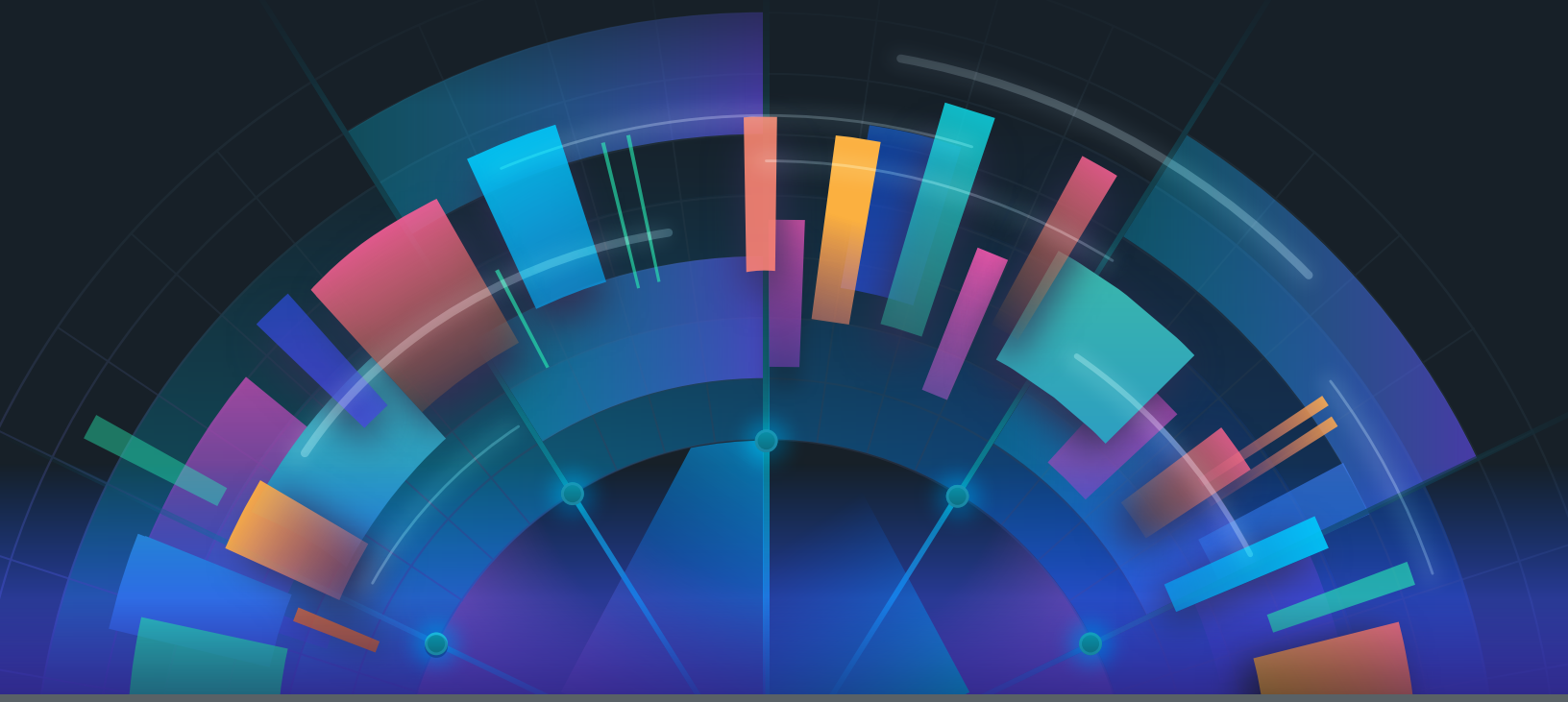
DZone

THE 2018 DZONE GUIDE TO

Big Data

STREAM PROCESSING, STATISTICS, & SCALABILITY

VOLUME V



BROUGHT TO YOU IN PARTNERSHIP WITH

data iku

HPCC SYSTEMS®

Dear Reader,

I first heard the term "Big Data" almost a decade ago. At that time, it looked like it was nothing new, and our databases would just be upgraded to handle some more data. No big deal. But soon, it became clear that traditional databases were not designed to handle Big Data. The term "Big Data" has more dimensions than just "some more data." It encompasses both structured and unstructured data, fast moving and historical data. Now, with these elements added to the data, some of the other problems such as data contextualization, data validity, noise, and abnormality in the data became more prominent. Since then, Big Data technologies has gone through several phases of development and transformation, and they are gradually maturing. A term that was considered as a fad and a technology ecosystem that was considered a luxury are slowly establishing themselves as necessary needs for today's business activities. Big Data is the new competitive advantage and it matters for our businesses.

The more we progress and the more automation we implement, data is always going to transform and grow. Blockchain technologies, Cloud, and IoT are adding new dimensions to the Big Data trend. Hats off the developers who are continually innovating and creating new Big Data Storage and Analytics applications to derive value out of this data. The fast-paced development has made it easier for us to tame fast-growing massive data and integrate our existing Enterprise IT infrastructure with these new data sources. These successes are driven by both Enterprises and Open Source communities. Without open source projects like Apache Hadoop, Apache Spark, and Kafka, to name a few, the landscape would have be entirely different. The use of Machine Learning and Data visualization methods packaged for analyzing Big Data is also making life easier for analysts and management. However, we still hear the failure of analytics projects more often than the successes. There are several reasons why. So, we bring you this guide, where these articles written by DZone contributors are going to provide you with more significant insights into these topics.

The Big Data guide is an attempt to help readers discover and help understand the current landscape of the Big Data ecosystem, where we stand, and what amazing insights and applications people are discovering in this space. We wish that everyone who reads this guide finds it worthy and informative. Happy reading!

By Sibanjan Das

BUSINESS ANALYTICS & DATA SCIENCE CONSULTANT & DZONE ZONE LEADER

Table of Contents

DZONE IS...

PRODUCTION
CHRIS SMITH, DIR. OF PRODUCTION
ANDRE POWELL, SR. PRODUCTION COORDINATOR
G. RYAN SPAIN, PRODUCTION COORDINATOR
ASHLEY SLATE, DESIGN DIR.
BILLY DAVIS, PRODUCTION ASSISSTANT

MARKETING
KELLET ATKINSON, DIR. OF MARKETING
LAUREN CURATOLA, MARKETING SPECIALIST
KRISTEN PAGÀN, MARKETING SPECIALIST
NATALIE IANNELLO, MARKETING SPECIALIST
JULIAN MORRIS, MARKETING SPECIALIST

BUSINESS
RICK ROSS, CEO
MATT SCHMIDT, PRESIDENT
JESSE DAVIS, EVP

SALES
MATT O'BRIAN, DIR. OF BUSINESS DEV.
ALEX CRAFTS, DIR. OF MAJOR ACCOUNTS
JIM HOWARD, SR ACCOUNT EXECUTIVE
JIM DYER, ACCOUNT EXECUTIVE
ANDREW BARKER, ACCOUNT EXECUTIVE
BRIAN ANDERSON, ACCOUNT EXECUTIVE
RYAN McCOOK, ACCOUNT EXECUTIVE
CHRIS BRUMFIELD, SALES MANAGER
TOM MARTIN, ACCOUNT MANAGER
JASON BUDDAY, ACCOUNT MANAGER

EDITORIAL
CAITLIN CANDELMO, DIR. OF CONTENT & COMMUNITY
MATT WERNER, PUBLICATIONS COORD.
MICHAEL THARRINGTON, CONTENT & COMMUNITY MANAGER
KARA PHELPS, CONTENT & COMMUNITY MANAGER
MIKE GATES, SR. CONTENT COORD.
SARAH DAVIS, CONTENT COORD.
TOM SMITH, RESEARCH ANALYST
JORDAN BAKER, CONTENT COORD.
ANNE MARIE GLEN, CONTENT COORD.
ANDRE LEE-MOYE, CONTENT COORD.

Executive Summary

BY MATT WERNER PUBLICATIONS COORDINATOR, DZONE

Classically, Big Data has been defined by three V's: Volume, or how much data you have; Velocity, or how fast data is collected; and Variety, or how heterogeneous the data set is. As movements like the Internet of Things provide constant streams of data from hardware, and AI initiatives require massive data sets to teach machines to think, the way in which Big Data is stored and utilized continues to change. To find out how developers are approaching these challenges, we asked 540 DZone members to tell us about what tools they're using to overcome them, and how their organizations measure successful implementations.

THE PAINS OF THE THREE V'S

DATA

Of several data sources, files give developers the most trouble when it comes to the volume and variety of data (47% and 56%, respectively), while 42% of respondents had major issues with the speed at which both server logs and sensor data were generated. The volume of server logs was also a major issue, with 46% of respondents citing it as a pain point.

IMPLICATIONS

As the Internet of Things takes more of a foothold in various industries, the difficulties in handling all three V's of Big Data will increase. More and more applications and organizations are generating files and documents rather than just values and numbers.

RECOMMENDATIONS

A good way to deal with the constant influx of data from remote hardware is to use solutions like Apache Kafka, an open source project built to handle the processing of data in real-time as it is collected. Currently, 61% of survey respondents are using Kafka, and we recently released our first Refcard on the topic. Using document store databases, such as MongoDB, are recommended to handle files, documents, and other semi-structured data.

DATA IN THE CLOUD

DATA

39% of survey respondents typically store data in the cloud, compared to 33% who store it on-premise and 23% who take a hybrid approach. Of those using the cloud or hybrid solutions, Amazon was by far the most popular vendor (70%) followed by Google Cloud (57%) and Microsoft Azure (39%).

IMPLICATIONS

The percentage of respondents using cloud solutions increased by 8% in 2018, while on-premise and hybrid storage decreased by 6% and 4%, respectively. As cloud solutions become more and more ubiquitous, the prospect of storing data on external hardware becomes more appealing as a way to decrease costs. Another reason for this increase may be in the abilities of some tools to process data, such as AWS Kinesis.

RECOMMENDATIONS

Not every organization may need a way to store Big Data if they do not yet have a strong use case for it. When a business strategy around Big Data is created, however, a cloud storage solution requires less up-front investment for smaller enterprises, though if your business deals in sensitive information, a hybrid solution would be a good compromise, or if you need insights as fast as possible you may need to invest in an on-premise solution to access data quickly.

BLINDED WITH DATA SCIENCE

DATA

The biggest challenges in data science are working with unsanitized or unclean data (64%), working within timelines (40%), and limited training and talent (39%).

IMPLICATIONS

The return on investment of analytics projects is mostly focused on the speed of decision-making (47%), the speed of data access (45%), and data integrity (31%). These KPIs all contribute to the difficulty of dealing with unclean data and project timelines. Insights are demanded quickly, but data scientists need to take time to ensure data quality is good so those insights are valuable.

RECOMMENDATIONS

Project timelines need to be able to accommodate the time it takes to prepare data for analysis, which can range from omitting sensitive information to deleting irrelevant values. One easy way to do this is to sanitize user inputs, keeping users from adding too much nonsense to your data. Our infographic on page 14 can give you a fun explanation of why unclean and unsanitized data can be such a hassle and why it's important to gleaning insights.

Key Research Findings

BY G. RYAN SPAIN PRODUCTION COORDINATOR, DZONE

DEMOGRAPHICS

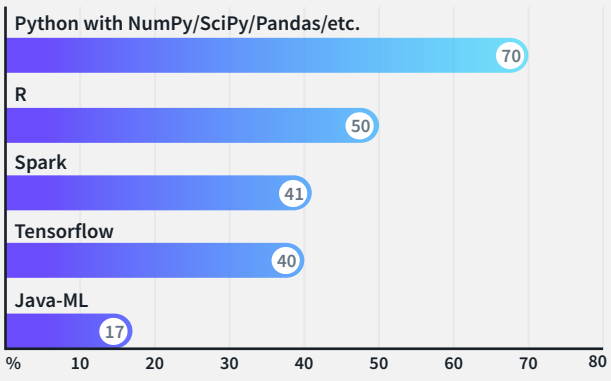
- 540 software professionals completed DZone's 2018 Big Data survey. Respondent demographics are as follows:
- 42% of respondents identify as developers or engineers, and 23% identify as developer team leads.
- 54% of respondents have 10 years of experience or more; 28% have 15 years or more.
- 39% of respondents work at companies headquartered in Europe; 30% work in companies headquartered in North America.
- 23% of respondents work at organizations with more than 10,000 employees; 22% work at organizations between 500 and 10,000 employees.
- 77% develop web applications or services; 48% develop enterprise business apps; and 21% develop native mobile applications.
- 78% work at companies using the Java ecosystem; 62% at companies that use JavaScript; and 37% at companies that use Python. 56% of respondents use Java as their primary language at work.

PYTHON

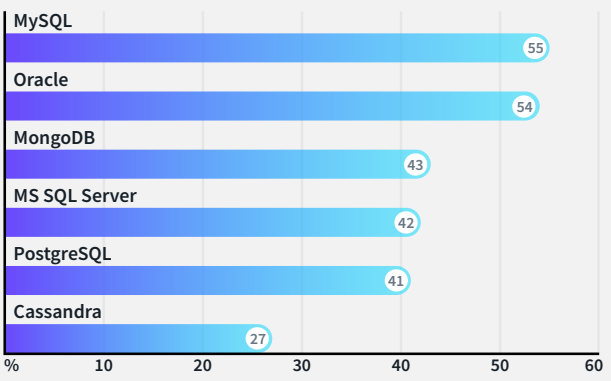
Python has been moving slowly towards the title of "most popular language for data science" for years now, and the language to beat has been R. R has been extraordinarily popular for data-heavy programming for some time as an open source implementation of S, a language specifically designed for statistical analysis. And while R still maintains popularity (in the TIOBE index, it moved from the 16th place ranking in January 2017 to the 8th place ranking in 2018), Python's use in data science and data mining projects has been steadily increasing. Last year, respondents to DZone's Big Data survey revealed that Python had overcome R as the predominant language used for data science, though its lead over R was statistically insignificant, and therefore didn't quite make it to "champion status." This was mentioned in last year's research findings as being consistent with trends in other available research on Python and R's use in data science: R is still popular for data/statistical analysis, but Python has been catching up.

This year, DZone's Big Data survey showed a significant difference between the use of R and Python for data science projects: R usage decreased by 10%, from 60% to 50%, among survey respondents in the last year, while Python increased 6%, from 64% to 70%. This means 20% more respondents this year use Python for data science than respondents who use R. While Python was not created specifically for data analysis, its dynamic typing, easy-to-learn syntax, and ever-increasing base of libraries has made it an ideal candidate for developers to start delving into data science and analysis more comfortably than they may have been able to in the past.

GRAPH 01. What languages/libraries/frameworks do you use for data science and machine learning?



GRAPH 02. What database management systems do you use in production?



THE THREE VS

The concept of "Big Data" has been a difficult one to define. The sheer amount (volume) of data that is able to be stored in a single hard disk drive, solid state drive, secure digital memory card, etc., continues to improve, and hardware technology has grown fast enough that I still remember buying a computer with a 10 gigabyte hard drive and being told from the salesperson at the electronics store "you'll never need more computer storage again." With new data storage options (cloud and hybrid storage, for example), storing large volumes of data isn't such a hard thing to overcome, though it still requires some planning to do. The complications added by "Big Data" include dealing not only with data volume, but also data variety (how many different types of data you have to deal with) and data velocity (how fast the data is being added).

Beyond that, "Big Data" is complicated by the fact that just storing this data is not enough; to get anything from the data being collected, it not only needs to be stored, but also needs to be analyzed. 76% of our survey respondents said they have to deal with large quantities of data (volume), while 46% said they have to work with high-velocity data and 45% said they have to work with highly variable data.

Each of these "Big Data" categories come with its own set of challenges. The most challenging data sources for those dealing with high-volume data were files (47%) and server logs (46%), and the most challenging data types were relational (51%) and semi-structured (e.g. JSON, XML; 39%). For those dealing with high velocity data, server logs and sensors/remote hardware (both 42%) were the most challenging data sources, and semi-structured (36%) and complex data (e.g. graph, hierarchical; 30%) were the most challenging data types. Finally, regarding data variety, the biggest data source challenges came from files (56%). Server logs, sensor/remote hardware data, ERP and other enterprise systems, user-generated data, and supply-chain/logistics/other procurement data all fell between 28% and 32% of responses labeling these tasks as "challenging."

DATABASE MANAGEMENT SYSTEMS

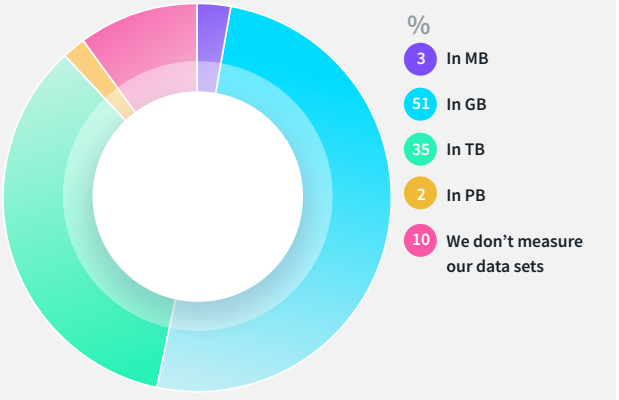
Database popularity has been (according to db-engines.com) moving in fairly steady patterns for the past several years. 2012/2013 saw some combat between MS SQL Server and MySQL before MySQL overcame Microsoft's DB, and Oracle's database stood above them all (although dropping significantly since 2016, at times in danger of being surpassed by MySQL). Among our Big Data survey respondents, MySQL is still the most popular DBMS from 2017, though its popularity has dropped (61% use in production in 2017 vs. 55% use in production in 2018). Oracle's use in production, on the other hand, has increased from 48% in 2017 to 54% in 2018. Other DBMS trend changes in production include an increase in respondents' usage of PostgreSQL from 35% to 41%, and a decrease in respondents' usage of MS SQL Server from 49% to 42%.

Finally, when asked which databases respondents used specifically for their "Big Data needs," the most common response was the NoSQL DBMS MongoDB, having 11% more respondents than the next DBMS for "Big Data," Oracle, at 29%. While we've seen advanced file systems (like Hadoop) start making an impact on Big Data collection and analysis, non-relational databases seem to also be showing their worth for dealing with data beyond the standard levels of volume, velocity, and variety.

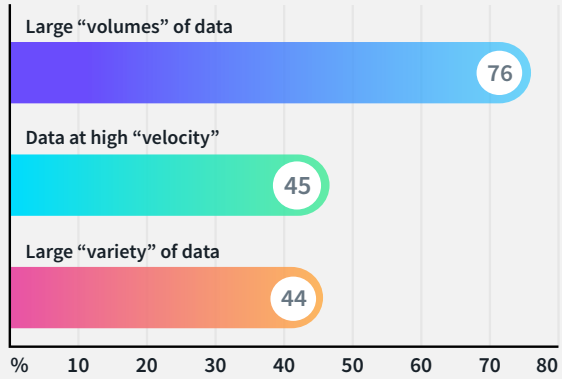
DATA IN THE CLOUD

Respondents typically working with data "in the cloud" rather than on-premise or in a hybrid manner has increased since last year's survey. Those who work with data in the cloud (particularly respondents who answered that they have data science experience) increased from 31% in 2017 to 39% in 2018. Meanwhile, respondents saying they typically deal with data on-premise or in a hybrid format decreased from last year's responses by 6% and 4%, respectively. While an increase in the adoption of data living in the cloud is unsuprising, given general development trends toward overall cloud usage, the growth in cloud data specifically for Big Data needs is minor compared to cloud adoption in other areas we have researched, such as Continuous Delivery. This is likely due to the fact that truly "big" data is often easier and faster to work with the closer it is.

GRAPH 03. How do you measure data sets?



GRAPH 04. What kind of big data do you work with?



Take Big Data to the Next Level with Blockchain Networks

BY ARJUNA CHALA
SR. DIRECTOR OF SPECIAL PROJECTS, HPCC SYSTEMS

QUICK VIEW

01. Blockchains are a form of electronic ledger that simply and securely store transaction data on a peer-to-peer network

02. Thanks to their architecture, Blockchains allow transactions between individuals without requiring a middleman to verify them

03. Blockchains will be used by multiple vertical markets to streamline transactions across ecosystems

04. Widespread use of cloud computing and Big Data are driving the adoption of Blockchains

Bitcoin has taken the financial markets by storm, and the world's first decentralized digital currency could have a similar effect on the big data market as well. It's not Bitcoin's financial significance that I'm referring to, but rather its technological significance. Specifically, Bitcoin was developed to use a peer-to-peer network technology called a blockchain. A blockchain is an electronic ledger containing blocks of related records which are distributed on a peer-to-peer computer network, where every computer in the network has a copy of the ledger. Any new entries into the ledger can only take place after a set of computers agree that the transaction is valid. For example, in a Bitcoin application where person A sells a Bitcoin to person B, the validation process confirms that person A does actually have a Bitcoin to sell by auditing person A's previous transactions in the ledger. Most importantly, multiple peers in the blockchain network execute this validation process independently so the transaction is approved by consensus rather than relying on a single entity.

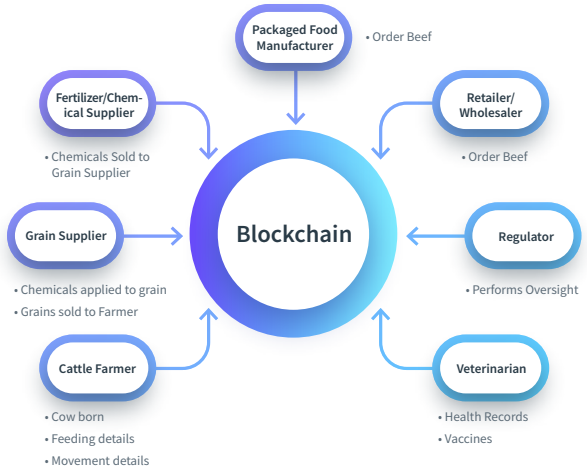
Blockchains can also automate much of the work typically handled by middlemen or escrow services. Middlemen (usually lawyers, notaries, or financial institutions) serve as trusted third parties that review and enforce the terms of a contract or agreement between two parties. While they provide a valuable service that helps ensure the validity of a transaction, a middleman's responsibility for verifying information can delay the closing of a deal. For example, an escrow company will hold the funds for a real estate purchase until it can verify that funds are indeed available and that the property title can be transferred from the seller to the purchaser; it's a process that can add days and weeks to the sale process. But if the purchase were made via a blockchain, the services of an escrow company wouldn't be needed. Since the blockchain can validate the transaction independently, both parties involved can rest assured that the transaction won't be authorized until the network achieves consensus that all terms of the transaction have been met. This greatly accelerates the transaction's closing time and eliminates the cost associated with having a middleman review the transaction. In essence, blockchains enable "smart contracts" which are digital agreements with terms and penalties agreed upon in advance so the blockchain can then enforce them automatically.

In addition to providing high availability semantics by using a peer-to-peer architecture, blockchains also provide excellent data security. Security in a blockchain can be a challenging concept to grasp at first: exactly how secure can a system be if everyone involved in it has access to the complete transaction records of every other participant? First, the fact that all participants in the blockchain have access to every transaction record means that in order to commit fraud, each and every participants' transaction logs would need to be compromised and altered. In a peer-to-peer blockchain network that could potentially include millions of client computers spread across the globe, that kind of record tampering isn't possible. Combine the sheer number of peers in a blockchain with the fact that the technology uses private/public keys to encrypt transaction records, and it becomes readily apparent why blockchains are positioned to be the new standard for performing sophisticated, transaction-based workflows with extremely high levels of integrity.

That kind of technology standard can provide significant benefits to an entire industry, not just one company. But it will require a leap of faith among players in vertical markets, as it calls for companies (even competitors) to transact in a shared environment. While at first glance that concept might appear counterintuitive, companies can benefit by pooling and sharing access to their data. Financial organizations are already sharing data about fraudulent transactions in order to better defend against them, and cybersecurity companies are doing the same with the data they gather from successful network breaches. Essentially, it's "a rising tide raises all ships"-type philosophy that allows companies to leverage each other's data to improve the reliability and veracity of transactions between all players in a supply chain.

Let me illustrate this by way of a vertical market use case example: agriculture. Like nearly every industry, agriculture companies are adopting big data systems to store and analyze the massive amounts of data their organizations generate in order find ways to improve their products and processes. What if that operational data were uploaded to a blockchain network that included data from other agricultural companies? Vendors in the supply chain could track raw materials and products as they move through the supply chain, and the benefits flow from there. Say a restaurant chain serving organic food wants to make sure all of its vegetable suppliers use organic fertilizer; they could check the transaction blockchain to confirm their suppliers are actually purchasing and applying organic fertilizer on the crops they grow, harvest, and sell. Or what if a pest or fungus begins to attack a certain crop? With the data available through a blockchain, analytics could quickly trace the spread of the pest through the ecosystem to find the source and take appropriate action. It could also help streamline compliance issues as industry watchdogs (government agencies, for example) could access the blockchain to verify participants are complying with industry standards and regulations.

Food Provenance Blockchain Network



The diagram illustrates how a blockchain network can verify all participants in a food supply chain are providing required information and meeting their responsibilities to other participants in the network.

Setting up a blockchain transaction network will take some effort among industry players. All parties using the blockchain need to reach consensus on what kind of data is shared (for instance, data on pricing or proprietary IP would likely be excluded from the blockchain), as well as what format the data is in when shared. But once that's established, no one player in the blockchain would be responsible for maintaining a server on behalf of the whole industry (not to mention taking on the op-ex costs and liability concerns involved in maintaining a big data network). Furthermore, blockchain technology could be relatively easy to implement on top of existing IT infrastructures thanks to the development of cloud technology. Cloud computing by definition uses a distributed network of resources, much like the distributed network of peer computers used to form a blockchain. Companies already using the cloud will be less likely to struggle with the fear, confusion, and doubt that often accompany the adoption of new technologies like blockchain.

Indeed, the blockchain model could serve as a catalyst for increased adoption of cloud and big data platforms. In industries where a blockchain network is used to store transaction data, companies not connected to the blockchain or not using big data analytics will be at a competitive disadvantage to those that do. As market trends change and alter the dynamics of the supply chain, non-connected companies won't be able to spot those trends and adjust their strategies as quickly as connected competitors can. Moreover, they won't be able to process transactions as quickly and accurately as other companies connected to the blockchain, a fact that's bound to be exploited by competitors that do use blockchain.

Time and time again, we've seen how the adoption of big data analytics can business benefits and efficiencies that were previously unknowable. And if that big data analytics platform has access to even more data thanks to a blockchain, the potential of big data analytics to positively change business operations grows even more compelling.

ARJUNA CHALA is the Sr. Director of Special Projects for the HPCC Systems® platform at LexisNexis Risk Solutions®. With almost 20 years of experience in software design, Arjuna leads the development of next generation big data capabilities including creating tools around exploratory data analysis, data streaming and business intelligence. Arjuna has a BS in Computer Science from RVCE, Bangalore University.

HPCC Systems: A powerful, open source Big Data analytics platform.

Two integrated clusters, a standards-based web services platform, and a declarative programming language form the basis of this comprehensive, massively scalable Big Data solution.

Open source. Easy to use. Proven.

Platform Features:

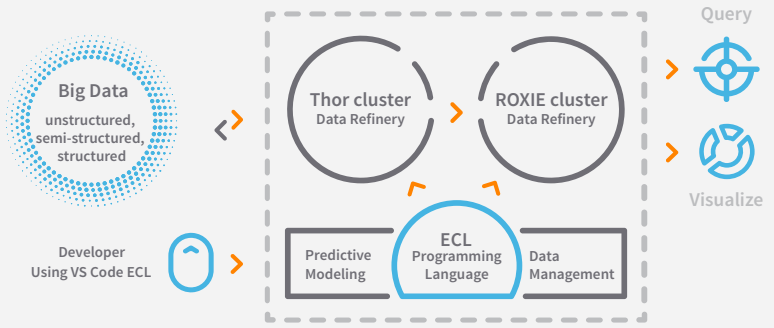
ETL
Extract, Transform, and Load your data using a powerful programming language (ECL) specifically developed to work with data.

Data Management Tools
Data profiling, data cleansing, snapshot data updates and consolidation, job scheduling, and automation are some of the key features.

Query and Search
An indexed based search engine to perform real-time queries. SOAP, XML, REST, and SQL are all supported interfaces.

Predictive Modeling Tools
In place (supporting distributed linear algebra) predictive modeling functionality to perform Linear Regression, Logistic Regression, Decision Trees, and Random Forests.

HPCC Systems



"With other Big Data technologies, we need to use many different open-source modules; it's a lot of work to make them work together. With HPCC Systems, there is just one language, ECL, that can do almost everything."

Mike Yang
Principal Technology Architect, Infosys

Visit: hpccsystems.com

LexisNexis® RISK SOLUTIONS

HPCC SYSTEMS®

Designed for Efficiency, HPCC Systems Makes the Data Lake Easy to Manage

All data today tends to be Big Data. Companies of all sizes are searching for platforms and tools to manage their rapidly growing data sets. However, despite the number of Big Data platforms available, cleaning, standardizing, normalizing, and analyzing data through the data maturity process remains a challenge for most Data Lake implementations.

Unlike other platforms, the open source HPCC Systems platform has an advantage of being built by one of the largest data aggregating companies in the world. LexisNexis Risk Solutions built HPCC Systems on the premise that big data is actually a solution and not a problem. Harvesting large data from thousands of sources helps in incorporating a learning-based approach to handling data. Data programs on the platform use ECL, an easy to use, data-centric programming language that's optimized for data processing and queries. In other words, ECL makes managing data easy. ECL's Data Flow programming paradigm helps programmers concentrate on solving the business problem and without worrying about the complexities of parallel programing.

The HPCC Systems Platform consists of two components, Thor, a data refinery capable of processing billions of records per second, and ROXIE, a web services engine capable of supporting thousands of users with sub-second response times. Together they provide an end-to-end solution for big data processing and analytics that supports rapid development and growth.

The HPCC Systems platform also includes a large library of built-in modules of common data manipulation tasks so users can begin analyzing their data immediately. The platform offers exceptional performance and scalability.

WRITTEN BY ARJUNA CHALA
SENIOR DIRECTOR AT HPCC SYSTEMS SPECIAL PROJECTS

PARTNER SPOTLIGHT

HPCC Systems

End-to-end Big Data in a massively scalable supercomputing platform. Open source. Easy to use. Proven.

HPCC SYSTEMS®

CATEGORY
End-to-end Big Data

RELEASE SCHEDULE
Major release annually, minor releases throughout the year.

OPEN SOURCE?
Yes

CASE STUDY

IProagrica is a leader in the agriculture industry, driving growth and improving efficiency by delivering high-value insight and data, critical tools, and advanced technology solutions. Proagrica needed a massively scalable data refinery environment for ingesting and transforming structured and unstructured data from its various data lakes, including extremely complicated and diverse data sets. Additional requirements included being able deliver quality/clean data from multiple data sets, provide data security, and deliver real-time analytics. Proagrica found the answer to their big data needs with HPCC Systems: a proven and enterprise-tested platform for manipulating, transforming, querying, and warehousing big data.

STRENGTHS

- ETL Engine: Extract and transform your data using a powerful scripting language
- Query Engine: An index-based search engine to perform real-time queries
- Data Management Tools: Data profiling and cleansing, job scheduling, and automation
- Predictive Modeling Tools: Linear/logistic regression, decision trees, random forests
- SOAP, XML, REST, and SQL are all supported interfaces

NOTABLE USERS

- LexisNexis Risk Solutions
- RELX Group
- Infosys
- ClearFunnel
- CPL Online

WEBSITE hpccsystems.com **TWITTER** @hpccsystems **BLOG** hpccsystems.com/blog

Solving Data Integration at Stitch Fix

BY LIZ BENNETT
DATA PLATFORM ENGINEER AT STITCH FIX

QUICK VIEW

01. Data integration is an important capability that data-driven companies should have a good solution for

02. Good data integration greatly decreases the effort necessary to reliably move data from A to B

03. Good data integration makes it much easier for data scientists to explore all the data a company has

04. Recent advancements in data infrastructure have made it much simpler to build a robust data integration solution

About a year ago, my colleagues and I began work on vastly improving the way data is collected and transported at Stitch Fix. We are members of the infrastructure cohort of the Data Platform team — the engineers that support Stitch Fix's army of data scientists — and the solution we built was a self-service data integration platform tailored to the needs of the data scientists. It was a total redesign of much of our data infrastructure, and it is revolutionizing the way data scientists interact with our data. This article is about:

1. Why data integration is an important capability that data-driven companies should be thinking about.
2. How our platform solved our data integration needs.
3. An overview of how we built the platform itself.

Before delving in, I'll define *data integration* (sometimes known as *data flow*) as the mechanisms that transport data from A to B, such as batch load jobs, REST APIs, message buses, or even rsync. This definition of data integration encompasses many things that are not often considered data integration mechanisms, but by defining this broad category of fundamentally similar things, it's possible to devise singular solutions to solve many disparate problems. More on this later.

From the outset of the project, one requirement was very clear: *fix the logging*, as in, make it easier to collect and use our **log data**. How did we start with the requirement of improving our logging and end up with a data integration platform? To answer this, I defer to Jay Kreps. In his masterpiece **I <3 Logs**, he talks in depth about logs and their myriad of uses, specifically:

1. *Data integration:* Making all of an organization's data easily available in all its storage and processing systems.
2. *Real-time data processing:* Computing derived data streams.
3. *Distributed system design:* How practical systems can be simplified with a log-centric design.

Centralized logging infrastructure is crucial for achieving all of this. However, logging infrastructure alone is not useful to data scientists — deploying a Kafka cluster and calling it a day would do them little good. After many discussions, we decided to build new logging infrastructure *and* build a data integration platform on top of it to deliver the most value to the data scientists.

By this point, we knew what we wanted to build. We still had a pile of technologies to research and months of work ahead of us, but we had decided on a name: The Data Highway.

DESIGNING THE DATA HIGHWAY

First, we decided on three goals for the Data Highway:

1. Simplify, scale up, and centralize the logging infrastructure
2. Provide a robust, flexible, and extensible data integration platform
3. Allow for self-service and ease of use for data scientists

I'll delve into each goal to clarify what we wanted to accomplish.

STANDARDIZE THE LOGGING INFRASTRUCTURE

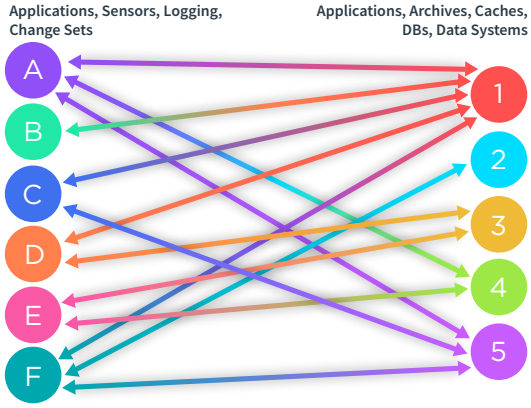
My team had already experimented with a few different approaches to logging. The logging infrastructure we were currently running was built using AWS Kinesis and Lambda functions, and it was developing some issues: it was difficult to integrate with non-AWS products, the lambda functions were proliferating and becoming too burdensome to maintain, we were running into throttling issues with Kinesis, and the system was too complex to allow for self-service for the data scientists. Furthermore, various ad hoc solutions to logging had sprouted up across the org,

and we had no best practices around how to log. Replacing the legacy logging infrastructure and cleaning up logging tech debt were priority number one.

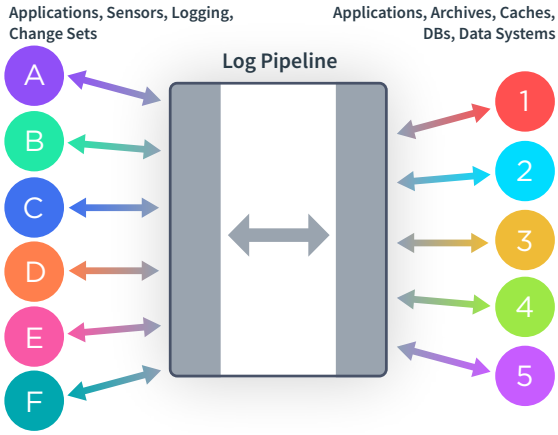
EXTENSIBLE DATA INTEGRATION

We wanted to make it trivial to get data to and from any application or system within Stitch Fix. How? By providing a centralized, standardized, robust, and extensible mechanism for transporting data from A to B.

To illustrate how this would work, first imagine you are a data scientist and you need data from the customer-facing landing page. Without a standard way to get that data, you'd probably build a one-off solution to transport the data. As companies grow, these one-off solutions proliferate. Eventually, this becomes a tangled mess that is exponentially more burdensome to maintain as more and more systems and applications are developed within the company.



The way out of this mire is to implement a centralized pipeline for transporting data. In this world, anything that produces or consumes data simply integrates with the main pipeline. This solution involves fewer connectors, and since each connector resembles one another, it's possible to develop an extensible abstraction layer to handle the integrations with the central pipeline.



This abstraction layer is what we strove to build with our data integration platform. Each connector would run together in the same infrastructure. They'd share common code and would be deployed, monitored, and maintained together. When new kinds of systems are rolled out, only one kind of connector would need to be developed to integrate the new system with the rest of the company. This cohesion of data makes it possible to have a high-level view of all the data within a company and makes mechanisms like data validation, stripping PII, streaming analytics, and real-time transforms much easier to build. This is tremendously valuable, particularly for data-driven companies.

SELF-SERVICE AND EASY TO USE

This is one of the central tenets of the data org at Stitch Fix. Our data scientists are wholly responsible for doing everything from data collection to cleaning, ETL, analysis, visualization, and reporting. We data platform engineers are never directly involved in that work. Instead, we focus on developing tooling to streamline and abstract the engineering elements of the data scientists' work. We are like a miniature AWS. Our products are finely crafted data science tools, and our customers are Stitch Fix's data scientists. This affords the data scientists complete autonomy to deliver impactful projects, and the data science teams can scale and grow without being limited by the cycles of the engineers. This relationship is harmonious and enables the modestly sized 20+ person data platform team to support all 80+ data scientists with relative ease.

So, to summarize, the Data Highway needed to be a centralized data integration platform backed by resilient logging infrastructure. It also needed to be simple and 100% self-service. Lastly, it needed to be robust, so the data scientists couldn't break it. This sounds daunting, but it was surprisingly achievable thanks to recent advancements in logging and data integration infrastructure.

THE IMPLEMENTATION

The Data Highway would have three major components:

1. The underlying logging infrastructure.
2. The abstraction layer to create the data integration platform.
3. The interface that data scientists would interact with.

I'll describe, on a high level, how each component is implemented and briefly discuss why we chose the technologies that we did.

THE LOGGING INFRASTRUCTURE

Currently, the most frequently adopted solutions for logging infrastructure are Kafka, Kinesis, and Google PubSub. We ruled out Google PubSub because integrating our AWS-based stack with a GCP product would be too costly. That left Kinesis vs. Kafka.

In the end, we chose Kafka despite the up-front cost of deploying and maintaining our own Kafka infrastructure. Although the ease of simply clicking a few buttons in Kinesis was tempting, the scales tipped in favor of Kafka because 1) we wanted something that would integrate well with AWS products and non-AWS products alike, 2) we wanted to improve our involvement in the open-source community, 3) we wanted fine grained control over our deployment, and 4) as far as data infrastructure goes, Kafka is quite mild-mannered, so we were optimistic that the maintenance burden would be manageable.

THE DATA INTEGRATION LAYER

For the data integration layer on top of Kafka, we considered a few options, including Apache Beam, Storm, Logstash, and Kafka Connect.

Apache Beam is a Swiss army knife for building data pipelines, and it has a feature called **Pipeline I/O**, which comes with a suite of built-in "transforms." This was close to what we wanted, but Beam, at the time, was still a very young project, and it is more oriented towards transformations and stream processing. The data integration features seemed like an add-ons feature.

Apache Storm was a viable option, but it would be tricky to build the self-service interface on top of it, since the data flow pipelines have to be written as code, compiled, and deployed. A system that relies on configuration as code would be better.

Logstash is handy because it has **so many integrations** — more than any other option we looked at. It's lightweight and easy to deploy. For many years, Logstash lacked important resiliency features, such as support for distributed clusters of workers and disk-based buffering of events. As of version 6.0, both of these features are now implemented, so Logstash could work well for a data integration layer. It's main drawback for our use case is that the data flows are configured using static files written in Logstash's (somewhat finicky) DSL. This would make it challenging to build the self-service interface.

Finally, we investigated **Kafka Connect**. Kafka Connect's goal is to make the integration of systems as easy, resilient, and flexible as possible. It is built using the standard Kafka consumer and producer, so it has auto load balancing, it's simple to adjust processing capacity, and it has strong delivery guarantees. It can run in a distributed cluster — so failovers are seamless. It is stateless because it stores all its state in Kafka itself. There is a large array of **open-source connectors** that can be downloaded and plugged in to your deployment. If you have special requirements for a connector, the API for writing your own connectors is simple and expressive, making it straightforward to write custom integrations. The most compelling feature is Kafka Connect's REST interface for configuring and managing connectors. This provided an easy way for us to build our self-service interface. So, we chose Kafka Connect to power the data integration layer, and it has worked well for us.

There were still a few gaps left to fill for this layer of the Data Highway. Kafka Connect works great when it is *pulling* data from other systems. Some of our data is *push*-based, however, like data sent using HTTP and syslog. So, we wrote lightweight services dedicated to ingesting that data. Also, some of the open-source Kafka Connect connectors didn't quite fit our use cases, so we wrote a couple of our own custom connectors.

One of the more challenging aspects of this piece of the Data Highway was monitoring and operational visibility. To solve this, we built a mechanism that sends special ***tracer bullet*** messages into every source that the Data Highway ingests data from. The tracer bullets then flow to every system the Data Highway can write data to. The tracer bullets flow continuously, and alerts fire when they are delayed. The tracer bullets provide a good high-level overview of the health of the system and are particularly helpful in our staging environment.

THE INTERFACE FOR DATA SCIENTISTS

The last thing to build was the Data Manager and Visualizer (a.k.a. the DMV, pun intended), which would be the self-service interface for data scientists to interact with the platform. We have a skilled frontend engineer on the data platform who is dedicated to building GUIs for our tools. Working together, we built the DMV, which shows every topic in the Data Highway, displays sample events per topic, and provides web forms for setting up new sources of data or configuring new sinks for the data to flow to.

In theory, the DMV could directly call the Kafka Connect APIs, but we chose to build a small abstraction layer, called CalTrans, in between Kafka Connect and the DMV. CalTrans converts simple and expressive REST requests into the bulky requests that Kafka Connect needs. CalTrans handles environment-specific default configs and hardcodes things that make sense to be configurable within Kafka Connect but that are static for our use cases. Lastly, CalTrans can talk to other systems besides Kafka Connect, so if some validation needs to be performed or some kind of miscellaneous administrative task needs to be completed upon configuring connectors, CalTrans performs that logic.

CONCLUSION

Hopefully, this has piqued your interest in data integration and given you some ideas for how to build your own data integration platform. Since building this system, I've developed an intense curiosity for these kinds of standardized integration systems and I see them everywhere now: USB ports, power outlets, airlines, DNS, train tracks, even systems of measurement. They are all simple, often regulated standardizations that enable many disparate things to interoperate seamlessly. Imagine how much potential could be unlocked, particularly by the hands of data scientists, if all the data a company had was as convenient to access as a power outlet? I believe proper data integration is the secret to unlocking this potential, and at Stitch Fix, this revolution is coming to fruition.

LIZ BENNETT has specialized in the realms of infrastructure and Big Data since 2013. She got her start on the News Feed team at LinkedIn, then moved on to Loggly, a SaaS-based log management company, where she was an integral contributor to their logging infrastructure. Now, she is a member of Stitch Fix's Data Platform team, where, in close collaboration with data scientists, she is leading efforts to evolve Stitch Fix's streaming and data integration infrastructure.

10 TIPS FOR ENSURING

YOUR NEXT DATA ANALYTICS PROJECT IS A SUCCESS

Data analytics can help you predict the future and make decisions with relative certainty. Yet, analytics has its challenges and can quickly drain your resources without leading to actionable results.

Here are 10 tips to ensure the success of your next analytics project.

BY WOLF RUZICKA, CHAIRMAN AT EASTBANC TECHNOLOGIES

1. ITERATE

Avoid an all-in, big bang approach – it's risky and costly. Try a faster results-based route that embraces agile and focuses on achieving Minimal Viable Predictions. Focus on the #1 problem you want to solve. Assemble data that correlates with that problem. Work backwards from there to identify other related data and follow the breadcrumbs that lead to actionable outcomes.

2. INVOLVE THE BUSINESS

Involve business stakeholders early. Only they will be able to assess the usefulness of the results, and their feedback will be key to correct course.

3. BUILD A ROBUST SYSTEM

If you build a fragile system, it can overshadow your entire project. The various layers of the system and their integrations need to be robust enough to minimize issues and frustration.

4. START WITH HISTORICAL DATA

It's tempting to plow into real-time data, however, historical data is easier to work with. Then, as you get into real-time, consider what it means for your business, is it the difference between three milliseconds or three days?

5. DO A REALITY CHECK

You can't have it all. You don't have the right tools, data, or skill set. Understand what's feasible and what's not. Don't overpromise, think first and commit later.

6. BE PATIENT

In 99% of cases, your first analytics iteration won't be meaningful. Weekly iterations, with feedback from stakeholders can lead to a result in two months. Failure is part of the process, learn from it. Focus on the right data, not just Big Data. The most valuable business insights are derived from surprisingly small data sets, which also minimizes risk.

7. CHOOSE BETWEEN REPORTING VS. AD-HOC DATA ANALYSIS

Consider the maturity of your efforts as you decide whether to conduct reporting (easily achieved with tools like Microsoft PowerBI) or more sophisticated ad-hoc analysis of real-time data, which requires more specialized tools and skills.

8. HOW WILL YOU VISUALIZE?

Bear in mind your users' unique data visualization preferences. Give them a sandbox, so they can determine what matters to them.

9. DON'T SKIP DATA GOVERNANCE

Can you trust your data? How accurate, clean, or well defined is it? Data cleaning is one of the most important aspects of analytics. Dirty and unreliable data leads to inconclusive results.

10. FEAR AND FIEFDOM

Data owners fear they'll lose relevance or control of data, if they are forced to share data sets with others. Counteract fiefdom through evangelism. Show how sharing data creates value across the board.

Big Data Realization with Sanitation

According to 63% of survey respondents, unsanitized data is by far the biggest challenge for data science work. Unsanitized or "dirty" data can create huge problems for analytics projects by introducing irrelevant or incorrect data to the database. In addition to the difficulties of dealing with dirty data, DZone readers have also had issues with the volume of complex data types like documents and server logs as well as the speed at which user-generated data is collected. Given these difficulties, we wanted to touch on the importance of cleaning and sanitizing data, and why the benefits are so important to any organization pursuing Big Data analytics.

Data sanitation and cleansing refers to the detection and removal of unwanted information from a data set. This can be sensitive data, like passwords, corrupted, or incorrect information. Without clean data, you may lack standardization and end up with less useful insights. Data sanitation is even more difficult when dealing with documents and log files, two data types that give developers a great deal of frustration.

Luckily, there are several ways to sanitize data. One common technique is to sanitize your inputs. Building off of the previous example, rather than allowing users to enter a US state themselves, they can select a choice from a dropdown menu. Other techniques include filters that are applied as data is collected that ensure the data follows a specific structure or removes duplicate information.

One of the biggest indicators of success for an analytics project is the speed at which decisions are made (48% of survey respondents). With clean data, less time can be spent making decisions and more time can be spent acting on them. You can move even faster if you use automation, or if user inputs are sanitized. It helps ensure that you're recording accurate and correct information and that you're not wasting storage space on irrelevant information.

Why Developers Should Bet Big on Streaming

BY JONAS BONÉR
CTO AND CO-FOUNDER, LIGHTBEND

QUICK VIEW

01. Streaming is the backbone of the most interesting AI/ML use cases today.

02. In the latest wave of big data, events are streamed continuously with the presumption that they are unbounded, so systems need to process it on the fly, as it arrives.

03. With streaming, the fundamental shift is moving from "data at rest" to "data in motion" — from batched to real-time.

04. The convergence of streaming and microservices is bringing all the power of streaming and "data in motion" into the microservices themselves — both as a communication protocol and as a persistence solution.

To many developers, real-time use cases are viewed as the exclusive playground of mega Internet players like Amazon and Netflix with their infinite resources, gigantic customer bases, and armies of PhD engineers. (Real-time means different things to different people. Here we define it as "process on arrival" to ensure low latency and fast time to value.)

I'm here to tell you that real-time is no longer an elite opportunity. Most enterprises today are dealing with dynamics that are pushing real-time requirements. Consider all the data that your web and mobile users are generating and requesting. How can you gather all this data, transform it, mine insight, and feed it back to your users in real-time (or as close to as possible) — delivering efficiencies in conversions, retention, and automation of common operational processes? The volume of data being generated by your users on the web, mobile, and (increasingly) IoT devices is growing vastly every year, and this is an opportunity that your company can't afford to miss out on.

YOU CAN'T DO REAL-TIME WITHOUT STREAMING

Streaming is the backbone of the most interesting machine learning and artificial intelligence use cases in the enterprise today. Let's discuss how we got from the early big data origins to where we are today with streaming: fast data.

BIG DATA WAVE ONE: HADOOP/BATCH

The first wave of big data was the Hadoop ecosystem, with HDFS, MapReduce, and friends (i.e. Hive, Tez, Mahout, Pig, YARN, HBase, Avro, ZooKeeper, and Flume). Store your data in HDFS and then perform batch processing of that data overnight — with hours of latency, i.e. hours of lead time to get insight and intelligence out of your data.

In these early days, big data was equal to Hadoop. But the original components of Hadoop were rooted in batch-mode and offline processing approaches commonplace for decades. This first wave of Hadoop/Batch had a lot in common with how most search engines worked in the early days, where data is captured to storage and then processed periodically with batch jobs.

Worth mentioning is that Apache Spark is one of the projects to successfully challenge the traditional batch model with what they called mini batching, which maintains the model of batching, but with increased batch frequency and thereby lower latency.

BIG DATA WAVE TWO: LAMBDA ARCHITECTURE

In the second wave, we saw that the need to react in real time to "data in motion" — to capture the live data, process it, and feed the result back into the running system within seconds-long (even sub-seconds) response time — had become increasingly important.

This need instigated hybrid architectures such as the lambda architecture, which had two layers — the speed layer for real-time online processing and the batch layer for more comprehensive

offline processing — where the result from the real-time processing in the "speed layer" was later merged with the "batch layer."

This model solved some of the immediate need for reacting quickly to (at least a subset of) the data. The downside was that it added needless complexity with the maintenance of two independent models and data processing pipelines, as well as an automated data merge in the end.

BIG DATA WAVE THREE: FULL EMBRACE OF STREAMING

During the second wave we started to realize that streaming could do most of our processing in real-time. This led to the third (at the time of writing: current) wave of big data: the move to a full embrace of streaming (sometimes referred to as the kappa architecture, as an evolution of lambda architecture).

In this new model, events are streamed continuously, with the presumption that they are unbounded — that they may never end — so systems can no longer wait to receive "all the data," and instead need to process it on the fly, as it arrives. This calls for new techniques: it's no longer possible to get a comprehensive view on all the data before processing it, but you need to define your window of processing — how you should group the incoming events and where to "draw the line" before processing the group — and distinguish between event time (when it's happening) versus processing time (when it's being processed).

With streaming, the fundamental shift is moving from "data at rest" to "data in motion" — from batched to real-time.

SIGNS THAT STREAMING DATA HAS ARRIVED

If you're looking for more evidence that streaming is winning out as the preferred method for how applications and systems interact with real-time data, here are some additional signs of that process:

STREAMING HAS A THRIVING ECOSYSTEM

It's a buyer's market for stream processing engines: Flink, Spark Streaming, Akka Streams, Kafka Streams, Storm, Cloud Dataflow (Google), Pulsar (Yahoo), Pravega (EMC), and others. It's a sign of maturity for streaming that more engines are being released, each with their specific use-cases, focus, and edge.

ARCHITECTURES FOR REAL-TIME STREAMING AND MICROSERVICES ARE CONVERGING

As we see more microservices-based systems grow to be dominated by data, their architectures are starting to look like big pipelines of streaming data. The convergence of streaming and microservices is bringing all the power of streaming and "data in motion" into the microservices themselves — both as a communication protocol as well as a persistence solution (using event logging) — including both client-to-service and service-to-service communication. Thinking in streams-as-values also forms the basis for designing your microservices around domain events, and so-called events-first domain-driven design.

STREAMING IS BECOMING THE NEW INTEGRATION

A growing sentiment within the industry is the need to rethink Enterprise Integration (EIP) in terms of streams as a first-class concept — specifically, to be able to think about streams as values, manipulate streams, join streams, or split them apart in a fully asynchronous and non-blocking fashion, with flow-control (backpressure) throughout. An emerging standard for this is the Reactive Streams initiative (now included in JDK9 as the Flow API).

TIME TO INVEST IN STREAMING

When you look at your career as a developer, a lot of the defining moments are based on how early you recognize an opportunity, embrace a new set of technologies, and where you decide to invest your time. Choosing tools of the trade and where to focus energies are the biggest bets you place as a developer.

Putting streaming at the top of your to-do list (to get invested in ASAP) is a bet you should absolutely consider making in 2018— for your career, your project, your business, and for fun.

It's easy to forget, but the web used to be a new concept, too — and now, not only does every company embrace it, but the web became so popular that web applications became the new standard and largely killed off the desktop software market. That's the sort of transformative impact that real-time streaming will have on application and system design long-term. Streaming is not a trend, it's a requirement — and by embracing it in your system design now, you will get ahead on the real-time adoption curve and take advantage of some of the most interesting new opportunities available to modern developers.

JONAS BONÉR is the founder and CTO of Lightbend, inventor of the Akka project, initiator and co-author of the Reactive Manifesto, and a Java Champion. Bonér is passionate about technology, learning new things, and helping others to work better and more effectively.

ENJOY EVERY STEP OF YOUR DATA JOURNEY WITH DATAIKU.

From analytics at scale to enterprise AI, Dataiku connects people, technologies, and processes to remove roadblocks and help businesses along their data journey. With a centralized, controlled data environment that's a common ground for both experts and explorers as well as a repository of best practices, Dataiku provides a shortcut to deployment and model management that will spark change for a data-powered company.

data iku

WWW.DATAIKU.COM

On the Data Value Problem & Data Science Tools

Today's businesses collect more data from more sources than ever before. From weblogs to transactional data, the Internet of Things (IoT), and everything in between, I'd say it's fair to assume that no company is struggling with data quantity.

But I do think many companies, whether they like to admit it or not, have a data value problem. That is, they struggle to gain real business value from all of the data (or any of the data) they're collecting.

These challenges point to the need to invest (or reinvest) in data teams via:

- **Staff**: Coders and non-coders alike need a way to meaningfully contribute to data projects.
- **Structure**: Clear, reproducible workflows, and a way for team leaders to monitor data projects.
- **Efficiency**: A faster way to clean and wrangle data.
- **Automation**: To alleviate the inefficiencies of rebuilding and retraining models.
- **Deployment Strategy**: Efficient means to deploy data projects into production quickly.

Data science tools or platforms are the underlying framework that allows data teams to be more productive in all of these areas. They allow for easy (but controlled) access to data, keep all work centralized (and thus reproducible), and facilitate critical collaboration not only among similar profiles but across different — and typically complementary — ones as well (data scientists, business/data analysts, IT, etc.).

Perhaps most importantly, data science platforms open up the door to true data innovation when teams don't have to spend precious time on administrative, organizational, or repetitive tasks, and it is through this data innovation that enterprises can truly start to gain value from their raw data.

WRITTEN BY FLORIAN DOUETTEAU
CEO & CO-FOUNDER, DATAIKU

PARTNER SPOTLIGHT

Dataiku

Dataiku moves businesses along their data journey to AI, removing roadblocks while providing the structure and stability needed to get there.

CATEGORY

Software Platform for Advanced Analytics, Machine Learning, and AI.

RELEASE SCHEDULE

Two major releases per year with minor releases in between

OPEN SOURCE?

No, but leverages open-source solutions.

CASE STUDY

In an effort to continue to grow their business in existing and new markets, DAZN - a subscription sports streaming service - wanted a fast, low-maintenance way to enable their small data team to run predictive analytics and machine learning projects at scale. DAZN knew that in order to accomplish their goals quickly, they would need simple technologies in the cloud. They turned to Amazon Web Services (AWS) and Dataiku in combination for their simplicity in setup, connection, integration, and usability. With AWS and Dataiku, the small data team built and now manages more than 30 models in parallel, all without needing to do any coding so that the processes are completely accessible to non-technical team members. Now, each Data Team member is 2.5x more efficient in deploying models. AWS and Dataiku have noticeably shifted the data culture at DAZN and have brought innovations in advanced analytics and machine learning throughout the company.

STRENGTHS

- Scalable
- For all data team members
- Supports open-source
- Brings reproducibility

NOTABLE CUSTOMERS

- GE
- SEPHORA
- KUKA
- Unilever
- Samsung

WEBSITE dataiku.com **TWITTER** @dataiku **BLOG** blog.dataiku.com

Introduction to Basic Statistics Measurements

BY SUNIL KAPPAL
ADVANCED ANALYTICS CONSULTANT

QUICK VIEW

01. Learn about the most common statistical methods that can help make data-driven decisions.

02. Understand the fundamental concepts of statistics in an easy-to-follow, jargon-free way.

03. Discover how to calculate variance and standard deviation and learn why they are useful figures.

Statistics consists of a body of methods for collecting and analyzing data (Agresti & Finlay, 1997).

It is clear from the above definition that statistics is not only about the tabulation or visual representation of data. It is the science of deriving insights from data, and can be numerical (quantitative) or categorical (qualitative) in nature. Briefly, this science can be used to answer questions like:

- How much data is enough to perform a statistical analysis?
- What kind of data will require what sort of treatment?

Methods to draw the golden nuggets out of the data:

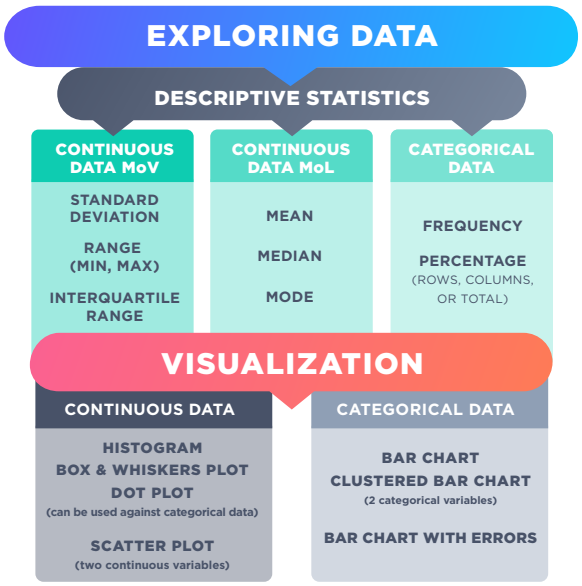
- Summarizing and exploring the data to understand the spread of the data, its central tendency, and its measure of association using various descriptive statistical methods.
- Drawing inferences from, forecasting, and generalizing the patterns displayed by the data to make some sort of conclusion.
- Furthermore, statistics is the art and science of dealing with events and phenomenon that are not certain in nature. They are used in every field of science.

The goal of statistics is to gain an understanding of data. Any data analysis should have the following steps:

1. Identify the research problem.
2. Define the sample and population.
3. Collect the data.
4. Perform descriptive and inferential analysis.
5. Use statistical research methods.
6. Solve the research problem.

DESCRIPTIVE STATISTICS

Descriptive statistics are summary statistics that quantitatively describe data using measures of central tendency and measures of variability or dispersion. The table below depicts the most commonly used descriptive statistics and visualization methods:



Descriptive statistics provide data summaries of the sample, along with the observations that have been made with regards to the sample data. Such summaries can be presented in the

form of summary statistics (refer to the above visual for summary statistics types by data type) or easy-to-decipher graphs.

It is worth mentioning here that descriptive statistics are mostly used to summarize values and may not be sufficient to make conclusions about the entire population or to infer or predict data patterns.

Until now, we have just looked at the descriptive statistics that can be used to explore data by data type. This section of the article will help readers appreciate the nuances involved in using summary statistics.

It is worth mentioning here that descriptive statistics are mostly used to summarize values and may not be sufficient to make conclusions about the entire population or to infer or predict data patterns.

MEANS AND AVERAGES

The term "mean" or "average" is one of the many summary statistics that can be used to describe the central tendency of the sample data. Computing this statistic is pretty straightforward: add all the values and divide the sum by the number of values to get the mean or average of the sample.

Example: The mean (or average) is (1+1+2+3+4)/5 = 2.2.

MEDIAN

A median is the value separating the higher half of a sample dataset from the lower half. The median can also be expressed as another way of finding the average of the sample data by sorting the number list from low to high and then finding the middle digit within the number list.

Example (one number in the middle):

Number list = 3, 2, 4, 1, 1.

Step 1: Sort the number list low to high = 1, 1, 2, 3, 4.

Step 2: Find the middle digit = 1, 1, **2**, 3, 4.

Median = 2.

In the previous section, the mean/average for the same dataset it was 2.2. However, the central tendency for the dataset using the median statistic is 2.

Example (two numbers in the middle):

With an even amount of numbers, things get little tricky. In this case, we have to identify the middle pair of numbers and then find the value that is halfway between them. This can be easily done by adding them together and dividing them by two.

Let's look at an example where we have fourteen numbers, and we don't have just one middle number; we have a pair of middle numbers.

Number list = 3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29.

Step 1: Sort the number list from low to high = 3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56.

As stated previously, we don't have just one middle number; we have a pair of middle numbers:

3, 5, 7, 12, 13, 14, **21, 23**, 23, 23, 23, 29, 40, 56.

In the above example, the middle numbers are 21 and 23. To find the value halfway between them, add them together and divide by 2: 21 + 23 = 44 ÷ 2 = 22.

Median = 22.

(Note that 22 is not in the number list, but that is okay because half of the numbers in the list are less than 22 and half of the numbers are greater than 22.)

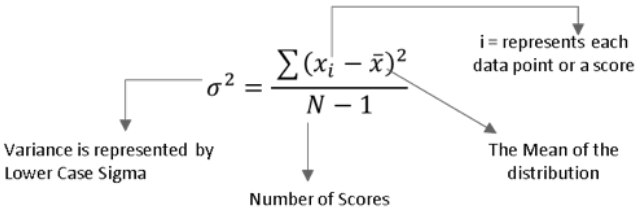
VARIANCE/DELTA AND STANDARD DEVIATION

In my twenty years of experience, I have seen people use these two terms interchangeably when summarizing a dataset — which, in my opinion, is not only incorrect but also dangerous. You can call me a purist, but this is the distinction that we have to make to understand the difference between these two statistics.

Just to clear the air, I will take a step back and try to define these two terms in simple English and in statistical terms.

Variance can be defined as the average of the squared differences from the mean. Variance can be the difference between an expected result and an actual result, such as between a budget and an actual expenditure.

VARIANCE FORMULA DEMYSTIFIED



I know that the above formula can be pretty daunting. Therefore, I will list out the steps to calculate the variance in an easy-to-understand manner:

1. Work out the mean.
2. Then, for each number, subtract the mean and square the results (squared difference).
3. Work out the averages of those squared values.

If you're still unclear about all the math, let's look at it visually to understand how to calculate the variance using a dataset.

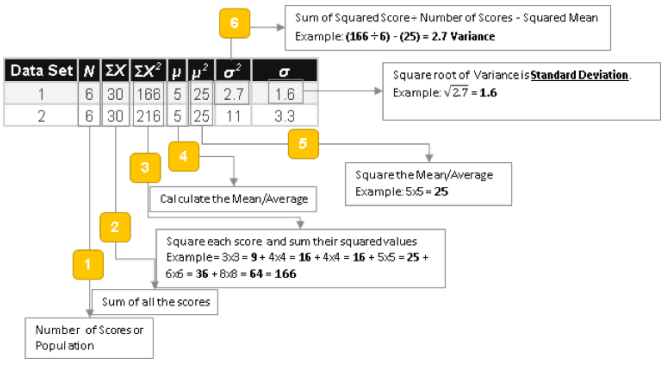
Note: The square root of variance is called the standard deviation. Just like variance, standard deviation is used to describe the spread. Statistics like standard deviation can be more meaningful when expressed in the same units as the mean, whereas the variance is expressed in squared units.

CALCULATING VARIANCE

As an example, let's look at two distributions and understand the step-by-step approach to calculating variance:

DATA SET 1	DATA SET 2
3	1
4	2
4	4
5	5
6	7
8	11

How it works:



The above example helps us appreciate the intricacies involved in calculating the variance statistics, vis a vis stating that the variance or delta is the same thing.

Delta can be defined as a change or a difference in percentage, where we simply subtract the historic value from the most recent value and divide it with the recent value to get a Delta percent (some people call it variance percent, as well).

Example: (255 – 234)/234 = 9% (Delta %), and if we do not divide this with the recent number, we will get the pure delta = 21.

Variance can be defined as the average of the squared differences from the mean. Variance can be the difference between an expected result and an actual result, such as between a budget and an actual expenditure.

STANDARD DEVIATION EXPLAINED

Standard deviation is defined as "the deviation of the values or data from an average mean."

Standard deviation helps us know how the values of a particular dataset are dispersed. A lower standard deviation concludes that the values are very close to their average, whereas higher values mean that the values are far from the mean value. Standard deviation values can never be negative.

Interpreting σ, standard deviation, using Chebyshev's rule for any population:

- At least 75% of the observations will lie within 2σ of μ.
- At least 89% of the observations will lie within 3σ of μ.
- At least 100(1 – 1/m2)% of the observations will lie within mxσ of μ.

Things to remember about standard deviation:

- Use it when comparing unlike measures.
- It is the most common measure of spread.
- Standard deviation is the square root of the variance.

SUNIL KAPPAL works as an Advanced Analytics Consultant. He has more than 20 years of experience in the fields of data analytics, business intelligence, statistical modeling, predictive models, and Six Sigma methodologies. Apart from delivering guest lectures and presentations on advanced analytics techniques, Sunil also writes guest blog posts for leading IT and analytics blogs. His book Statistics for Rookies has received a lot of appreciation by industry leaders like Carla Genry and Kirk Bourne. Sunil has also authored various technical papers on the use of speech analytics to identify fraud by combining Naïve Bayes classification methods in conjunction with Benford's Law, a mathematical technique to identify financial fraud.

DIVING DEEPER

INTO BIG DATA

#BIG DATA TWITTER ACCOUNTS

 @BigDataGal

 @KirkDBorne

 @drob

 @Ronald_vanLoon

 @craigbrownphd

 @marcusborba

 @kinlane

 @karpathy

 @HansMichielscom

 @schmarzo

BIG DATA ZONES

Big Data Zone
dzone.com/big-data
The Big Data/Analytics Zone is a prime resource and community for Big Data professionals of all types. We're on top of all the best tips and news for Hadoop, R, and data visualization technologies. Not only that, but we also give you advice from data science experts on how to understand and present that data.

Database Zone
dzone.com/database
The Database Zone is DZone's portal for following the news and trends of the database ecosystems, which include relational (SQL) and nonrelational (NoSQL) solutions such as MySQL, PostgreSQL, SQL Server, NuoDB, Neo4j, MongoDB, CouchDB, Cassandra and many others.

AI Zone
dzone.com/ai
The Artificial Intelligence (AI) Zone features all aspects of AI pertaining to Machine Learning, Natural Language Processing, and Cognitive Computing. The AI Zone goes beyond the buzz and provides practical applications of chatbots, deep learning, knowledge engineering, and neural networks.

BIG DATA REFCARDZ

Apache Kafka
Download this new Refcard to get started with Apache Kafka. Learn more about its background, steps to run the quickstart script, components, and how to connect REST APIs.

Recommendations Using Redis
In this Refcard, learn to develop a simple recommendation system with Redis, based on user-indicated interests and collaborative filtering. Use data structures to easily create your system, learn how to use commands, and optimize your system for real-time recommendations in production.

Querying Graphs With Neo4j
Armed only with a pattern and a set of starting points, graph databases explore the larger neighborhoods around these initial entities — collecting and aggregating information from millions of nodes and relationships — but leaving the billions outside the search perimeter untouched.

BIG DATA BOOKS

Data Analytics Made Accessible by Dr. Anil Maheshwari
In this free e-book version updated for 2018, learn about use cases and real-world stories in the world of data science.

Too Big to Ignore: The Business Case for Big Data by Phil Simon
A daily podcast that covers test automation, A/B testing, testing in DevOps, and more.

Data Smart by John W. Foreman
How exactly do you do data science and what does it mean? Get familiar with this sometimes-intimidating prospect.

BIG DATA PODCASTS

IBM Analytics Insights Podcasts:
Get the latest in big data and analytics, and learn about their implications for the enterprise from a range of experts in multiple industries.

Data Skeptic:
Learn about high-level concepts in data science and listen to interviews with big data researchers and practitioners.

Linear Digressions:
Get insights on real-life data science applications, data science career paths, and reports on the state of data science.

Executive Insights on the State of Big Data

QUICK VIEW

01. Know the specific business problem you are trying to solve to have a successful big data business strategy. Start small and scale.

02. Focus on delivering value quickly by aggregating data from different sources and in different formats to provide unique insights that add business value.

03. While it takes a while to get a big data initiative up and running, once it is you can see a huge reduction in expenses and increases to the bottom line.

BY TOM SMITH
RESEARCH ANALYST AT DZONE, INC.

To gather insights on the state of Big Data in 2018, we talked to 22 executives from 21 companies who are helping clients manage and optimize their data to drive business value. Here's who we spoke to:

RESPONDENTS

- **Emma McGrattan**, S.V.P. of Engineering, Actian
- **Neena Pemmaraju**, VP, Products, Alluxio Inc.
- **Tibi Popp**, Co-founder & CTO, Archive360
- **Laura Pressman**, Marketing Manager, Automated Insights
- **Sébastien Vugier**, SVP, Ecosystem Engagement & Vertical Solutions, Axway
- **Kostas Tzoumas**, Co-founder & CEO, Data Artisans
- **Shehan Akmeemana**, CTO, Data Dynamics
- **Peter Smails**, V.P. of Marketing & Business Development, Datos IO
- **Tomer Shiran**, Founder and CEO, Dremio
- **Kelly Stirman**, CMO, Dremio
- **Ali Hodroj**, V.P. Products & Strategy, GigaSpaces
- **Flavio Villanustre**, CISO & V.P. of Technology, HPCC Systems
- **Fangjin Yang**, Co-founder & CEO, Imply
- **Murthy Mathiprakasam**, Dir. of Product Marketing, Informatica
- **Iran Hutchinson**, Product Mgr. & Big Data Analytics Software/ Systems Architect, InterSystems
- **Dipti Borkar**, V.P. of Products, Kinetica
- **Adnan Mahmud**, Founder & CEO, LiveStories
- **Jack Norris**, S.V.P. Data & Applications, MapR
- **Derek Smith**, Co-founder & CEO, Naveego
- **Ken Tsai**, Global V.P., Head of Cloud Platform & Data Mgmt., SAP
- **Clarke Patterson**, Head of Product Marketing, StreamSets
- **Seeta Somagani**, Solutions Architect, VoltDB

KEY FINDINGS

1. There are several keys to having a successful big data strategy: **1) know the business problem you are trying to solve; 2) governance and operations; 3) strategy and structure: 4) speed of delivery; and, 5) people**.

Start with key use cases that can benefit from big data technology before moving to broad, organization-wide projects. Be driven by a pragmatic approach, look at the requirements and the problems you want to solve. Focus on outcomes and results.

Data operations must ensure that data is moving across the enterprise securely and transparently. Back-up, recovery, and protection are critical with the growth of ransomware and the criticality of data for business. Data governance becomes extremely important since data is sensitive and must be guarded appropriately.

Identify the data fabric strategy you are going to use to pursue new technologies – multi-cloud, hybrid processes, and microservices. Build common practices and architecture framework around the concept of a data lake. Provide fast-data processing and real-time analytics. Empower management to make informed decisions in real time.

There is a talent factor which must be considered to make any initiavtive successful. Have the right people in place that understand both the technology and the business goals. Provide resources for non-technical people to clean, work with, and garner insights from data.

2. We asked how companies can get more out of big data, since it seems like they are not seeing the return or business value that was initially anticipated. Several responses spoke to the difficulty, complexity, and time required to implement a big data initiative. This is consistent with the wisdom expressed by the late economist Rudi Dornbusch, things take longer to happen than you think they will, but then they happen much faster than you thought they ever could.

The keys for companies to get more out of big data more quickly is to: **1) focus on delivering value with data quickly; 2) use the cloud and new toolsets to accelerate the process; and, 3) identify specific business problems to solve**.

Unify and process data in real-time from disparate data stores to provide unique insights. Place greater near-term priority on carrying out investigations with business models that deepen insights into market segments and geographies.

Recognize data resources for their highest uses and ingest into business operations to take more intelligent action to drive topline revenue growth. This requires real-time operations, analytics, and transactional integration.

3. While the uptake of big data projects may be slow, based on the case studies shared by our respondents, **it is being successfully implemented in at least 10 vertical industries** with financial services, healthcare, and retail being the most prevalent. There are myriad applications including: 1) reduced fraud through more precise detection; 2) more targeted, relevant sales and marketing activities; 3) improved customer experience and proactive churn detection; 4) analysis of IoT data using machine learning; and, 5) improved compliance with regulations through proactive identification of non-compliant activity.

4. The most common issues preventing companies from realizing the benefits of big data are: **1) inability to evolve from legacy technology; 2) insufficient knowledge of big data and skillset; 3) failure to define the business problem to solve; and, 4) data quality and management**.

Unwillingness to embrace the cloud and not realizing is it not feasible to keep supporting legacy enterprise systems is a major issue. Some businesses want to use existing infrastructure rather than setting up the right backbone infrastructure with storage, transport, compute, and failover capabilities.

The people aspect is real, as expertise is necessary to understand the best technology needed to get the most from data. Companies don't know where to start, and they need someone on board who knows how to approach big data projects.

You need to start with a defined purpose in mind. Identify the application and the use cases and work from there. Understand that big data analytics are sets of tools and technologies that must be selected and applied for measurable outcomes.

Inability to get heads around the data and the need move the data from storage to compute and back again as needed is the last issue. Preparing and cleaning the data can take weeks, which leaves insufficient time for analytics. Data lakes become data swamps thanks to data that is inaccurate, incomplete, and without context.

5. The biggest opportunities in the continued evolution of big data are **using artificial intelligence (AI), machine learning (ML), and cognitive learning (CL) to provide higher level services that drive business value**. Healthcare will use AI/ML for disease diagnosis detecting patterns humans never could. There will be more opportunities to use AI/ML to augment business resources while providing self-service to more users. Businesses will succeed by using AI/ML to make customers' lives simpler and easier. There will be a natural evolution in AI/ML voice interfaces that reduce the friction with which people interact with machines.

Companies will be more responsive to customers in real-time. All data types and sources will be integrated to provide real-time intelligence, and that data will be the number one business driver in every industry. Data protection and privacy by design will be fully integrated. There will be greater speed and reuse of data management processes with greater trust in the data and less manpower required to manage it.

6. The only concern regarding the state of big data that was mentioned by just four of 21 respondents was security. **Security becomes more important as we become more reliant on data and it becomes more distributed**. Security is a function of human failure to follow best practices. Ultimately this will be automated to reduce threats. While security is a big deal, some feel it cannot be regulated. Consumers may stop doing business with companies who are deemed to be unsecure or unethical.

7. As usual, our respondents suggested a breadth of topics about what developers must be knowledgeable. **Understanding the business problem you are working to solve was most frequently mentioned, followed by an understanding of deployment architectures and data**.

Understand the use case and figure out the best solution stack to achieve your goals. Have a clear understanding of the range of business objectives within the company and how those align with the capabilities of various technologies, as well as the business value of the datasets you are working with.

Be cognizant of the cloud, microservices, geographic distribution, and security. Leverage the data fabric to simplify processes, and learn about open source options for AI/ML, such as Apache Spark.

Understand the basic data vocabulary of structure, dimension, and variables. Know that data is decentralized and distributed by nature. Know how to work with data at scale, how to handle the concurrency of multiple users. Understand how the data ecosystem works. While it's rare to find individuals with a perfect combination of skills, certain toolsets and systems can alleviate the need for serious programming experience, help with the data modelling part and even reduce the reliance on deep understanding of the mathematical models behind the predictions.

TOM SMITH is a Research Analyst at DZone who excels at gathering insights from analytics—both quantitative and qualitative—to drive business results. His passion is sharing information of value to help people succeed. In his spare time, you can find him either eating at Chipotle or working out at the gym.

Solutions Directory

This directory of Big Data and analytics frameworks, languages, platforms, and services provides comprehensive, factual comparisons of data gathered from third-party sources and the tool creators' organizations. Solutions in the directory are selected based on several impartial criteria, including solution maturity, technical innovativeness, relevance, and data availability.

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
1010data	Insights Platform	Data management, analysis, modeling, reporting, visualization, and RAD apps	Available by request	1010data.com/products/insights-platform/analysis-modeling
Actian	Vector	DBMS, column store, analytics platform	30 days	actian.com/analytic-database/vector-smp-analytic-database
Aginity	Aginity Amp	Data analytics management platform	Demo available by request	aginity.com/amp-overview
Alation	Alation	Enterprise data collaboration and analytics platform	Demo available by request	alation.com/product
Alluxio Open Foundation	Alluxio	Distributed storage system across all store types	Open source	alluxio.org
Alpine Data	Alpine Chorus 6	Data science, ETL, predictive analytics, and execution workflow design and management	Demo available by request	alpinedata.com/product/
Alteryx	Alteryx Analytics Platform	ETL, predictive analytics, spatial analytics, automated workflows, reporting, and visualization	Available by request	alteryx.com/products/alteryx-designer
Amazon Web Services	Amazon Kinesis	Stream data ingestion, storage, query, and analytics PaaS	N/A	aws.amazon.com/kinesis
Amazon Web Services	Amazon Machine Learning	Machine learning algorithms-as-a-service, ETL, data visualization, modeling and management APIs, batch and realtime predictive analytics	N/A	aws.amazon.com/machine-learning
Apache Foundation	Ambari	Hadoop cluster provisioning, management, and monitoring	Open source	ambari.apache.org
Apache Foundation	Apex	Stream and batch processing on YARN	Open source	apex.apache.org
Apache Foundation	Avro	Data serialization system (data structure, binary format, container, RPC)	Open source	avro.apache.org
Apache Foundation	Beam	Programming model for batch and streaming data processing	Open source	beam.apache.org
Apache Foundation	Crunch	Java library for writing, testing, running MapReduce pipelines	Open source	crunch.apache.org

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
Apache Foundation	Drill	Distributed queries on multiple data stores and formats	Open source	drill.apache.org
Apache Foundation	Falcon	Data governance engine for Hadoop clusters	Open source	falcon.apache.org
Apache Foundation	Flink	Streaming dataflow engine for Java	Open source	flink.apache.org
Apache Foundation	Flume	Streaming data ingestion for Hadoop	Open source	flume.apache.org
Apache Foundation	Giraph	Iterative distributed graph processing framework	Open source	giraph.apache.org
Apache Foundation	GraphX	Graph and collection processing on Spark	Open source	spark.apache.org/graphx
Apache Foundation	GridMix	Benchmark for Hadoop clusters	Open source	hadoop.apache.org/docs/r1.2.1/gridmix.html
Apache Foundation	Hadoop	MapReduce implementation	Open source	hadoop.apache.org
Apache Foundation	Hama	Bulk synchronous parallel (BSP) implementation for big data analytics	Open source	hama.apache.org
Apache Foundation	HAWQ	Massively parallel SQL on Hadoop	Open source	hawq.incubator.apache.org
Apache Foundation	HDFS	Distributed file system (Java-based, used by Hadoop)	Open source	hadoop.apache.org
Apache Foundation	Hive	Data warehousing framework on YARN	Open source	hive.apache.org
Apache Foundation	Ignite	In-memory data fabric	Open source	ignite.apache.org
Apache Foundation	Impala	Distributed SQL on YARN	Open source	impala.apache.org
Apache Foundation	Kafka	Distributed pub-sub messaging	Open source	kafka.apache.org
Apache Foundation	MADlib	Big data machine learning in SQL	Open source	madlib.apache.org/
Apache Foundation	Mahout	Machine learning and data mining on Hadoop	Open source	mahout.apache.org
Apache Foundation	Mesos	Distributed systems kernel (all compute resources abstracted)	Open source	mesos.apache.org
Apache Foundation	Oozie	Workflow scheduler (DAGs) for Hadoop	Open source	oozie.apache.org
Apache Foundation	ORC	Columnar storage format	Open source	orc.apache.org
Apache Foundation	Parquet	Columnar storage format	Open source	parquet.apache.org

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
Apache Foundation	Phoenix	OLTP and operational analytics for Apache Hadoop	Open source	phoenix.apache.org
Apache Foundation	Pig	Turns high-level data analysis language into MapReduce programs	Open source	pig.apache.org
Apache Foundation	Samza	Distributed stream processing framework	Open source	samza.apache.org
Apache Foundation	Spark	General-purpose cluster computing framework	Open source	spark.apache.org
Apache Foundation	Spark Streaming	Discretized stream processing with Spark's RDDs	Open source	spark.apache.org/streaming
Apache Foundation	Sqoop	Bulk data transfer between Hadoop and structured datastores	Open source	sqoop.apache.org
Apache Foundation	Storm	Distributed realtime (streaming) computing framework	Open source	storm.apache.org
Apache Foundation	Tez	Dataflow (DAG) framework on YARN	Open source	tez.apache.org
Apache Foundation	Thrift	Data serialization framework (full-stack)	Open source	thrift.apache.org
Apache Foundation	YARN	Resource manager (distinguishes global and per-app resource management)	Open source	hadoop.apache.org/docs/r2.7.1/hadoop-yarn/hadoop-yarn-site/YARN.html
Apache Foundation	Zeppelin	Interactive data visualization	Open source	zeppelin.apache.org
Apache Foundation	ZooKeeper	Coordination and state management	Open source	zookeeper.apache.org
Attunity	Attunity Visibility	Data warehouse and Hadoop data usage analytics	Demo available by request	attunity.com/products/visibility
Attunity	Attunity Replicate	Data replication, ingestion, and streaming platform	Available by request	attunity.com/products/replicate
BigML	BigML	Predictive analytics server and development platform	N/A	bigml.com
Bitam	Artus	Business intelligence platform	Available by request	bitam.com/artus
Board	BOARD All in One	BI, analytics, and corporate performance management platform	Demo available by request	board.com/en/product
CAPSENTA	Ultrawrap	Database wrapper for lightweight data integration	Available by request	capsenta.com
Cask Data	Cask	Containers (i.e. data, programming, application) on Hadoop for data lakes	Demo available by request	cask.co
Cask Data	Cask Data App Platform	Analytics platform for YARN with containers on Hadoop, visual data pipelining, and data lake metadata management	Free tier available	cask.co/products/cdap

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
Cazena	Cazena	Cloud-based data science platform	Demo available by request	cazena.com/what-is-cazena
Chart.js	Chart.js	Simple JavaScript charting library	Open source	chartjs.org
Cirro	Cirro Data Cloud	Database management system for cloud databases	Demo available	cirro.com/#/#product
Cisco	Cisco Edge Fog Fabric	IoT and streaming data analytics	N/A	cisco.com/c/en/us/products/analytics-automation-software/edge-analytics-fabric
Cloudera	Cloudera Enterprise Data Hub	Predictive analytics, analytic database, and Hadoop distribution	Demo available by request	cloudera.com/products/enterprise-data-hub.html
Confluent	Confluent Platform	Data integration, streaming data platform	Free tier available	confluent.io/product
D3.js	D3.js	Declarative-flavored JavaScript visualization library	Open source	d3js.org
Databricks	Databricks	Data science (ingestion, processing, collaboration, exploration, and visualization) on Spark	14 days	databricks.com/product/databricks
Dataguise	Dataguise DgSecure	Big data security monitoring	Available by request	dataguise.com
Dataiku	Dataiku DSS	Collaborative data science platform	14 days	dataiku.com/dss/features/connectivity
Datameer	Datameer	BI, data integration, ETL, and data visualization on Hadoop	Available by request	datameer.com/product/product-overview
DataRobot	DataRobot	Machine learning model-building platform	Demo available by request	datarobot.com/product
DataRPM	DataRPM	Cognitive predictive maintenance for industrial IoT	Available by request	datarpm.com/platform
DataTorrent	DataTorrent RTS	Stream and batch (based on Apache Apex) application development platform	Free tier available	datatorrent.com/products-services/datatorrent-rts
DataWatch	DataWatch Monarch	Data extraction and wrangling, self-service analytics, streaming visualization	30 days	datawatch.com/our-platform/monarch
Disco Project	Disco	MapReduce framework for Python	Open source	discoproject.org
Domo	Domo	Data integration, preparation, and visualization	Available by request	domo.com/product
Druid	Druid	Columnar distributed data store w/realtime queries	Open source	druid.io/
Eclipse Foundation	BIRT	Visualization and reporting library for Java	Open source	eclipse.org/birt/
EngineRoom.io	EngineRoom	Geospatial, data transformation and discovery, modeling, predictive analytics, and visualization	N/A	engineroom.io

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
EnThought	SciPy	Scientific computing ecosystem (multi-dimensional arrays, interactive console, plotting, symbolic math, data analysis) for Python	Open source	scipy.org
Exaptive	Exaptive	RAD and application marketplace for data science	Free tier available	exaptive.com
Exasol	Exasol	In-memory analytics database	Free tier available	exasol.com/en/products
Facebook	Presto	Distributed interactive SQL on HDFS	Open source	prestodb.io
Fair Isaac Corporation	FICO Decision Management Suite	Data integration, analytics, and decision management	N/A	fico.com/en/analytics/decision-management-suite
GFS2 Group	GFS	(Global File System) Shared-disk file system for Linux clusters	Open source	sourceware.org/cluster/gfs/
GoodData	GoodData Platform	Data distribution, visualization, analytics (R, MAQL), BI, and warehousing	N/A	gooddata.com/platform
Google	Protocol Buffers	Data serialization format and compiler	Open source	developers.google.com/protocol-buffers/docs/overview
Google	TensorFlow	An open-source software library for machine intelligence	Open source	tensorflow.org
Graphviz	Graphviz	Graph visualization toolkit	Open source	graphviz.org
H2O.ai	H2O.ai	Stats, machine learning, and math runtime for big data	Free tier available	h2o.ai
H2O.ai	H2O	Open-source prediction engine on Hadoop and Spark	Open source	h2o.ai
Hitachi Group	Pentaho	Data integration layer for big data analytics	30 days	pentaho.com/product/product-overview
Hortonworks	Hortonworks Data Platform	Hadoop distribution based on YARN	N/A	hortonworks.com/products/data-platforms/hdp
Hortonworks	Hortonworks DataFlow	Streaming data collection, curation, analytics, and delivery	N/A	hortonworks.com/products/data-platforms/hdf
IBM	IBM BigInsights	Scalable data processing and analytics on Hadoop and Spark	Available by request	ibm.com/analytics/us/en/technology/biginsights
IBM	IBM Streaming Analytics	Streaming data application development and analytics platform	Available by request	ibm.com/cloud/streaming-analytics
IBM	IBM InfoSphere Information Server	Data integration, data quality, and data governance	Available by request	ibm.com/analytics/information-server
Ignite	Infobright DB	Column-oriented store with semantic indexing and approximation engine for analytics	N/A	ignitetech.com/solutions/information-technology/infobrightdb
Infor	Birst	Enterprise and embedded BI and analytics platform	Available by request	birst.com/product

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
Informatica	Enterprise Data Lake	Collaborative, centralized data lake, data governance	N/A	informatica.com
Informatica	Big Data Management	Data integration platform on Hadoop	N/A	informatica.com
Informatica	Relate 360	Big Data analytics, visualization, search, and BI	N/A	informatica.com
Informatica	Big Data Streaming	Event processing and streaming data management for IoT	N/A	informatica.com
Information Builders	WebFOCUS	BI and analytics	Demo available by request	informationbuilders.com/products/intelligence
Information Builders	Omni-Gen	Data management, quality, integration platform	Available by request	informationbuilders.com/products/omni
Intersystems	IRIS	Data mangement, interoperability, and analytics	N/A	intersystems.com/products/intersystems-iris/#technology
Java-ML	Java-ML	Various machine learning algorithms for Java	N/A	java-ml.sourceforge.net
Jinfonet	JReport	Visualization, embedded analytics for web apps	Available by request	jinfonet.com/product
JUNG Framework	JUNG Framework	Graph framework for Java and data modeling, analyzing, and visualizing	Open source	jung.sourceforge.net
Kognitio	Kognitio Analytical Platform	In-memory, MPP, SQL and NoSQL analytics on Hadoop	Free tier available	kognitio.com/products/kognitio-on-hadoop
Lavastorm	Lavastorm Server	Data preparation, analytics application development platform	Free tier available	lavastorm.com/product/explore-lavastorm-server
LexisNexis	LexisNexis Customer Information Management	Data management and migration	N/A	risk.lexisnexis.com/corporations-and-non-profits/customer-information-management
LexisNexis	HPCC Platform	Data management, predictive analytics, and Big Data workflow	Open source	hpccsystems.com
Liaison Technologies	Liaiason Alloy	Data management and integration	Demo available by request	liaison.com/liaison-alloy-platform
Lightbend	Lightbend Reactive Platform	JVM application development platform with Spark	Free tier available	lightbend.com/products/reactive-platform
LinkedIn	Pinot	Real-time OLAP distributed data store	Open source	github.com/linkedin/pinot
LISA Lab	Theano	Python library for multi-dimensional array processing w/GPU optimizations	Open source	deeplearning.net/software/theano
Loggly	Loggly	Cloud log management and analytics	30 days	loggly.com/product
Logi Analytics	Logi Analytics Platform	Embedded BI and data discovery	Demo available by request	logianalytics.com/analytics-platform

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
Looker	Looker Business Intelligence	Data analytics and business intelligence platform	Demo available by request	looker.com/product/business-intelligence
Looker	Looker Embedded Analytics	Embedded analytics, data exploration, and data delivery	Demo available by request	looker.com/product/embedded-analytics
MapR	MapR Event Streams	Global publish-subscribe event streaming system	Free tier available	mapr.com/products/mapr-streams
MapR	MapR Analytics and Machine Learning Engines	Real-time analytics and machine learning at scale	Free tier available	mapr.com/products/analytics-ml
MapR	MapR Converged Data Platform	Big Data platform on enterprise-grade Hadoop distribution with integrated open-source tools (Spark, Hive, Impala, Solr, etc.), NoSQL (document and wide column) DBMS	Free tier available	mapr.com/products/mapr-converged-data-platform
Micro Focus	ArcSight Data Platform	Data collection and log management platform	Available by request	software.microfocus.com/en-us/products/siem-data-collection-log-management-platform/overview
Micro Focus	IDOL	Machine learning, enterprise search, and analytics platform	N/A	software.microfocus.com/en-us/products/information-data-analytics-idol/overview
Micro Focus	Vertica	Distributed analtyics database and SQL analytics on Hadoop	Free tier available	vertica.com/overview
Microsoft	SSRS	SQL Server reporting (server-side)	Free tier available	msdn.microsoft.com/en-us/library/ms159106.aspx
Microsoft	Azure Machine Learning Studio	Predictive analytics and machine learning development platform	Free tier available	azure.microsoft.com/en-us/services/machine-learning-studio
Microsoft	Power BI	Business intelligence platform	Free tier available	powerbi.microsoft.com
MicroStrategy	Advanced Analytics	Predictive analytics, native analytical functions, data mining	Free tier available	microstrategy.com/us/products/capabilities/advanced-analytics
New Relic	New Relic Insights	Real-time application performance analytics	Demo available by request	newrelic.com/insights
NumFocus	Julia	Dynamic programming language for scientific computing	Open source	julialang.org
NumFocus	Matplotlib	Plotting library on top of NumPy (like parts of MATLAB)	Open source	matplotlib.org
NumFocus	NumPy	Mathematical computing library (i.e. multi-dimensional arrays, linear algebra, Fourier transforms) for Python	Open source	numpy.org
NumFocus	Pandas	Data analysis and modeling for Python	Open source	pandas.pydata.org
Objectivity	ThingSpan	Graph analytics platform with Spark and HDFS integration	Free tier available	objectivity.com/products/thingspan

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
OpenText	OpenText Big Data Analytics	Analytics and visualization with analyatics server	45 days	opentext.com/what-we-do/products/analytics/opentext-big-data-analytics
OpenTSDB Authors	OpenTSDB	Time-series database on Hadoop	Open source	github.com/OpenTSDB/opentsdb/releases
Oracle	Big Data Discovery	Big Data analytics and visualization platform on Spark	Demo available	oracle.com/big-data/big-data-discovery
Oracle	R Advanced Analytics for Hadoop	R interface for manipulating data on Hadoop	N/A	oracle.com/technetwork/database/database-technologies/bdc/r-advanalytics-for-hadoop/overview
Palantir	Gotham	Cluster data store, on-the-fly data integration, search, in-memory DBMS, ontology, and distributed key-value store	N/A	palantir.com/palantir-gotham
Palantir	Foundry	Data integration platform	N/A	palantir.com/palantir-foundry
Panoply	Panoply	Data management and analytics platform	21 days	panoply.io
Panorama Software	Necto	Business intelligence, visualization, and data management	Available by request	panorama.com/necto
Paxata	Paxata Adaptive Information Platform	Data integration, preparation, exploration, visualization on Spark	Demo available by request	paxata.com/product/paxata-adaptive-information-platform
Pepperdata	Pepperdata Cluster Analyzer	Big data performance analytics	Demo available by request	pepperdata.com/products/cluster-analyzer
Pivotal	Pivotal Greenplum	Open-source data warehouse and analytics	Open source	pivotal.io/pivotal-greenplum
Pivotal	Spring Cloud Data Flow	Cloud platform for building streaming and batch data pipelines and analytics	N/A	cloud.spring.io/spring-cloud-dataflow
Prognoz	Prognoz Platform	BI and analytics (OLAP, time series, predictive)	Free tier available	prognoz.com/platform
Progress Software	DataDirect Connectors	Data integration: many-source, multi-interface (ODBC, JDBC, ADO.NET, OData), multi-deployment	Available by request	progress.com/datadirect-connectors
Project Jupyter	Jupyter	Interactive data visualization and scientific computing on Spark and Hadoop	Open source	jupyter.org
Pyramid Analytics	BI Office	Data discovery and analytics platform	Free tier available	pyramidanalytics.com/pages/bi-office.aspx
Qlik	Qlik Sense	Data visualization, integration, and search	Free tier available	qlik.com/us/products/qlik-sense
Qlik	Qlik Analytics Platform	Data visualization platform	Free tier available	qlik.com/us/products/qlik-analytics-platform

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
Qlik	QlikView	Business intelligence application platform	Free tier available	qlik.com/us/products/qlikview
Qubole	Qubole Data Service	Data engines for Hive, Spark, Hadoop, Pig, Cascading, Presto on AWS, Azure, Google Cloud	Free tier available	qubole.com
Rapid7	InsightOps	Log management and analytics	Available by request	logentries.com
RapidMiner	RapidMiner Studio	Predictive analytics workflow and model builder	Free tier available	rapidminer.com/products/studio
RapidMiner	RapidMiner Radoop	Predictive analytics on Hadoop and Spark with R and Python support	Free tier available	rapidminer.com/products/radoop
Red Hat	Ceph	Distributed object and block store and file system	Open source	ceph.com/get
RedPoint	RedPoint Data Management	Data management, quality, integration (also on Hadoop)	Demo available by request	redpoint.net/products/data-management-solutions
SAP	SAP HANA	In-memory, column-oriented, relational DBMS (cloud or on-premise) with text search, analytics, stream processing, R integration, and graph processing	Free tier available	sap.com/products/hana.html
SAS	SAS Platform	Analytics, BI, data management, and deep statistical programming	Available by request	sas.com/en_us/software/sas9.html
Sencha	InfoVis Toolkit	JavaScript visualization library	Open source	philogb.github.io/jit
Sisense	Sisense	Analytics, BI, visualization, and reporting	Available by request	sisense.com/product
Skytree	Skytree	Machine learning platform with self-service options	Available by request	skytree.net/products
Software AG	Terracotta In-Memory Data Management by Software AG	In-memory data management, job scheduler, Ehcache implementation, and enterprise messaging	Available by request	terracotta.org
Splunk	Splunk Enterprise	Operational intelligence for machine-generated data	60 days	splunk.com/en_us/products/splunk-enterprise.html
Stitch	Stitch	ETL-as-a-service	Free tier available	stitchdata.com
StreamSets	Dataflow Performance Manager	Data management and analytics platform	N/A	streamsets.com/products/dpm
Sumo Logic	Sumo Logic	Log and time-series management and analytics	30 days	sumologic.com
Tableau	Tableau	Interactive data visualization for BI	Available by request	tableau.com
Tableau	Tableau Desktop	Visualization, analytics, exloration (with self-service, server, hosted options)	14 days	tableau.com/products/desktop
Talend	Talend Data Fabric	Real-time or batch data management platform	N/A	talend.com/products/data-fabric

COMPANY	PRODUCT	DESCRIPTION	FREE TRIAL	WEBSITE
Talend	Talend Open Studio	ELT and ETL on Hadoop with open-source components	N/A	talend.com/download/talend-open-studio
Tamr	Tamr	Data management, sanitation, analytics, and BI	Demo available by request	tamr.com/product
Targit	Targit Decision Suite	BI, analytics, discovery front-end with self-service options	Demo available by request	targit.com/en/software/decision-suite
Teradata	Teradata	Data warehousing, analytics, data lake, SQL on Hadoop and Cassandra, big data appliances, R integration, and workload management	N/A	teradata.com/Solutions
The R Foundation	R	Language and environment for statistical computing and graphics	N/A	r-project.org
Thoughtspot	Thoughtspot	Relational search engine	Demo available by request	thoughtspot.com/product
TIBCO	TIBCO Data Virtualization	ETL, data virtualization, and integration platform	N/A	tibco.com/products/tibco-data-virtualization
TIBCO	Jaspersoft	BI, analytics (OLAP, in-memory), ETL, data integration (relational and non-relational), reporting, and visualization	Free tier available	jaspersoft.com/business-intelligence-solutions
TIBCO	TIBCO Spotfire Platform	Data mining and visualization	30 days	spotfire.tibco.com
Treasure Data	Treasure Data	Analytics infrastructure as a service	Demo available by request	treasuredata.com
Trifacta	Trifacta Wrangler	Data wrangling, exploration, and visualization on Hadoop	N/A	trifacta.com/products/wrangler
University of Waikato	Weka	Machine learning and data mining for Java	Open source	cs.waikato.ac.nz/ml/weka
Unravel	Unravel	Predictive analytics and machine learning performance monitoring	Available by request	unraveldata.com/optimize-troubleshoot-and-analyze-big-data-performance
Waterline Data	Waterline Data	Data marketplace (inventory, catalogue with self-service) on Hadoop	Demo available by request	waterlinedata.com/product-overview
Wolfram	Wolfram Language	Knowledge-based programming language with many domain-specific libraries	Available by request	wolfram.com/language
Workday	Workday Prism Analytics	Data preparation, discovery, and analytics on Hadoop and Spark	N/A	workday.com/en-us/applications/prism-analytics.html
Xplenty	Cascading	Platform to develop big data applications on Hadoop	Open source	cascading.org
YCSB	YCSB	General-purpose benchmarking spec	Open source	github.com/brianfrankcooper/YCSB/wiki/Getting-Started
Yellowfin	Yellowfin	Business intelligence and data visualization	Available by request	yellowfinbi.com/platform
Zaloni	Zaloni	Enterprise data lake management	Demo available	zaloni.com/platform
Zoomdata	Zoomdata	Analytics, visualization, and BI with self-service on Hadoop, Spark, many data stores	Available by request	zoomdata.com

GLOSSARY

ALGORITHM
A series of instructions used to solve a mathematical problem.

APACHE HADOOP
An open-source tool to process and store large distributed data sets across machines by using MapReduce.

APACHE SPARK
An open-source Big Data processing engine that runs on top of Apache Hadoop, Mesos, or the cloud.

ARTIFICIAL INTELLIGENCE
The ability of a machine to recognize its environment, act in a rational way, and maximize its ability to solve problems.

BACKPROPAGATION
The process by which a neural network will auto-adjust its parameters until it can consistently produce the desired outcome with sufficient reliability.

BIG DATA
A common term for large amounts of data. To be qualified as big data, data must be coming into the system at a high velocity, with large variation, or at high volumes.

BUSINESS INTELLIGENCE
The process of visualizing and analyzing business data for the purpose of making actionable and informed decisions.

CLUSTER
A subset of data that share particular characteristics. Can also refer to several machines that work together to solve a single problem.

COLLABORATIVE FILTERING
A technique used by recommender systems to identify information or patterns in how several users may rate objects.

DATA FLOW MANAGEMENT
The specialized process of ingesting raw device data, while managing the flow of thousands of producers and consumers. Then performing basic data enrichment, analysis in stream, aggregation, splitting, schema translation, format conversion and other initial steps to prepare the data for further business processing.

DATA GOVERNANCE
The process of managing the availability, usability, integrity, and security of data within a Data Lake.

DATA LAKE
A storage repository that holds raw data in its native format.

DATA PREPARATION
The process of collecting, cleaning, and consolidating data into one file or data table, primarily for use in analysis.

DATA PROCESSING
The process of retrieving, transforming, analyzing, or classifying information by a machine.

DATA SCIENCE
A field that explores repeatable processes and methods to derive insights from data.

MACHINE LEARNING
An AI that is able to learn by being exposed to new data, rather than being specifically programmed.

MAPREDUCE
A data processing model that filters and sorts data in the Map stage, then performs a function on that data and returns an output in the Reduce stage.

NOSQL DATABASE
Short for "not only SQL", or any database that uses a system to store and search data other than just using tables and structured query languages.

PARSE
To divide data, such as a string, into smaller parts for analysis.

PERSISTENT STORAGE
A non-changing place, such as a disk, where data is saved after the process that created it has ended.

R
An open-source language primarily used for data visualization and predictive analytics.

REAL-TIME STREAM PROCESSING
A model for analyzing sequences of data by using machines in parallel, though with reduced functionality.

RELATIONAL DATABASE MANAGEMENT SYSTEM (RDBMS)
A system that manages, captures, and analyzes data that is grouped based on shared attributes called relations.

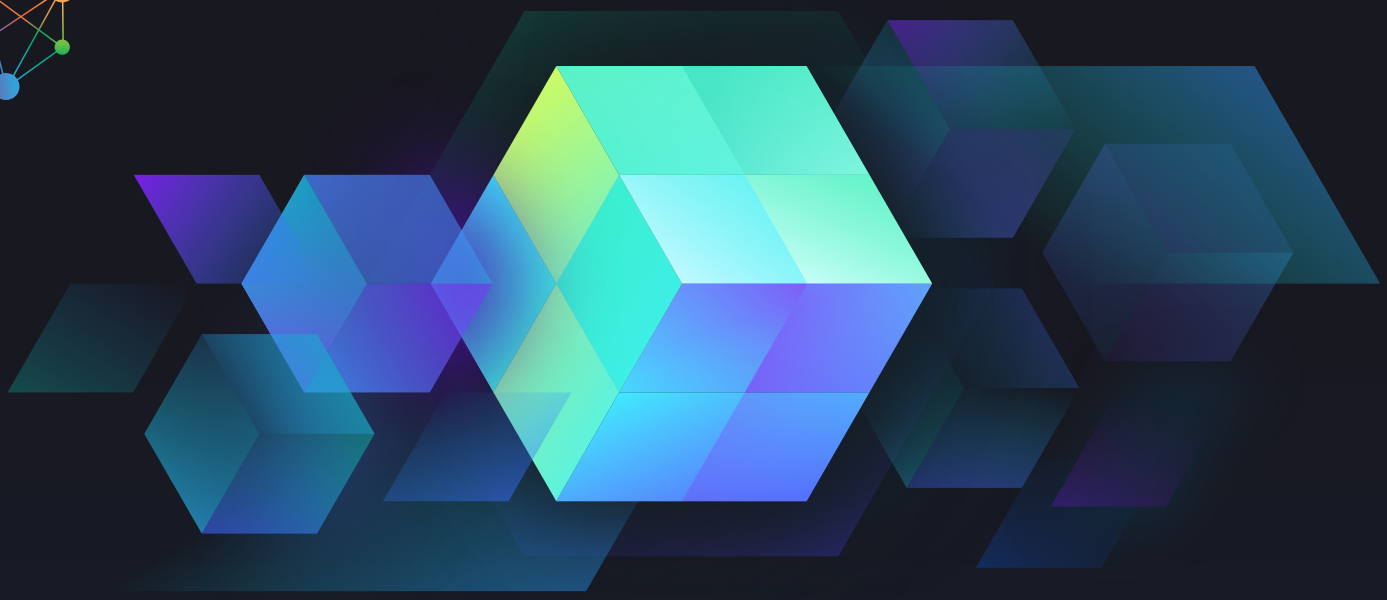
SMART DATA
Digital information that is formatted so it can be acted upon at the collection point before being sent to a downstream analytics platform for further data consolidation and analytics.

STRUCTURED DATA
Information with a high degree of organization.

STRUCTURED QUERY LANGUAGE (SQL)
The standard language used to retrieve information from a relational database.

UNSTRUCTURED DATA
Data that either does not have a pre-defined data model or is not organized in a pre-defined manner.

ZONES
Distinct areas within a Data Lake that serve specific, well-defined purposes.



INTRODUCING THE

Microservices Zone

Start Implementing Your Microservice Architecture and Scale Your Applications

Whether you are breaking down existing monolithic legacy applications or starting from scratch, see how to overcome common challenges with microservices.

Keep a pulse on the industry with topics like: Java Microservices Tutorials and Code Examples, Best Practices for a REST-Based Microservices Architecture, and Patterns and Anti-Patterns to Get Started.

Visit the Zone

REST-BASED MICROSERVICE ARCHITECTURE

PATTERNS AND ANTI-PATTERNS

INTER-SERVICE COMMUNICATION

APPLICATIONS FOR CONTAINERS