SQL-For-Hadoop: Critical Technology For Customer-Obsessed Firms

Which Type Of SQL-For-Hadoop Will Help You Turn Customer Obsession Into Reality?

by Brian Hopkins and Mike Gualtieri

August 3, 2015

Why Read This Report

To help your firm become customer-obsessed, you must supply your business partners with access to the data they need to understand customers. Although they're attracted to Hadoop's low cost and flexibility, enterprise architects still approach it cautiously because they fear it's only for firms with big data science teams or Hadoop-literate analytics software code-slingers. But robust technology that enables SQL to access Hadoop has shattered this myth. Forrester examined more than a dozen Hadoop SQL solutions organized by three distinct approaches. Enterprise architecture (EA) professionals should read this report to understand the architectures of these solutions and what to look for as they plan to meet business needs with big data platforms and support emerging systems of insight.

Key Takeaways

SQL-For-Hadoop Gives Customer-Obsessed Firms An Advantage

To become customer-obsessed, EA pros must understand customers' behaviors, needs, and goals by leveraging the big data now at their disposal. SQL-for-Hadoop democratizes data by making it more available for marketing, sales, product, and risk teams, which can use the analytical tools they want and the querying skills they already have.

Three Flavors Of SQL-For-Hadoop Offer Something For Everybody

SQL-for-Hadoop options are plentiful. Forrester categorized more than a dozen solutions into three broad types. Pure solutions are based on open source and offer flexibility. Boosted solutions from database vendors offer performance and SQL compliance. Database-plus solutions add Hadoop extensions to mature analytic databases.

Several Architecture Factors Must Influence Your Future State

Enterprise architects must consider Hadoop distribution support, SQL compliance, metadata locality, cluster requirements, support for data formats and other sources, and how solutions fit into vendor platform strategies.

SQL-For-Hadoop: Critical Technology For Customer-Obsessed Firms

Which Type Of SQL-For-Hadoop Will Help You Turn Customer Obsession Into Reality?

by Brian Hopkins and Mike Gualtieri
with Leslie Owens, Elizabeth Cullen, and Diane Lynch

August 3, 2015

Table Of Contents

Notes & Resources

Forrester interviewed 11 vendor companies, including Actian, Cloudera, Databricks, Hortonworks, HP, IBM, MapR Technologies, Microsoft, Oracle, Pivotal Software, and Teradata.

Related Research Documents

Digital Insights Are The New Currency Of Business

Hadoop Ecosystem Overview, Q4 2014

SQL-For-Hadoop: 14 Capable Solutions Reviewed

FORRESTER®

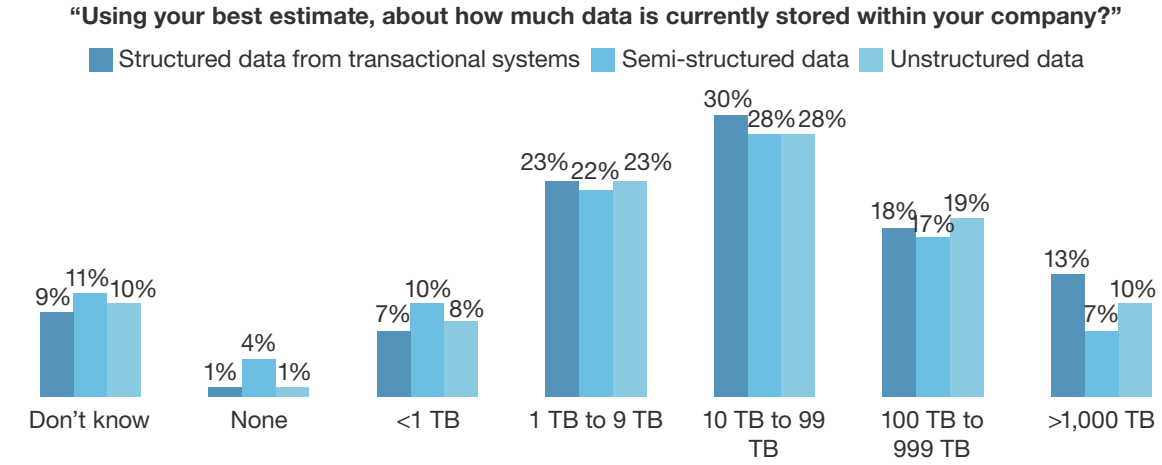
Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
+1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com

SQL-For-Hadoop Democratizes Data For Customer-Obsessed Firms

Hopes are high that big data will help firms make use of the huge amount of data they have collected as they seek insights about customers (see Figure 1). In fact, 74% of data and analytics decision-makers expect their firms to be using big data technology for customer service analytics by Q2 2016.[1]

EA pros who fail to open up all this data and make it democratically available to more people in their business will be at a disadvantage in their efforts to be customer-obsessed. They simply won't be able to find insight in all that data as easily as their competitors will.

FIGURE 1 Firms Are Collecting Huge Amounts Of Data, And They Need Unified Access



Base: 1,805 global data and analytics technology decision-makers

Source: Forrester's Global Business Technographics® Data And Analytics Survey, 2015

SQL-For-Hadoop Solutions Make Data More Available

The idea that Hadoop is only for deeply technical people like developers and data scientists is a myth — in fact, 22% of data and analytics decision-makers say that a lack of technical skills is one of the biggest challenges with big data, and only 19% say the same for a lack of business competency.[2] An SQL-for-Hadoop solution helps democratize data by making it more cost-effective and more available to your business, using the business intelligence (BI) tools and skills you already have — something Forrester refers to as "BI on Hadoop."[3]

Forrester defines SQL-for-Hadoop solutions as:

> *Software that allows users or applications to work with Hadoop data using structured query language (SQL) that is nearly or fully compliant with the American National Standards Institute (ANSI).*

The technology will help your business more easily:

- **Unify customer data economically with large-scale batch processing.** SQL has long been a staple tool for data manipulation in extract, transform, and load (ETL) pipelines, but the cost of performance has been an issue for previous-generation data integration tools. Firms are turning to Hadoop as a low-cost way to do massive-scale data pipeline operations that can create more-detailed customer views or find fraud using more data (see Figure 2). For example, a financial services company used Apache Hive to scour 25 billion transactions per week, looking for new fraud patterns.
- **Prepare massive data sets for predictive analytics to predict customer behavior.** Prepping training data sets for predictive analytics is costly and time-consuming.[4] High-performance SQL-for-Hadoop solutions let firms break through once-insurmountable technical barriers. For example, a fitness retailer uses Actian Vortex to prepare, ingest, and analyze more than 2 billion rows of data from retail stores, fitness tracking tools, and devices to develop cross-sell and upsell opportunity models.
- **Allow analysts to spot customer issues quickly through interactive analytics.** Business analysts searching for insight and data scientists conducting early-stage data exploration have an unquenchable thirst for fast data. SQL-for-Hadoop solutions help satisfy their need. For example, a communication services provider used Apache Drill to enable ad hoc SQL analytics on network data to improve customer experiences, using Tableau and Tibco Spotfire on massive amounts of extremely diverse, high-velocity data formatted in JSON.

FIGURE 2 SQL-For-Hadoop Solutions Help Your Firms Turn Big Data Into Business Insight

58%
will have implemented Hadoop by Q2 2016.

89%
will also have implemented SQL-for-Hadoop by Q2 2016*

Query type	Typical in use cases	Typical data sizes	Typical concurrent queries per cluster	Typical response times
Batch queries Advanced analytics, data exploration, and data pipelines	Unified customer view, predictive model preparation, customer segmentation, and churn analytics	Tens of TB to PBs Billions of transactions per day	One to 20	Minutes to hours
Application data access Applications with SLAs tap data in Hadoop.	Real-time, personalized customer engagement applications and fraud detection	Zero to 10 TB Tens of millions of transactions per day	Low hundreds	Subseconds to seconds
Ad hoc Self-service business intelligence and data discovery	Customer-correlation analysis and customer profitability analysis	One to 100 TB Hundreds of millions of transactions per day	Dozens to hundreds	Seconds to minutes

Base: 1,805 global data and analytics technology decision-makers
*Base: 1,050 global data and analytics technology decision-makers whose firms will have implemented Hadoop by Q2 2016
Source: Forrester's Global Business Technographics® Data And Analytics Survey, 2015

SQL-For-Hadoop Solutions Come In Three Flavors

The odd but beneficial combination of SQL and Hadoop has created an explosion of solutions fueled by open source with Apache projects at the center. Apache Drill, Hive, Impala, Phoenix, Spark SQL, and now Presto (Apache license) are all available via open source. The data warehouse vendors have been busy as well, porting their analytic database technologies to Hadoop or extending their software with external table support for data stored in Hadoop.

To help enterprise architects understand the maze of technical options, Forrester canvassed the market, interviewed vendors, studied their offerings, and referenced customer implementations. We identified more than a dozen solutions with which enterprise architects and application development professionals should be familiar. Our "SQL-For-Hadoop: 14 Capable Solutions Reviewed" Forrester report provides in-depth analysis on the solutions identified.[5]

Enterprise architects must first understand the three general approaches into which these solutions fall: pure, boosted, and database-plus (see Figure 3).[6]

FIGURE 3 There Are Three Types Of SQL-For-Hadoop Solutions

	Pure	Boosted	Database-plus
What is it?	SQL access of data stored in Hadoop with open source solutions that leverage Hadoop and potentially other open source distributed data stores	SQL access of data stored in Hadoop using database technology ported to run in Hadoop, both with and without additional helper node.	SQL access to Hadoop as external table of another database system or data virtualization server
Vendors	Hadoop-centric: • Hive (Hortonworks) • Impala (Cloudera) • Phoenix Hadoop-capable: • Drill (MapR) • Presto (Facebook) • Spark SQL (Databricks)	With Helper Node: • Actian Vortex • IBM Big SQL • Pivotal HAWQ No Helper Node: • HP Vertica For SQL On Hadoop	Data warehouse plus Hadoop: • HP Vertica For SQL On Hadoop • IBM PureData for Analytics • Microsoft PolyBase • Oracle Big Data SQL • Teradata QueryGrid Data virtualization with Hadoop: • All data virtualization vendors
Advantages	Open source innovation; flexibility in format choices	SQL compliance; complex query performance	Security and SQL compliance; less additional software to manage
Hadoop Distribution and Cloud Support	Vendors support different Hadoop distributions. Only MapR comes with Drill, Hive, and Impala. Some solutions require additional software (e.g., Spark or Zookeeper).	Vendors support different Hadoop distributions aligned to their distribution affinities (e.g., IBM and Big Insights, Pivotal and HD Insights). HP and Actian support all distributions. Cloud options vary by vendor.	Vendor software or appliance and vendor-optimized Hadoop.

Pure Open Source Solutions Are Built On Or Extend Hadoop

The mother of all SQL-for-Hadoop solutions is Apache Hive. It was the first Hadoop tool to provide query access to data stored in Hadoop, but many other open source solutions are accelerating to meet the burgeoning demand, including Apache Drill, Impala, and Presto. Pure Hadoop-for-SQL solutions share common characteristics:

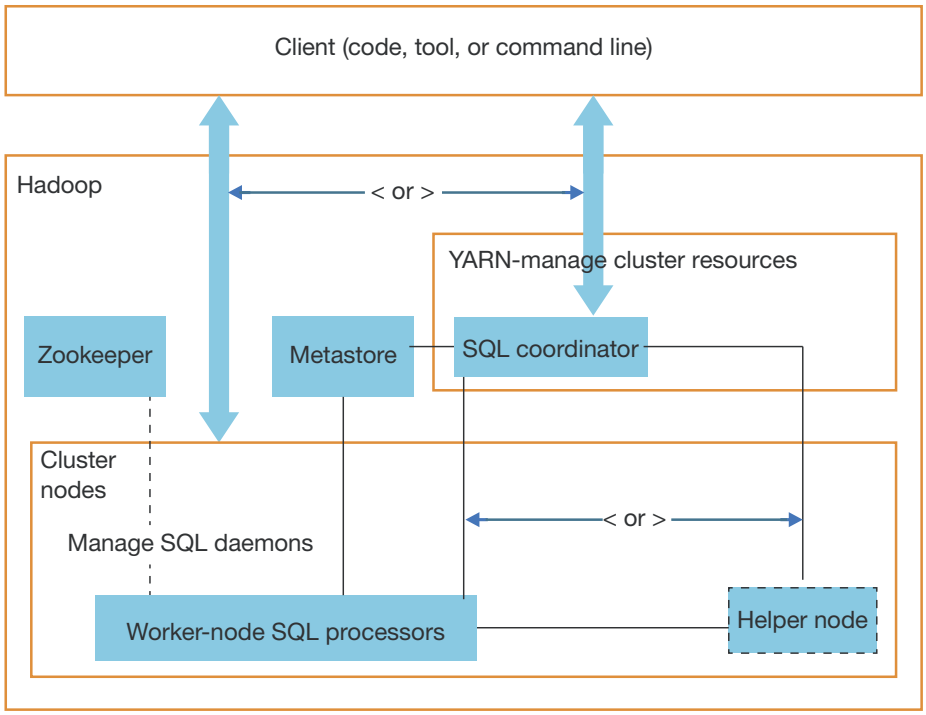
- **They are open source, with very active communities.** All the pure solutions originated as open source, which means they needed and now have active developer communities. For example, in addition to Hortonworks, Apache Hive has active code-committing developers from Cloudera, Facebook, Intel, Microsoft, and Yahoo. Spark SQL is part of Apache Spark, which boasts the largest developer community of any Apache project. Other, newer projects have fewer participants, but most of the solutions are growing rapidly in support.
- **They support a wide variety of data formats.** If you need to query data stored in diverse formats, pure SQL-for-Hadoop solutions offer advantages. For example, Apache Drill was specifically designed to work without first defining a schema, making it ideal for exploring large, diverse data sets in difficult formats like JSON records.[7] Hive is also flexible — it works with many formats with comparable performance. All the pure solutions support Parquet and are strengthening their support for the optimized row columnar (ORC) format.[8] This is important to architects because the special formats required by some boosted solutions require transformation steps, additional software components, and, most importantly, time.
- **They are still gaining maturity in ANSI compliance, performance, and concurrency.** These open source tools were built by developer communities to meet specific needs, not from existing mature SQL processing code bases. Because of this, they are generally less compliant than solutions from vendors that ported their mature databases to run in Hadoop, and they may also have problems maintaining performance while simultaneously running a large number of complex queries. The pure solutions generally recognize the need to improve compliance and concurrency and are making big strides very quickly. Enterprise architects must carefully watch pure solutions before discounting them based on SQL, performance, and concurrency.
- **Many have specific Hadoop distribution and format affinities.** The most popular pure solutions generally have specific distribution affinities based on their origin. For example, all distributions support Hive, but Hortonworks is working hard to optimize its distribution to run especially fast queries against the ORC format. Cloudera's distribution is equally optimized for Parquet with Impala, and MapR Technologies has the best Apache Drill performance. Support for the solution depends on the distribution as well; for example, Hortonworks doesn't support Apache Drill or Impala, and Cloudera doesn't support Drill. Only MapR Technologies supports all three.

Pure SQL-for-Hadoop solutions come in two different subtypes (see Figure 4):

- **Solutions implemented specifically for Hadoop, leveraging Hive metastore and YARN.** Two pure solutions, Hive and Impala, were specifically developed to run against Hadoop, so they exploit the latest that Hadoop has to offer. For example, their latest versions are designed to run in YARN containers for resource management, but they differ in metadata management.[9] They also fully exploit the Hive metastore and HCatalog interface. This is significant for those planning on several SQL-for-Hadoop solutions because others do not so easily leverage Hadoop innovations.
- **Solutions that run in but are not limited to Hadoop.** Apache Spark does not require Hadoop as much of the data is cached directly in memory, but many reference implementations Forrester reviewed use Hadoop distributed file system (HDFS) for disk-based data storage and YARN for shared cluster resource management. Other solutions, such as Apache Drill and Presto, were implemented primarily for Hadoop but have connectors or connection toolkits that can extend them to other data sources easily. Both Apache Drill and Presto will eventually directly support YARN, which will improve their performance for clusters with different workloads.[10]

The most important factor in selecting one of these solutions? Among Apache Drill, Hive, and Impala, it's your choice of Hadoop vendor. If you haven't chosen a vendor, make its SQL-for-Hadoop strategy an important part of your selection criteria. Implement the other solutions on a case-by-case basis to fit specific needs.

FIGURE 4 Pure SQL-For-Hadoop And Boosted SQL-For-Hadoop Architecture



Boosted Solutions Feature Vendor Database Code Ported To Hadoop

Long-time vendors of database and data warehouse technology take Hadoop very seriously. They all recognize the value it brings to enterprises and the synergy they must create with their existing database and data warehouse solutions. Many major vendors are leveraging their powerful, mature analytics database technology to create what Forrester calls "boosted SQL-For-Hadoop," characterized by:

› **Mature SQL engines ported to run on Hadoop without an intermediary database.** For example, Actian Vectorwise, HP Vertica, IBM DB2, and Pivotal Greenplum Database engines have been ported to run in YARN containers. These solutions generally boast better performance than the pure open source solutions, but actual performance will vary depending on data formats, volumes, and query needs. The degree to which you have already implemented a vendor's analytic database is a big factor when considering solutions.[11]

› **A broader big data platform stack.** Each of these solutions is part of a vendor's broader big data platform stack and offers additional advantages when implemented along with the other platform components.[12] For example, Actian DataFlow, part of the Actian Analytics Platform, is used to expedite the preparation and ingestion of data into its optimized data format. Similarly, HP sells HP Haven, and IBM has IBM BigInsights. The extent to which your firm aligns to or has already bought into a vendor's platform strategy is another strong consideration in selecting the right solution.

- **ANSI SQL to run even complicated queries with varying degrees of push-down.** Boosted SQL-for-Hadoop solution vendors all claim a high degree of compliance to an ANSI SQL standard. For example, Actian claims full SQL '92 compliance; Pivotal Software claims compliance up to 2003; and IBM claims full 2011compliance. In addition, Actian, HP, IBM, and Pivotal Software reference the percentage of TPC-DS queries they can run.[13]
- **Special file formats to boost performance.** Vendors of all types attempt to boost performance by employing special optimized formats; boosted solutions are no different. For example, Actian's special format is required to enable its "vectorized query execution" architecture.[14] Similarly, HP and Pivotal Software boast improved performance when their optimized formats are used. This is changing, however. For instance, HP is beefing up ORC performance.
- **Varying support for Hadoop distributions.** The vendors have taken a very different approach to Hadoop support. Actian and HP support all three major distributions, while IBM currently supports just one — IBM BigInsights. Pivotal Software supports Hortonworks and Pivotal HD. Soon, IBM and Pivotal Software will support any distribution certified to the Open Data Platform (ODP) alliance core.[15]

The architecture of these solutions is similar to pure SQL-for-Hadoop, except that:

- **Boosted solutions may or may not use YARN.** Hadoop is not just for SQL. Any number of other heterogeneous workloads, such as MapReduce and other applications, may also be running on the cluster. SQL-for-Hadoop solutions that use YARN can play nicely with other applications by allowing the cluster manager to priority the resources of the cluster. Boosted solutions that don't use YARN may overuse the cluster resources and make it much harder to manage heterogeneous workloads.
- **A helper node in HDFS may be colocated with NameNode.** Most vendors needed a place to port their analytics database code to and chose a helper-node configuration. IBM and Pivotal Software recommend colocating this helper software on the same server as the Hadoop NameNode; Actian recommends a dedicated server in the cluster. Helper nodes often orchestrate the query across the Hadoop cluster and/or act as a query origin point to pre-optimize the query to improve performance.
- **Schema definition strategies vary, but most can leverage Hive.** The Hive metastore is the de facto Hadoop standard for defining schemas. When boosted vendors can read and/or write the Hive metastore, the enterprises that also use a pure solution like Hive or Spark SQL can share schema information. Because the boosted SQL-for-Hadoop vendors are best primarily for high-performance, complex SQL queries, most firms will need at least Hive and possibly other solutions, such as Spark SQL. Accordingly, sharing metadata between solutions is important, and most of the boosted vendors allow you to read from the Hive metastore, if not write to it.

What's most important in choosing a solution? It's your implementation of the rest of the vendor's big data stack and the extent to which you have already implemented the vendor's non-Hadoop analytic database solutions.

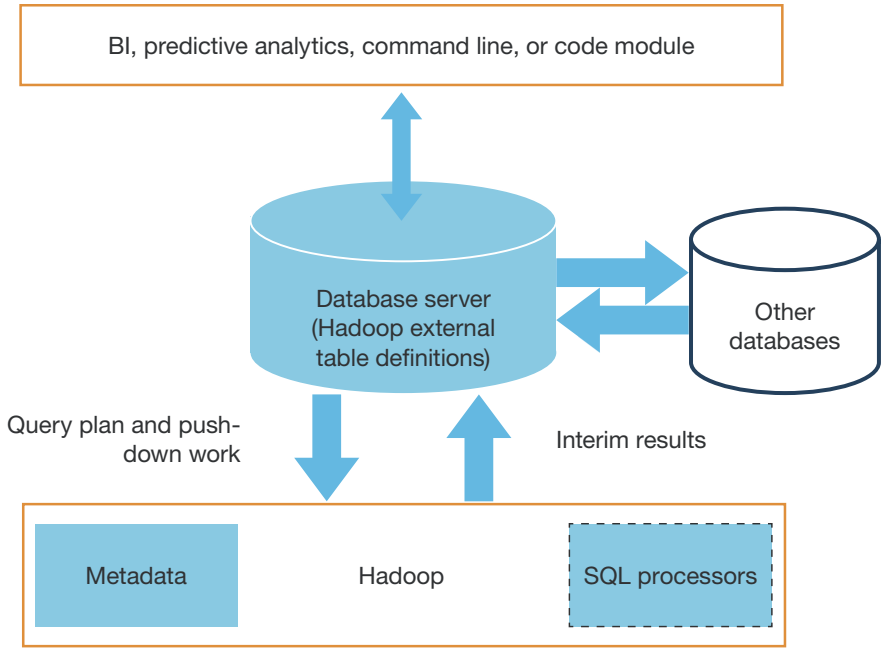
Database-Plus Solutions Add Hadoop To Their Federated Query Capability

Database-plus-Hadoop solutions manage query planning and execution, schema definition, and security in a vendor's database offering (see Figure 5). There are two variants, which are logically similar but very different in implementation — data-warehouse-plus-Hadoop and data virtualization. The data-warehouse-plus-Hadoop vendors include IBM, Microsoft, Oracle, and Teradata. They have added Hadoop as external tables. Data virtualization vendors include Cisco, Denodo Technologies, IBM, Oracle, and SAP; these platforms have slowly added Hadoop as a potential data source, although SQL support varies. When considering either type of solution, think about:

- **The version of the master database or virtualization server required.** Hadoop support comes packaged with the vendor's solution in most cases; it is not an add-on module. But you need to have the right version. For example, Teradata QueryGrid comes with only Teradata Database version 14+. IBM, Microsoft, and Oracle have similar version restrictions. Recognize the opportunity — adding SQL-for-Hadoop capability may provide the justification you need to fund that upgrade project.
- **The Hadoop distributions supported.** Each variant and vendor has a different strategy for Hadoop distribution support. For example, Oracle Big Data SQL is optimized to work with Oracle's Hadoop appliance and certified to work with Cloudera's distribution if the user doesn't have the appliance. On the other hand, Teradata has certified its solution with all three major distributions and joined the ODP initiative. Cisco's virtualization product supports all three major distributions, whereas Denodo Technologies is certified only with Hortonworks.
- **How they scale and how much push-down processing they do.** The cost for performance is a big factor in your architectural planning, as scaling out appliances can be expensive. Alternatively, expanding Hadoop can be cheap. A big factor affecting the cost is the amount of query processing that solutions push down to Hadoop, and capabilities vary. For example, Oracle Big Data SQL and Teradata QueryGrid both boast a lot of push-down, but their strategies are very different. Oracle Big Data pushes its SQL engine code into Hadoop. Teradata leverages Hive for push-down processing and recently announced a partnership commitment for Presto.[16]
- **The SQL support and schema definition.** These solutions prefer to maintain centralized control of SQL compliance and schema definition, but approaches vary. For example, Oracle's strategy is to own the query, so its solution boasts the same SQL compliance and security as Exadata; Oracle and Teradata can access Hive for table definition, but Microsoft currently cannot. Teradata supports all the SQL languages of its connected databases, adding flexibility, whereas Oracle conforms everything to its SQL, ensuring uniform query performance.[17]
- **The other supported data sources.** One of the key features of the database-plus-Hadoop architecture is the variety of data sources that can be federated in a single query. Generally, the virtualization solutions are much better at broad database support than the data warehouse solutions, but the warehouse solutions have better performance for a narrower set of sources.

Vendor strategies vary widely for the warehouse vendors as well. For example, Oracle supports only Oracle products and Cloudera, but both Microsoft and Teradata are adding Oracle to their solutions, along with Hadoop and several NoSQL databases.

FIGURE 5 The Database-Plus-Hadoop Architecture Adds Hadoop To The Supported Data Sources



Recommendations

Include SQL-For-Hadoop In Your Digital Insights Architecture

Enterprise architects will likely need several SQL-for-Hadoop solutions to enhance their big data capabilities as they build out an insights fabric to found their digital insights architecture. Select the right solution to ensure that your architecture can support important insights discovery operations such as:

- **Filling data lakes and processing data streams.** You'll need to work with the data that is being ingested into your data lake or being processed through your streaming architecture. SQL-for-Hadoop offers solutions for both. For example, more firms are implementing Spark SQL as part of their stream data-processing workflow, along with Apache Storm and Kafka.
- **Interactively exploring data.** Customer support, marketing, and fraud insights teams want to exploit new data sources, looking for outliers and unusual correlations, but often the data is unmodeled and of unknown value. SQL-for-Hadoop is an ideal solution to let these teams explore fresh new data without loading into an expensive database.
- **Preparing large predictive analytics-training data sets.** Predictive analytics has become a core technology that supports algorithmic decision-making, but models must be fed with data and trained to deliver valuable insights. SQL-for-Hadoop lets data scientists more easily work with larger, more diverse data sets as they work through the predictive analytics life cycle.
- **Quickly assembling custom customer views.** Your business will likely move too fast to keep up, even with a flexible data lake accessible with SQL-on-Hadoop. Marketers, for example, will constantly want new customer data views combining data from warehouses, their own databases, and Hadoop. Use database-plus solutions, such as those from Cisco, Microsoft, or Teradata, to quickly make all the needed data available in an integrated table for SQL access.

Engage With An Analyst

Gain greater confidence in your decisions by working with Forrester thought leaders to apply our research to your specific business and technology initiatives.

Analyst Inquiry

Ask a question related to our research; a Forrester analyst will help you put it into practice and take the next step. Schedule a 30-minute phone session with the analyst or opt for a response via email.

Learn more about inquiry, including tips for getting the most out of your discussion.

Analyst Advisory

Put research into practice with in-depth analysis of your specific business and technology challenges. Engagements include custom advisory calls, strategy days, workshops, speeches, and webinars.

Learn about interactive advisory sessions and how we can support your initiatives.

Supplemental Material

Survey Methodology

Forrester's Global Business Technographics® Data And Analytics Survey, 2015 is an online survey fielded in January through March 2015 of 3,005 business and technology decision-makers located in Australia, Brazil, Canada, China, France, Germany, India, New Zealand, the UK, and the US from companies with 100 or more employees.

Forrester's Business Technographics provides demand-side insight into the priorities, investments, and customer journeys of business and technology decision-makers and the workforce across the globe. Forrester collects data insights from qualified respondents in 10 countries spanning the Americas, Europe, and Asia. Business Technographics uses only superior data sources and advanced data-cleaning techniques to ensure the highest data quality.

Endnotes

[1] Source: Forrester's Global Business Technographics Data And Analytics Survey, 2015.

[2] We permitted respondents to select up to three challenges. Source: Forrester's Global Business Technographics Data And Analytics Survey, 2015.

[3] Forrester differentiates business intelligence (BI) on Hadoop from SQL-for-Hadoop. BI on Hadoop is the use of BI technology solutions to analyze data in Hadoop. These tools may leverage SQL-for-Hadoop technology as a query mechanism. To learn more about BI on Hadoop, see the "The Forrester Wave™: Enterprise Business Intelligence Platforms, Q1 2015" Forrester report.

[4] For an evaluation of the 13 leading big data predictive analytics solution providers and to see how each vendor performed, see the "The Forrester Wave™: Big Data Predictive Analytics Solutions, Q2 2015" Forrester report.

[5] In this research, Forrester has identified and reviewed open source and commercial SQL engines for Hadoop for AD&D professionals to learn about the maturity and sweet spot for each and choose the best for their enterprises' needs. You may need to choose more than one to satisfy all of your requirements. See the "SQL-For-Hadoop: 14 Capable Solutions Reviewed" Forrester report.

[6] As enterprise architects look at how to deliver a trusted, real-time, integrated, and secure data platform to support applications, they look at data virtualization. In the three years since Forrester's last evaluation, data virtualization vendors have improved their security, scalability, big data, data discovery, data quality, and cloud capabilities. Forrester has evaluated nine vendors — Cisco Systems, Denodo Technologies, IBM, Informatica, Microsoft, Oracle, Red Hat, SAP, and SAS Institute — on 60 criteria for current offering, strategy and market presence. See the "The Forrester Wave™: Enterprise Data Virtualization, Q1 2015" Forrester report.

[7] There are performance implications for schema-on-read solutions because they have to inspect the schema of each record. Apache Drill must do this for JSON but will read the schema only once for predefined relational data.

[8] Cloudera and Hortonworks are busy working to support both ORC and Parquet formats in both Hive and Impala; however, Forrester expects that Hortonworks will still work better with ORC via Tez and that Impala will still perform better with Parquet.

[9] YARN stands for "yet another resource negotiator." It is part of the Hadoop 2.x core distribution. For a more detail explanation of YARN, see the "Hadoop Ecosystem Overview, Q4 2014" Forrester report.

[10] Drill expects to enable YARN in approximately version 1.3. Cask demonstrated a way to run Presto through YARN but had to develop Apache Twill to do it. Source: Alvin Wang, "Running Presto over Apache Twill," Cask, April 3, 2014 (http://blog.cask.co/2014/04/running-presto-over-apache-twill/).

[11] The most common use case is shifting workloads between vendor's supported databases and Hadoop.

[12] Example solutions include, Actian Analytics Platform, IBM BigInsights, and Pivotal Big Data Suite.

[13] The Transaction Processing Performance Council (TPC) publishes the TPC benchmark for decision support systems — Standard Specification, or TPC-DS. It's a decision support benchmark that provides a set of recommended SQL queries that analytics systems (read decision support) must be able to run. It is often used by SQL-for-Hadoop vendors as opposed to ANSI SQL as a way to prove the compliance and usefulness of their solutions for running the general types of queries most firms need for analytics. Source: "Active TPC Benchmarks," TPC (http://www.tpc.org/information/benchmarks.asp).

[14] Vectorized query execution was originally developed as a research project at Centrum Wiskunde & Informatica, the national research institute for mathematics and computer science in the Netherlands. Source: Peter Boncz, Marcin Zukowski, and Niels Nes, "MonetDB/X100: Hyper-Pipelining Query Execution," CWI, 2005 (http://oai.cwi.nl/oai/asset/16497/16497B.pdf).

[15] Hadoop is a must-have for large enterprises, but adopting it can be a challenge. The newly founded Open Data Platform (ODP) initiative promises to enable "big data solutions to flourish atop a common core platform." For analysis on the goals of the ODP initiative, the challenges and opportunities it will bring to the market, and the consequences that need to be considered when planning and executing on a Hadoop strategy, see the "Brief: Can Hadoop's Enterprise Loose Ends Be Tied By The Open Data Platform Initiative?" Forrester report.

[16] For example, Teradata pushes query execution down to Hadoop by leveraging Hive, whereas Oracle employs the proprietary Smart Scan technology that comes with its Big Data Appliance to optimize where the processing happens. While these are both technically push-down, they are very different approaches with different performance characteristics. The difference in vendor philosophy is evident in these choices. Oracle wants to own the query and can guarantee a high degree of SQL compliance, performance, and security by limiting clients to its special optimized Hadoop. Teradata wants to unify everything and so lives with handing off a lot of query processing to Hive.

[17] SQL support will make a big difference to your users, so plan carefully, with a deep understanding of your needs. For example, the Teradata solution puts the onus of conforming the SQL statement to the lowest common denominator across technologies on the user; the Oracle solution handles support and translation in the software but supports few data sources.

FORRESTER®
CHALLENGE THINKING. LEAD CHANGE.

We work with business and technology leaders to develop customer-obsessed strategies that drive growth.

PRODUCTS AND SERVICES

- Core research and tools
- Data and analytics
- Peer collaboration
- Analyst engagement
- Consulting
- Events

Forrester's research and insights are tailored to your role and critical business initiatives.

ROLES WE SERVE

Marketing & Strategy Professionals

CMO
B2B Marketing
B2C Marketing
Customer Experience
Customer Insights
eBusiness & Channel Strategy

Technology Management Professionals

CIO
Application Development & Delivery
- Enterprise Architecture

Infrastructure & Operations
Security & Risk
Sourcing & Vendor Management

Technology Industry Professionals

Analyst Relations

CLIENT SUPPORT

For information on hard-copy or electronic reprints, please contact Client Support at +1 866-367-7378, +1 617-613-5730, or clientsupport@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.