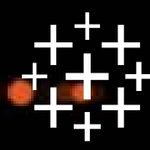
TOP TEN

Big Data

TRENDS FOR 2017

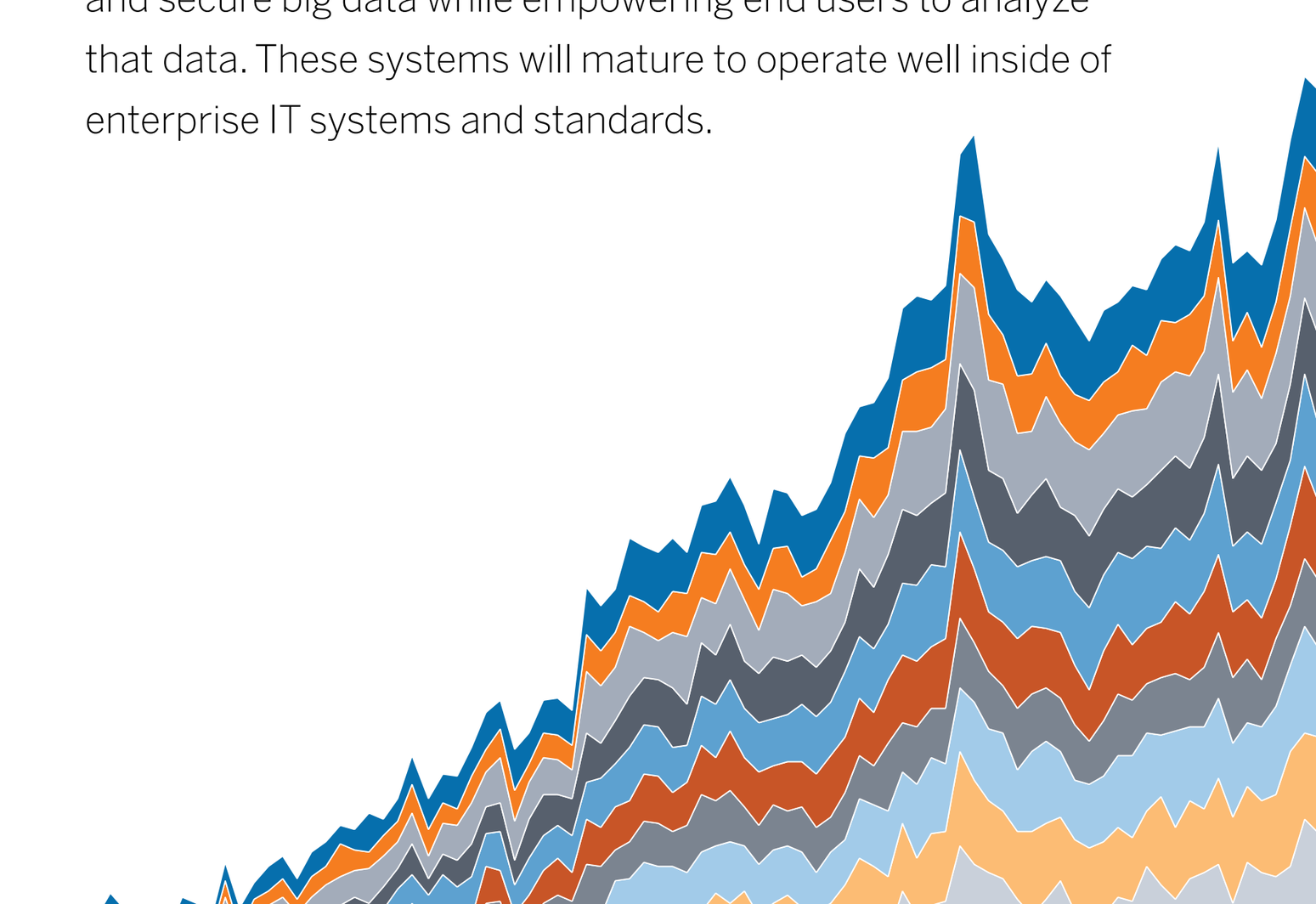


tableau

Each year at Tableau, we start a conversation about what's happening in the industry.
The discussion drives our list of the top big-data trends for the following year. These are our predictions for 2017.

Top 10 Big Data Trends for 2017

2016 was a landmark year for big data with more organizations storing, processing, and extracting value from data of all forms and sizes. In 2017, systems that support large volumes of both structured and unstructured data will continue to rise. The market will demand platforms that help data custodians govern and secure big data while empowering end users to analyze that data. These systems will mature to operate well inside of enterprise IT systems and standards.



1

Big data becomes fast and approachable: Options expand to speed up Hadoop

Sure, you can perform machine learning and conduct sentiment analysis on Hadoop, but the first question people often ask is: How fast is the interactive SQL? SQL, after all, is the conduit to business users who want to use Hadoop data for faster, more repeatable KPI dashboards as well as exploratory analysis.

This need for speed has fueled the adoption of faster databases like Exasol and MemSQL, Hadoop-based stores like Kudu, and technologies that enable faster queries. Using SQL-on-Hadoop engines (Apache Impala, Hive LLAP, Presto, Phoenix, and Drill) and OLAP-on-Hadoop technologies (AtScale, Jethro Data, and Kyvos Insights), these query accelerators are further blurring the lines between traditional warehouses and the world of big data.

FURTHER READING: AtScale BI on Hadoop benchmark Q4 2016

2

Big data no longer just Hadoop: Purpose-built tools for Hadoop become obsolete

In previous years, we saw several technologies rise with the big-data wave to fulfill the need for analytics on Hadoop. But enterprises with complex, heterogeneous environments no longer want to adopt a siloed BI access point just for one data source (Hadoop). Answers to their questions are buried in a host of sources ranging from systems of record to cloud warehouses, to structured and unstructured data from both Hadoop and non-Hadoop sources. (Incidentally, even relational databases are becoming big data-ready. SQL Server 2016, for instance, recently added JSON support.)

In 2017, customers will demand analytics on all data. Platforms that are data- and source-agnostic will thrive while those that are purpose-built for Hadoop and fail to deploy across use cases will fall by the wayside. The exit of Platfora serves as an early indicator of this trend.

FURTHER READING: Uncommon sense: The big data warehouse

3

Organizations leverage data lakes from the get-go to drive value

A data lake is like a man-made reservoir. First you dam the end (build a cluster), then you let it fill up with water (data). Once you establish the lake, you start using the water (data) for various purposes like generating electricity, drinking, and recreating (predictive analytics, ML, cyber security, etc.).

Up until now, hydrating the lake has been an end in itself. In 2017, that will change as the business justification for Hadoop tightens. Organizations will demand repeatable and agile use of the lake for quicker answers. They'll carefully consider business outcomes before investing in personnel, data, and infrastructure. This will foster a stronger partnership between the business and IT. And self-service platforms will gain deeper recognition as the tool for harnessing big-data assets.

FURTHER READING: Maximizing data value with a data lake

4

Architectures mature to reject one-size-fits all frameworks

Hadoop is no longer just a batch-processing platform for data-science use cases. It has become a multi-purpose engine for ad hoc analysis. It's even being used for operational reporting on day-to-day workloads—the kind traditionally handled by data warehouses.

In 2017, organizations will respond to these hybrid needs by pursuing use case-specific architecture design. They'll research a host of factors including user personas, questions, volumes, frequency of access, speed of data, and level of aggregation before committing to a data strategy. These modern-reference architectures will be needs-driven. They'll combine the best self-service data-prep tools, Hadoop Core, and end-user analytics platforms in ways that can be reconfigured as those needs evolve. The flexibility of these architectures will ultimately drive technology choices.

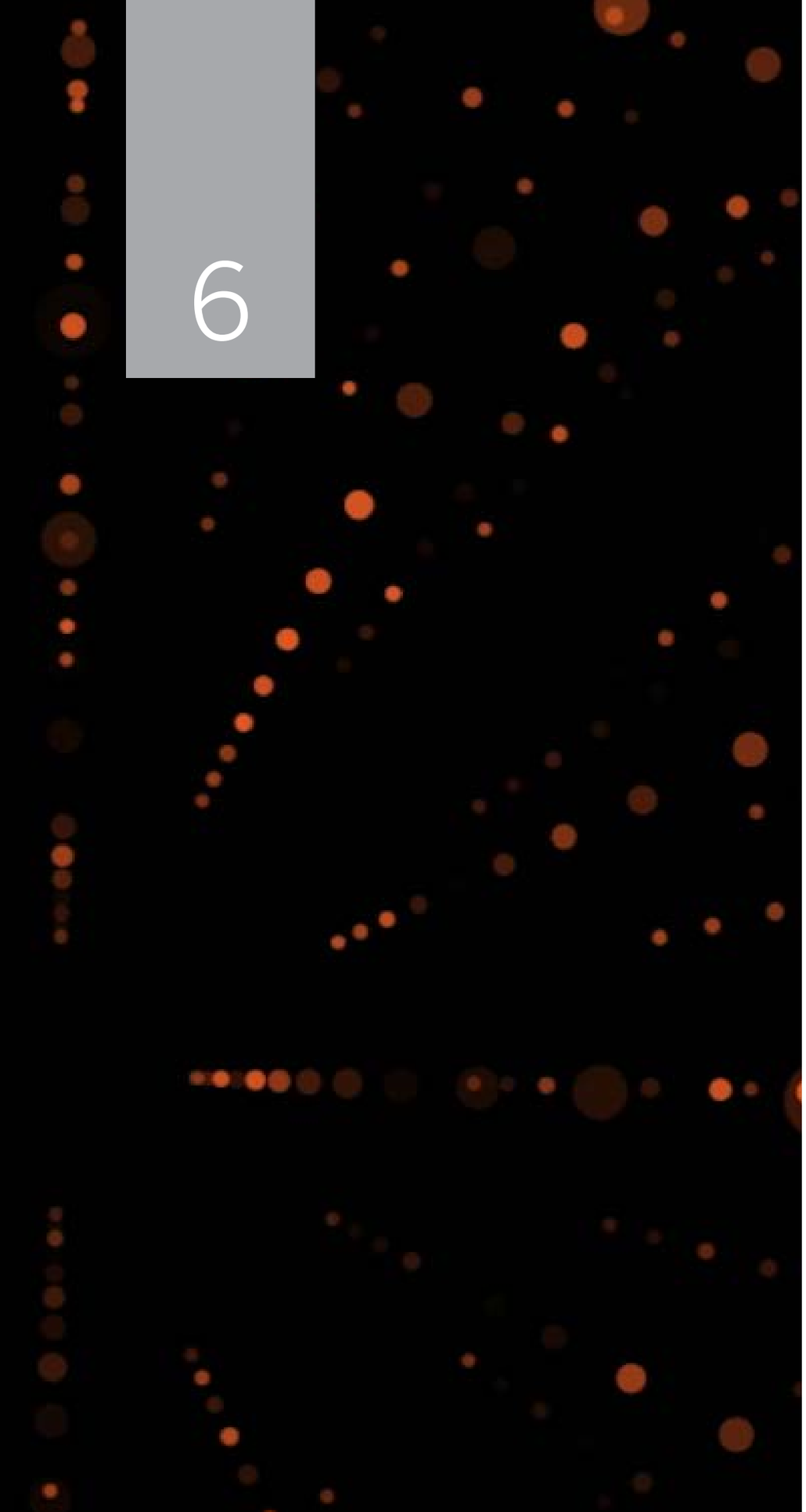
FURTHER READING: The cold/warm/hot framework and how it applies to your Hadoop strategy

5

Variety, not volume or velocity, drives big-data investments

Gartner defines big data as the three Vs: high-volume, high-velocity, high-variety information assets. While all three Vs are growing, variety is becoming the single biggest driver of big-data investments, as seen in the results of a recent survey by New Vantage Partners. This trend will continue to grow as firms seek to integrate more sources and focus on the "long tail" of big data. From schema-free JSON to nested types in other databases (relational and NoSQL), to non-flat data (Avro, Parquet, XML), data formats are multiplying and connectors are becoming crucial. In 2017, analytics platforms will be evaluated based on their ability to provide live direct connectivity to these disparate sources.

FURTHER READING: Variety, not volume, is driving big data initiatives



6

Spark and machine learning light up big data

Apache Spark, once a component of the Hadoop ecosystem, is now becoming the big-data platform of choice for enterprises. In a survey of data architects, IT managers, and BI analysts, nearly 70% of the respondents favored Spark over incumbent MapReduce, which is batch-oriented and doesn't lend itself to interactive applications or real-time stream processing.

These big-compute-on-big-data capabilities have elevated platforms featuring computation-intensive machine learning, AI, and graph algorithms. Microsoft Azure ML in particular has taken off thanks to its beginner-friendliness and easy integration with existing Microsoft platforms. Opening up ML to the masses will lead to the creation of more models and applications generating petabytes of data. As machines learn and systems get smart, all eyes will be on self-service software providers to see how they make this data approachable to the end user.

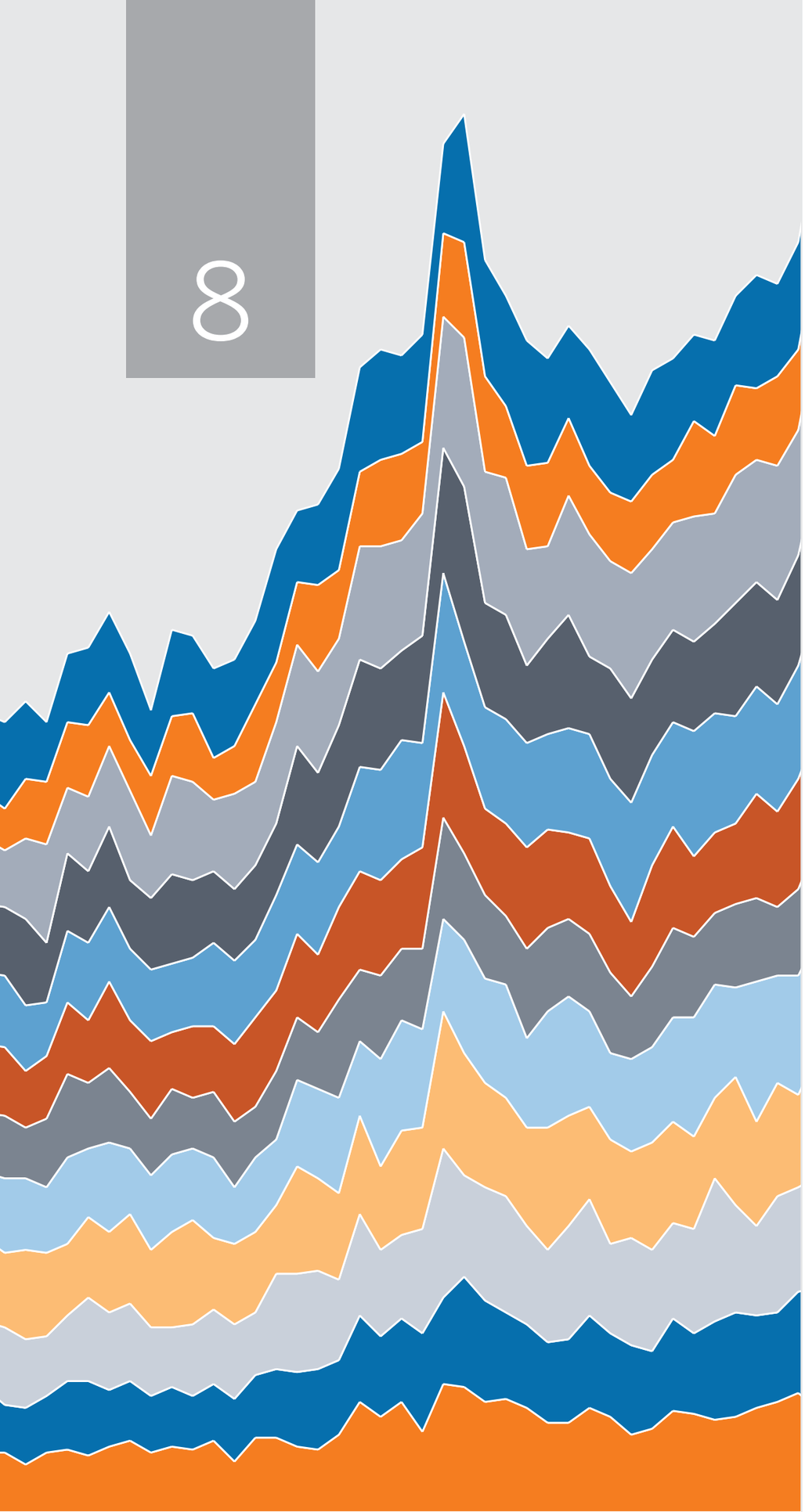
FURTHER READING: Why you should use Spark for machine learning

7

The convergence of IoT, cloud, and big data create new opportunities for self-service analytics

It seems that everything in 2017 will have a sensor that sends information back to the mothership. IoT is generating massive volumes of structured and unstructured data, and an increasing share of this data is being deployed on cloud services. The data is often heterogeneous and lives across multiple relational and non-relational systems, from Hadoop clusters to NoSQL databases. While innovations in storage and managed services have sped up the capture process, accessing and understanding the data itself still pose a significant last-mile challenge. As a result, demand is growing for analytical tools that seamlessly connect to and combine a wide variety of cloud-hosted data sources. Such tools enable businesses to explore and visualize any type of data stored anywhere, helping them discover hidden opportunity in their IoT investment.

FURTHER READING: Tableau on solving IoT's last-mile challenge



8

Self-service data prep becomes mainstream as end users begin to shape big data

Making Hadoop data accessible to business users is one of the biggest challenges of our time. The rise of self-service analytics platforms has improved this journey. But business users want to further reduce the time and complexity of preparing data for analysis, which is especially important when dealing with a variety of data types and formats.

Agile self-service data-prep tools not only allow Hadoop data to be prepped at the source but also make the data available as snapshots for faster and easier exploration. We've seen a host of innovation in this space from companies focused on end-user data prep for big data such as Alteryx, Trifacta, and Paxata. These tools are lowering the barriers to entry for late Hadoop adopters and laggards and will continue to gain traction in 2017.

FURTHER READING: Why self-service prep is a killer app for big data

9

Big data grows up: Hadoop adds to enterprise standards

We're seeing a growing trend of Hadoop becoming a core part of the enterprise IT landscape. And in 2017, we'll see more investments in the security and governance components surrounding enterprise systems. Apache Sentry provides a system for enforcing fine-grained, role-based authorization to data and metadata stored on a Hadoop cluster. Apache Atlas, created as part of the data governance initiative, empowers organizations to apply consistent data classification across the data ecosystem. Apache Ranger provides centralized security administration for Hadoop.

Customers are starting to expect these types of capabilities from their enterprise-grade RDBMS platforms. These capabilities are moving to the forefront of emerging big-data technologies, thereby eliminating yet another barrier to enterprise adoption.

FURTHER READING: The phases of Hadoop maturity: Where exactly is it going?

10

Rise of metadata catalogs helps people find analysis-worthy big data

For a long time, companies threw away data because they had too much to process. With Hadoop, they can process lots of data, but the data isn't generally organized in a way that can be found.

Metadata catalogs can help users discover and understand relevant data worth analyzing using self-service tools. This gap in customer need is being filled by companies like Alation and Waterline which use machine learning to automate the work of finding data in Hadoop. They catalog files using tags, uncover relationships between data assets, and even provide query suggestions via searchable UIs. This helps both data consumers and data stewards reduce the time it takes to trust, find, and accurately query the data. In 2017, we'll see more awareness and demand for self-service discovery, which will grow as a natural extension of self-service analytics.

FURTHER READING: Data catalogs as a strategic requirement for data lakes

About Tableau

Integrating data visualization into your retail programs and processes is easier than you think.

Tableau Software helps people see and understand data no matter how big it is, or how many systems it is stored in. Quickly connect, blend, visualize and share data dashboards with a seamless experience from the PC to the iPad. Create and publish dashboards with automatic data updates, and share them with colleagues, partners or customers—no programming skills required. Begin a free trial today.

TABLEAU.COM/TRIAL