Bloor

Market Review

Market Report Paper by Bloor

Author **Philip Howard**

Publish date **September 2020**

SQL on Hadoop

> **Hadoop is not as popular today as it once was. In part this is because... Hadoop has failed to live up to its own hype... This has left many organisations with a significant investment in Hadoop infrastructure and expertise but without the expected benefits... A relevant use case is therefore to use appropriate technologies... to upgrade your existing Hadoop stack, so that you can actually get the benefits you were looking for in the first place.**

Executive summary

This Market Review represents an update to Bloor Research's 2018 Market Report for SQL on Hadoop Engines (see *https://www.bloorresearch.com/research/sql-engines-hadoop/*). This was based on research conducted during 2017 and that report is now somewhat out-of-date. In that paper we identified half-a-dozen major use cases and several minor ones. However, the market has moved on. Some of those use cases (for example, archival, a minor use case) have migrated to cloud object storage, while others have converged or moved to other platforms. At least one new use case has emerged. Similarly, from a vendor perspective, the biggest differences between providers in 2017/8 were the use cases they targeted and the extent of their SQL support. In many cases the latter was limited: many products, both proprietary and open source, could not run a complete set of benchmark queries or could only do so with limited scalability. This too has changed so that we would now not expect these to be issues.

This review, unlike a Market Update, does not include a comparison of different products. What it does do, after a discussion of use cases, is to distinguish between different approaches to SQL implementation on Hadoop - there are three essentially different methods – discuss how these map to the various uses cases and which vendors have adopted which approach. For this version of our Market Review we have also added a section on Actian Vector for Hadoop.

“This Market Review represents an update to Bloor Research's 2018 Market Report for SQL on Hadoop Engines.”

Use cases

> “...there remain significant use cases for SQL on Hadoop.”

Hadoop does not quite have the same appeal today as it did in 2017/8. The rise of cloud object storage on the one hand, combined with the growth in cloud-based data warehouse offerings (all the traditional vendors along with various pure-play offerings) on the other, has meant that some use cases are at least contested by other approaches. We have also seen the rise of cloud-based ELT (extract, load and transform) tools from the likes of FiveTran and Matillion, along with those of more traditional data integration vendors, eat into the market for Hadoop as a data integration platform. However, regardless of this evolution there remain significant use cases for SQL on Hadoop.

When we looked at the market three years ago, we identified six different major use cases for SQL on Hadoop though, as we noted at the time, some of these overlap with 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, the market has evolved we now believe that there are five major use cases, as follows:

1. **Hybrid transactional analytic processing (HTAP)**, also sometimes known as hybrid operational analytic processing (HOAP), and by other acronyms. These environments combine transactional/ operational processing with real-time analytics. A typical environment would be IoT, where sensor data is being ingested and processed in real-time while supporting predictive maintenance algorithms.

2. **Online analytic processing (OLAP).** May be either multi-dimensional OLAP (MOLAP) or relational OLAP (ROLAP). Some OLAP style use cases may be subsumed into use case 1.

3. **Supporting 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. Such environments may also support OLAP.

4. **As platforms to support the discovery and training of datasets that support machine learning and AI.** Typically, but not necessarily, relevant machine and deep learning algorithms will be deployed on other platforms.

5. **To support data virtualisation**, allowing analytic queries to be run against multiple data sources, with the Hadoop engine used to synthesise query results. This approach reduces or eliminates the need to replicate data within the analytic environment while providing a single point of access to that data, thereby reducing security concerns.

Vendors

Since we published our previous Market Update on this subject there have been some significant changes to the relevant vendor community. Most notably, Cloudera and HortonWorks have merged, and MapR has disappeared into the maw that is HPE. Further, Esgyn (the company providing commercial support for Apache Trafodion) is now focusing almost entirely on the Chinese market and we understand that its US offices have closed or, at least, been de-populated. In addition, we understand that Kognitio is now focusing more on the cloud and containerised versions of its product and, although its SQL on Hadoop version is still available, it is no longer the main focus of the company's marketing. Finally, at just about the time we published our previous Market Update, Starburst was spun out of Teradata as a provider of commercial support for Apache Presto. In the context of the report it is notable that although Starburst is offered as a SQL on Hadoop engine this is not the company's primary target market. Like Kognitio it is focused elsewhere, which we will discuss further in the next section.

Of the remaining vendors we can consider their offerings as belonging to one of three categories:

A. General-purpose offerings that consist of a layer on top of Hadoop. These are typically ports of SQL engines from conventional database and data warehousing environments. For example, IBM BigSQL. This has the advantage of offering the same syntax as the Db2 range of databases but is otherwise a purely generic offering. An exception that is not a port of something pre-existing is offered by Cloudera.

B. Products that have been constructed as a specialised layer on top of Hadoop. That is, products that have been designed for a specific purpose rather than merely offering generic SQL capability. These include AtScale, Kyvos Insights, Splice Machine and Jethro Data.

C. Products that have been specifically engineered into, as opposed to on top of, Hadoop. These actually fall into two sub-categories:

a. Products that have been specifically designed, from the outset, for Hadoop. The most notable of these is Apache Presto and its commercial supporter Starburst. Varada is a new entrant to the market that is also based on Presto. It is arguable that Esgyn and Trafodion also fall into this camp even though their origins go back to Tandem Non-Stop systems.

b. Products ported from other environments, but which have been re-engineered specifically for Hadoop. Offerings in this category include Actian Vector for Hadoop and, at least in theory, Kognitio. Companies in this group, like those in A, often provide SQL commonality across their platforms. Thus, for example, Actian Vector for Hadoop uses the same version of SQL as Actian Avalanche, which is the company's cloud-based data warehousing solution.

There is one further point to consider. We have mentioned that Hadoop is not as popular today as it once was. In part this is because of alternative approaches and, in part, it is because Hadoop has failed to live up to its own hype. And this applies as much to SQL on Hadoop implementations using technologies like Impala and Hive as it does to Hadoop per se. This has left many organisations with a significant investment in Hadoop infrastructure and expertise but without the expected benefits. What we might consider an orthogonal use case (compared to those listed above) is therefore to use appropriate technologies from the various vendor types shown, to upgrade your existing Hadoop stack, so that you can actually get the benefits you were looking for in the first place.

“Since we published our previous Market Update on this subject there have been some significant changes to the relevant vendor community.”

Vendors and use cases

While in theory you can pretty much do anything with any product, different vendors and their products frequently focus on different and specific use cases, and these specialities are illustrated in the following table.

	Generic	Layered	Specially engineered	Re-engineered
HTAP / HOAP	All	Splice Machine*	Esgyn	Actian*
OLAP		All	Esgyn	
EDW	All	Splice Machine	Esgyn	Actian*, Kognito
Machine learning	All	Splice Machine		Actian, Kognito
Data virtualisation			Presto/Starburst and Varada	

Note that while we have shown Splice Machine as supporting HTAP/HOAP, its emphasis is more on leveraging transactional data for predictive analytics than embedding analytics into operational applications. With respect to Actian you should also note that Actian X is another HTAP solution from Actian, likely to be preferred when the emphasis is on transactions rather than analytics.

Conclusion

Some clear recommendations fall out of a consideration of these various technologies when mapped against typical use cases. The first is that if you are considering data virtualisation then Starburst and Presto is where you want to look, though we should add that there are various other non-Hadoop vendors that play in this space, including multiple graph database providers, traditional data warehousing suppliers and pure-play companies such as Denodo.

The second obvious conclusion is that if you only want an OLAP style solution the you would probably be best to look at one of the companies that specialise in this space.

The more interesting question is what to do if you want a more general-purpose solution encompassing any or all of HTAP, data warehousing (including, possibly OLAP) and support for machine learning. We would ignore Esgyn unless you are in China, and probably Kognitio since SQL on Hadoop is no longer a focus. This leaves various generic products, Splice Machine and Actian. There might be reasons for selecting, say IBM BigSQL, if you want SQL compatibility with other IBM databases – and similarly for other comparable vendors – but most likely your choices will come down to Cloudera, Splice Machine and Actian. Of these, Actian's Vector for Hadoop is perhaps the most interesting, thanks to its deep integration into the Hadoop architecture, which we discuss in the next section.

Actian's Vector for Hadoop is perhaps the most interesting, thanks to its deep integration into the Hadoop architecture.

FURTHER INFORMATION
Further information about this subject is available from
www.bloorresearch.com/update/2599

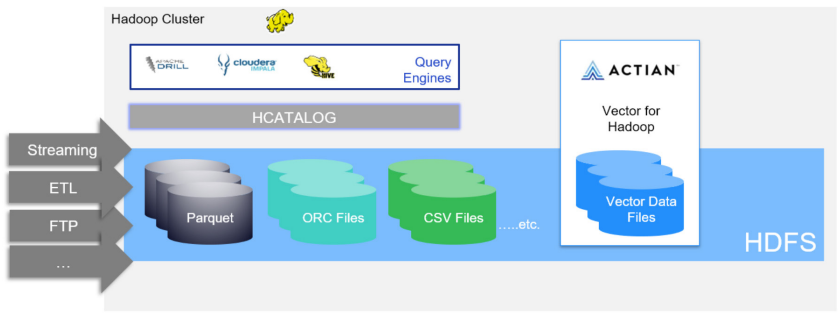
Actian Vector for Hadoop

Actian Vector for Hadoop is what was previously Actian's Vector product ported to the Hadoop platform as a SQL engine. However, Vector uses a symmetric multi-processing (SMP) based architecture that's scales up rather than out. Vector for Hadoop, on the other hand is, by necessity, a massively parallel processing (MPP) solution that uses Hadoop for clustering purposes. So, this represents more than just a port from one environment to another and involved substantial re-engineering. While the principle components of the product are based on the Vector product, it also leverages the query planner and optimiser from what used to be known as Ingres RDBMS but is now called Actian X. While there is access to Parquet and ORC files, which are treated as external tables, Actian has developed a proprietary storage mechanism on top of the Hadoop distributed file system (HDFS). Apache YARN is supported as is HCatalog – see ***Figure 1*** – and available security is both row and role based. Spark is supported natively for scaling on multi-node clusters and in addition to supporting access Hadoop file formats such as Parquet and ORC, it will allow you to perform functions such as SQL joins across different table types. Support is also provided for Spark SQL and Spark R applications.

Vector for Hadoop, like Vector is an in-memory, columnar database, whose development was based on academic research at CWI, the Dutch National Research Institute for Mathematics and Computer Science.

As its name suggests, one of its major differentiators is the vectorised processing it supports – illustrated in ***Figure 2*** – which exploits Intel's vector instruction set (hence the Vector name) to process more data elements per instruction (SIMD: single instruction, multiple data) and optimises for L1 and L2 cache.

Figure 1 – Apache YARN is supported as is HCatalog

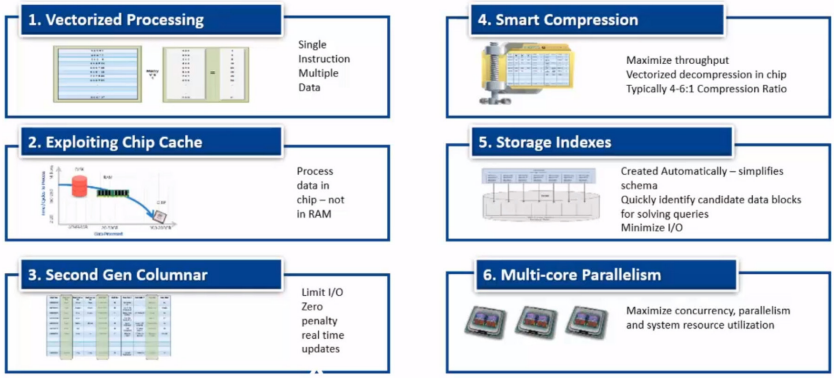


Another major feature is the exploitation of in-memory positional delta trees (*"second gen columnar"* in ***Figure 2***). This patented technology is worth commenting on further with the following being an extract from the original academic research paper on this topic: *"our goal is that read-only queries always see the latest database state yet are not (significantly) slowed down by the update processing. To this end, we propose the Positional Delta Tree (PDT), that is designed to minimize the overhead of on-the-fly merging of differential updates into (index) scans on stale disk-based data."* In other words, this is about improved query performance regardless of whatever updates are being made, while preserving consistency. It is also characteristic of Actian's approach, in that it leverages a number of patented technologies to maximise performance., with other relevant and supported technologies highlighted in ***Figure 2***.

Ultimately, Actian Vector for Hadoop is intended to provide a (very) high performance, enterprise grade analytics platform for companies that want to leverage SQL on top of a low cost, scalable solution, regardless of the potential use case. As mentioned previously, it uses the same version of SQL for its Actian Avalanche Cloud Data Warehouse as a Service (DWaaS) platform, which means that you can easily port Hadoop projects into Avalanche if you need to. Thanks to support for Azure HDInsight as a deployment platform this same portability is available across both on-premises and cloud-based implementations of Vector for Hadoop. Of course, you can also deploy Vector for Hadoop on customer self-hosted cloud environments.

Ultimately, Actian Vector for Hadoop is intended to provide a (very) high performance, enterprise grade analytics platform for companies that want to leverage SQL on top of a low cost, scalable solution, regardless of the potential use case.

Figure 2 – The exploitation of in-memory positional delta trees



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 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**