2017 Trends to Watch: Big Data

Machine learning, IoT, and streaming grab the spotlight

Publication Date: 21 Nov 2016 | Product code: IT0014-003164

Tony Baer

Summary

Catalyst

There were no individual breakout headlines in 2016 to compare with the sudden introduction of Spark in 2015, which overshadowed much of last year. But our 2016 predictions have been borne out. A rising tide has lifted Hadoop (although not yet to profitability for vendors); those who reported revenues cited run rates averaging 40–45% higher year over year. But there are also new, alternative paths to big data, such as Spark-based services in the cloud, that reinforce the idea that big data is not automatically synonymous with Hadoop. We have also seen machine learning proliferate, from consumer services to enterprise applications and tooling; for instance, machine learning has become table stakes for data preparation and other tools related to managing curation of data for data lakes. And we have seen a significant uptick in client queries on implementing data lakes.

Ovum view

The breakout use case for big data will be fast data. The Internet of Things (IoT) is increasing the urgency for enterprises to embrace real-time streaming analytics, as use cases from mobile devices and sensors become compelling to a wide range of industry sectors. Machine learning, which has garnered its share of hype, will continue to grow; but in most cases, machine learning will be embedded in applications and services rather than custom-developed because few organizations outside the Global 2000 (or digital online businesses) will have data scientists on their staff. But for those that do, getting those data scientists connected will become a top priority through the use of new collaboration tools. Meanwhile, the availability of cloud-based Spark and related machine learning and IoT services will provide alternatives for enterprises considering Hadoop. Finally, for enterprises that have gotten to the stage of planning data lakes, security and data preparation will be the first technologies they employ to make their data lakes governed and transparent.

Key messages

- Machine learning will be the biggest disruptor for big data analytics in 2017.
- Making data science a team sport becomes a top priority.
- IoT pushes real-time streaming analytics to the front burner.
- The cloud sharpens Hadoop-Spark "co-opetition."
- Security and data preparation will drive data lake governance.

Recommendations

Recommendations for enterprises

Big data was once the shiny new thing; machine learning (a form of artificial intelligence) has taken its place. But do not necessarily feel pressured to recruit high-priced (although now, more abundant) data scientists; chances are that the customer-360 application, predictive maintenance solution, or intrusion detection system that your organization is considering implementing will include machine learning. If your organization plans to recruit, or already has data scientists on staff, your challenge is to ensure that their work will not get bottled up on their laptop; collaboration will become the order of the day in 2017. While machine learning has drawn the buzz, in 2017, the more immediate uptake will be for real-time streaming solutions. We have tracked the (re)emergence of fast data, which has arrived owing to the confluence of technology and the urgency of use cases involving IoT and mobile. There is open competition for a new generation of streaming engines. The market and state of technology are immature – the market has not yet winnowed down to any "winners." Couch your bets by either implementing standalone applications, or incorporating a buffer (e.g., a PubSub engine) that will decouple the streaming engine from the underlying application. And finally, if your organization has decided to build a data lake, be sure to pay attention to data preparation and security – governance starts with knowing what data is in the lake and how it should be secured.

Recommendations for vendors

Across the big data vendor ecosystem, the overriding challenges are to simplify and integrate the pieces. For Hadoop vendors, the cloud is a "frenemy": it is the place to host managed services that can put much of the complexity in a black box, but it is also the place where incumbent cloud providers (e.g., Amazon) have home court advantage with their own managed services. Spark is also your "frenemy" because it runs on Hadoop, but can also run standalone in a cluster that can run almost "bare metal" (with just a JVM, with a lighter-weight cluster manager like Mesos) or packaged by the cloud provider. For vendors in the machine-learning space, 2017 will be the year that the big players – especially IBM – threaten the emerging long tail with unified approaches. Your best defense and offense is offering tools that are quick and inexpensive to implement, and provide hooks that touch business users; integration with BI tools can reduce friction by eliminating the need for your most difficult target audience to learn new interfaces. As for streaming players, 2017 will be a free-for-all – we expect that SQL support will become table stakes, but will only be the first step toward wider battles for securing ecosystem support.

Business trends and technology enablers

The hot growth area of information management

Here is why we cover big data: Earliest on the maturity curve compared to established technologies in information management, big data, according to Ovum research, posts growth rates approaching 50%, compared to 5–10% for more mature areas such as business intelligence, data management, enterprise content management, and enterprise performance management. It is little surprise that the number of big data queries from Ovum clients continues to grow year over year.

Ovum conservatively estimates that the big data market will total $1.7bn in 2016, and will grow to more than $9bn by 2020, comprising 10% of the overall market for information management tooling. While the volume of sales is and will continue to be North America-dominated, the rest of world will also be posting marginally higher growth rates.

The big elephant and the cloud

Because Hadoop has proven difficult to implement, a year ago, we forecast that the next wave of adopters would increasingly embrace cloud and appliance adoption because they would not have the same technology practitioner depth and expertise as the pioneering adopters.

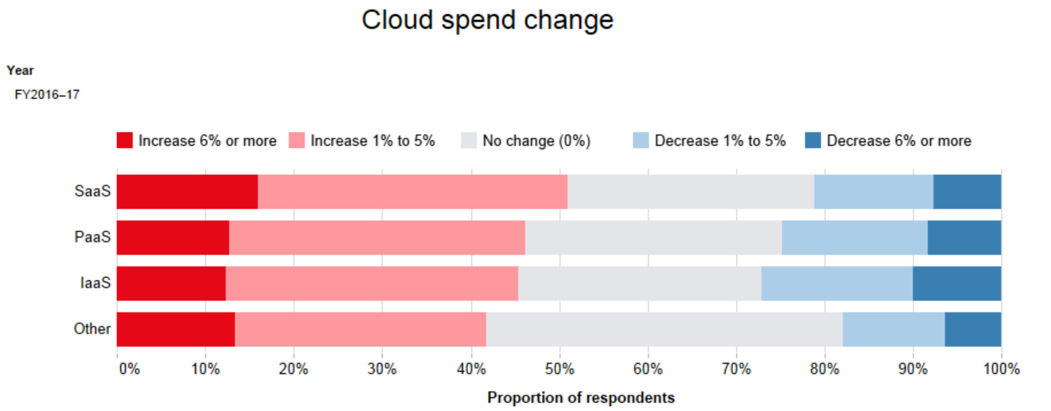
To put this in perspective, we will consider cloud adoption at a broad level first. According to Ovum's latest global enterprise insights research (published in October 2016), roughly 40–45% of respondents plan to grow their cloud spending in 2017 over previous-year levels, as shown in Figure 1; significantly, that is a 5–10% jump over our previous year's numbers. Admittedly, the Ovum numbers are not a measure of overall cloud adoption or penetration, but they provide a key indicator as to whether cloud spending has hit an inflection point where it is becoming the norm for most enterprises.

This appears consistent with what we are seeing with adoption of Hadoop in the cloud, based on our observations of

- Amazon EMR customers, who comprise about 15% of the market;
- Cloudera, MapR, and Hortonworks, who typically report about 15–25% of their customers (mostly early adopters) deploying whole or in part to the cloud;
- SaaS applications (e.g., Workday Analytics) that embed Hadoop, comprising another 5%; and
- Point services (e.g., Spark, machine learning) from all major cloud providers, and from AI platforms like IBM's Watson, accounting for another 5%.

We believe that in the next couple years big data cloud deployment will hit the 50% threshold – encompassing Hadoop and non-Hadoop cloud-based services, such as for machine learning and Spark.

Figure 1: Cloud spend change, 2016–17



Copyright: Ovum

Sample size: 1,538
Question: How do you expect your spending to increase or decrease from 2015 to 2016 and 2016 to 2017 on the above cloud areas?
Vertical: All. Sub-vertical: All. Country: All. Enterprise size: All.

Source: Ovum

Where there's smoke, there's fire

Lots of shiny new things

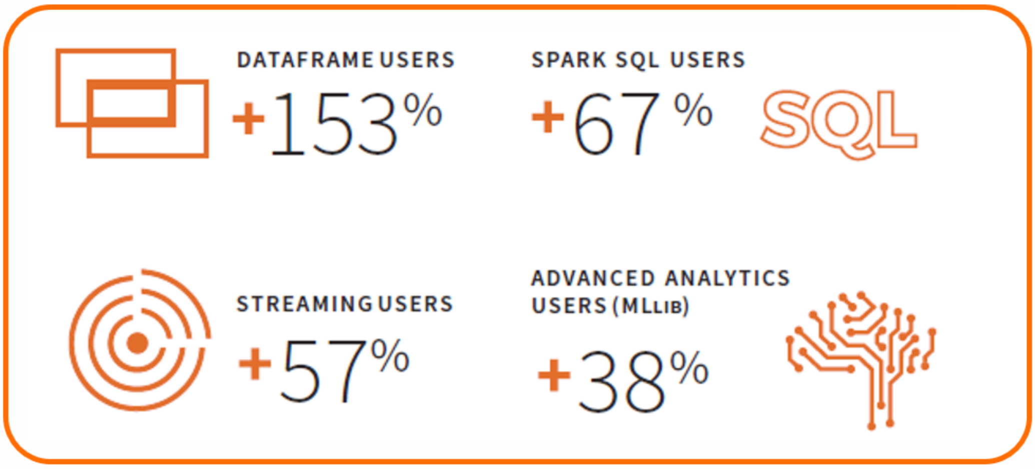
This report covers lots of domains, practices, and technologies that are considered by different stakeholders as shiny new things.

- To upper management, it is all about big data that could enable them to deliver new products and services, engage more intimately with their customers, or operate with superior risk mitigation and security. There is a sense that market leaders are getting the first crack at solidifying unfair advantage.

- To data miners and deep analytics programmers, the role of data scientist – and the open source languages (e.g., R, Python, Scala) – have that allure. So do the domains of artificial intelligence and machine learning.
- To data engineers, the people who used to perform ETL whose role has broadened to munging data sets and deploying analytics programs, it is working with new tools for wrangling data and writing (or readying) the analytic programs that will actually run.
- To business end users, it is the new generation of self-service data discovery and exploration tools that allow them to ask questions that were not feasible within the confines of traditional BI tools and the data contained in data marts.

The biggest shiny new thing is machine learning and artificial intelligence, which we cover later in this report. But keep in mind that more mundane workloads (SQL) and highly compelling real-time workloads from the IoT have made Spark SQL and Spark Streaming more popular processes among Spark programmers (themselves, a pretty advanced cohort), as shown in Figure 2. Nonetheless, we still view machine learning as the chief disruptor for big data in 2017.

Figure 2: Apache Spark's fastest growing areas in 2016



Source: Databricks

Machine learning is the big disruptor

It is already there, but you might not realize it

Under the covers, machine learning is already becoming ubiquitous. For consumers, machine learning is already omnipresent in their everyday online lives, from shopping at Amazon to searching Google and watching entertainment from sites like Netflix. And it is also coming true in business. A survey by Narrative Science, a natural language-based analytics provider, revealed that many enterprises are taking advantage of machine learning and artificial intelligence capabilities, but do not necessarily know it. Only 38% reported using those capabilities to automate routine tasks, yet 88% said they used analytic tools that incorporated predictive analytics, automated written reporting and communications, and voice recognition and response. There is no shortage of vendor machine learning/AI activity.

Focusing on data management, last year we successfully forecast that machine learning would become table stakes for data curation and data-wrangling tools. The incorporation of machine learning has been widespread. While the pioneers of commercialized machine learning came from the start-up community, every major household IT player has invested, or is currently investing, in tools and solutions involving machine learning. Among them:

- IBM is focusing on its Watson cognitive computing business, branding it as "the AI platform for business."
- Amazon, Microsoft, Google, HPE, and IBM (along with Databricks) offer cloud-based machine-learning PaaS services.
- Oracle and SAP are acquiring, or have already acquired, start-up companies offering machine learning for marketing-oriented enterprise solutions.
- Machine learning has become table stakes for the new breed of data preparation tooling used for managing data inventory in data lakes – from start-ups to established players like IBM, Informatica, and Oracle.

Analytic applications embedding machine learning are becoming the norm

There is a growing list of vertical applications that are embedding machine learning. Ovum's 2015 report *Machine Learning in Business Use Cases* listed several dozen start-ups that are embedding deep learning for applications that are heavily oriented toward vision and speech recognition. In 2016, the technology industry newsletter *Venture Beat* compiled an even broader survey showing a wide range of use cases from marketing to customer support; industrial IoT; sales optimization; adtech; investment finance; security, fraud, and threat detection; education; transport logistics; legal; healthcare; and others. Commercial examples include security, fraud detection, and risk management (Graphistry, Bitsight); precision agriculture (Blue River); adtech (Metamarkets); student lending (Earnest); IoT (PTC Coldlight); customer intelligence (Clarabridge, Preact, Sentient Technology, Vidora); Nauto (autonomous vehicles); precision medicine (Deep Genomics, IBM Watson Health); and data preparation and enrichment for data lakes (Alation, IBM Dataconnect, Informatica Rev, Oracle Big Data Preparation, Paxata, Tamr, Trifacta). BI tools (SAP BusinessObjects) are also already starting to embed machine learning/AI to provide "guided experiences."

Surprise: Demand for data scientists is actually soft

Believe it or not (we initially did not), there is evidence of a softening demand for data scientists; data from Indeed.com shows flat demand over the past four years (see Figure 3). One of the coauthors of the now-famous 2012 *Harvard Business Review* article that birthed the meme of the data scientist being the "sexiest job of the 21st century" now states that there is no shortage of sexy candidates. In a recent *Wall Street Journal* post published in the CIO Journal section, Thomas Davenport, distinguished business professor at Babson College, stated that college and university data science programs are now rising to the task of filling the pipeline of data science candidates.

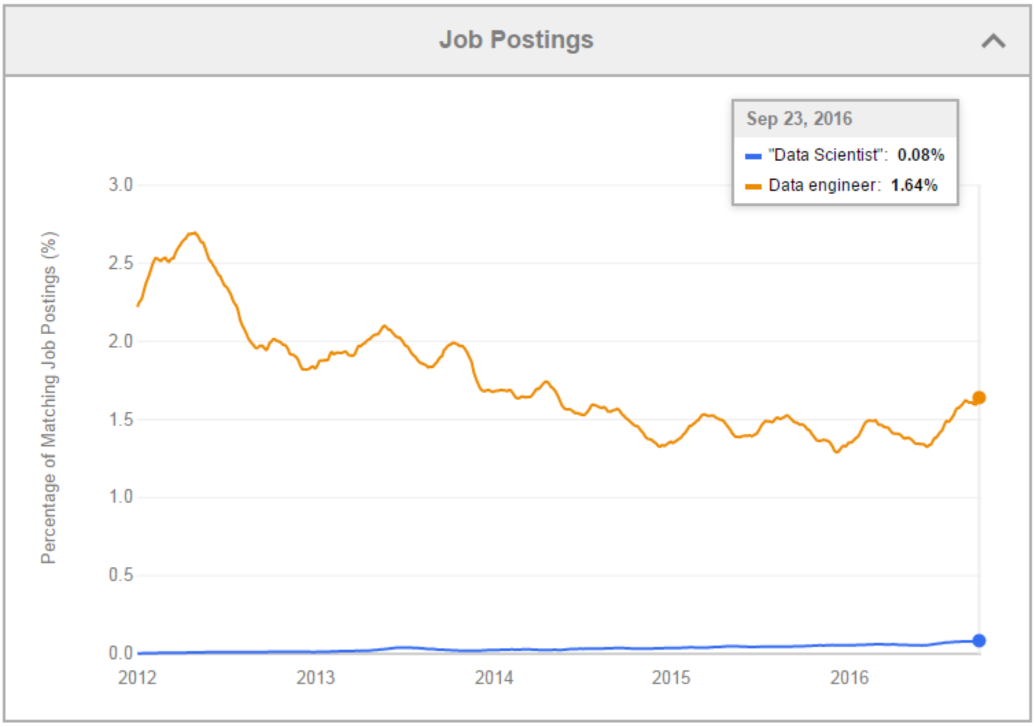
There is no question that there are more graduates coming into the market with formal or informal data science credentials, and most are likely getting hired. Who is recruiting these prospects? In all likelihood, excluding online digital businesses, relatively few enterprises outside the Global 2000 are absorbing them, and few would have any idea of how to use data scientists. Outside the Global 2000, the organizations that are retaining data scientists are those where data science initiatives are coming from specific business units. Otherwise, for the mass of organizations that rely on packaged analytics, the need is not for data scientists per se, but applications or tools that apply data science under the hood.

Making data science a team sport becomes top priority

A disconnect is impeding the work of data scientists

Making data scientists productive is just half the equation. They form and test hypotheses, but they typically do not select data sets, provision clusters, and optimize their algorithms for production. That is the job of the data engineer. (As we just cited, data engineers are actually in higher demand.) The real need is getting data scientists and data engineers better connected to ensure that the models that the data scientist has written and tested on his or her laptop gets deployed properly with the right data sets on the cluster (which is the data engineer's expertise).

Figure 3: Demand for data scientists versus data engineers



Source: Indeed.com

Tools rapidly emerging to connect data scientists with the business

Visit any big data conference, and you will see a growing spectrum of providers offering tooling that automates some aspect of the lifecycle of developing, deploying, and running machine-learning algorithms. The common thread running through these tools is the goal of freeing the data scientist or data engineer from having to reinvent the wheel by writing algorithms from scratch. Selected examples include:

- IBM's Watson Data Platform, which will (when all pieces get released) encompass workspaces for data scientists, data engineers, business analysts, and application developers, represents one of the most ambitious attempts to integrate data science into the business.
- Automation tools like DataRobot, which automatically run your data against multiple algorithms to determine the best fit for the problem.
- Alteryx, which provides a hybrid experience of a self-service data visualization and analytics tool, with a back-end development environment (R is supported) for developing standard machine-learning analytics programs.
- Dataiku, which resembles an integrated analytics tool that includes connectors to external data sources, visual data preparation and data transformation, a choice of roughly 30 pre-packaged algorithms that the user chooses, and model versioning.
- Domino Data Labs, which manages the lifecycle and deployment of machine-learning projects.

Providers like Alteryx and Dataiku reveal what we expect to be the dominant motif for machine learning in 2017: it must be a team sport that supports collaboration between data scientists and business analysts. The overlying trend will be toward collaboration environments where business analysts and data scientists can share workflows in the planning, deployment, and execution of machine-learning models. In part, this is attributable to the realization that enterprises will not gain the full value of machine learning if the models remain inside the heads of data scientists, and that aiming for a market that includes business analysts not only expands the addressable audience, but targets the audience that controls the budgets.

IoT pushes real-time streaming analytics to the front burner

Hallelujah! Streaming analytics are reborn

Analyzing data in motion is nothing new – event-processing programs have been around for nearly 20 years. In the Ovum report, *Fast Data 2015–16: The Rebirth of Streaming Analytics,* we outlined the perfect storm that has transformed real-time streaming from a niche technology to one with broad, cross-industry appeal. These include: open source technology, which lowered barriers to entry for both technology providers and customers; scalable commodity infrastructure, which made the processing of large torrents of real-time data in motion economically and technically feasible; and the explosion in bandwidth and smart sensor technology.

The reason for all this activity is the demand created by emerging IoT use cases; this is where real-time sense, analyze, and respond has spurred technology vendors to pick up where niche CEP (Complex Event Processing) left off. According to Cisco, by 2019, two-thirds of all IP traffic will originate from non-PC devices. And with the vast majority of devices sitting outside data centers, we expect that most streaming applications will be deployed in the cloud.

Today, there is a growing array of choices that are competing for new workloads. Eventually, we expect that the market will winnow down to three to four streaming platforms. But the platforms are works in progress; for instance, Spark Streaming and Amazon Kinesis Analytics only recently introduced support for SQL time-windowing queries, while Heron (a redo of technology based on Storm) was just open sourced. Given the early state of the market, we do *not* expect the market to winnow down in 2017; we expect it will take 24–36 months for streaming engines to mature and IoT implementations to attain critical mass. An updated view of the streaming market landscape is shown in Figure 4.

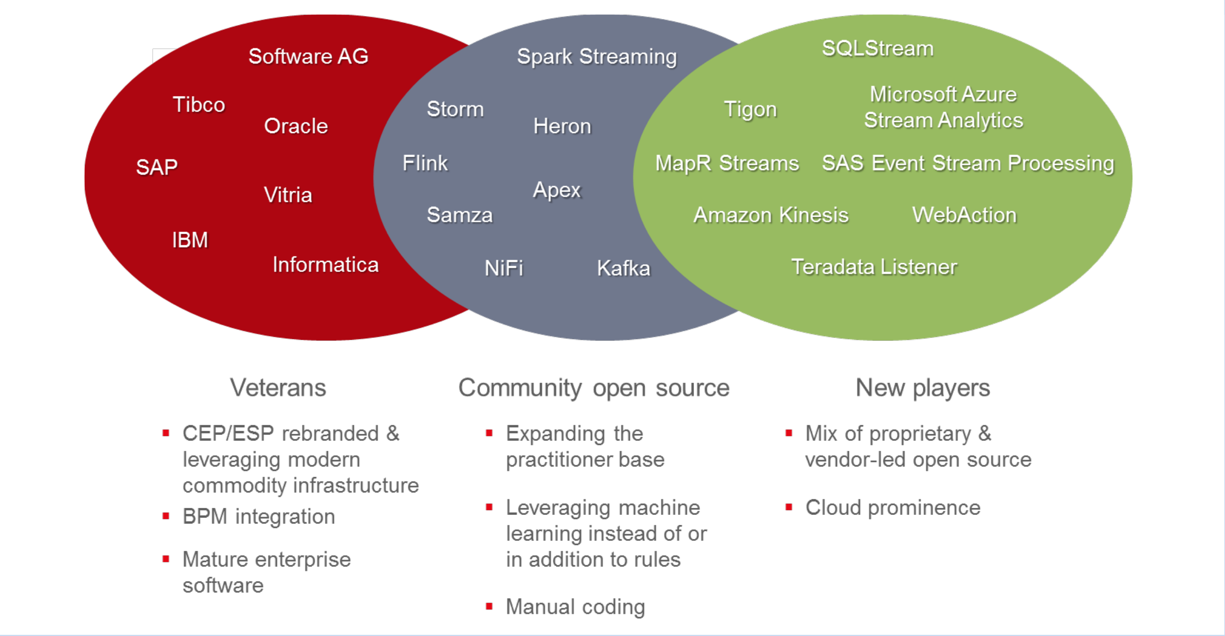
SQL is the hotspot. Sound familiar?

We expect SQL support will become a checklist item for streaming analytics in 2017. As with SQL on Hadoop, the driver is obvious – go where the developers and the BI/analytic tools already are. Extending SQL query to streaming makes good on the vision for real-time analytics. This is not to disrespect Java or Scala, which can be more efficient for executing some forms of complex queries that would be cumbersome in SQL.

The success of SQLStream, which is integrated with Teradata, and most recently, OEM'ed by Amazon as the heart of the last piece of its Kinesis streaming technology platform, is a good indicator of demand for SQL support. SQLStream is hardly alone; consider the following evidence:

- Apache Storm has a SQL capability that can query data surfaced through the Trident microbatching mechanism.
- With Spark 2.0, Spark Streaming is being reengineered, and as part of that reengineering, is adding support of "Structured Streaming" that extends Spark SQL to it.
- The Apache Flink community has begun an effort to add a SQL interface through a new Table API.

Figure 4: Streaming and message-queuing engines



Veterans

- CEP/ESP rebranded & leveraging modern commodity infrastructure
- BPM integration
- Mature enterprise software

Community open source

- Expanding the practitioner base
- Leveraging machine learning instead of or in addition to rules
- Manual coding

New players

- Mix of proprietary & vendor-led open source
- Cloud prominence

Source: Ovum

The (re)emergence of PubSub

PubSub is a well-established model that, in essence, buffers recipients from senders. While direct feeds from sender to target are used when the requirement is for strict real-time processing, PubSub provides more flexible, scalable alternatives that buffer recipients from senders, enabling wider distribution of data or content. Traditionally, this required complex middleware such as Enterprise Service Busses (ESBs) and message brokers that were notoriously difficult to configure. But new, lighter-weight alternatives that support distributed scale-out architectures are reviving interest in PubSub; they can work with streaming engines, or without them. All they need is a source, such as a change-data capture log of a database.

The best known, Kafka, has drawn wide support thanks to an efficient, highly scalable architecture that offers a much smaller footprint than predecessors such as RabbitMQ or ActiveMQ. Hortonworks, MapR, and Teradata are also backing alternative engines for managing data flow and distribution.

As streaming applications grow more widespread and requirements get more complex, a larger proportion of them will require a PubSub or similar data flow engine to provide the flexibility to broadcast or distribute content. You are going to hear a lot more about these engines in 2017; many organizations will take their first steps to streaming using them.

The cloud sharpens Hadoop-Spark co-opetition

Yin and yang, or just yang

The fact that Spark can run with or without Hadoop has prompted debate on whether Spark would replace Hadoop. Forget about the fact that comparing the two is apples and oranges: Spark is only a compute engine, while Hadoop stores data and runs many compute engines, including Spark. Spark can execute standalone on a bare metal-like cluster with only the JVM, or under a lighter-weight cluster manager like Mesos. The debate rages because, if you eliminate the overhead of a general-purpose data-processing and storage engine (and in Hadoop's case, YARN), Spark should run far more efficiently. The drawback, however, is that standalone Spark clusters lack the security or data governance features of Hadoop.

The cloud sparks Spark/Hadoop realignment

Ovum believes that, for on-premise implementations, creating standalone Spark clusters will be more trouble than it is worth, unless:

- The organization already has extrinsic security and data governance tools that can form an umbrella over the dedicated Spark cluster; or
- The organization does not care about securing or governing data sent to Spark clusters for execution.

The cloud, however, is a different story. Spark services in the cloud (along with similar services delivering machine learning and IoT processing that in many cases also use Spark under the covers) from providers like Databricks, AWS, or Azure run standalone, drawing data typically from cloud object storage. Cloud providers can provide the cluster management, storage, and security features missing from Spark that would be cumbersome for enterprises to attempt on their own.

We do not believe that Spark will kill off Hadoop in 2017. But the cloud will force a realignment where Hadoop gets positioned as the low-cost storage and general-purpose compute engine for the data lake (not excluding Spark processing), while cloud-based services become the primary home for dedicated Spark computing.

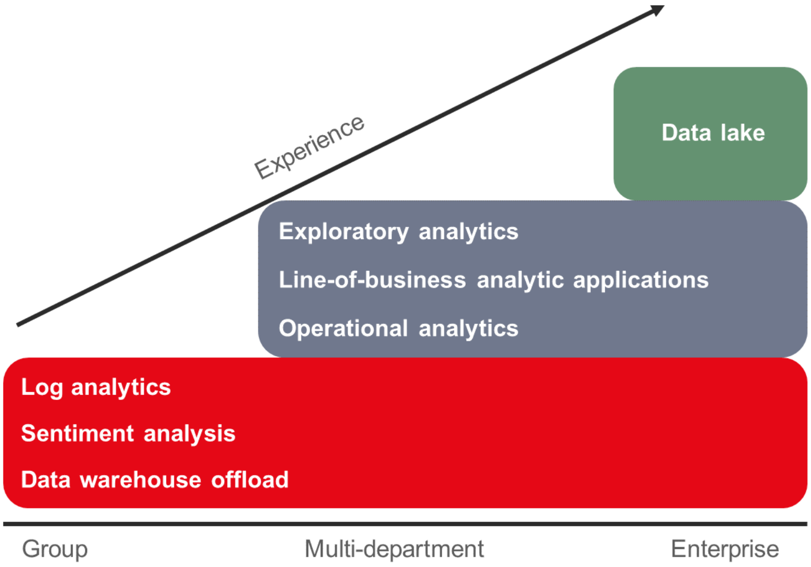
Competition from dedicated services will in turn increase pressure for market consolidation in Hadoop. The privately held players, Cloudera and MapR, have gone on record with their expectations of becoming cash flow neutral in 2017; if their momentum holds, we expect at least one (if not both) to IPO next year. On the other hand, for publicly held Hortonworks, 2017 will be a make-or-break year where the company must either stanch the losses or put itself up for acquisition.

Security, data prep to drive data lake governance

Data lakes entered the agenda last year

As of a year ago, when the Hadoop installed base numbered roughly 2,000 companies, we forecast that early majority implementers would proceed to the stage of data lake adoption over the course of 2016 (see Figure 5). While we do not have scientific data, in our conversations with industry sources and in the rate of client queries that we have fielded, we have found a distinct uptick of interest and questions about data lake adoption. The question of course is, once you start building a data lake, how do you keep it from degenerating to anarchy.

Figure 5: Data lake is a later stage of Hadoop adoption



Source: Ovum

Governance depends on the target audience, use case

We are seeing organizations beginning at different starting points. For instance:

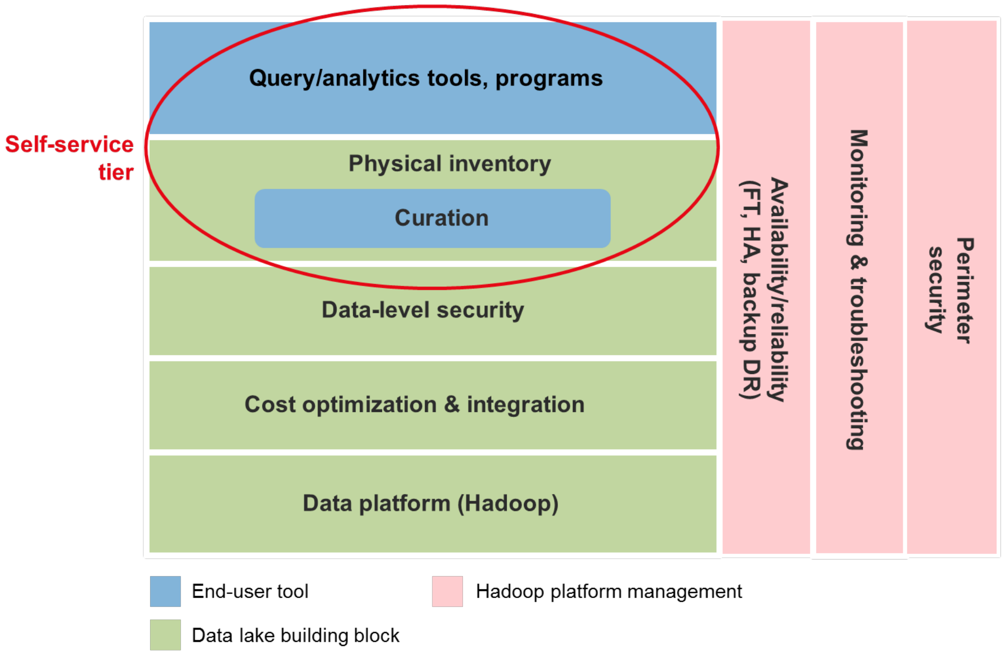
- A federal agency is building a diverse data lake containing grant data and patents from an Oracle database, plus journal articles in XML format. As the user base is only about a dozen named users within a specific organization, governance is primitive, relying on the X.500

protocol for authentication. If other agencies piggyback on the data lake infrastructure, the organization will consider logically segmenting data to support role-based access controls.

- A financial institution storing consumer credit records unintentionally found itself building a data lake. It used Hadoop as an inexpensive near-line repository for analytic data from multiple sources. As the Hadoop cluster grew, with more data ingested, the user base grew (approximately 200 people now access the system). The organization recruited a power user to act as "product owner" of the data lake, and began delineating data as transient (to be purged), persistent, and archival for offloading to tape. Its current plans are to build a catalog.
- An insurance B2B underwriting hub works with data that is mostly owned by business clients. It is codifying data preparation practices to make them repeatable.

Data lake governance includes a wide range of tasks and disciplines, as shown in the Ovum reference architecture (see Figure 6). The common points of pain for data lake adopters are related to the inventorying and securing of data. Data preparation is a logical first step for organizations that are seeking to eliminate reliance on standalone Excel spreadsheets. As this capability has become widely available in offerings, ranging from data integration providers to functionality that is part of analytic and data science tools, we expect significant uptake in 2017 – as organizations seek to replace manual Excel processes. We also expect more focus this year on securing access; down the road, as organizations accumulate data assets, they will seek to automate cataloging and ingest functions.

Figure 6: Data lake governance reference architecture


Self-service tier
Query/analytics tools, programs
Physical inventory
Curation
Data-level security
Cost optimization & integration
Data platform (Hadoop)
Availability/reliability (FT, HA, backup DR)
Monitoring & troubleshooting
Perimeter security
End-user tool
Hadoop platform management
Data lake building block

Source: Ovum

Appendix

Methodology

This report was compiled based on extensive consultation with enterprises, technology vendors, and consultants.

Further reading

Developing a Strategy for Data Lake Governance, IT0014-003113 (May 2016)

Fast Data 2015–16: The Rebirth of Streaming Analytics, IT0014-003064 (October 2015)

Machine Learning in Business Use Cases, IT0022-000335 (April 2015)

"IBM's data science IDE – not your father's analytic tool," IT0014-003128 (June 2016)

"Amazon streaming analytics hits high gear," IT0014-003151 (August 2016)

"Is the Google Cloud Platform ready for prime time?" IT0014-003111 (April 2016)

"Amazon launches machine learning service for developers on AWS," IT0022-000361 (May 2015)

Author

Tony Baer, Principal Analyst, Information Management

tony.baer@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

The contents of this product are protected by international copyright laws, database rights and other intellectual property rights. The owner of these rights is Informa Telecoms and Media Limited, our affiliates or other third party licensors. All product and company names and logos contained within or appearing on this product are the trademarks, service marks or trading names of their respective owners, including Informa Telecoms and Media Limited. This product may not be copied, reproduced, distributed or transmitted in any form or by any means without the prior permission of Informa Telecoms and Media Limited.

Whilst reasonable efforts have been made to ensure that the information and content of this product was correct as at the date of first publication, neither Informa Telecoms and Media Limited nor any person engaged or employed by Informa Telecoms and Media Limited accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts and figures as no liability can be accepted in this regard – readers assume full responsibility and risk accordingly for their use of such information and content.

Any views and/or opinions expressed in this product by individual authors or contributors are their personal views and/or opinions and do not necessarily reflect the views and/or opinions of Informa Telecoms and Media Limited.

Ovum

TMT intelligence | informa

CONTACT US

www.ovum.com

analystsupport@ovum.com

INTERNATIONAL OFFICES

Beijing

Dubai

Hong Kong

Hyderabad

Johannesburg

London

Melbourne

New York

San Francisco

Sao Paulo

Tokyo