June 30, 2016

REPORT: BIG IDEA

Democratize Big Data: How to Bring Order and Accessibility to Data Lakes

Success Depends on Adequate Governance, Cataloging and Hands-On Access to Data in Hadoop

Doug Henschen

Vice President and Principal Analyst

Content Editor: R "Ray" Wang

Copy Editor: Maria Shao

Layout Editor: Aubrey Coggins

TABLE OF CONTENTS

EXECUTIVE SUMMARY 3

BROADER ACCESS DRIVES INSIGHTS AND ACTIONS 4

HADOOP EMERGES AS A CORPORATE STANDARD FOR BIG DATA MANAGEMENT AND INSIGHT 5

DATA LAKE SUCCESS DEPENDS ON A MATURE APPROACH 7

SEEK EASE OF USE, REPEATABILITY AND AUTOMATION 8

LOOK TO NEXT-GENERATION VENDORS TO FILL DATA LAKE GAPS 12

CONSIDER INCUMBENTS FOR BROADER NEEDS 16

RECOMMENDATION: TARGET THE CENTER OF ANALYTICAL GRAVITY 19

RECOMMENDATION: CONSIDER CONSULTANTS AND SYSTEMS INTEGRATORS 21

TAKEAWAY: PREVENT DATA SWAMPS AND MAKE BIG DATA ACCESSIBLE 23

ANALYST BIO 24

ABOUT CONSTELLATION RESEARCH 25

EXECUTIVE SUMMARY

This report explores key trends in the popularization of Hadoop and strategies being employed to manage data lakes and make them more accessible. Companies have embraced the concept of the data lake or data hub that spans analytical and data-driven application needs. But gaps remain in the maturity and capability of the Hadoop stack, leaving organizations to struggle with how to ingest, cleanse, transform, discover, enrich, blend, catalog and govern data in data lakes.

If the data lake concept is to succeed, Constellation Research believes organizations need three key capabilities:

1. Data management and governance
2. Data cataloging and metadata management
3. Self-service discovery and data preparation

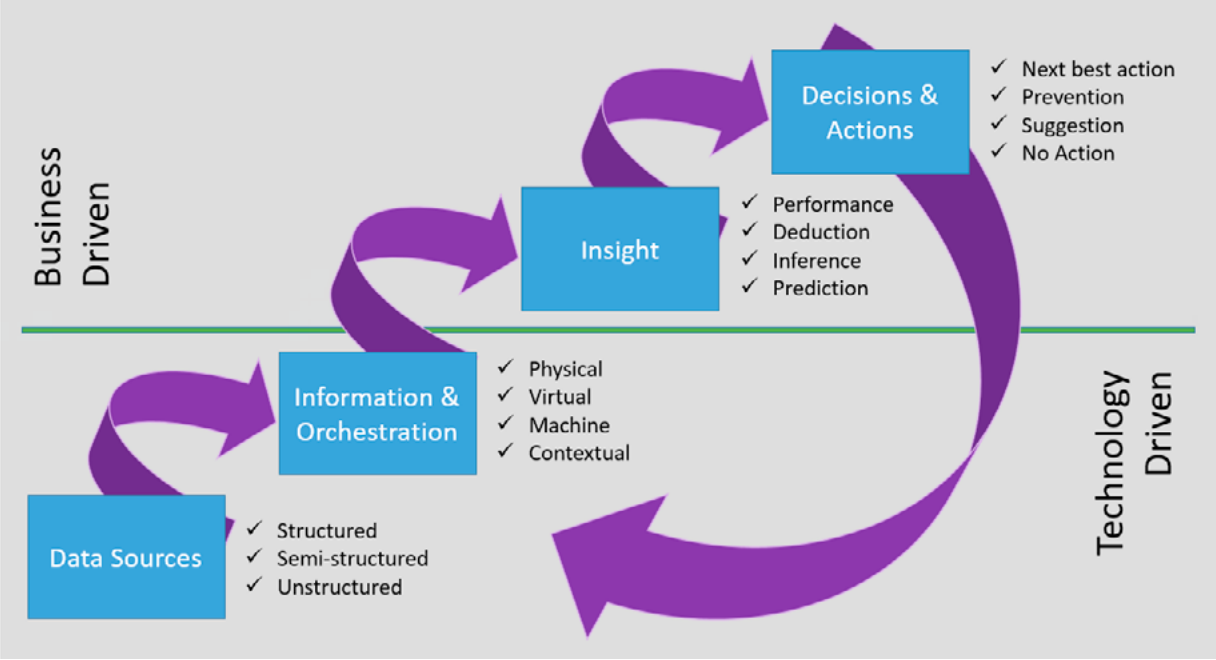
This report examines these three capabilities and the mix of tools available from Hadoop distributions and from next-generation and incumbent data management vendors. In particular, it looks at categories of tools and suites that not only abstract the complexities of Hadoop but also open up access to enterprise IT and business professionals. These users expect unified, enterprise-grade interfaces, self-service data discovery and prep capabilities, and prebuilt integrations to popular business intelligence and analytics tools.

BROADER ACCESS DRIVES INSIGHTS AND ACTIONS

Innovators and industry leaders enable data-driven decisions across their organizations. However, companies must do preparatory work before data-driven decisions can be achieved.

First, organizations must ingest and organize a vast variety of data originating from numerous sources, including and increasingly from external sources. This data often arrives in different formats, which is fine for a modern platform such as Hadoop, but context and metadata must still be maintained, ingested, enriched and applied to support the orchestration of information and alignment with business processes. Over time, insight and more metadata are gained as queries and algorithms are applied to the information to identify correlations, trends and patterns. These insights then guide decisions and actions (see Figure 1). When the orchestration and insight steps are closed off to small groups of experts, the entire process takes longer. The

Figure 1. Better Orchestration and Business User Access Drive Insights and Actions in Constellation's Data-to-Decision Framework.



Source: Constellation Research

broader and more collaborative the access, the more quickly insights can be discovered and decisions can be driven by the data.

Technology Optimization and Innovation is about investing in innovation and strategic advantage while lessening the cost of providing ongoing support. New economic realities necessitate that organizations become smarter in adopting new technologies that can deliver business value while reducing the cost of delivering IT services. Hadoop is viewed by some as a lower-cost alternative to relational databases and storage for managing information at scale. That's one benefit, but a key role of the data lake is to support exploration and correlation of varied data types to uncover new insights that drive innovation and competitive advantage.

HADOOP EMERGES AS A CORPORATE STANDARD FOR BIG DATA MANAGEMENT AND INSIGHT

The sheer scale and variety of data now generated and used in business have created demand for new platforms for managing information. Relational databases are not going away, but they are too costly and too rigid to handle data in all its varieties and at today's massive scale. Apache Hadoop has emerged as the leading platform for Big Data analysis, thanks to its ability to handle high volumes and diverse velocities and varieties of data in a cost-effective way. NoSQL databases (a separate topic not addressed in this report) have emerged as the leading alternative to relational databases for running high-scale transactional applications.

More than 3,000 organizations are now paying customers of the three big Hadoop software distributors: Cloudera, Hortonworks and MapR. Tens of thousands more are using either Hadoop cloud services, such as Amazon Elastic

MapReduce and Microsoft Azure HDInsight, or are experimenting with open source distributions on-premises or in the cloud.

Hadoop is now routinely showing up in the enterprise, and it's a given at big banks, insurers, retailers, telecommunications companies and other large, data-intensive organizations. Nonetheless, Hadoop is far from mature. Giant internet firms such as Yahoo and Facebook have led the way on Hadoop innovation, but with their deep engineering and data science teams, these firms are tolerant of low-level management tools and coding-intensive data-prep and data-analysis methods. More mainstream organizations often lack the skills and resources to spin-up and maintain code-intensive systems.

Within these more mainstream enterprises, early Hadoop use cases have often been limited to cost savings tied to data warehouse and storage optimization. The data warehouses aren't being eliminated, but organizations are moving older data out of these comparatively expensive platforms built on relational databases and are putting that data into Hadoop for archival storage and, in more sophisticated cases, analytics. These strategies require a disciplined approach to metadata management for retrieval or operational use.

In other cases, firms are reducing the amount of new data that is loaded into the warehouse. They're using Hadoop as a lower-cost alternative to extract, transform and load (ETL) data processing. The raw data at scale stays in Hadoop for exploratory analytics while structured data is refined from the raw data and moved into the warehouse or served up through integrations or through APIs as data services. Metadata and data lineage are required to maintain connectivity to Hadoop and the provenance of the data.

As Hadoop deployments mature, the number of use cases and data-driven applications starts to expand. This expansion goes hand-in-hand with growth in the number and variety of data sets in Hadoop and it results in rising complexity in data-management needs. It is at this critical point when a project that may have started as Hadoop cluster experimentation must evolve into a data lake. Veterans have

been here before; it's similar to the difference between a data warehouse and the underlying database management system.

DATA LAKE SUCCESS DEPENDS ON A MATURE APPROACH

The rough idea of the data lake is to serve as the first destination for data in all its forms, including structured transactional records and unstructured and semi-structured data types such as log files, clickstreams, email, images, social streams and text documents. Some label unstructured and semi-structured as "new" data types, but most have been around a long time. We just couldn't afford to retain or analyze this information—until now.

The data lake is not synonymous with Hadoop. A lake should be part of the total analytic ecosystem. Thus, it must be integrated with and support incumbent data-management infrastructure, including data warehouses, data marts and the applications that feed and use the data lake.

Data lakes can handle all forms of data, including structured data, but they are not a replacement for an enterprise data warehouse that supports predictable production queries and reports against well-structured data. The value in the data lake is in exploring and blending data and using the power of data at scale to find correlations, model behaviors, predict outcomes, make recommendations, and trigger smarter decisions. Breakthrough ad hoc discoveries can also be turned into automated data pipelines that can feed data-driven applications and trigger smart actions.

The key challenge is that a Hadoop deployment does not magically turn into a data lake. As the number of use cases and data diversity increases over time, a data lake can turn into a swamp if you fail to plan and implement a well-ordered data architecture.

If Hadoop-based data lakes are to succeed, you'll need to ingest and retain raw data in a landing zone with enough metadata tagging to know what it is, when it originated or changed and where it's from. You'll want zones for refined data that has been cleansed and

normalized for broad use. You'll want zones for application-specific data that you develop by aggregating, transforming and enriching data from multiple sources. And you'll want zones for data experimentation. Finally, for governance reasons, you'll need to be able to track audit trails, versions and data lineage as required by regulations that apply to your industry and organization.

To support such an architecture, organizations need a robust and mature set of capabilities for data ingestion, transformation, profiling, cataloging, governance, exploration, discovery and preparation. The Apache Hadoop community is working on all of the above as well as more fundamental challenges, including Hadoop systems management and security-and-access controls. But many of the tools and projects at the core of Hadoop are coding-intensive and unfamiliar to enterprise IT and data-management professionals. What's more, Hadoop-native tools may only address Hadoop, whereas data pipelines often span multiple platforms, including NoSQL databases and relational sources and targets as well as Hadoop.

Fortunately, we're seeing a rich ecosystem emerging around Hadoop to address data lakes and larger analytical ecosystems. Options include both open source and commercial software designed to integrate with and complement Hadoop. Some tools also work with other Apache Spark, NoSQL and NewSQL databases and conventional relational databases. This report addresses emerging data management, governance, profiling, cataloging, discovery and self-service prep options that are democratizing the data lake (see Figure 2). In the pages that follow, Constellation Research will examine each category, associated vendors, what to look for, and strategic buying considerations.

SEEK EASE OF USE, REPEATABILITY AND AUTOMATION

Data lakes can turn into data swamps as the volume and variety of data retained and managed grows. That's why it's so important to have a well-planned data architecture. You can't just dump everything into the lake and

Figure 2. Three Ways to Better Manage and Democratize Data Lakes.

Source: Constellation Research

trust that you'll be able to find or effectively use the information.

A well-planned data architecture will include the ability to ingest and retain raw data with enough metadata to illuminate the substance and provenance of the data. You'll want to establish a zone for refined data that has been cleansed and normalized to serve as standard data (a single source of truth on customers, for example) that can be used in many applications. And you'll want to establish a zone for application-specific data that might have been transformed, aggregated and enriched for specific use cases. For governance reasons, you may need to show audit trails and track data lineage as required by regulations that apply to specific industries.

There are multiple projects within the Apache Hadoop framework that address fundamental challenges, including data access, security controls, data profiling, cataloging and data lineage tracking. Unfortunately, some of these projects are fractured along distributor lines, a fracturing that contributes to gaps and differences in data lake management.

Cloudera, for example, touts Apache Sentry for unified role-based security access control and RecordService for the centralized enforcement of these permissions across

the Hadoop platform. Cloudera Manager provides perimeter security, with Active Directory and Kerberos integration. Cloudera Navigator provides data governance and data management capabilities that offer better visibility and validation into data available, including audit, column-level lineage, discovery, metadata management, and policy enforcement. Cloudera also provides encryption across all data and key management as part of its platform.

Hortonworks touts Ranger, Knox and Atlas as its mix of security and data-governance software. Ranger addresses centralized, policy-based authorization, authentication, audit and security controls across the Hadoop stack, while Knox provides perimeter security. Hortonworks introduced Apache Atlas in 2015 as a metadata service for accessing and exchanging data inside and outside of Hadoop. Atlas addresses data-lineage tracking across Squoop, Hive, Falcon, Storm and Kafka as well as metadata management and classification by asset type and business language, giving organizations a better understanding of data inside the data lake. Classification-based security enables data stewards to apply tags such as PII or PCI to data assets such as columns, tables or databases.

MapR follows yet another path on data security and governance and delivers these capabilities at the platform level. The company replaces the Hadoop Distributed File System (HDFS) with a commercial POSIX file system with read/write capabilities as well as other data organization constructs such as "volumes" to govern data. Volumes facilitate the best practice of organizing data in different zones and MapR touts its approach as offering higher scalability and easier management than Hadoop distributions that rely on HDFS. The POSIX file system also supports Linux pluggable authentication and flexible access control modules known as Access Control Expressions (ACEs).

The basics of data-access control and security are addressed by all Hadoop distributions, but at this writing, available governance tools are not complete. For example, governance has to account for metadata for data linkage, master data referencing, and data quality.

Hadoop vendors rely on third-party vendors for these capabilities, providing APIs to their Hadoop-native tools. What's more, the components typically used for data processing and transformation, namely MapReduce and the Apache Spark Core, are comparatively low-level and coding-intensive, requiring specialized expertise.

To address the critical gaps in capabilities required to manage the data lake, commercial vendors are extending and complementing the tools and components native to Hadoop. Two camps of vendors have emerged -- next-generation firms and incumbent data-management firms -- that are attempting to democratize the data lake by making it easier for enterprise data management teams to ingest, cleanse, combine, transform, catalog and govern data in data lakes. Many are also attempting to make data accessible to a broader community of users, including data analysts and data-savvy business users by adding their own versions of the style of self-service data exploration, discovery and prep tools pioneered by next-generation vendors.

The next-generation data management vendors, which have all emerged within the last five years, are mostly focused on supporting Hadoop and helping to manage data lakes. Incumbent data management vendors have served SQL-centric, relational needs for a decade or more, but most have either added technologies or reengineered their technologies to work with Hadoop, NoSQL databases and a wider variety of high-scale data. Both camps promise to simplify and democratize the data lake through three crucial traits:

- **Ease of use:** The next-generation and incumbent data management vendors both promise a simplified and unified data-management experience that opens up the data lake to more professionals by abstracting them from the complexities of coding and myriad, low-level interfaces. This reduces the need for specialized skills and, in some cases, introduces self-service capabilities for business users. Many of these vendors expose data catalogs and self-service data-prep capabilities directly to

data analysts and data-savvy business users to broaden access to the data lake.

- **Repeatability:** Both camps help professionals build data pipelines for ingesting data, capturing metadata, applying data-quality rules and executing parsing, filtering and transformation steps in a repeatable way. This ensures consistency and eliminates repetitive, labor-intensive tasks. Self-service data-prep tools also contribute to repeatability, enabling data analysts and business users to find new insights and combinations of data that can be adapted to production data pipelines.

- **Automation:** Beyond creating repeatable workflows with job-scheduling features, next-generation and incumbent data management vendors often automate tasks such as data ingestion, integration, quality filtering, data transformation, data profiling and data cataloging. Some tools also apply machine learning, natural language processing and semantic technologies to automatically discover, categorize, prepare and recommend data sets.

A lake should be part of the total analytic ecosystem. Thus, it must be integrated with and support incumbent data-management infrastructure, including data warehouses and the applications that feed and use the data lake.

The Technologies

- Data ingestion
- Data transformation
- Workflow
- Data cataloging
- Metadata management
- Data lineage
- Data sampling and profiling
- Self-service data prep

LOOK TO NEXT-GENERATION VENDORS TO FILL DATA LAKE GAPS

As the data lake concept has taken root, organizations lacking Hadoop skills and experience have longed for easier-to-use and more unified tools for data lake management. Next-generation vendors were the first to identify the gaps in open-source distributions and many have created tools, like data catalogs and self-service prep tools, that have since become must-have data lake capabilities. Some of these vendors and tool sets are Hadoop-centric while others also address data-management needs across relational platforms.

- **Data management and governance:** Data ingestion, transformation, profiling, cataloging, quality and metadata management are all core capabilities required for data lake management. The most comprehensive offerings from next-generation vendors are from Podium Data and Zaloni. Both vendors also supplement Hadoop-native security and data-lineage capabilities and both added self-service data-prep modules last year.

- **Data cataloging and metadata management:** Data catalogs and metadata enable users to see what data sets are available, what they contain, where they're from, how they're used and more. These capabilities are provided by Podium Data and Zaloni, but the focused, best-of-breed players in this space are Alation, Collibra and Waterline Data. These vendors continuously scan available data, capture and enhance metadata and support settings, alerts and notifications on data policies. The technology also supplements Hadoop-native data-lineage and security capabilities.

Alation is particularly adept at cataloging structured data and can scan relational sources such as Amazon Redshift, IBM Netezza, MySQL and Teradata. Alation has a SQL-centric understanding of keys, indexes and query usage patterns. It's also certified with Cloudera and Hortonworks, with Hadoop implementations being primarily Hive focused, with support for both Impala and Tez.

Collibra is aimed at data governance and stewardship, providing a business semantics glossary and collaborative tools for managing data structures, hierarchies and mapping.

Waterline is a Hadoop-centric vendor that uses machine learning and semantic analysis technology to automatically profile and add metadata to structured data and to unstructured and semi-structured data sets including clickstreams, log files, JSON data, Avro and Parquet formats.

- **Self-service discovery and data preparation:** This is a popular route to democratization of data because it's about letting data analysts and data-savvy business users profile, explore and statistically sample data, typically with the help of machine learning, natural language processing, and semantic analysis technologies. The next step is self-service cleansing, joining, enrichment, quality, mastering and transformation of data without help from IT. Prep work is usually supported by collaborative features and repeatable workflows. These tools also provide data access to personal and corporate business intelligence, data visualization and analytical tools.

  Pioneers in this category include Datameer, Paxata, Tamr and Trifacta, all of which have evolved their technologies to work with Apache Spark. Paxata and Trifacta are close competitors while Datameer and Tamr differ in that they also offer analytical capabilities. Datameer offers general-purpose Big Data visualization and analysis tools while Tamr offers focused customer and clinical data integration, procurement and media analytics. Podium Data and Zaloni joined the self-service push last year with their releases of Podium Prepare and Zaloni Mica.

Constellation Research Analysis

No two organizations are exactly alike, so priorities for one firm may be less important to another. These differences can relate to the nature of the business and variety of data in use. Firms differ in their talent pools and ability to attract new expertise. They also

differ in sophistication and current technology investments. Finally, companies differ in their ability and appetite to spread data-driven decision-making to business users. Keeping these variables in mind, Constellation Research believes that next-generation vendors should be considered in the following contexts:

- **Breadth of capabilities:** In the best-of-breed versus suite wars, Constellation Research believes that suites usually win…eventually. But when new categories like data cataloging and self-service data prep are still emerging, the innovators tend to have differentiating features and functions while the suite vendors are still playing catch up. Another consideration is that organizations don't always need the breadth of capabilities offered by a suite. Those considering Podium and Zaloni, for example, might also consider the ingestion, transformation, cleansing and workflow capabilities of an incumbent vendor that they may already use (see below). What's more, having a suite does not mean you have to use all of its components. A company might want to add a best-of-breed data cataloging or self-service data prep tool if the suite vendor is far behind.

- **Ease of use:** "Easy to use" is a relative term, with some products being easy to use and intuitive to data scientists and others being straightforward to data-savvy business users. Before looking at products, first consider what classes of users and how many users will need to use the tools. This applies to both administrative and management tools and data-access, exploration and prep tools. Complexity and lack of familiarity kill adoption and create bottlenecks. Make sure would-be users try (and perhaps even train) before you buy.

- **Enterprise-grade capabilities:** Consider the maturity of products to be deployed in complex, shared services environments with strict service level agreements and governance and compliance requirements. Next-generation tools must support collaborative data workflows and governance activities while also meeting scalability, multi-tenancy, and automated data-delivery capabilities.

- **SQL-centric versus Big-Data-centric needs:** Even Hadoop vendors like Cloudera, which offers a high-scale data mart alternative with its Impala SQL engine for Hadoop, acknowledge that a data lake is no replacement for an enterprise data warehouse (EDW). On the other hand, an EDW is not well suited to flexible, schema-on-read analysis in which users can explore and find value in variable and semi-structured data such as clickstreams, log files, sensor data, mobile app data, social data and so on.

A key question is whether your data lake will drive predominantly SQL-centric or exploratory, Big-Data-centric analysis. Next-generation tools are usually more adept at handling big, messy data sets while incumbent data management tools typically shine in SQL-centric settings. Realistically, many enterprises lack and are finding it hard to hire people with experience with Big Data analysis techniques, so they're falling back on familiar, SQL-centric analyses. Nonetheless, Constellation Research believes companies should set their sights on supporting both styles of analysis. Thus, the next-generation vendors should be considered and incumbent vendors assessed in terms of how far they've come in developing next-generation capabilities.

CONSIDER INCUMBENTS FOR BROADER NEEDS

Well-established data management vendors, including IBM, Informatica, Oracle, Pentaho (now a Hitachi Group company), SnapLogic, Syncsort and Talend, have long focused on simplifying, streamlining and automating data management tasks. All are evolving their suites to work with modern Big Data platforms including Hadoop, Apache Spark and NoSQL databases. The question is how far they've progressed in their evolutionary journey and how they might fit with your data lake goals and larger analytics strategy.

Incumbents are worthy of consideration, particularly if you're using one of these suites and your people are already familiar with its pre-built connectors to source systems, visual

data-flow orchestration interfaces, out-of-the-box data-parsing and data-transformation routines, job scheduling modules and services-based data-delivery capabilities. All of the above save data-management professionals time and deliver on the promise of ease of use, repeatability and automation. Some of these vendors have truly comprehensive suites incorporating data quality and master data management capabilities, and these, too, are being evolved to address Big Data and data lake needs.

Most of the vendors mentioned above emerged in the relational database era while a few, including IBM and Syncsort, got started in the mainframe era. Several have also added their answers to the data cataloging and self-service data prep modules popularized by next-generation vendors.

Constellation Research Analysis

The incumbents, by definition, have the advantage of incumbency. If you're already invested with a particular vendor and plan to stick with it for your SQL-centric needs, it can't hurt you to consider that vendor's Hadoop and data lake-focused capabilities. If those capabilities build on and have the look and feel of the rest of suite, you'll at least have product familiarity on your side and training needs might be reduced. Here are four other evaluation criteria to consider:

- **Breadth of capabilities:** While some suites are focused tightly on data ingestion, transformation and enrichment to data-delivery – the data lake equivalent of ETL -- others offer more comprehensive capabilities including data quality, master data management, data cataloging and self-service data prep. In some cases, suites have been assembled through multiple acquisitions, so consider whether they present consistent user, deployment and administrative experiences.

Some vendors, notably SnapLogic, have recently re-architected to deliver their technology software-as-a-service style, simplifying and speeding deployment and ongoing administration. Still other vendors, notably Syncsort, have deep capabilities

for offloading and integrating mainframe workloads to the data lake. Consider short-term and long-term needs and choose the portfolio and deployment style that best fit your needs.

- **Hadoop compatibility:** Most, but not all, incumbent vendors have adapted their suites to run on Hadoop. This step is crucial, as it's the key to handling data at Hadoop scale with all the advantages of distributed, massively parallel processing. Develop a detailed understanding of how the technology runs on and works with Hadoop, including compatibility with YARN, native security, governance modules, MapReduce processing, open-source data-connection and data-movement tools such as Hive and HCatalog as well as Hadoop-centric data formats such as Parquet and ORC. Also inquire about licensing terms and any requirements to run software or data-processing steps on servers that are separate from the Hadoop cluster. Finally, investigate data-pipeline capacity constraints when moving data between Hadoop and other platforms. You don't want a monolithic, hard-to-change, bottlenecked system constraining your data lake.

- **Spark compatibility:** Apache Spark has become a standard component of Hadoop distributions, thanks to its versatility and in-memory processing power. The most advanced incumbent vendors can not only connect to Spark but invoke Spark as part of data processing pipelines. Some systems can automatically assign workloads to the best-fit processing engine for a given job, whether that's MapReduce for high-scale batch processing or Spark for low-latency, in-memory processing of smaller batches or streaming data.

- **Data cataloging and self-service data prep:** Incumbent data management vendors have typically focused on the needs of IT and data professionals, but some have responded to the demand for data-analyst and business-user compatible data cataloging and data-prep tools. The key questions are just how business-user friendly and capable these tools are. Are they better suited to advanced data analysts or are they truly usable by

reasonably data-savvy business users? It's important for would-be users to try before the company buys. Further, can self-service data-prep workflows be put directly into production or are they representations of workflows that have to be built out and connected to data by IT types? If it's the latter, IT bottlenecks may remain.

RECOMMENDATION: TARGET THE CENTER OF ANALYTICAL GRAVITY

Hadoop vendors talk about the "center of data gravity" moving to Hadoop, but Constellation Research believes you also have to consider the center of analytical gravity. That is, on what analytical platforms – Hadoop, Spark, relational databases, embedded apps, etc. – will companies derive the greatest return on insight and will that change?

When companies are starting from scratch, it's much easier to make technology selections. Internet startups building from scratch to harness Big Data, for example, can start with open source distributions and add commercial or home-grown tools as needed. Greenfield scenarios like this are rare.

Within many companies, particularly mainstream firms, Hadoop and open source platforms in general have tended to enter through the back door. Pilots and proof-of-concept projects were initiated in isolation, with small teams assembled apart from mainstream data management staffs.

Times are changing. As Hadoop has evolved into a corporate standard, as use cases have multiplied, and as the data lake concept has taken root, data management and data analysis must become more of a single, cohesive ecosystem. That does not mean that one platform takes over. Rather, the data lake must support evolved analytical tools and evolved applications and application infrastructure. Ideally, it should also support a wide spectrum of analytic talent, from spreadsheet warriors to data scientists.

Data marts might move to the Hadoop environment, or they might remain in relational environments for many years to come.

Advanced analytical analyses might move to Hadoop or Spark, or they might move to distributed database or in-memory grids. The question is where will the bulk of the analysis take place and what are the breakthrough insights and production workloads of tomorrow?

SQL has been around a long time and it was never terribly effective at cracking variable and semi-structured data types like clickstreams, log files, sensor data, mobile app data, social data and so on. The data lake's ability to handle varied data types at high scale has changed the game, driving dynamic, 360-degree customer views and breakthrough insights based on correlations of social, mobile, machine and transactional data.

The data lake's ability to handle varied data at high scale has changed the game, driving dynamic, 360-degree customer views and breakthrough insights based on correlations of social, mobile, machine and transactional data.

While it's important for the data lake to support SQL-centric analysis, don't think of it as a modern, high-scale replacement for the data warehouse. If you're not tapping new data and coming up with new insights, you're not doing it right.

Two Other Important Trends to Consider: Cloud and Streaming

There are two other important trends to weigh when considering data lake architecture and management technologies:

- **Cloud deployment options:** Cloud is the fastest-growing deployment mode for Hadoop, for databases, and for applications and infrastructure. It's important to consider private-cloud and public-cloud deployment options and cloud-service compatibility. A starting-point question is whether a tool or suite is available as software as a service. If not, are certain components available in the cloud? If it's a software-only offering, does it have a micro-services architecture and RESTful APIs that facilitate cloud deployment?

- **Streaming support:** Among the hottest trends emerging in 2016 in the data-to-decisions domain is support for streaming data processing and analysis. Marketers and e-commerce vendors have been among the pioneers of real-time use cases because presenting the right content, offer or recommendation at the right time can make a huge dollars-and-cents difference in campaign and sales performance.

  New streaming use cases are popping up all over the place, with connected cars, smart oil fields, smart utilities and precision medicine being popular examples. A key point is that historical data, such as that held in a data lake and in data warehouses, is invariably needed to bring context to the real-time insights. Therefore, streaming data management and streaming data delivery capabilities -- to apps and real-time analytical tools – should be data lake capabilities.

RECOMMENDATION: CONSIDER CONSULTANTS AND SYSTEMS INTEGRATORS

How might your firm take advantage of Big Data? Experienced consultants and systems integrators tend to start with the big questions, such as what are the business objectives and competitive challenges? Next, what insights might lead to a deeper understanding of customers, optimization of current practices, monetization of data or, even better, breakthrough products or services that might yield entirely new revenue streams? What's the total inventory of data available -- even if

it's not currently used -- and how might it drive breakthroughs and innovations?

Big-picture thinking and experience are reasons organizations consider the services of a consulting firm or systems integrator such as Accenture, Deloitte, IBM, CapGemini, Infosys, Wipro or a smaller firm such as Persistent Systems or Teradata's Think Big Analytics unit. The firms tend to have analytics practices and experienced Big Data deployment teams that can draw on lessons learned across scores, if not hundreds, of deployments. What's more, these consultancies and integrators are experienced at defining and implementing data architecture strategies.

In some cases, integrators support Hadoop and data lake deployments with technology and automation capabilities as well as frameworks, blueprints and best-practice methodologies. Infosys, for example, has the Infosys Information Platform, which offers a unified interface to both Hadoop and Spark, supporting data pipelines that might make use of MapReduce, Spark and Infosys' own machine learning technologies. Infosys has partnerships with both Waterline and Trifacta.

In another example, Think Big Analytics focuses mainly on consulting services for building and maintaining data lakes and data science with open source technologies. Think Big offers tools such as a data lake framework for simplifying ingestion and governance using diverse processing tools and Dashboard Engine that opens up access to data through REST, SQL and MDX interfaces with support for session-based analysis, time-series analysis, dashboarding and reporting through mainstream tools such as Tableau and MicroStrategy.

Consultants and systems integrators often have partnerships and experience deploying open-source distributions and incumbent commercial offerings, but check whether they have experience with next-generation options, which aren't as well known. Some next-generation vendors contend that consultants and integrators have a bias toward incumbents or pure open-source software because of the consultants' experience set.

Also helpful is consultant and systems integrator experience with optimization of data warehouses and offloading of archival and data-processing workloads to Hadoop. They can help firms assess current-state and plan future-state deployments, identifying opportunities for innovation and determining the appropriate center of analytical gravity.

TAKEAWAY: PREVENT DATA SWAMPS AND MAKE BIG DATA ACCESSIBLE

If the data lake is to succeed, you can't just ingest data and create new data sets indiscriminately. You'll need to create zones for landing raw data, cleansing and normalizing standard data, and aggregating and enhancing application-specific data. In short, you need a mature approach to data lake management and governance.

Unfortunately, the capabilities and technologies of the Hadoop stack are incomplete and many tools are low-level, coding intensive and unfamiliar to enterprise data management teams. Commercial vendors are closing the gaps with:

1. Data management and governance tools that are accessible to enterprise IT
2. Data cataloging and metadata management tools that are accessible to IT, analysts and data stewards
3. Self-service discovery and data-prep tools that are accessible to analysts and data-savvy business users.

The choice of tools and suites that might best complement the Hadoop stack will depend on the data, business goals, sophistication, and analytical focus of the organization. But the success of the data lake depends on a mature approach and technology-empowered data management, analysis and business professionals.

Doug Henschen

Vice President and Principal Analyst

Doug Henschen is Vice President and Principal Analyst at Constellation Research, Inc., focusing on data-driven decision making. His Data-to-Decisions research examines how organizations employ data analysis to reimagine their business models and gain a deeper understanding of their customers. Data insights also figure into tech optimization and innovation in human-to-machine and machine-to-machine business processes in manufacturing, retailing and services industries.

Henschen's research acknowledges the fact that innovative applications of data analysis require a multi-disciplinary approach, starting with information and orchestration technologies, continuing through business intelligence, data visualization, and analytics, and moving into NoSQL and Big Data analysis, third-party data enrichment, and decision management technologies. Insight-driven business models and innovations are of interest to the entire C-suite.

Previously, Henschen led analytics, Big Data, business intelligence, optimization, and smart applications research and news coverage at InformationWeek. His experiences include leadership in analytics, business intelligence, database, data warehousing, and decision-support research and analysis for Intelligent Enterprise. Further, Henschen led business process management and enterprise content management research and analysis at Transform magazine. At DM News, he led the coverage of database marketing and digital marketing trends and news.

@DHenschen | www.constellationr.com/users/ doug-henschen

www.linkedin.com/in/doughenschen

ABOUT CONSTELLATION RESEARCH

Constellation Research is an award-winning, Silicon Valley-based research and advisory firm that helps organizations navigate the challenges of digital disruption through business models transformation and the judicious application of disruptive technologies. Unlike the legacy analyst firms, Constellation Research is disrupting how research is accessed, what topics are covered and how clients can partner with a research firm to achieve success. Over 350 clients have joined from an ecosystem of buyers, partners, solution providers, C-suite, boards of directors and vendor clients. Our mission is to identify, validate and share insights with our clients.

Organizational Highlights

- Named Institute of Industry Analyst Relations (IIAR) New Analyst Firm of the Year in 2011 and #1 Independent Analyst Firm for 2014 and 2015.
- Experienced research team with an average of 25 years of practitioner, management and industry experience.
- Organizers of the Constellation Connected Enterprise – an innovation summit and best practices knowledge-sharing retreat for business leaders.
- Founders of Constellation Executive Network, a membership organization for digital leaders seeking to learn from market leaders and fast followers.

www.ConstellationR.com

@ConstellationR

info@ConstellationR.com

sales@ConstellationR.com

San Francisco | Belfast | Boston | Colorado Springs | Cupertino | Denver | London | New York | Northern Virginia
Palo Alto | Pune | Sacramento | Santa Monica | Sydney | Toronto | Washington, D.C

constellation RESEARCH