Bloor

MarketReport

Market Report Paper by Bloor
Author **Philip Howard**
Publish date **December 2017**

SQL Engines on Hadoop

> “
> It is clear that Impala, LLAP, Hive, Spark and so on, perform significantly worse than products from vendors with a history in database technology.
> ”

Author **Philip Howard**

Executive summary

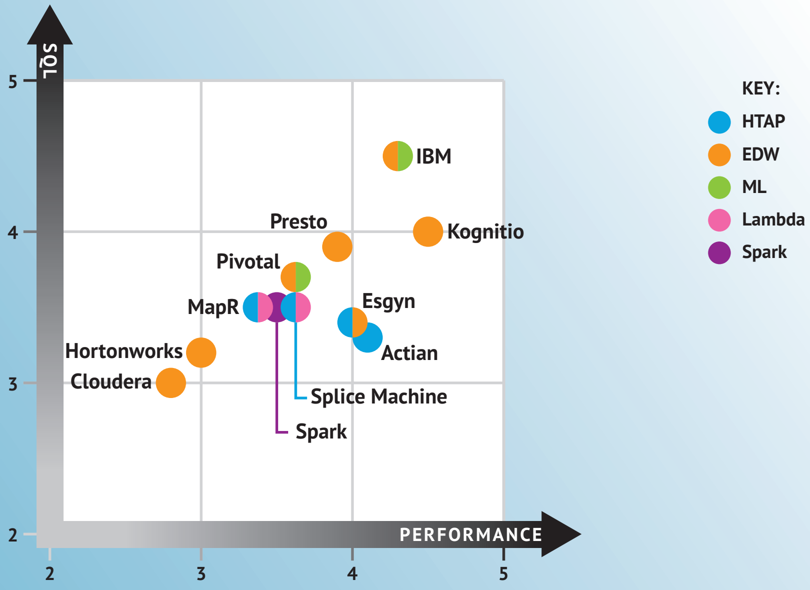
Hadoop is used for a lot of different purposes and one major subset of the overall Hadoop market is to run SQL against Hadoop. This might seem contrary to Hadoop's NoSQL roots, but the truth is that there are lots of existing investments in SQL applications that companies want to preserve; all the leading business intelligence and analytics platforms run using SQL; and SQL skills, capabilities and developers are readily available, which is often not the case for other languages. However, the market for SQL engines on Hadoop is not mono-cultural. There are multiple use cases for deploying SQL on Hadoop and there are more than twenty different SQL on Hadoop platforms. Mapping the latter to the former is not a trivial task, as different offerings are optimised for some purposes but not others.

The key differentiators between products are the use cases they support, their performance and the level of SQL they offer. While all of these are discussed in detail in this paper it is worth briefly explaining that SQL support has two aspects: the version supported (ANSI standard 1992, 1999, 2003, 2011 and so on) plus the robustness of the engine at supporting SQL queries running with multiple concurrent thread and at scale.

Figure 1 illustrates an abbreviated version of the results of our research. This shows various leading vendors, and our estimates of their product's positioning relative to performance and SQL support. Use cases are shown by the colour of each bubble but for practical reasons this means that no vendor/product is shown for more than two use cases, which is why we describe **Figure 1** as abbreviated. Thus, for example, we are using "EDW" as shorthand for products that support both transactional lookups and complex analytics, which are otherwise individual use cases. Also, it excludes vendors targeting OLAP, as the leaders in this market – Jethro Data and Kyvos Insights – have distinct approaches that are not easily compared.

> **"The key differentiators between products are the use cases they support, their performance and the level of SQL they offer."**

Figure 1 –
Use cases by performance and SQL support. Use cases include Hybrid Transactional and Analytic Processing (HTAP), a merger of the transactional look-ups and complex analytics (EDW: enterprise data warehouse), combined batch and real-time/streaming analytics (Lambda architectures), and machine learning (ML). OLAP and some other use cases are omitted.



Use cases

We have identified six different use cases for SQL on Hadoop. Some of these overlap one another and there will also be instances where a user wants more than one of these use cases running on the same cluster. However, we believe that the examples detailed provide the bedrock for making decisions about potential solutions.

The main use cases we have identified, in no particular order, are:

1. Transactional look-ups. This will often be combined with other use cases.
2. Hybrid transactional analytic processing (HTAP).
3. Complex queries against large datasets. Typically involving many users. We might describe this as “traditional data warehousing” and, certainly, there are vendors aiming to replace enterprise data warehouses (EDW) via this use case. Often combined with transactional look-ups.
4. Online analytic processing (OLAP). May be either multi-dimensional OLAP (MOLAP) or relational OLAP (ROLAP).
5. To support machine (and deep) learning.
6. A “collapsed” lambda (or kappa) architecture designed to support both batch and real-time (streaming) analytics. Will often be combined with either or both of OLAP and machine learning,

There are several other uses cases where you might want to use SQL on Hadoop but, often enough, Hadoop on its own will be enough. These use cases include extract, load and transform (ELT) and archival, as well as (ad hoc) data preparation. The last of these was identified as a use case by one of the vendors, although none of the suppliers – including the identifier - we have spoken to, have claimed to target it. The same applies to data discovery and similar use cases where you would probably be better off to rely on an information/data catalogue running on your data lake. One vendor also suggested a use case as an operational data store.

> “We have identified six different use cases for SQL on Hadoop. Some of these overlap one another and there will also be instances where a user wants more than one of these use cases running on the same cluster.”

Offerings

Products in this market tend to fall into one of six categories and in the following lists we have highlighted those products we examine in more detail in this report. The groupings consist of:

- Pure-play open source projects. This category includes Hive, HBase, Tajo, Phoenix, while Ignite and Spark. See also the OLAP-based projects below. All of these are Apache projects. Of the less well-known offerings Phoenix supports on-line transaction processing (OLTP) running against HBase; Ignite is an in-memory computing platform that is commercially supported (and was originally developed) by GridGain. It is typically used either as a Hadoop accelerator and/or to provide immediate consistency. Tajo is a big data warehouse. There have been no new releases of Tajo for 18 months, so we suspect that it is defunct.
- Vendor supported open-source projects. This group includes **Drill** (supported by MapR), **Presto** (Teradata), HAWQ (**Pivotal**) and Trafodion (**Esgyn**). All of these, again, are Apache projects. Also in this category are Impala (Cloudera) and Hive + LLAP (Hortonworks – live long and process - previously known as Stinger). Note that Drill does not have to run on Hadoop.
- Traditional data warehousing products that have been used as the basis for SQL on Hadoop platforms. These include IBM Db2 (**Big SQL**), Oracle, Vertica, **Pivotal HDB** (HAWQ: effectively a port of Greenplum), **Kognitio** (which is free-to-use) and **Actian VectorH**. VectorH is the odd one out here because Actian Vector is a symmetric multi-processing (SMP) solution that has been developed into a massively parallel processing (MPP) environment. All the other products were MPP-based originally.
- Other MPP-based solutions. This category consists of Transwarp and **Esgyn**. The latter is is a descendant of Tandem NonStop, HP Neoview and other HPE-based warehousing developments.
- Specialist offerings. Mostly these are targeted at OLAP environments. In this category are Apache Kylin (MOLAP) and Apache Lens (ROLAP) as well as **Kyvos Insights** and **Jethro Data**. **Splice Machine** is also in this category but has rather broader capabilities (see later). AtScale will compete with products in this category but is a “BI on Hadoop” engine rather than a SQL on Hadoop platform: as such it is not discussed further here.
- Others that are often referred to as SQL on Hadoop engines, but which are not. Included in this category are Splout SQL, which is really about data serving, and Concurrent Lingual, which is used for application development. Druid, which started life as an MDX engine (and which now has limited SQL support) is another data serving product with OLAP capabilities. Apache Calcite is a general-purpose SQL optimiser but not an engine per se. None of the products in this group are discussed in this report.

In the vendor/product section of this report we include short descriptions of many, though not all, of the proprietary products (open source or otherwise), with the exception of Oracle, Vertica and Transwarp, none of which responded to our requests for information. While the omission of Oracle and Vertica is no great loss (a straight line can be drawn across from their traditional products), we would have liked to include details about Transwarp.

> **“Traditional data warehousing products have been used as the basis for SQL on Hadoop platforms. These include IBM Db2 (Big SQL), Oracle, Vertica, Pivotal HDB (HAWQ: effectively a port of Greenplum), Kognitio (which is free-to-use) and Actian VectorH.”**

Performance benchmarks

A great many vendors in this space have conducted and published benchmarks. Some of these have been validated by third parties, some of them have been conducted by third parties, but the majority have not involved any independent authorities. Although TPC (transaction processing council) tests have typically been the basis for these benchmarks, none of them have been authenticated by TPC. The individual product descriptions that follow outline the results of the various benchmarks that have been performed by different vendors. We will therefore confine ourselves here to general comments.

The first point that we would like to note is that TPC-DS (Decision Support) tests are not just an indicator of performance but also of SQL support. TPC-H, on the other hand, is based on SQL 92, which is hardly up-to-date. We are disappointed with Actian, therefore, that it is focused on TPC-H and not TPC-DS.

The second point is that many tests are done using relatively small datasets and a single processing thread, when what you are really want is multiple users running against large sets of data. IBM, for example, has demonstrated that while Spark is perfectly capable of running all TPC-DS queries at small scale it breaks down as you scale up.

Thirdly, some vendors, notably Hortonworks and Cloudera, both of which have been guilty of publishing partial results. For example, just selecting (no doubt the best ones) 15 of the 99 TPC-DS test to report on.

To conclude this section – while not all products have been benchmarked and some have been benchmarked against different standards – it is clear that Impala, LLAP, Hive, Spark and so on, perform significantly worse than products from vendors with a history in database technology. Moreover, it is much more likely that companies in the latter category will be able to support all of your queries and run them successfully: the level of SQL support from the pure-play, Cloudera or Hortonworks products, tends to be limited.

While on the subject of SQL support, it is worth commenting that the level of support for ANSI standard SQL varies widely. IBM – not just in Big SQL and Db2, but across its product range – is much the most advanced vendor in this respect. Conversely, there are a number of products whose ANSI support dates back to the last century.

> “The level of support for ANSI standard SQL varies widely. IBM – not just in Big SQL and Db2, but across its product range – is much the most advanced vendor in this respect.
> There are a number of products whose ANSI support dates to the last century.”

Product suitability

While performance may be a major determinant in buying decisions, it is only relevant when comparing apples with apples, and the products covered in this paper constitute an entire fruit bowl. In this section we therefore match products to use cases.

1. Transactional look-ups. This will often be combined with use case 3 (see above). Various products, often in conjunction with HBase are suitable here. Notable contenders would be IBM Big SQL (with HBase), Splice Machine (which incorporates HBase into its Lambda architecture), Esgyn and Actian VectorH.
2. Hybrid transactional analytic processing (HTAP). This is a major focus area for both Esgyn and MapR Drill. Splice Machine is also a suitable contender here, though its emphasis is slightly different (more on leveraging transactional data for predictive analytics than embedding analytics into operational applications). The InterSystems IRIS Data Platform also competes here though it is not based on Hadoop (but is a clustered solution) as do others.
3. Complex queries against large datasets. Typically involving many users. We might describe this as "traditional data warehousing" and, certainly, there are vendors aiming to replace enterprise data warehouses (EDW) via this use case. Often combined with transactional look-ups. All the "ported" data warehouses play in this space, as do Actian VectorH, Esgyn and Splice Machine. Kognitio comes out well in the benchmark studies we have investigated.
4. Online analytic processing (OLAP). May be either multi-dimensional OLAP (MOLAP) or relational OLAP (ROLAP). The vendor-based products in this area are much stronger than any open source offerings. Kyvos Insights, Jethro Data and Splice Machine are the vendors to consider.
5. To support machine (and deep) learning. Pivotal, IBM and Splice Machine are the companies most active in this area, but where IBM relies on Spark MLlib, Pivotal is a major contributor to the Apache MADLib project. Splice Machine ships with MLlib.
6. Lambda architectures. Both MapR and Splice Machine are in the business of "collapsing" lambda architectures to support batch and streaming analytics into a single platform. In the case of Splice Machine, a Spark processing engine is embedded into the platform. In this context it is worth commenting that independent benchmarking has found that tight integration with Spark results in a 11x performance improvement compared to simply connecting to Spark. We would expect an embedded engine to do even better.

There are two other use cases worth commenting on. Esgyn has identified operational data stores as a target use case. More interestingly, MapR Drill supports queries against semi-structured data such as JSON, as well as structured data. It has extended its SQL support to allow this. Competitors to MapR for this sort of functionality tend to come from other environments: the InterSystems IRIS data platform, for example, encompasses the same capabilities and extends to unstructured data.

> “While performance may be a major determinant in buying decisions, it is only relevant when comparing apples with apples, and the products covered in this paper constitute an entire fruit bowl.”

Conclusion

SQL on Hadoop is all about horses for courses and, in this paper, we have discussed both the horses and the courses. Table 1 highlights our results. Readers should recognise that you can do OLAP, for example, with any EDW product, but the likes of Kyvos and Jethro will typically provide better performance, hence our recommendations. We have also suggested some SQL but not Hadoop-based vendors that you might like to consider as alternatives to the SQL on Hadoop products, though these are not intended to represent an exhaustive list. Specifically, we have concentrated on scale-out clustered solutions and have omitted products such as IBM Informix or SAP HANA, both of which target HTAP (for example), because they employ architectures that are a long way removed from Hadoop.

Use case	Transactional	HTAP	Complex	OLAP	M/L	Lambda	Mixed data
Recommended	IBM Big SQL Actian VectorH	Esgyn MapR Drill Splice Machine	Kognitio IBM Big SQL Presto	Jethro Data Kyvos	Pivotal IBM Big SQL	Splice Machine MapR Drill	MapR Drill
Others	Esgyn		All EDW	Splice Machine All EDW	Splice Machine Some EDW		
Non-Hadoop		InterSystems		AtScale Druid			InterSystems

Kognitio

Kognitio was founded in 1987 as a data warehousing vendor. At that time, the company was known as WhiteCross Systems. In the early 90s the company introduced what we would today call a database appliance, running on proprietary hardware. Key features are that, from the outset, the database employed a massively parallel architecture and that processing was in-memory. Indeed, given the recent hype about in-memory databases, Kognitio could reasonably claim to be the progenitor of this market.

The history of the Kognitio data warehousing product has been one of gradually moving away from its proprietary roots as the industry has caught up with its requirements. For example, in the mid-90s the company adopted standard industry chipsets, in the mid-2000s it moved to blade computing and away from its own hardware and, most recently (in 2016), the company ported Kognitio onto Hadoop (which involved changing the storage model so that it would work with Apache YARN), though the stand-alone version of the product is still available. And, moreover, the company has made the database free-to-use (with, optionally, paid-for support).

Kognitio targets complex queries against large datasets. For example, one of its clients has 10,000 Tableau dashboards running against a 9 PB database Hadoop cluster (all updating within seven seconds) and over a hundred individual analytic queries (some of which are complex). Needless to say, given the company's longevity, Kognitio has sophisticated optimiser, workload management, high availability, load balancing and so forth. It supports parallel processing for any supported language, such as R or Python, as well as SQL. Support is provided for ORC, JSON and Parquet. There is a query streaming capability that caters for situations where you don't have enough memory

From a performance perspective the company ran a series of benchmarks (validated by Enterprise Management Associates) comparing Kognitio 8.1.50, Impala 2.6.0 and Spark 2.0. In each case the same 12 node cluster was used, running the TPC-DS query set for both a single query stream and ten concurrent query streams. In both cases the data volume was set at 1TB. For the single query stream, Kognitio was the only product to complete all 99 queries and it was fastest on 92 of them (Impala fastest for six queries, Spark for one). For the ten query streams test, Kognitio was "long running" for four queries, Impala failed on more than a quarter of all the queries (mostly because it does not support the appropriate SQL syntax) and Spark did not complete fifteen of the queries. Of the 95 queries completed, Kognitio was fastest on all but eight of these. As an example of comparative performance, for single stream queries that were completed by all vendors (70), Kognitio completed in slightly more than 13 minutes, Impala took longer than 50 minutes and we gave up counting Spark when it had taken longer than Impala for just the first thirteen of the seventy queries.

Kognitio
3a Waterside Park, Cookham Road
Bracknell RG12 1RB, United Kingdom

www.kognitio.com

Strengths

- Because Kognitio has always been an in-memory database, it has been expensive. However, that is no longer the case, as memory prices have come down. Moreover, with the company adopting a free-to-use licensing model, Kognitio should be much better placed than it has been historically.

Threats

- Like many other (but not all) vendors in the enterprise data warehousing (EDW) space, Kognitio has ported its product onto Hadoop. This means that the company will be up against many of the same suppliers that it has historically competed against. Unfortunately, these vendors are all larger and better known than Kognitio and, while this shouldn't be a deciding factor, it often is.

Recommended for

Mixed workload environments that combine transactional look-ups with complex analytics.

About the author

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

This document is copyright **© 2017 Bloor**. No part of this publication may be reproduced by any method whatsoever without the prior consent of Bloor Research.

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.

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

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**