with HortonWorks, which was completed in January 2019. The company is expecting revenues of around $780m for the financial year ending January 2020.

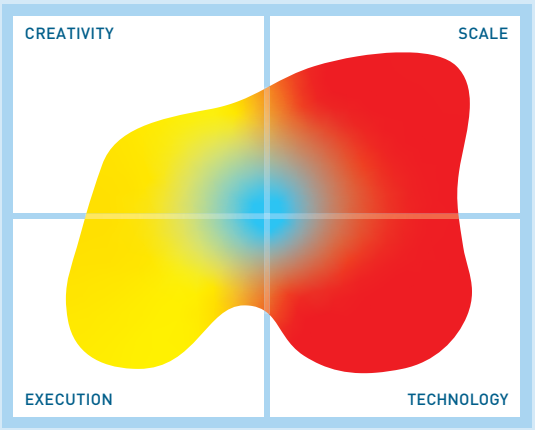
> "Thanks to the Cloudera platform, we can serve our customers much better and faster, we can respond to regulatory compliance much better and faster, and we can improve our fraud detection capabilities."
> KBTG (Kasikorn Bank)

While historically known as a commercial provider of a Hadoop distribution the company is now marketing itself as the *"Enterprise Data Cloud Company"*. This doesn't mean that it is eschewing its heritage but that it is focusing on the provisioning of data in the cloud, whereby users do not have to worry about any complexity involved in provisioning big data and data warehouse clusters.

Figure 1 – Marketecture, showing capabilities of Cloudera Data Warehouse

What is it?

The Cloudera Data Warehouse is based on the Cloudera Data Platform (CDP), which is effectively a convergence of the Hadoop platforms that were previously offered by Cloudera and HortonWorks individually. The new Cloudera has adopted the open source principle previously advocated by HortonWorks in that everything within the CDP



The image in this Mutable Quadrant is derived from 13 high level metrics, the more the image covers a section the better. Execution metrics relate to the company, Technology to the product, Creativity to both technical and business innovation and Scale covers the potential business and market impact.

platform is now available with an open source license where previously some Cloudera products were treated as proprietary. The company has also adopted a cloud-first development process whereby new features are first available in the cloud (AWS or Azure) and only subsequently for on-premises implementations. For example, Cloudera ML (machine learning) Experience and the Cloudera Data Catalog are both available for in-cloud deployments today (2019) but will only be available on-premises sometime during the first half of 2020.

Figure 1 is a marketecture diagram showing the capabilities provided by CDP of which the Cloudera Data Warehouse is one use case. The whole environment involves more than 30 different open source (Apache) projects, especially in the security and governance and analytics layers, but it would be tedious to call each of these out by name (but see ***Figure 2***). Suffice it to say that, as ***Figure 1*** illustrates, the environment is comprehensive.

What does it do?

Figure 2 shows the various elements of the platform that are specific to the Cloudera Data Warehouse environment.

Performance	4.5	Data science support	5
Architecture	3.5	Data federation/virtualisation	4.5
Ease of use/administration	4.5	Geospatial and other datatype support	5

> "We've created a platform that provides our scientists with insights that can shorten delivery timelines, reduce costs, expand reach, increase safety, and, in the end, improve, extend, and save lives."
> GlaxoSmithKline

As can be seen, a major feature of the Cloudera Data Warehouse is that it is, from a storage perspective, much more than an HDFS platform. Other notable capabilities which you may not be able to directly infer from ***Figure 2***, include support for time-series and geospatial data (important in Internet of Things contexts, as is support for Apache NiFi and Kafka); auto-scaling (cluster shrinks or expands depending on workload, and compute is separated from storage), auto-suspension and auto-resumption; workload isolation so that differing tasks do not compete with each other for resources (different compute engines can run against the same data); and data and metadata caching to improve performance so that subsequent queries run faster than the first, even when the query is similar rather than identical or when you have a different query but running against the same data. There are also migration capabilities for users moving to the cloud from on-premises implementations.

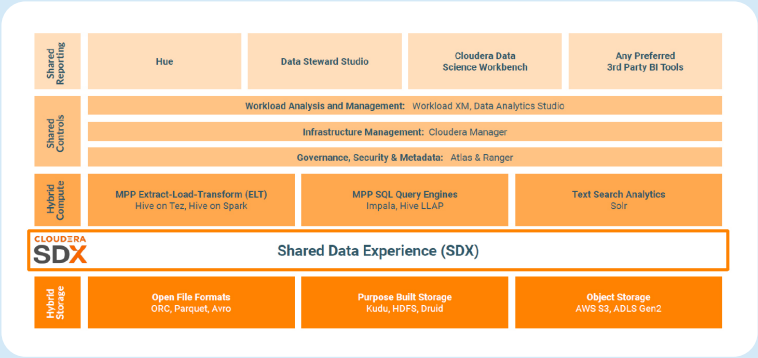


***Figure 2** – Elements of platform specific to Cloudera Data Warehouse*

It is also worth commenting on Data Steward Studio, which provides a user interface for data stewards that includes profiling and classification capabilities as well as the ability to discover sensitive data using both machine learning and natural language processing. Policy management and auditing are also provided. Finally, Cloudera Machine Learning (CML) offers a self-service, data science and machine learning development environment with on-demand access to (governed) business data, auto-scaling computing resources and users' preferred libraries, frameworks and IDEs for the Python, R and Scala ecosystem. The Cloudera Data Science Workbench is CML's on-premises equivalent.

Why should you care?

Hadoop clusters are complex to implement and manage so moving to a cloud environment where that complexity is removed, makes a lot of sense, so you don't have to worry about resource management or, for that matter, governance and security, because these are built-in. Moreover, Cloudera Data Warehouse isn't just about Hadoop: if you want to use Amazon S3 or Azure BLOB storage instead of, or in addition to HDFS, then you can do that, and/or you can leverage other Apache database engines.

Going beyond this purely architectural perspective, Cloudera Data Warehouse is unusual in that it also provides a data catalog, the Data Steward Studio, the Data Science Workbench and Apache Hue (which is a SQL editor). Most competitive data warehouses, in the cloud or otherwise, do not offer all of these complementary technologies and it is arguable that Cloudera Data Warehouse is a misnomer: the product is much more than just a data warehouse.

However, those are technical arguments. The other major benefit that Cloudera offers is that it is completely based on open source projects. Whether for licensing reasons or simply because of preference, this will be a major argument in favour of Cloudera for many decision makers.

The Bottom Line

We are going to have to stop thinking about Cloudera as a Hadoop company and start considering it as a general-purpose, data and database oriented organisation that is focused on open source and the cloud. That will be a powerful argument in its favour for many people.

FOR FURTHER INFORMATION AND RESEARCH CLICK HERE