BIG DATA 50

The hottest big data startups of 2014

Jeff Vance

Startup50

Table of Contents

A note from Feedzai i

Big Data Startup Landscape – Overview ii

About the Author iv

Introduction – the Big Data Boom 1

Notes on Methodology & the origin of the Big Data 50 2

The Big Data 50 5

Big Data for Security 5

Feedzai 5

Fortscale 7

Vorstack 8

Poised for Explosive Growth 9

Entrigna 9

Nuevora 11

Roambi 13

Machine Learning Mavens 14

Oxdata 14

Ayasdi 15

BigML 16

Context Relevant 17

Skytree 18

Sumo Logic 19

Built for Speed 20

DataTorrent 20

MemSQL 22

ParStream 23

Tidemark 24

VoltDB 25

Analytics for Sales, Marketing, and Social Media 26

BloomReach *27*

NGDATA *28*

Ninja Metrics *29*

PlaceIQ *31*

Pursway *32*

SumAll *33*

Teikametrics *34*

Billion-Dollar Baby 35

Cloudera *35*

The NoSQL Crowd 36

Aerospike *36*

Couchbase *38*

DataStax *39*

ScaleArc *41*

Splice Machine *42*

Hadoop Darlings 44

Altiscale *44*

Continuuity *45*

Hortonworks *46*

MapR Technologies *48*

Qubole *49*

Platfora *50*

Xplenty *51*

Making Big Data Easy 52

Alpine Data Labs *52*

import.io *54*

Predixion Software *55*

SiSense *56*

Focusing on Performance Management 58

CloudPhysics *58*

Concurrent *59*

Search, Warehousing, and “People Analytics” 61

LucidWorks *61*

MammothDB *62*

VoloMetrix *63*

Data Connections and Preparation 65

Datameer *65*

Paxata *66*

Tamr *67*

Trifacta *68*

A Note from Feedzai

At Feedzai, we're excited to be a part of the 2014 Big Data 50, and to offer this copy of the Report because it's data that powers our business, and ultimately, most businesses today. We have seen quotes that say we as a society are creating 2.5 quintillion bytes of data every day! This data is on our computers and on our phones; it is information stored in large and small businesses alike. The data is a combination of photos, music, social posts, digital transactions, GPS signals, and on and on, and it certainly is BIG and can be unwieldy.

We at Feedzai believe every business can unlock the power of big data, and we have used machine learning techniques and data science to understand it. Our specific focus is on commerce and helping to prevent fraud. We are the only fraud prevention solution available for omnichannel commerce, and we have found a way to analyze and use machine learning to help you anticipate and detect fraud earlier.

We know that big data analysis cannot be done by humans alone, so we rely on machines to boost human intelligence. But, machine learning has many connotations, so we'd like to explain Feedzai's point of view on machine learning and why we use it.

Machine learning is defined as "intelligent systems that autonomously learn, predict and act using data." It moves beyond that set of pre-cooked rules and self-detects subtle patterns in data – much faster than humans can. In the case of fraud, our core expertise, transaction patterns change very quickly, and they have a long tail distribution. The nature of the long tail makes it uneconomical to focus on a few incidents because shopping behaviors can cross many channels – from in-store, to online, to mobile - therefore, inhibiting you to see true patterns. By leveraging machine learning, we can analyze data more efficiently, build statistical models quickly, and most importantly, in the case of fraud prevention, react to new criminal behaviors faster.

Machine learning is certainly used by many companies today and will continue to be essential to create efficiencies in the future. At Feedzai, machine learning has helped us to build our business and provide financial institutions and retailers with a more advanced solution to fight fraud. To date, Feedzai has protected over $500B worth of commerce . . . and that number will only continue to grow.

Big Data Startup Landscape – Overview

The Big Data space is heating up – to the point that startup Cloudera has raked in more than $1 billion in VC funding. That's not a typo; that's 1 billion with a "b." The Cloudera funding is telling. It's an absurd amount of money, but more than three quarters of that funding came from Intel, which shifted its Big Data focus away from its in-house efforts to a close working relationship with Cloudera.

If Intel believes the value chain of computing is on the brink of upheaval, well, their track record means we should all take them seriously.

The people who make their living prognosticating about tech trends also see big things ahead for Big Data. IDC predicts that the market for Big Data technologies will reach $32.4 billion by 2017, or about six times the growth rate of the overall information and communication technology market.

Wikibon, meanwhile, believes the overall Big Data market will exceed $50 billion by 2017, while Gartner forecasts that Big Data will drive $34 billion in IT spending this year, increasing to $232 billion by 2016. Gartner also predicts that by 2015 65 percent of packaged analytic applications with "advanced analytics" will include embedded Hadoop.

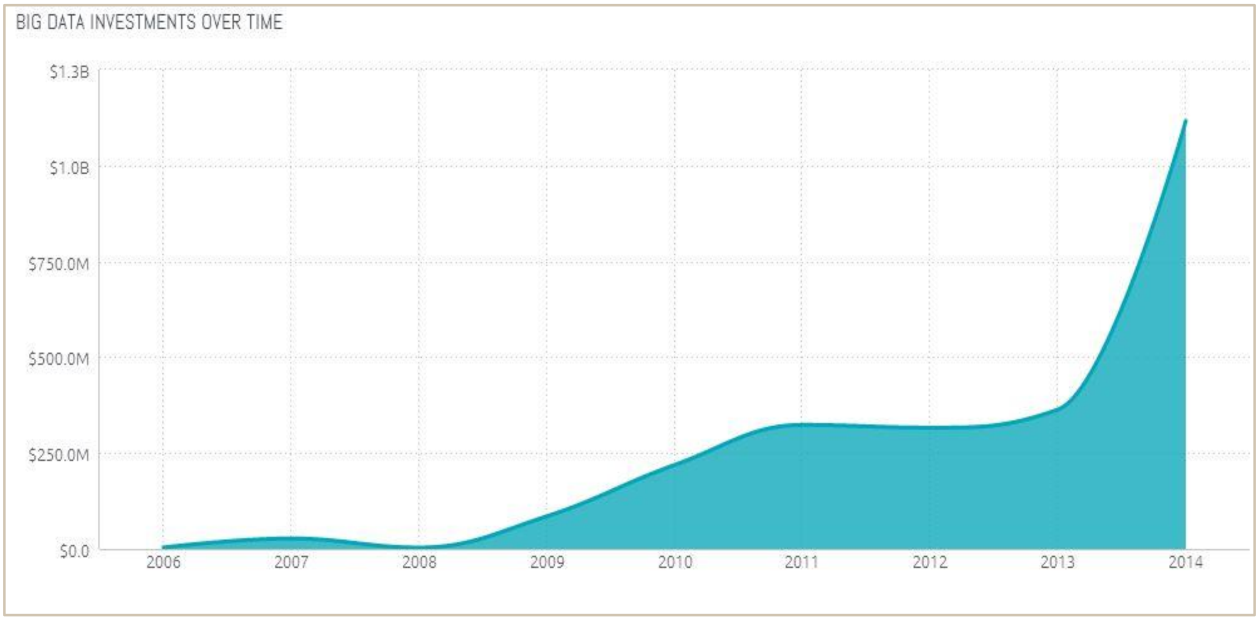


Figure 1. SiSense and CrunchBase track Big Data investments

Big Data 50 startup SiSense has but its analytics tools to work to crunch CrunchBase's data on Big Data investments. Figure 1 shows the spike in investments just from 2013 to today, with the Cloudera investment leading the pack.

The Big Data 50 startups who earned their way into this report have a big head start, aggressively pursuing those Big Data billions. In fact, many of the startups in this report are pushing at the boundaries of Big Data, making it very likely (and I almost never say this) that analysts are underestimating the potential of this market.

About the Author

Jeff Vance began covering startups just before the dotcom crash. In 1999, he joined Pinestream Communications, where he edited startup-focused publications targeted at the investment community.

He set out on his own in 2004, establishing himself as a freelance writer who contributed feature articles and editorials to such publications as *Forbes.com*, *Network World*, *CIO, Datamation, Wi-Fi Planet, Cloudbook* and many others.

Later that same year, he founded Sandstorm Media, a copywriting and content marketing firm.

Jeff founded Startup50.com after he wrote two related articles about mobile startups for *Network World* and *CIO*. “I put out a call for entries on HARO and received more than 150 recommendations,” Jeff recalls. “As part of the process of narrowing down the field, I put up a ‘Final Forty’ list on the Sandstorm Media blog. The response to that was overwhelming. Emails flooded my inbox. New startups popped out of the weeds, and several VCs called me to ask if I could pass on the rest of the 150. Other VCs asked if I publish my selection criteria as an informal RFP-type document, and to this day, that page remains one of the most heavily trafficked ones on the site.”

Startup50.com was launched as platform to connect with startups, cover them in ways that traditional publications wouldn’t, and as a way to run journalism/publishing 2.0 experiments.

In all truth, Startup50.com is still an experiment, and Jeff says he plans to keep it that way. Publishing is a mess, traditional PR is on the decline, and most startup founders suffer from a stilted point of view that results from smelling their own exhaust all day, every day. If you keep telling the world you can do no wrong, you start to believe it.

Startup50.com is a corrective, a no-B.S. site that focuses on one thing and one thing only: cool startups and the characteristics that separate the successful ones from the science projects.

Introduction – the Big Data Boom

Rosy forecasts that predict huge markets, such as Wikibon's belief that the overall Big Data market will exceed $50 billion by 2017, must be taken with a heaping helping of salt. After all, research firms won't sell reports that predict $1 million markets.

However, this is a rare case where the analysts may actually be erring on the conservative side. Big Data has seen little of the backlash that you still hear about cloud, and what little skepticism is out there sounds to my ears like he old-school baseball scouts in ***Moneyball***, who can't wrap their heads around data-driven decision making and who are being left in the past as a result.

There is one point these doubters make that merits attention, however: you can't completely remove human decision making from the process. Smart business leaders, technologists, and even MLB GMs know that, and removing human decisions has never been the point. The point has always been to overcome the limitations of the human mind, not to replace it.

Neuroscience, psychology, and behavioral economics have fairly well proven that humans are terrible at making certain types of decisions, especially those that require predictions about future events. We too heavily weight the evidence we've saw last. We are overconfident about our own prognosticating abilities, and once we've made a prediction, we tend to seek information that confirms what we already believe and dismiss anything that challenges our beliefs.

That's not a very good decision-making process if you want to run a data-driven business, but it's the de facto standard practice.

What the startups in the Big Data 50 are seeking to do is to find is ways to measure things that we've only rough estimated before, to measure those things granularly, and to bring, eventually, human brain power into the process at the right time, armed with the best information available, so we make decisions based on real-world conditions and not our guts.

There will always be things that defy analytics, fuzzy concepts like workplace culture, or motivation, or even basic altruism (although some Big Data startups are already trying to measure and analyze those things), but the realm of the unmeasurable is receding. I'm sure other concerns will rush in to fill the void, but, whatever the case, Big Data is ushering in yet another era of rapid change.

The startups in the Big Data 50 report are in the vanguard of that change. These startups are disrupting such varied industries as traditional business intelligence, brick-mortar retail, pharmaceutical research, social marketing, and on and on and on. Keep an eye on them. They're providing a sneak peek of what the world will look like in five or ten years.

Notes on Methodology & the origin of the Big Data 50

The Big Data 50 has been a year or so in the making. I know what you're thinking:
"What took you so long?"

Well, I have a confession: I'm pretty data-savvy for a journalist, but I'm no data scientist.

I don't have a numbers background. I was a liberal arts major (and even spent time earning that coveted, but already anachronistic MFA degree in creative writing) and have been a working journalist for most of my adult life.

Long story short, I'm no Nate Silver.

As I compiled several Big Data and Hadoop startup roundups for ***Network World***, ***CIO***, ***Computerworld***, and ***Datamation***, I'd feel rather like an imposter when I was featured on lists of Top Big Data influencers like this one.

I also realized I should take these data-driven lessons to heart, so I embarked on a crash course that would help me transform my business through better analytics.

My first data experiment was with voting. I figured I could improve my selection process for my startup roundups by pulling readers into the process and letting them vote.

That led to carefully crafted forms that startups had to fill out before I'd even look at them. Employing techniques from psychology and behavioral economics, I found was to get startups to honestly evaluate themselves and their peers – in other words, getting them to reveal things they wouldn't tell you if you asked them directly.

Soon, these experiments were successful enough that they needed their own home, which prompted me to launch my own startup-obsessed site, ***Startup50.com***, and start laying the foundation for the Big Data 50.

Over the course of the past couple of years, I've evaluated more than 200 Big Data startups. Many of these I dismissed out of hand. These tended to be science projects, wishful-thinking credit-card startups, or startups that used the term "Big Data" far too liberally.

As I fine-tuned my process, startups had to fill out long, detailed questionnaires. They had to reveal detailed information about funding, end users, and future growth projects (I didn't make them open their books or anything, but I substituted other data points for rough verification).

Along the way, I learned all sorts of things about these startups, many of them things the startups probably didn't realize they were revealing. To use a concept that's becoming popular in books like ***Decisive*** and ***Think like a Freak***, I did my best to have this garden of startups weed itself.

Some of these techniques are amazingly simple. For instance, if a startup couldn't respond in a timely fashion to my emails, if they made promises they couldn't deliver, if they retained a PR agency full of people who couldn't write two coherent consecutive sentences – those startups weeded themselves out.

They most likely would have been booted in the past too, but now that happens automatically and with no intervention by me.

I also turned some of my own pet peeves into weed-yourself tripwires. For instance, any *pre-product* startup referring to itself as "a leading provider" gets, if not bumped, at least devalued (I have to cut them a little slack since this practice is so common). But if your startup is on the fence, ranked informally in the high forties, using that one stupid term could be enough to weed you out.

You may call that harsh; I call it turning a nuisance into a sneaky time saver.

Once the weeding was done, the factors I weighed most heavily were funding, the track record/experience of the founders, market positioning, named customers, social media influence, and an ability (one rare for startups and PR pros alike) to not be a pain in my ass (yep, that's a data point, and a very, very important one).

A final filter I looked through as I evaluated each startup was a simplified version of Richard Koch's (author of the ***80/20 Principle***) Star Principle, which he developed to guide his investment strategies. Koch's Star Principle, boiled way down, is that he only invests in companies that are leaders in their niche, and which are positioned in fast-growing niches.

Here’s a good overview from author Perry Marshall’s review of Koch’s book:

> *When I visited Richard Koch at this home in Portugal, I told him he should take this book off the market, buy up all the used copies and re-package it for $20,000. Why? Because it explains the essence of the strategy that grew his wealth from $2 million to $200 million in less than 20 years.*
>
> *Essentially, this book says “Why do I bat 50% when most Venture Capitalists bat 5%? The secret is the Star Principle.” Only bet on #1 players, and only play in markets that are growing at 10% per year or more. Disqualify everything else. And if you’re not the #1 player, redefine the market so you ARE the #1 player.*
>
> *Now of course Richard goes into much greater detail than that, but that’s the gist of it. I realized when I read it, that all the major home runs I've hit in my career did in fact follow that formula. I’d just never articulated it as well or as clearly as Richard had.*

Of course, for startups, this formula must be tweaked. I mean, I already said that if a pre-product startup claims to be a “leaving provider,” they’re almost certainly bumped, but there are other stand-ins to lean on, such as top-flight named customers, heavy-duty VC funding from the right VCs, and evidence that the startup has tied its marketing and PR efforts directly to its sales funnel in a data-driven and at least partially automated way.

I’m sure you’ll dispute some of the startups I included and pine for some that I did not. That’s natural. Feel free to email me either way to let me know what you think.

And please remember, **these startups are not ranked. They are grouped with other startups with whom they most logically fit**.

The Big Data 50

Big Data for Security

These startups are turning data into security insights.

Feedzai

What they do: Feedzai uses real-time, machined-based learning to help companies prevent fraud.

Headquarters: San Mateo, CA

CEO: Nuno Sebastião. Prior to Feedzai, he led the development of the European Space Agency's satellite simulation infrastructure.

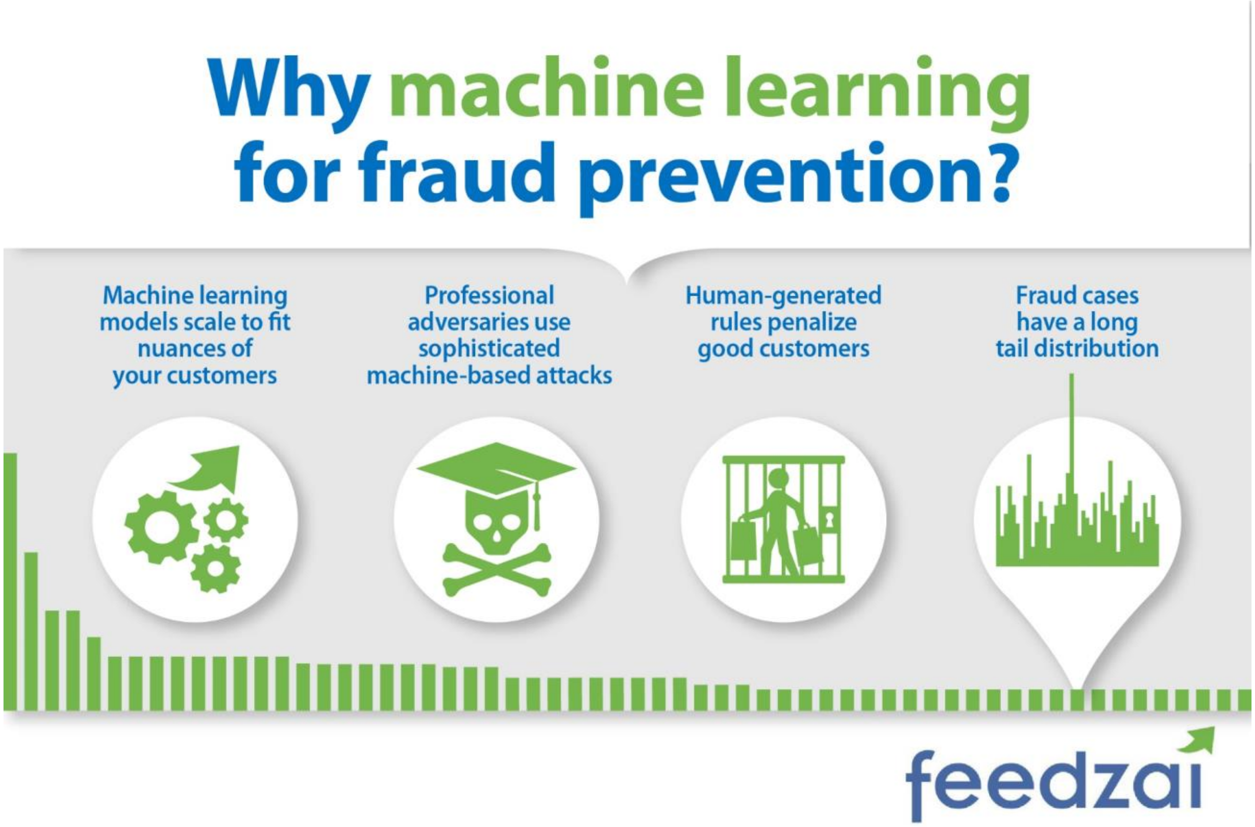
Founded: 2013

Funding: Feedzai has raised $4.3 million from SAP Ventures, Data Collective, and other international investors.

Why they're one of the 50: It's no great revelation that online fraud is a major problem. However, its impact is often underestimated. For instance, the Target breach could end up costing as much as $680 million, according to the Ponemon Institute.

Feedzai claims that it can detect fraud in any commerce transaction, whether the credit card is present or not, in real-time. Feedzai combines artificial intelligence (AI) to build more robust predictive models and analyze consumer behavior in a way that mitigates risk, protects consumers and companies from fraud, and preserves consumer trust.

Feedzai's software attempts to understand the way consumers behave when they make purchases anywhere, online or off. Feedzai says that its fraud detection system aggregates both online and offline purchases for each consumer over a longer time-frame, which results in earlier, more reliable detection rates.


Why machine learning for fraud prevention?
Machine learning models scale to fit nuances of your customers
Professional adversaries use sophisticated machine-based attacks
Human-generated rules penalize good customers
Fraud cases have a long tail distribution
feedzai

The software uses data to create profiles for each customer, merchant, location, and POS device, with up to a 3-year history of data behind each one. Profiles are updated for each consumer after every transaction. As a result, Feedzai claims to be able to detect fraud up to 10 days earlier than traditional methods and expose up to 60 percent more fraudulent transactions.

Clients include Coca-Cola, Logica, Vodafone, Ericsson, Payment Solutions, and Servebase Credit Card Solutions.

Competitive Landscape: Competitors include SiftScience, Signifyd, Kount, and Retail Decisions (ReD).

Fortscale

What they do: Transform enterprise Big Data into actionable security insights.

Headquarters: Currently Tel Aviv, but they plan to move HQ to Silicon Valley and retain the Tel Aviv site for R&D

CEO: Idan Tendler, who previously built and led the Cyber Security Business Group of Elbit Systems

Founded: 2012

Funding: The startup just closed a $10M Series A from led by Intel Capital and Blumberg Capital. Seed/angel investors, including the Swarth Group, also participated. This brings total funding to date to $12 million.

Why they're one of the 50: Fortscale's mission is to turn enterprise Big Data into "User Intelligence, "making users' profiles and behavior visible and easy to investigate. Fortscale's solution offers enterprises a proactive, intelligence-driven approach to cyber security based on Big Data analytics to help defend against the scourge of targeted attacks and low-and-slow attacks that threaten intellectual property and financial assets.

Fortscale claims that is able to enhance SOC teams and security analysts' capabilities to that of advanced data scientists, while leveraging their existing infrastructure and know how. Fortscale's focus on an enterprise's users allows security teams to gain insights about malicious or rogue users, to pinpoint high-risk user behavior, and to monitor access activity.

The goal is to produce "User Intelligence" based on extracted data from existing Big Data repositories (e.g. SIEM) by leveraging machine learning algorithms, canned reports, visualization tools, and queries. Fortscale analyzes historical log data and run peer analysis on it. The result is focusing the security analysts' efforts on the most important events, leads and threats, enabling them to make better, faster decisions.

Competitive Landscape: The security space is already crowded, and while I like this Big Data approach, others are rushing in here as well. For the time being, Big Data security is a land grab, but the land is getting snatched up pretty quickly. Competitors include Sumo Logic, CloudPhysics, Palantir, Splunk, and others.

Vorstack

What they do: Provide a data-driven cyber-threat security solution.

Headquarters: San Mateo, CA

CEO: Joe Eandi. Prior to founding Vorstack, Eandi served in various executive roles at LiveOps for seven years, and before that he held numerous legal positions at Inktomi.

Founded: 2011

Funding: $5.2 million. The financing, announced in late April, was led by Glenn McGonnigle and Tom Noonan of TechOperators and included participation from EMC Ventures, as well as funding from previous angel investor Aligned Partners.

Why they're one of the 50: The security information deluge from disparate external and internal sources is driving organizations to ignore and not take action on real and viable data breaches. Too many security warnings – many with false positives or are not relevant to the organization – create a workload burden on the limited internal resources, which can obscure truly significant threats. The lack of well-structured risk assessment processes and the dependence on human intervention makes discovery and remediation of the relevant cyber threats inefficient – with potentially devastating results.

Vorstack argues that organized cyber-crime requires organized anti-crime. The Vorstack platform provides the automation and control over threat and fraud intelligence enterprises need to identify and securely share cyber-threat intelligence.

Vorstack helps enterprises implement a risk assessment process that collects potential cyber-threat information from internal and external sources. Vorstack then helps the business organize the collected information. It excludes what is not relevant to the organization, and then analyzes the collected data, providing alerts on relevant threats and delivering remediation information.

Vorstack argues that it increases cyber-threat security through automation and the *collaboration* of data. Vorstack says that it's the sharing of data that is the key. What companies see from a threat perspective is shared, and that decreases the time it takes to identify a problem.

Competitive Landscape: Vorstack will compete with ThreatConnect, Internet Identity, FireEye, Looking Glass, and many others.

Poised for Explosive Growth

These startups are relatively new to the scene, but have made big strides in a short amount of time. The more entrenched startups in this report should watch their backs, since these folks may be creeping up on them:

Entrigna

What they do: Provide analytical tools that help users gain insights and develop predictions from streams of real-time, high-velocity data.

Headquarters: Chicago, IL

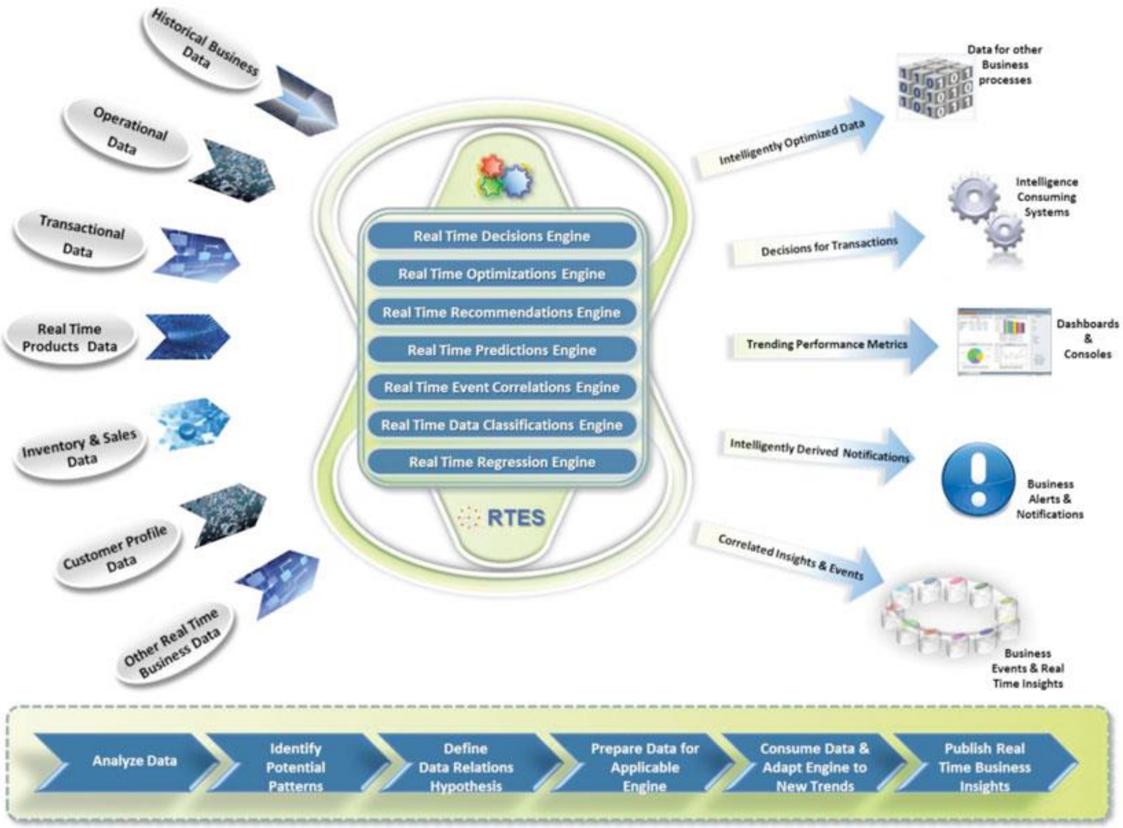
CEO: Murali Kashaboina. He previously served as Managing Director, Enterprise Architecture at United Airlines and played a key role in the merger integration of United and Continental Airlines.

Founded: 2012

Funding: The startup is currently financed by ~$1.1 million in client contracts and through the founding teams’ own investments. Entrigna says that it is in the process of raising a minimum of

$5 million from angel investors and is about to close two more client contracts that will provide an additional $1.2 million.

Why they're one of the 50: With the tremendous increase in computing power and decrease in memory and storage costs, today's businesses are weighed down with a deluge of disconnected data, much of it highly relevant to effective decision-making. This data is associated with business applications, processes, transactions, operations, customers, customer insights, products, product insights, policies, systems, business-partnerships, competition – the list goes on and on.


Historical Business Data
Operational Data
Transactional Data
Real Time Products Data
Inventory & Sales Data
Customer Profile Data
Other Real Time Business Data
Real Time Decisions Engine
Real Time Optimizations Engine
Real Time Recommendations Engine
Real Time Predictions Engine
Real Time Event Correlations Engine
Real Time Data Classifications Engine
Real Time Regression Engine
RTES
Intelligently Optimized Data
Data for other Business processes
Decisions for Transactions
Intelligence Consuming Systems
Trending Performance Metrics
Dashboards & Consoles
Intelligently Derived Notifications
Business Alerts & Notifications
Correlated Insights & Events
Business Events & Real Time Insights
Analyze Data
Identify Potential Patterns
Define Data Relations Hypothesis
Prepare Data for Applicable Engine
Consume Data & Adapt Engine to New Trends
Publish Real Time Business Insights

Since this data exists in many different formats, gaining a unified view of anything is difficult, if not impossible. Much of this data is also highly volatile and often contains time-sensitive business intelligence.

If detected real-time, such "in-the-moment" business intelligence can be used to dynamically determine an optimal course of action.

To tackle this problem, Entrigna developed a real-time decisions platform called *RTES* (Real Time Expert System). RTES enables real-time decisions by offering "decision frameworks" packaged together in one system. The system relies on a combination of machine learning, predictive analytics, business rules, complex event processing, optimization, and artificial intelligence in order to derive real-time actionable business decisions.

Essentially, the RTES platform exposes such decision frameworks as built-in modularized services that can be combined and applied on an organization's business data on a real-time basis. Then, users can identify intelligent patterns that can lead to real-time business insights.

Customers include Jisu, Opargo, Pacific Gas & Electric, Sonata Software, and Decision Analytics International.

Competitive Landscape: Entrigna will compete against legacy data mining and BI tools, such as those from IBM and SAS. Most of the Big Data startups leaning on machine learning are focused on specific verticals, such as fraud detection for credit card companies. However, other general purpose, machine-learning based startups include Ayasdi, Feedzai, Skytree, and Sumo Logic.

Nuevora

What they do: Provide Big Data analytics applications.

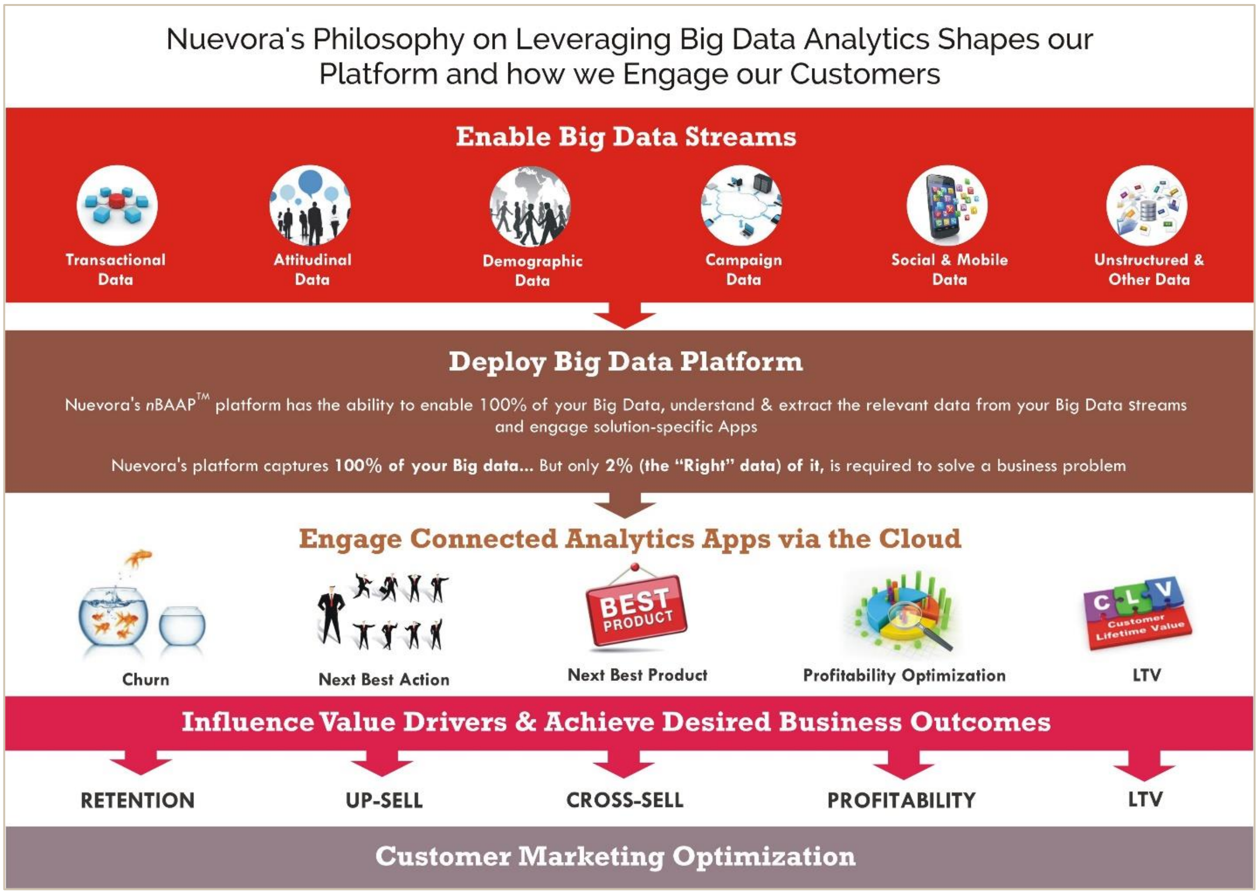
Headquarters: San Ramon, CA

CEO: Phani Nagarjuna, who most recently served as EVP of Products and Business Development for OneCommand, which provides a SaaS-based CRM and Loyalty Automation Platform for the auto retail industry.

Founded: 2011

Funding: $3 million in early funding from Fortisure Ventures.

Why they're one of the 50: Nuevora has set its sights on one of Big Data's early growth areas: marketing and customer engagement. Nuevora's nBAAP (Big Data Analytics & Apps) Platform features purpose-built analytics apps based on best-practices-driven predictive algorithms. nBAAP is based on three key Big Data technologies: Hadoop (data processing), R (predictive analytics), and Tableau (visualizations).



Nuevora's Philosophy on Leveraging Big Data Analytics Shapes our Platform and how we Engage our Customers

Enable Big Data Streams

Transactional Data

Attitudinal Data

Demographic Data

Campaign Data

Social & Mobile Data

Unstructured & Other Data

Deploy Big Data Platform

Nuevora's *n*BAAP™ platform has the ability to enable 100% of your Big Data, understand & extract the relevant data from your Big Data streams and engage solution-specific Apps

Nuevora's platform captures **100% of your Big data...** But only **2% (the "Right" data) of it,** is required to solve a business problem

Engage Connected Analytics Apps via the Cloud

Churn

Next Best Action

Next Best Product

Profitability Optimization

LTV

Influence Value Drivers & Achieve Desired Business Outcomes

RETENTION

UP-SELL

CROSS-SELL

PROFITABILITY

LTV

Customer Marketing Optimization

On top of all of this, Nuevora's algorithms work on disparate sources of data (transactional, social media, mobile, campaigns) to quickly identify patterns and predictors in order to tie specific goals to individual marketing tactics.

The platform includes pre-built apps for the customer marketing business process – acquisition, retention, up-sell, cross-sell, profitability, and customer lifetime value (LTV). With only "last-mile" configurations required for individual customer situations, Nuevora's apps empower organizations to anticipate their customers' behaviors.

Nuevora gives end users the ability to continually recalibrate their predictions using a "closed-loop recalibration engine," which helps organizations keep up with only the most pertinent insights based on the latest data.

Competitive Landscape: When Nuevora assesses the competitive landscape, it zeroes in on big consulting firms, such as Accenture, and other predictive analytics companies, such as Alpine Data Labs.

However, since pretty much every marketing platform under the sun now includes some sort of analytics engine, I also expect them to compete with the major marketing automation providers, such as ExactTarget (which uses Pentaho for its Big Data analytics).

Roambi

What they do: Develop mobile reporting applications that allow business users to access their data on the go.

Headquarters: San Diego, CA

CEO: Santiago Bercerra. Previously, he founded Infommersion, which he later sold to Business Objects in 2005. He was also the founder and CEO of Graphical Information, a software company that he sold to Oracle.

Founded: 2008

Funding: $48 million total, including $30 million from Sequoia Capital.

Why they're one of the 50: One thing the Big Data world has been slow to catch up with is mobile. Even data scientists want their data on smartphones and tablets these days.

Roambi offers two BI/data analytics/mobile reporting apps designed for iOS mobile smartphones and tablets. (The Android versions will be available soon.) Roambi Analytics is a mobile reporting/analytics tool that takes data from any source and transforms it into "views" to help you understand those numbers. Roambi Flow helps you tell the story the analytics reveal, creating multi-touch online documents that provide the context behind the numbers and tell the stories the numbers reveal.

The cloud-based platform powering Roambi's apps was designed to support deployments that match the scale of mobile devices inside the largest enterprises. Roambi's architecture includes full offline capabilities and enterprise-grade security features, such as encryption, remote wipe capabilities, SSO integration, and file expiration.

Roambi also features pre-built integrations to various applications, including Salesforce.com and Box.

Customers include ADP, AirBnB, Bramer Bank, Telefonica, ABInBev, the Phoenix Suns, Qualcomm, and the Sydney Airport.

Competitive Landscape: Roambi will compete with Tableau, QlikView, and others.

Machine Learning Mavens

These startups apply machine learning to tackle complex problems.

Oxdata

What they do: Provide an open source platform, H2O, which applies algorithms and machine learning to turn Hadoop into a Big Data statistical analysis engine.

Headquarters: Mountain View, CA

CEO: SriSatish Ambati. Before 0xdata, co-founded Platfora and was the Director of Engineering at DataStax.

Founded: 2012

Funding: $ 1.7 million from Nexus Venture Partners and angel investors.

Why they're one of the 50: Big Data analysis, for all of its attention, is still difficult, especially with a fragmented technology landscape. Data analysts tend not to code to MapReduce. Meanwhile, modeling and machine-learning algorithms do not scale on commodity hardware.

With 0xdata's H2O platform, enterprises can use all of their data (instead of sampling) in real-time for better predictions. Data scientists can take both simple and sophisticated models to production from the same interactive platform used for modeling. H2O is also used as an algorithms library for "making Hadoop do math."

By keeping the familiar R API, 0xdata argues that its machine learning and statistical functions are more user friendly that competing solutions. 0xdata also uses fine-grain parallelism from fork/join over a distributed array to get consistent performance. All of this complexity is hidden from the end user and a simple html interface gives them the power of the algorithms.

Early named customers include Trulia, Vendavo, and Rushcard.

Competitive Landscape: Competitors include SAS, Revolution Analytics, Skytree, and Apache Mahout.

Ayasdi

What they do: Apply Big Data analysis in order to solve complex problems, including finding cures for cancers and other diseases, exploring new energy sources, and preventing terrorism and financial fraud.

Headquarters: Menlo Park, CA

CEO: Gurjeet Singh, who was previously a Research Scientist at Stanford.

Founded: The startup was founded in 2008 but stayed in stealth-mode until its launch in January 2013.

Funding: Ayasdi has raised $43.4 million in VC funding from FLOODGATE, Khosla Ventures, Institutional Venture Partners, GE Ventures, and Citi Ventures. The company also received $1.2 million in DARPA and NSF grants.

Why they're one of the 50: According to Ayasdi, since the creation of SQL in the 1980s, data analysts have tried to find insights by asking questions and writing queries. The query-based approach has two fundamental flaws. First, all queries are based on human assumptions and biases. Second, query results only reveal slices of data and do not show relationships between similar groups of data. While this method can uncover clues about how to solve problems, it is a game of chance that usually results in weeks, months, and years of iterative guesswork.

Ayasdi believes a better approach is to look at the “shape” of the data. Ayasdi argues that large data sets have a distinct shape, or topology, and that shape has significant meaning. Ayasdi claims to help companies determine that shape in minutes so they can automatically discover insights from their data without ever having to ask questions, formulate queries, or write code.

Ayasdi’s Insight Discovery platform uses Topological Data Analysis (TDA) in tandem with machine learning techniques to enable data scientists, domain experts, and business analysts to optimize their data without coding.

Customers include GE, Citi, Merck, USDA, Mt Sinai Hospital, the Miami Heat, and the CDC.

Competitive Landscape: The machine learning space is wide open. Ayasdi will compete against IBM’s Watson, SAS, Entrigna and Skytree.

BigML

What they do: BigML’s hosted machine learning platform for advanced analytics is designed to appeal to users of all skill sets, from marketing analysts on up to application developers and data scientists.

Headquarters: Corvallis, OR

CEO: Francisco Martin, who was most recently Founder/CEO of Strands, an artificial intelligence and recommendation technology provider.

Founded: 2011

Funding: ~$2.5 million in angel/seed funding.

Why they’re one of the 50: Machine learning has long been an expensive and highly specialized activity. BigML intends to democratize machine learning by simplify the process, so pretty much everyone can use it, and by offering a flexible pricing model that can appeal both to individual users as well as large organizations. As a result, predictive analytics can be applied to a larger range of use cases.

According to Predictive Impacts, the Predictive Analytics market will reach $1.8 billion by the end of 2014.

BigML's hosted interface is targeted at the average business user, but it is underpinned by a RESTful API that gives programmers the freedom to create predictive models, ensembles, and clusters that can be incorporated directly into applications and services. So, the average business user can simply drag data files (.csv) into the interface and start developing predictive models that show likely outcomes for various scenarios, while the data scientist can use BigML to quickly process massive amounts of data with advanced algorithms – all at a fraction of the cost of traditional analytics packages.

BigML says it has more than 10,000 registered users, and says it has about 200 paying customers. CSC and Groupon are two **early named customers.**

Competitive Landscape: Competition will come from two camps: incumbents and startups. The traditional heavyweights include SAS, SPSS and IBM. Startups such as Skytree Analytics, Alteryx, and Precog will also compete with BigML.

Context Relevant

What they do: Develop predictive analytic software solutions.

Headquarters: Seattle, WA

CEO: Stephen Purpura. He previously served as the Chief Security Officer of MSFDC, an early Internet bill payment system.

Founded: March 2012

Funding: Total funding to date is $28 million raised in two round. The latest round (a $21 million Series B) closed in May and was led by Foundation 8, which was joined by previous investors Madrona Venture Group, Vulcan Capital, and Bloomberg Beta. Several prominent Seattle-area angels also joined in the round.

Why they're one of the 50: Until recently, predictive analytics and machine learning have been limited to large sophisticated organizations, such as Wall Street firms. While open-source tools give IT a way to reduce some costs associated with these tools, they still require large teams to unlock the value in an organization's data.

As a Context Relevant spokesperson wrote to me, "A world in which every business question requires a large new project with a budget to match is not sustainable." True enough.

To address this problem, Context Relevant has developed software that makes it possible for non-specialists to apply the power of machine learning in order to answer questions about everything from valuing derivatives to identifying the optimal webpage to present to a prospective customer. Context Relevant's machine learning technology automatically learns relationships within data and locates the information of greatest utility, shifting the focus of analytics from process to results, which, the company argues, radically reduces time-to-insight.

Context Relevant uses real-time data from HDFS, SQL, web logs, CRM systems, market data, and social media to output analyses and projections on how a business might be impacted in the future. The application also uses "behavioral libraries," which analyze interactions specifically for finance, web personalization, and online travel.

Competitive Landscape: Context Relevant will compete with incumbents, such as SAS, as well as with such startups as Pivotal, DataTorrent, and Skytree.

Skytree

SKYTREE. THE MACHINE LEARNING COMPANY

What they do: Develop machine-learning platforms for Big Data analytics.

Headquarters: San Jose, CA

CEO: Martin Hack, who previously served as a director of marketing for GreenBorder Technologies (acquired by Google) and as a product line manager for SonicWALL.

Founded: 2012

Funding: Skytree secured $18 million in Series A funding in April 2013. U.S. Venture Partners led the round and was joined by a new investor syndicate that includes UPS and Scott McNealy, co-founder and former CEO of Sun Microsystems and Chairman of Wayin. Additional investors include Javelin Venture Partners and Osage University Partners. To date, Skytree has raised a total of $19.6 million.

Why they're one of the 50: According to Skytree, advanced analytics, contrary to popular belief, "is not a meat grinder into which you can dump data in one end and expect nuggets of wisdom to come out of the other end."

Skytree has created a general purpose platform that allows data scientists to focus on what matters most, which Skytree says is Mean Time to Insights (MTI), and focus on what they are good at: building and deploying analytic models rather than coding algorithms.

Skytree is certified with major Hadoop distributions Cloudera, Hortonworks, and MapR. Skytree also supports YARN to deliver analytics on complex and fast-changing datasets stored in the cloud. Their architecture can be installed in the cloud, across a cluster, or locally.

Skytree argues that machine learning is the key that unlocks an entire treasure trove of predictions, customer recommendations, and anomaly detections that most people don't even know are possible. Machine learning solves that problem by unleashing algorithms on massive amounts of data and finding patterns that data scientists didn't even know existed.

Customers include Adconion Media Group, Brookfield, CANFAR, eHarmony, SETI Institute, and USGA.

Competitive Landscape: Skytree says that most of the competition they run into is either from roll-your-own solutions or from legacy BI platforms from the likes of SAS and IBM, which potential customers may simply choose to stick with. Other startups in this space include Ayasdi and Entrigna.

Sumo Logic

What they do: Apply machine learning to data center operations, using data analysis to pinpoint anomalies, predict and uncover potentially disruptive events, and identify vulnerabilities.

Headquarters: Redwood City, CA

CEO: Vance Loiselle, formerly VP of Global Services at BMC. He joined BMC via the acquisition of BladeLogic, which he co-founded. BMC acquired BladeLogic for $800 million.

Founded: 2010

Funding: $80 million in funding from Sequoia Capital, Accel Partners, Greylock Partners, and Sutter Hill Ventures.

Why they’re one of the 50: Sumo Logic claims to address the “unknown” problem of machine data: how do you get insights about data that you don’t know anything about, or, worse, what do you do when you don’t even know what you should be looking for?

Sumo Logic argues that managing machine data – the output of every application, website, server, and supporting IT infrastructure component in the enterprise – is the starting point for IT data analysis. Many IT departments hope they will be able to improve system or application availability, prevent downtime, detect fraud, and identify important changes in customer and application behavior by studying machine logs. However, traditional log management tools rely on pre-determined rules and thus fail to help users proactively discover events they don’t anticipate.

Sumo Logic’s Anomaly Detection attempts to solve this pain point by enabling enterprises to automatically detect events in streams of machine data, generating previously undiscoverable insights within a company’s entire IT and security infrastructure and allowing remediation before an issue impacts key business services.

Sumo Logic uses pattern-recognition technology to distill hundreds of thousands of log messages into a page or two of patterns, dramatically reducing the time it takes to find a root cause of an operational or security issue.

Customers include Netflix, McGraw-Hill, Orange, Pagerduty, and Medallia.

Competitive Landscape: Sumo Logic will compete with the likes of CloudPhysics, Splunk, and open-source alternatives like Elasticsearch and Kibana.

Built for Speed

These startups place a premium on speed, giving customers the ability to gain real-time insights from their data.

DataTorrent

What they do: Enable companies to take action in real time via a real-time stream processing platform built on Hadoop.

Headquarters: Santa Clara, CA

CEO: Phu Hoang, who was previously a founding member of the engineering team at Yahoo!, where he served as Executive VP of Engineering.

Founded: 2012

Funding: The startup closed an $8 million Series A round in June 2013. August Capital led the round and was joined by AME Cloud Ventures. The company previously secured $750K in seed funding from Morado Ventures and Farzad Nazem.

Why they're one of the 50: DataTorrent argues that we'll soon start thinking about latency issues when we think about Big Data solutions. A whole sub-sector has sprung up pushing the Fast Data movement, in fact.

DataTorrent RTS enables enterprises to take action in real-time as a result of high-performance complex processing of data as it is created. DataTorrent points out that "data is happening now, streaming-in from various sources – in real-time, all the time." Many organizations struggle to process, analyze, and act on this never-ending and ever-growing stream of information.

For some insights, by the time data is stored to disk, analyzed, and responded to, it's already too late. For instance, if a hacker compromises a credit card account and manages to make a few purchases, plenty of damage has already been done, even if that account is cut off within minutes. DataTorrent contends that an organization's ability to recognize and react to events instantaneously isn't just a business advantage. In today's word, it is a necessity.

In June 2014, DataTorrent released the G.A. version of DataTorrent RTS, a real-time streaming analytics engine built on Hadoop. Unlike traditional batch processing that can take hours DataTorrent claims that it can process more than 1 billion data events per second, which is the equivalent of processing 46 cumulative hours of streaming Twitter data in one second.

Competitive Landscape: DataTorrent's main competition comes from IBM InfoSphere Streams and SAS Event Stream Processing Engine.

MemSQL

What they do: Provide in-memory row store and flash optimized column store database technology for real-time and historical Big Data analytics.

Headquarters: San Francisco, CA

CEO: Eric Frenkiel. Before MemSQL, he worked at Facebook on partnership development.

Founded: 2011

Funding: The startup is backed by $45 million in funding from Accel Partners, Khosla Ventures, First Round Capital, and Data Collective. Their most recent funding was a $35 million Series B closed in January 2014.

Why they're one of the 50: Big Data and real-time analytics have the potential to profoundly impact the way organizations operate and how they engage with customers. However, there are challenges that prevent companies from fully extracting value from their data. Legacy database technologies are prone to latency, require complex and expensive architectures, and rely on slow disk-based technology.

The result is an outdated computing infrastructure that cannot handle the velocity and volume of data in the timeframe required of a true real-time solution.

MemSQL says that it solves this performance bottleneck with a distributed in-memory computing model that runs on cost-effective commodity servers. MemSQL's in-memory SQL database accelerates applications, powers real-time analytics, and combines structured and semi-structured data into a consolidated Big Data solution. MemSQL says that it empowers organizations to make data-driven decisions, which helps them to better engage customers, discover competitive advantages, and reduce costs.

Customers include Comcast, Zynga, Ziff Davis, Shutterstock, Novus, Samsung, CPXi, Narus, Mattr, Prodege, Tradelab, Clarity Services, and Kurtosys.

Competitive Landscape: Competitors include incumbents like SAP and Oracle, the open-source platform MongoDB, and startups such as Aerospike.

ParStream

What they do: Develop database technologies to enable real-time Big Data analytics.

Headquarters: Cupertino, CA

CEO: Peter Jensen. Before joining Parstream, Jensen was the CEO of StopTheHacker, which was acquired in 2013 by Cloudflare. Before StopTheHacker, Peter was VP of worldwide sales for Pancetera (acquired by Quantum) and Thinstall (acquired by VMware).

Founded: 2008

Funding: ParStream has secured $13.6 million in Series A and B funding from Khosla, Baker Capital, CrunchFund, Tola Capital and Data Collective.

Why they're one of the 50: Traditional databases just weren't designed for Big-Data-scale analytics, and they certainly aren't able to deliver those insights in real time. Traditional databases analyze data sequentially and aren't able to take advantage of advances in multi-core processing.

When I spoke with co-founder Michael Hummel at CTIA last year, he noted that memory is a big bottleneck for traditional databases. Meanwhile, the Big Data database darling, Hadoop, has trouble scaling efficiently.

Hummel argues that ParStream's database was purpose-built for speed. Whereas many database platforms exist for the purpose of storing and analyzing large quantities of data, ParStream was designed to deliver faster response times and to reduce Big Data storage infrastructure costs in the process.

ParStream enables "Fast Data" by using a distributed architecture that processes data in parallel. ParStream was specifically engineered to deliver both big data and fast data, enabled by a unique High Performance Compressed Index (HPCI). This removes the extra step and time required for decompression of data.

ParStream claims to provide sub-second response times on billions of data records while continuously importing new data.

Customers include CAKE, MPREIS, bd4travel, INRA, Searchmetrics, Ellisphere, the German Ministry of Economics, and DERTour.

Competitive Landscape: Competitors include SAP HANA, Apache platforms and Vertica Systems (HP).

Tidemark

What they do: Provide a cloud-based business planning and analytics solution.

Headquarters: Redwood City, CA

CEO: Christian Gheorghe. He previously founded Tian Software, which was acquired by OutlookSoft. After SAP acquired OutlookSoft, Christian served as SVP and CTO.

Founded: 2010

Funding: Tidemark has raised approximately $80 million in funding from Greylock Partners, Andreessen Horowitz, Redpoint Venutres, Tenaya Capital, and Silicon Valley Bank.

Why they're one of the 50: Most companies already have large volumes of data, and it just continues to pile up. However, most companies also struggle to understand how the business is performing, what's driving the performance, and what they can do about it. Much of the data people use to make decisions is no longer housed inside the walls of the business. Instead, it comes from external, unstructured sources like news outlets, Twitter, and other streams.

Legacy systems weren't designed to handle this type of data, so companies are stuck trying to manually assemble disparate data into a view that can be used by the business to steer the company. Unfortunately, this data becomes outdated almost instantly, and it isn't actionable.

To tackle this problem, Tidemark designed its system by focusing on actual end user pain points. Okay, every vendor says they have a customer focus, but that's often empty rhetoric. Prior to writing a single line of code for Tidemark's applications, CEO Christian Gheorghe and his co-founder, Tony Rizzo, dedicated nine months to surveying Fortune 1,000 companies to find out where they were struggling with their existing business intelligence and analytics solutions.

Armed with this information, Tidemark designed its cloud-based solution to financial and operational business planning, consolidation, and analytics to the entire enterprise. Tidemark does this by providing a series of intuitive analytical apps that replace manual, spreadsheet-based planning processes or inflexible legacy solutions. Unlike these legacy solutions, Tidemark gives enterprises access to real-time, in-context data about the company's financials. As a result, they improve business performance, reduce risk, and accelerate decision making.

Tidemark helps companies “run in the now” by giving every user access to a computational grid that is free of cubes and doesn’t limit data use by constraining volumes or dimensions. This allows data-driven decisions to be processed up to 10 times faster than typical cube-based approaches, which helps eliminate lag time and confusion over outdated information.

The solution also adds context to numbers, and helps business users manage data by “processes, not cubes.”

Customers include Acxiom, Brown University, Chiquita, Chuck E. Cheese, Hostess Brands, HubSpot, Netfix, ServiceSource, and University of Miami.

Competitive Landscape: Tidemark divides its competitors into two camps: the first consists of legacy on-premises vendors, such as SAP BPC, Oracle Hyperion, and IBM Cognos. The second are cloud-hosted providers, which include the likes of Adaptive Insights, Anaplan, and Host Analytics.

VoltDB

What they do: Provide an in-memory database management system for processing “Fast Data.”

Headquarters: Bedford, MA

CEO: Bruce Reading, who was formerly SVP and GM of Compuware

Founded: 2009

Funding: $18.7 million from Sigma Partners and Kepha Partners.

Why they’re one of the 50: Applications that require high-speed data ingestion along with real-time analytics and in-the-moment decisioning capabilities, such as IoT and M2M, present unique problems that traditional database systems and other Big Data architectures cannot easily solve.

VoltDB is tackling this problem with a "NewSQL" database solution designed to deliver both the ACID compliance of a relational database and the scalability of NoSQL. It is purpose-built to support any application – financial, mobile, energy, gaming, retail or otherwise – with high requirements for ingesting large quantities of information at a high rate (database throughput requirements reaching millions of operations per second).

It provides visibility into real-time data and enables the application to make automated decisions based upon the data. It does this by employing a closed loop process where analytics are fed back into the decision-making process, allowing data to inform the front end of the application that is acting on new data as it arrives, thus maximizing business value.

VoltDB was built specifically for speed. While many others in this space have focused on general database work or better reporting, VoltDB has focused entirely on the "Fast Data" problem that organizations face today and in the future. The Velocity of data is the largest factor driving the accumulation of data that is so often referred to as Big Data.

Customers include Airpush, Alcatel-Lucent, Arkalogic, BOLDstreet, CGI/Logica, Eagle Investments, Ericsson, HP, Mavizon, Neverblue, Neustar, Novatel Networks, Openet, Shopzilla, and Yahoo!

Competitive Landscape: The in-memory database space is heating up. VoltDB competes with solutions from incumbents, such as SAP HANA and Oracle TimesTen. They also compete with such startups as MemSQL and NouDB.

Analytics for Sales, Marketing, and Social Media

These startups analyze buying behaviors and consumer preferences.

BloomReach

What they do: Provide Big Data marketing applications.

Headquarters: Mountain View, CA

CEO: Raj De Datta, formerly Entrepreneur-in-Residence at Mohr-Davidow Ventures and Director of Product Marketing at Cisco.

Founded: 2009; remained in stealth mode until February 2012.

Funding: BloomReach has raised $41 million in three rounds of funding from Bain Capital Ventures, NEWA, and Lightspeed Venture Partners.

Why they're one of the 50: Forrester Research estimates that the U.S. e-commerce market will hit $370 billion by 2017; worldwide, the market already topped $1 trillion, according to eMarketer.

Connecting these consumers with the products and content that they want and need means that smart businesses end up capturing an ever larger slice of that market. Companies like Amazon, Blue Nile, and even Walmart already leverage large-scale data and tech advantages. To compete with these companies, smaller retailers need to reach their audiences with increasing precision and accuracy.

BloomReach's Organic Search combines web-wide intelligence and site-level content knowledge with machine learning and natural language processing to predict demand and dynamically adapt pages to match consumer behavior and intent. This helps companies capture up to 60 percent of net-new users. BloomReach also takes a data-driven approach to m-commerce, more accurately matching consumers with content and products. This increases revenue-per-site-visit by up to 40 percent, and drives sales across all shopping channels.

Customers include Guess, Deb Shops, and Neiman Marcus

Competitive Landscape: Big Data marketing platforms are popping up faster than weeds after the first spring rains. While behemoths like Google, Amazon, and IBM have similar technologies, they keep them in-house. Others providing similar services the rest of us can use include Kontera, DataSong, and Persado.

NGDATA

What they do: Provide a Big Data management solution.

Headquarters: Gent, Belgium; U.S. HQ: New York, NY

CEO: Luc Burgelman. Prior to NGDATA, he co-founded Porthus, which delivered cloud-based solutions, had an IPO in 2006 and was acquired by Descartes Systems Group in 2010. At that time, Burgelman became the EVP global marketing and product strategy.

Founded: 2009

Funding: The startup’s most recent funding came in September of 2013 when NGDATA closed a $3.3 million investment round led by Capricorn Venture Partners with participation from existing investors, including Sniper Investments and angel investors. Before that, in October 2012, NGDATA received $2.5 million from ING, Sniper Investments, Plug and Play Ventures, and U.S. angel investors.

Why they’re one of the 50: Consumer-focused companies, such as banks, media, and telcos, have actually collected more data about their consumers than companies like Google and Amazon. However, in spite of possessing so much data, these companies still don’t know their customers well.

There are hundreds of internal and external data sources, and each represent a small silo of information about the consumer. Traditional systems are inadequate to process massive unstructured data, scale storage elastically and locate actionable data quickly in large datasets.

Additionally, analyzing batch data does not provide an up-to-date consumer view, and aggregating data into a unified Hadoop-based system still does not provide a single view of the consumer because data exists in silos without the ability to match various interactions to a unique consumer.

NGDATA intends to solve these issues by delivering a combination of interactive Big Data management, machine-learning technologies and consumer intelligence in a single solution.

The company’s customer experience management solution, Lily, enables storing, indexing and analyzing massive data sets and provides a "Customer DNA" based on thousands of metrics about the behavior and context of the customer. Lily allows business users to quickly locate the most relevant data and garner deep insights. Through machine-learning, the solution learns from end-users’ profiles and their data interactions to provide business users with the ability to do real-time segmentation and better target clients with more personalized offerings.

Customers include AXA, De Persgroep, France Telecom, Orange, and Telenet.

Competitive Landscape: This is a tough space. NGDATA will compete with the likes of AgilOne, Continuuity, InsightsOne, Platfora, and SAS, to name only a few.

Ninja Metrics

What they do: Helps companies determine the “Social Value” of gamers, app users, and ticket buyers.

Headquarters: Redondo Beach, CA

CEO: Dmitri Williams. He previously ran the Virtual Worlds Observatory.

The author of more than 40 peer-reviewed articles on gamer psychology and large-scale data analysis, Dmitri's work has been featured on CNN, Fox, the Economist, the New York Times, and most major news outlets.

He has testified as an expert on video games and gamers before the U.S. Senate, and is a regular speaker at industry and academic conferences.

Founded: 2011

Funding: $2.8M from Harvard Business School Angels and Tech Coast Angels.

Why they're one of the 50: The entire business landscape now has access to automated predictive analytics engines that replace the work that used to be done by an expensive on-site team of data scientists. However, it's not always easy to gain real insights from the data, and it's especially hard to figure out which people in a social network have the most influence.

A new report from Juniper Research predicts that the mobile video games market will be worth more than $29 billion by 2016 – a 38% increase over this year's projected total of $20.9 billion.

Ninja Metric's analytics platform, the Katana engine, analyzes the actions of video gamers to determine the "Social Value" of a given player, or the players most likely to influence others to participate in and spend money on a game. Ninja Metrics' "Social Value" scoring helps developers better understand how gamers interact with their apps, while also helping marketers quickly monetize the everyday social connections of their customers.

With "Social Value" gaming companies can pinpoint the "social whales" in a system and see how they influence others to buy in-game items, level ups, and even purchase apps or, say, movie tickets. These insights are delivered via an automated, cloud-based tool that delivers daily metrics.

Customers include Gamzio, Imperia Online, Sleepy Giant, Nerd Kingdom, Beyond Gaming, and DDM.

Competitive Landscape: For now, Ninja Metrics is in a unique position. While there are tons of other analytics engines out there, none are specifically discussing Social Value/Social Whales.

PlaceIQ

What they do: Provide a location-based insights and consumer-targeting platform.

Headquarters: New York, NY

CEO: Duncan McCall, who formerly founded PublicEarth.

Founded: 2010

Funding: The startup is backed by $27 million raised in three round of funding from IA Ventures, Social Leverage, kbs+ Ventures, Neu Venture Capital, US Venture Partners, Valhalla Partners, Harmony Partners, and Iris Capital.

Why they're one of the 50: Mobile advertising and marketing present a unique challenge. The typical way companies try to understand consumer behavior online is through cookies. On smartphones and tablets, cookies don't have as much traction. Even if cookies are enabled in mobile browsers, they aren't terribly useful, since browsers are giving way to apps.

However, a potentially better replacement is location. Just as cookies track your journeys through the Web, marketers can glean demographic information from the actual physical locations you've visited.

PlaceIQ says that it "provides a multidimensional depiction of consumers across location and time." This allows brands to define audiences and intelligently communicate with those audiences to support greater ROI. PlaceIQ's platform analyzes customers based on where they have been, adding relevancy to a brand's marketing strategy and providing demographic insights that can help improve business strategy.

Customers include Mazda, Darden Restaurants, and Montana Tourism.

Competitive Landscape: The competition includes Verve Mobile, xAd, Placed, Sense Networks, jiWire, 4INFO, and Millennial Media.

Pursway

What they do: Pursway uses big data analytics and proprietary algorithms to help companies identify the customers who are most likely to influence how people in their social networks shop.

Headquarters: Herzliya, Israel; U.S. HQ: Waltham, MA

CEO: Dave Ellenberger, who previously served as CEO of 170 Systems.

Founded: 2009

Funding: $17 million from Battery Ventures and Globespan Capital Partners.

Why they're one of the 50: In an era of social-savvy, data-driven marketing initiatives, marketers are increasingly looking for ways to unlock the power of relationship-based marketing. Most consumer behavior is influenced by the opinions of people we know and trust – family, friends, and colleagues. While marketers have known this for quite a while, they have trouble acting on it.

Pursway's software is intended to improve customer acquisition, cross-selling opportunities, and retention. By imprinting a social graph onto existing customer and prospect data, identifying actual relationships between buyers, and identifying target customers who have a demonstrated influence over others' purchasing decisions, Pursway argues that it can help consumer-facing organizations close the gap between how businesses market and how people actually buy.

The core of Pursway's various services is Connect, a Big Data database that maps current relationships among more than 120 Million U.S. consumers. Drawing from thousands of open data sources, Connect matches entities to create a single network in which real human connections are identified. Connection types include places of residence, schools attended, places of work, professional activities, travel, and social media sharing.

MyPIVO, Pursway's dashboard, gives marketers a view into their customers' real-life relationships and their potential sales influence. Pursway's subscription-based scoring tools give marketers insight into the friendships, connections, and potential influence of existing customer and prospect pools to better target marketing messages, drive sales, and increase the ROI of any data-driven marketing campaign.

Customers include Sony, Orange, and Comcast.

Competitive Landscape: Competitors include Angoss, IBM, and SAS.

SumAll

What they do: Provide data analytics tools focused on delivering marketing, sales and social media insights.

Headquarters: New York, NY

CEO: Dane Atkinson. He was formerly CEO of Squarespace and founder/CEO of SenseNet.

Founded: 2011

Funding: SumAll is backed by several rounds of funding that total $13.5 million, which was raised from Battery Ventures, Wellington Partners, Matrix, General Catalyst, and venture debt from SVB.

Why they're one of the 50: The marketing vertical is one of the most aggressive sectors when it comes to adopting Big Data tools. However, many tools require that marketers basically start over from scratch, mining data from social media sites but ignoring their own massive volumes of data.

SumAll's product is an analytics tool that helps businesses make more money by using their own data. SumAll tries to break down various data silos, from those associated with legacy apps to those involved with social media.

SumAll brings all the disparate revenue, payment, social and organic traffic data into one place so users can see the interactions across their business and understand if a social campaign is driving traffic which is converting into traffic. SumAll can help businesses figure out, say, the value of a "like" on Facebook or the value of a website visit.

Customers include Siemens, Pandora, Starbucks, HBO, and TED.

Competitive Landscape: These aren't necessarily head-to-head comparisons, but SumAll will compete with Hootsuite, Nimble, Gooddata and Kissmetrics.

Teikametrics

What they do: Provide data analytics software for third-party Amazon sellers that enables better trading decisions and manages inventory in Fulfillment by Amazon (FBA).

Headquarters: Boston, MA

CEO: Alasdair Mclean-Foreman, who previously founded an e-commerce company in his dorm room, which grew into a multi-million dollar company selling high-end sporting goods. He also founded the weight loss and fitness company Traineo and has built and provided e-commerce solutions to large organizations including Newscorp, The Times of London, L'Oreal, and The New York Marathon.

Founded: 2011

Funding: $1.5M from early stage private investors.

Why they're one of the 50: Amazon is the new Walmart, having the ability to effect broad swaths of the retail economy through their policies and practices. As Amazon grows, there is a big opportunity for retailers to profit and Teikametrics' software is designed to help them do just that. Amazon is selling more and more brands on their own, and Teikametrics helps retailers pinpoint the brands that Amazon is not selling.

Amazon has created a unique platform for third-party sellers, specifically with Fulfillment by Amazon (FBA). Retailers have the opportunity to make millions on Amazon, but this is not possible without the right data.

Teikametrics' SaaS platform captures multi-variable data for each seller's transactions, creating a data-driven model that improves supply side efficiency and identifies new investment opportunities. By identifying the most profitable inventory and analyzing Amazon market conditions, sellers can make better trading decisions. The platform is always adjusting to the rapidly changing supply and demand on Amazon.com.

The software helps end users create a customized trading model specific to their needs. The principles of the software are similar to that of commodity and equities trading.

Customers include many traditional brick-and-mortar stores who have no choice but to move parts of their business to Amazon. Named customers include The Vitamin Shoppe, f.y.e., Smart Home, and Newbury Comics.

Competitors: Teikametrics will compete with Channel Advisor, Appeagle, Channel Max, Mercent, and Wiser.

Billion-Dollar Baby

Raise a billion dollars in funding, forge a tight partnership with Intel, and you get your own category.

Cloudera

What they do: Provide a Hadoop-based Big Data platform.

Headquarters: Palo Alto, CA

CEO: Tom Reilly. Prior to joining Cloudera in 2013, he was VP and GM of enterprise security at HP. He also served as CEO of enterprise security company ArcSight, which HP acquired in 2010.

Founded: 2008

Funding: Cloudera has raised over $1 billion in venture capital to date. (Yep, that's not a typo; that's $1 billion with a "B." I'm wondering what you are: "why dilute yourself that much," but founder Mike Olson has given a pretty concise answer: a close relationship with Intel, which provided much of the new funding.) Other investors include Accel Partners, Google Ventures, Greylock Partners, Ignition Partners, In-Q-Tel, Meritech Capital Partners, and T. Rowe Price.

Why they're one of the 50: Big Data is hot, and Cloudera pioneered the Hadoop-based Big Data space.

Cloudera lets users query all of their structured and unstructured data to gain a view beyond what's available from relational databases. Cloudera recently released Impala, a new open-source interactive query engine for Hadoop that enables interactive querying on massive data sets in real time.

Moreover, they're sitting on a giant pile of VC cash and have a top-notch management team. Cloudera is also the first major Big Data vendor to start investing heavily in a Big Data Achilles' heel: security.

In June 2014, Cloudera acquired Gazzang, a startup specializing in encryption software for Big Data environments.

Frankly, I thought long and hard about leaving Cloudera off this list – not because they don't belong, but because they've been doing well enough for long enough that I'm not sure that the label "startup" really fits that well anymore.

However, they pretty much proved the business case for Hadoop, and they're moving the space forward with the Gazzang acquisition, so, for the time being, anyway, I'd be foolish to exclude them.

Customers include Experian, FICO, National Cancer Institute, Nokia, and Western Union.

Competitive Landscape: Cloudera clearly has first-mover advantage, but competitors include EMC, Pivotal, Hortonworks and MapR. One of their earlier competitors, Intel, which had its own home-grown distribution, dropped it and adopted Cloudera instead.

The NoSQL Crowd

They've left SQL behind but haven't jumped on the Hadoop bandwagon.

Aerospike

What they do: Provide a hybrid-memory (DRAM and flash) NoSQL database for mission-critical, real-time Big-Data-driven apps.

Headquarters: Mountain View, CA

CEO: Joe Gottlieb. Before joining Aerospike, he served as president and CEO of Sensage, where he led it to a successful acquisition by KEYW.

Founded: 2009

Funding: They've raised three rounds of funding, a Series A from Alsop Louie and Tim Draper, a Series B led by NEA, and a $20M Series C led by NEA as well. (Aerospike has not released funding specifics for its first two rounds.)

Why they're one of the 50: Enterprises collect data from everywhere these days. Data generated from click streams, tweets, likes, comments, ratings, video, photos, blogs, devices, sensors, call detail records, subscriber data, enterprise sales data, demographic data and more all have value – value that could be unlocked if businesses find a way to collect and analyze this type of data.

Increasing competition is driving the need to engage consumers with real-time context-driven applications. This shift is putting pressure on existing enterprise infrastructure. The new pattern of read/write workloads, the need to combine low latency transactions with "hot" analytics is not being met by DRAM intensive caching tiers, in-memory databases, or first-gen NoSQL databases.

According to Market Research Media, the worldwide NoSQL market is expected to reach $3.4 billion by 2018, with a compound annual growth rate (CAGR) of 21 percent between 2013 and 2018. The NoSQL market is expected to generate $14 billion in revenues over the period 2013-2018.

Aerospike was built from scratch to push the limits of modern hardware – multi-core and multi-CPU servers and flash storage. Aerospike's NoSQL database relies on a Hybrid Memory System that treats Flash like memory, not a rotational drive. Aerospike claims that its in-memory Flash/ DRAM approach delivers the speed of DRAM with the price/performance of Flash. The result is that Aerospike clusters can deliver the same service with substantially fewer servers than DRAM-only systems. Fewer servers translate to better reliability and much lower costs.

Aerospike Smart Clusters automatically distribute data across a cluster. They make sure that transactions are ACID, data is never lost and service never goes down – not even for maintenance. When a server does go down, the system automatically fails over, moves data around and heals itself with no manual intervention.

Customers include eBay, Acuity, Federated Media Publishing, x+1, The Trade Desk, and madvertise.

Competitive Landscape: Aerospike will compete with Redis and Cassandra, as well as any number of DIY projects.

Couchbase

What they do: Provides NoSQL database technology.

Headquarters: Mountain View, CA

CEO: Bob Wiederhold. He formerly served as chairman and CEO of Transitive, which was acquired by IBM in 2008.

Founded: 2011

Funding: On June 26, 2014, Couchbase announced a $60 Million series E round of financing. The round was led by two new investors, WestSummit and Accel Growth Fund and brings the total amount of Couchbase funding to $115 Million. All existing venture capital investors also participated in this round of financing, including Adams Street Partners, Accel Partners, Mayfield Fund, North Bridge Venture Partners, Ignition Partners, and DoCoMo Capital.

Why they're one of the 50: The landscape for Big Data database technology is in flux. Hadoop and NoSQL seem to be the platforms most in favor, although plenty of organizations are still betting on SQL.

Couchbase is placing its bet on NoSQL. The startup argues that its NoSQL document-oriented database technology provides the scalability and flexible data modeling needed for Big Data-scale projects. Couchbase also claims to offer the first NoSQL database for mobile devices.

In May 2014, Couchbase launched Couchbase Mobile, an operational big data management platform that supports both cloud and edge based computing for better mobile analytics.

Mobile applications are expected to work anytime or anywhere, regardless of the limits of network availability or the complexity required to enable data synchronization. But mobile app developers are limited in the apps they can build because current mobile databases are all relational and, thus, are poorly suited to manage the many types of unstructured data generated by today's mobile users. Additionally, Couchbase argues that no productized solution has existed for the synchronization of unstructured data that is shared between the cloud and the device.

Couchbase Mobile delivers a solution that enables mobile app developers to meet the demand for always-available, always-responsive mobile applications. Couchbase Mobile includes three components: Couchbase Lite, Couchbase Sync Gateway, and Couchbase Server. Taken together, these products provide mobile application developers with a platform for creating robust and network-independent applications, simplifying the code needed to manage data synchronization and reducing time to market.

Customers include AOL, Cisco, Concur, LinkedIn, Orbitz, Salesforce.com, Zynga, Amadeus, McGraw-Hill Education, and Nielsen.

Competitive Landscape: Competitors include MongoDB and DataStax.

DataStax

What they do: Provide a NoSQL Big Data application platform.

Headquarters: San Mateo, CA

CEO: Billy Bosworth, who previously served as VP and GM of the Enterprise Data Unit at Quest Software.

Founded: 2010

Funding: DataStax's most recent round of funding was a Series D round for $45M, taking them to a total of $84M in funding. DataStax is backed by Lightspeed Venture Partners, Crosslink Capital, Meritech Capital Partners, Scale Venture Partners, DFJ Growth, and Next World Capital.

Why they're one of the 50: DataStax has a solid management team, a serious amount of VC funding, and are tackling a pressing pain point. DataStax argues (and most in the BD space tend to agree) that traditional relational databases from Oracle and the like are incapable of supporting the emerging real-time line of business applications. However, relational databases are far more expensive, less scalable and vulnerable to going offline during disasters.

DataStax enables companies to develop applications that can analyze data in real time, scale as usage increases, and avoid disaster by spanning multiple databases and the cloud.

DataStax's NoSQL database is a commercialized version of Apache Cassandra. It is architected to support real-time enterprise databases that require vast scalability, high-velocity performance, flexible schema design, and continuous availability. DataStax Enterprise combines the Cassandra database with enterprise search, analytics and security, and a visual monitoring and management tool called OpsCenter.

Customers include eBay, Netflix, Adobe, Intuit, Healthcare Anytime, and Ooyala.

Competitive Landscape: The competition is really tight for this particular subsector of the Big Data market, with DataStax competing against 10Gen, NuoDB, MongoDB, Couchbase, Hortonworks and others. This space is a land grab for the time being, but it won't stay that way forever.

ScaleArc

What they do: Provide database infrastructure software that simplifies the way database environments are deployed and managed.

Headquarters: Santa Clara, CA

CEO: Justin Barney, who previously served as ScaleArc's President and COO before being promoted to CEO. Prior to ScaleArc, he was VP of Sales for Software in the Americas at Juniper Networks, and he previously served as VP of Sales at Citrix Systems.

Founded: 2009

Funding: ScaleArc is backed by $18 million from Accel Partners, Trinity Ventures, Nexus Venture Partners, and angel investors.

Why they're one of the 50: According to ScaleArc, the growth of online and mobile applications is straining traditional database infrastructures. For companies doing business online, application availability and performance are key determinants of the customer experience and, ultimately, revenue.

However, companies struggle with the complex challenge of growing their database infrastructures to handle increasing demand, without negatively impacting the customer experience, or consuming resources that may be better used elsewhere. Traditional SQL environments are bogged down by an increasing volume of database queries from a growing number of applications that need access to structured data – leading to poor application performance and system outages.

And the problem is even worse for mobile applications, as performance takes an even bigger hit with increased latency.

ScaleArc argues that companies need a way to optimize SQL query traffic without extensive modifications to existing applications or databases. To improve performance, they need to offload existing databases without investing in costly new infrastructure. Finally, they need full visibility into SQL traffic to more efficiently troubleshoot and resolve issues before they become major problems that impact revenue.

ScaleArc's flagship product, iDB, is software that inserts transparently between applications and databases, requiring no modifications to applications or databases. ScaleArc claims that it can be deployed in about 15 minutes. Then, users gain visibility into all database traffic with granular real-time SQL analytics.

iDB provides instant scalability and higher availability for databases with dynamic clustering, load balancing and sharding capabilities, and it provides a transparent SQL-NoSQL hybrid caching engine, which lets any application use a NoSQL cache without any code changes or drivers.

Customers include Demand Media, Disney UTV, KIXEYE, Sazze (dealspl.us), Flipkart, Weather Decision Technologies and others.

Competitors landscape: ScaleArc will compete with the likes of ScaleBase and ParElastic.

Splice Machine

What they do: Provide a Hadoop-based RDBMS designed to scale real-time applications using commodity hardware without application rewrites.

Headquarters: San Francisco, CA

CEO: Monte Zweben, who previously worked at the NASA Ames Research Center where he served as the Deputy Branch Chief of the Artificial Intelligence Branch. He later founded and served as CEO of Blue Martini Software.

Founded: 2012

Funding: They are backed by $19 million in funding from Interwest Partners and Mohr Davidow Ventures.

Why they're one of the 50: Application and web developers have been moving away from traditional relational databases due to rapidly growing data volumes and evolving data types. New solutions are needed to solve scaling and schema issues. Splice Machine argues that even a few short months ago Hadoop, while viewed as a great place to store massive amounts of data, wasn't ready to power applications.

Now, with emerging database solutions, features that made RDBMS so popular for so long, such as ACID compliance, transactional integrity, and standard SQL, are available on top of the cost-effective and scalable Hadoop platform. This enables developers to get the best of both worlds in one general-purpose database platform.

Splice Machine provides all the benefits of NoSQL databases, such as auto-sharding, scalability, fault tolerance, and high availability, while retaining SQL, which is still the industry standard. Splice Machine optimizes complex queries to power real-time OLTP and OLAP applications at scale without rewriting existing SQL-based apps and BI tool integrations. By leveraging distributed computing, Splice Machine can scale from terabytes to petabytes by simply adding more commodity servers. Splice Machine is able to provide this scalability without sacrificing the SQL functionality or the ACID compliance that are cornerstones of an RDBMS.

Splice Machine allows users to replace Oracle and MySQL databases with a scale-out Hadoop RDBMS.

An early **named customer** is Harte Hanks.

Competitive Landscape: Competitors include Oracle, MemSQL, NuoDB, Datastax, and VoltDB.

Hadoop Darlings

Hadoop is nearly synonymous with Big Data these days, and these startups hope that one day they will be too.

Altiscale

What they do: Provide Hadoop-as-a-Service (HaaS).

Headquarters: Palo Alto, CA

CEO: Raymie Stata, who was previously CTO of Yahoo.

Founded: March 2012

Funding: Altiscale is backed by $12 million in Series A funding from General Catalyst and Sequoia Capital, along with investments from individual backers.

Why they're one of the 50: Hadoop has become almost synonymous with Big Data, yet the number of Hadoop experts available in the wild cannot hope to keep up with demand. Thus, the market for Hadoop-as-a-Service should rise in step with Big Data. In fact, according to TechNavio, the Hadoop-as-a-Service market will top $19 billion by 2016.

Altiscale's service is intended to abstract the complexity of Hadoop. Altiscale's engineers set up, run, and manage Hadoop environments for their customers, allowing customers to focus on their data and applications. When customers' needs change, services are scaled to fit – one of the core advantages of a cloud-based service.

Altiscale argues that they are "the only firm to actually provide a soup-to-nuts Hadoop deployment. By comparison, AWS forces companies to acquire, install, deploy, and manage a Hadoop implementation – something that takes a lot of time."

Customers include MarketShare, OpenTable, and Internet Archive.

Competitive Landscape: The Hadoop-as-a-Service space is heating up. Competitors comes from incumbents, such as Amazon Elastic MapReduce (EMR), Microsoft's Hadoop on Azure, and Rackspace's service based on Hortonworks' distribution. Altiscale will also compete directly with Hortonworks and with such startups as Cloudera, Mortar Data, Qubole, and Xpleny.

Continuuity

What they do: Provide a Hadoop-based Big Data application hosting platform.

Headquarters: Palo Alto, CA

CEO: Jonathan Gray, who was previously an HBase software engineer at Facebook.

Founded: 2011

Funding: $12.5 million from Battery Ventures, Ignition Partners, Andreessen Horowitz, Data Collective and Amplify Partners.

Why they're one of the 50: Continuuity has come up with a clever way to get around the dearth of Hadoop experts: they offer an application developer platform targeted at Java developers. The lower-level infrastructure is all abstracted away by the Continuuity platform.

The startup's flagship product, Reactor, is a Java-based integrated data and application framework that layers on top of Apache Hadoop, HBase, and other Hadoop ecosystem components. It surfaces capabilities of the infrastructure through simple Java and REST APIs, shielding end users from unnecessary complexity. Continuuity describes Reactor as a "Big Data Application Server for Hadoop." It abstracts all the complexities of Hadoop and enables any developer to build Big Data applications.

Continuuity's Loom service is a cluster management solution. Clusters created with Continuuity Loom utilize templates of any hardware and software stack, from simple standalone LAMP-stack servers and traditional application servers like JBoss to full Apache Hadoop clusters comprised of thousands of nodes. Clusters can be deployed across many cloud providers (Rackspace, Joyent, OpenStack) while utilizing common SCM tools (Chef and scripts).

In June, Continuuity entered into a partnership with AT&T Labs to develop and release into open source a new real-time data processing framework that will provide streaming analytics capabilities. Initially code-named jetStream, it will be made available to the market via open source in the third quarter of 2014.

Competitive Landscape: As of now, Continuuity is uniquely positioned. Indirect competitors come from the Hadoop-as-a-Service camp (AWS EMR, Altiscale, Infochimps, Mortar Data, etc.). One thing to keep an eye in is the CEO situation. Founding CEO Todd Papaioannou, who was previously VP and chief cloud architect at Yahoo!, left the company last year. Co-founder and previous CTO Jonathan Gray has taken over the CEO role. This is Gray's first role as a business leader.

Hortonworks

What they do: Provide an open-source Apache Hadoop-based Big Data solution.

Headquarters: Palo Alto, CA

CEO: Rob Bearden. He previously served as COO of both SpringSource and JBoss, and also served in senior roles at Oracle, where he directed a $1 billion sales organization, I2 and Manhattan Associates.

Founded: 2011

Funding: In March 2014, Hortonworks nailed down a massive $100 million round of funding. The oversubscribed funding round was led by funds managed by BlackRock and Passport Capital and joined by all existing investors, which include Tenaya Capital, Dragoneer Investment Group, Benchmark Capital, Index Ventures, and Yahoo! This brings total funding to date to more than $225 million.

Why they're one of the 50: Hortonworks argues that "there is no one-size-fits-all solution for Big Data. Every company has its own unique needs to be met. Hadoop is the only Big Data platform that is capable of adapting to each company's needs, but it does not – and cannot – do so in a vacuum. Hadoop works best in partnership with other big data applications, tools, frameworks and platforms so as to offer the most flexible solution for each user. And as other big data technologies mature and companies' needs grows, Hadoop must adapt to address those needs."

Hortonworks believes that community-driven open source is the fastest path to innovation and adoption of Hadoop. Through its work with the Apache Software Foundation and its Hadoop partnership ecosystem, Hortonworks has been able to launch enterprise features and projects into the public domain in partnership – not competition – with other players in the Big Data market.

As a Hortonworks spokesperson wrote to me, "It's not about getting the biggest slice of the pie – it's about making the pie itself as big as possible."

Hmm, that's a lesson U.S. economic policy makers would be wise to learn (but I'm not holding my breath). Investors are buying into this message, however, as evidenced by Hortonworks recent $100 million funding round.

Hortonworks also has, perhaps, the most impressive pedigree of any startup in the Big Data 50. Founded by 24 engineers from the original Yahoo! Hadoop development and operations team, Hortonworks argues (convincingly) that it has amassed more Hadoop experience under one roof than any other organization. Its team members are active participants and leaders in Hadoop development; designing, building, and testing the core of the Hadoop platform.

The Hortonworks Data Platform (HDP) deeply integrates with an organization's existing IT investments to create a modern data architecture. This gives enterprises the foundation for a data lake, which combines data from multiple silos. Organizations can access a unified pool of data from traditional data sources (RDBMS, OLTP, OLAP), data systems (EDW, MPP, RDBMS), and business applications (analytics/BI, customer applications, enterprise applications).

Hortonworks has strategic partnerships in place with Microsoft, Rackspace, Red Hat, SAP, Teradata, and dozens of other companies.

Customers include Spotify, Western Digital, eBay, Bloomberg, Kohl's, AT&T, TrueCar, Cardinal Health and UC Irvine.

Competitive Landscape: Competitors include roll-your-own Hadoop projects and other Hadoop startups, such as Cloudera, and MapR.

MapR Technologies

What they do: Provide a Hadoop distribution/NoSQL Big Data platform.

Headquarters: San Jose, CA

CEO: John Schroeder. He previously served as CEO of Calista Technologies, which was acquired by Microsoft. Before that, he was CEO of Rainfinity, which EMC purchased.

Founded: 2009

Funding: In June 2014, MapR Technologies raised $110 million in financing in a round led by Google Capital, with participation from Qualcomm Ventures and existing investors Lightspeed Venture Partners, Mayfield Fund, NEA, and Redpoint Ventures

Why they're one of the 50: MapR argues that Hadoop suffers from an insufficient high availability design that results in downtime and an inability to protect against the application and user errors that lead to lost data. Hadoop's distributed file system is designed to be "append only," which forces interactive applications to spend excessive time writing new files and results in a 150M file cluster limit.

MapR was founded to address Hadoop's limitations, transforming it into an enterprise-grade system that more organizations can actually use.

The new architecture is a high-performance data platform that supports full random read/write data access, real-time streaming, and can scale to 1 trillion files. MapR argues that this is a new distributed architecture that provides enterprise storage capabilities, such as snapshots and mirroring, but does not have the NameNode scaling drawbacks that plague clustered file systems. MapR also innovated at the shuffle layer to provide high-speed analytic processing.

Additionally, the MapR platform provides self-healing, automated failover, and an integrated in-Hadoop database for real-time capabilities.

Named customers include Ancestry.com, Rubicon, and comScore.

Competitive Landscape: Competitors include Cloudera, Pivotal, and Hortonworks.

Qubole

What they do: Offer Big Data-as-a-Service with a “true auto-scaling Hadoop cluster.”

Headquarters: Mountain View, CA

CEO: Ashish Thusoo, who ran Facebook’s data infrastructure team before co-founding Qubole. He also co-founded Apache Hive.

Founded: 2011

Funding: The startup is backed by $7 million, which includes Series A funding from Lightspeed Ventures and Charles River Ventures, as well as angel funding from Venky Harinarayan and Anand Rajaram.

Why they’re one of the 50: Since Hadoop is a relatively new technology, finding someone with the expertise necessary to run and maintain it can be a tall order. By providing a managed solution, Qubole hopes to make Hadoop an easy-to-use technology.

Qubole handles the initial setup and then maintains the clusters. Qubole's auto-scaling feature automatically spins up users' clusters when a job is started and automatically scales or contracts based on workload, cutting back on costs and management requirements.

An intuitive UI expands the reach of this service beyond data analysts to entire lines of businesses. Qubole contends that some customers have more than 60 percent of their employees using Qubole.

Customers include Pinterest, MediaMath, Nextdoor and Saavn.

Competitive Landscape: Qubole will compete with Altiscale, Amazon EMR, Treasure Data, and others.

Platfora

What they do: Provide a Big Data analytics platform built on Hadoop that is designed for business people, rather than data scientists.

Headquarters: San Mateo, CA

CEO: Ben Werther, who formerly served as VP of Products at DataStax and head of product at Greenplum.

Founded: 2011

Funding: $65 million to date. The latest round ($38 million Series C) was locked down in March. Tenaya Capital led the round, while Citi Ventures, Cisco, Allegis Capital, Andreessen Horowitz, Battery Ventures, Sutter Hill Ventures, and In-Q-Tel all participated.

Why they're one of the 50: As with many startups in this report, Platfora was founded in order to simplify Hadoop. While businesses have been rapidly adopting Apache Hadoop as a scalable and inexpensive solution to store massive amounts of data, they struggle to extract meaningful value from that data. The Platfora solution masks the complexity of Hadoop, which makes it easier for business analysts to leverage their organization's myriad data.

Platfora tries to simplify the data collection and analysis process, automatically transforming raw data in Hadoop into interactive, in-memory business intelligence, with no ETL or data warehousing required. Platfora provides a big data analytics platform designed for line of business users. Platfora gives business analysts visual, self-service analytical tools that help them navigate from events, actions, and behaviors to business facts.

Platfora claims to have the first scale-out in-memory Big Data analytics platform for Hadoop. Platfora's focus on simplifying Hadoop and Big Data analysis is becoming a more common goal of late, but they are an early mover in this respect.

Customers include Comcast, Disney, Edmunds.com, and the Washington Post.

Competitive Landscape: Platfora competes with the likes of Splunk, Tableau, IBM, SAP, SAS, Alpine Data, and Rapid-I.

Xplenty

What they do: Provide Hadoop-as-a-Service.

Headquarters: Tel Aviv, Israel

CEO: Yaniv Mor, who previously managed the NSW SQL Services practice at Red Rock Consulting.

Founded: 2012

Funding: Xplenty is backed by an undisclosed amount of seed funding from Magma Venture Capital, and it is currently in the midst of locking down a Series A round.

Why they're one of the 50: While Hadoop is being hyped like crazy these days, not all the hype is empty. It has indeed become the de facto infrastructure technology for Big Data. The trouble is that the development, implementation, and maintenance of Hadoop require a very specialized skill set.

Xplenty technology provides Hadoop processing on the cloud via a coding-free design environment, so businesses can quickly and easily benefit from the opportunities offered by Big Data without having to invest in hardware, software, or highly specialized personnel.

According to Xplenty, competing services still target developers, whereas Xplenty targets the data and Business Intelligence (BI) users who do not know how to write code, but who need to move data to a Big Data platform.

A drag-and-drop interface eliminates the need to write complex scripts or code of any kind. With its automatic server configuration feature, users can simply point to a data source, configure the data transformation tasks, and tell the platform where to write the results to. Xplenty's platform uses SQL terminology. Thus, for data analysts, the learning curve should be minimal.

Customers include Fiverr, WalkMe, Dealply Technologies, and Travel Global Systems.

Competitive Landscape: The main competition comes from Amazon's Elastic MapReduce (EMR). Other Hadoop-as-a-Service competitors include Altiscale, Mortar Data, Qubole, and recently Microsoft with Hadoop on Azure. Rackspace is about to launch its own Hadoop-as-a-Service offering based on Hortonworks' distribution.

Making Big Data Easy

These startups mask the complexity of Big Data, so non-data scientist can make data-driven decisions.

Alpine Data Labs

What they do: Provide a Hadoop-based data analysis platform.

Headquarters: San Francisco, CA

CEO: Joe Otto, formerly Sr. VP of Sales and Service at Greenplum.

Founded: 2011

Funding: $23.5 million in total funding, including $16 in Series B Funding, from Sierra Ventures, Mission Ventures, UMC Capital and Robert Bosch Venture Capital.

Why they're one of the 50: Most executives and managers don't have the time or skills to code in order to glean data insights, nor do they have the time to learn about complex new infrastructures like Hadoop. Rather, they want to see the big picture. The trouble is that complex advanced analytics and machine learning typically require scripting and coding expertise, which can limit access to data scientists.

Alpine Data mitigates this issue by making predictive analytics accessible via SaaS.

Alpine Data provides a visual drag-and-drop approach that allows data analysts (or any designated user) throughout an organization to work with large data sets, develop and refine models, and collaborate at scale without having to code. Data is analyzed in the live environment, without migrating or sampling, via a web app that can be locally hosted.

Alpine Data leverages the parallel processing power of Hadoop and MPP databases and implements data mining algorithms in MapReduce and SQL. Users interact with their data directly where it already sits. Then, they can design analytics workflows without worrying about data movement. All this is done in a web browser, and Alpine Data then translates these visual workflows into a sequence of in-database or MapReduce tasks.

Alpine Data Labs argues that most competing solutions are either desktop-based or a point solutions without any collaborative capability. In contrast, Alpine Data offers a "SharePoint-like" feel to it. On top of collaboration and search, it also provides modeling and machine learning under the same roof. Alpine is also part of the No-Data-Movement camp. Regardless if a company's data is in Hadoop or MPP Database, Alpine sends out instructions, via its In-Cluster Analytics, without ever moving data.

Customers include Sony, Havas Media, Scala, Visa, Xactly, NBC, Avast, Blackberry, and Morgan Stanley.

Competitive Landscape: Alpine will compete both with large incumbents (SAS, IBM, SPSS, and SAP) and such startups as Nuevora, Platfora, Skytree, Revolution Analytics, and Rapid-I.

import.io

What they do: Provide tools that help users turn any webpage into data without writing any code.

Headquarters: London, UK and San Francisco, CA

CEO: David White, who was previously Head of Technology Innovation at RBS.

Founded: 2012

Funding: $1.3 million from Wellington Partners Venture Capital and angel investors Louis Monier (founder of Alta Vista) and Emmanuel Javal.

Why they're one of the 50: Getting data from websites is difficult because it requires you to write a screen scraper, which is technically challenging and unreliable. Import.io helps users create a semantic overlay of the website using point-and-click, which builds an API to the site. Users can then access this data programatically and in real-time.

Import.io also lets you can crawl websites and interact with search boxes to access data from behind a form. Import.io argues that it is easy enough to use that anyone, regardless of technical skill, can get usable, programatically accessible data from publicly available websites.

In April 2014, import.io added a feature that lets you extract data from behind a login screen. (Simply input your username and password, and you're set.)

The service is currently free, but import.io will eventually begin charging large corporate clients who want custom datasets curated and maintained for them. "For example, we have created a jobs dashboard for a major recruitment company, which monitors thousands of jobs pages in real time and alerts their sales team whenever a new vacancy is posted," a company spokesperson wrote in an email to me.

Competitive Landscape: While there are large companies like Bloomberg that curate and sell datasets, the technology that helps people to build and access any web data is relatively new. Startups like Scraper Wiki, Kimono, and Outwit Hub will all compete in this space.

Predixion Software

What they do: Develop predictive analytics solutions.

Headquarters: San Juan Capistrano, CA

CEO: Simon Arkell, who previously spent 15 years working on software company mergers, acquisitions, and management for Gramercy Venture Advisors, Triton Pacific Capital Partners, and Versifi Technologies.

Founded: 2009

Funding: Predixion Software is backed by $33 million in total raised in three funding rounds. Backers include DFJ Frontier, Accenture, and GE.

Why they're one of the 50: In the past, only large enterprises could afford the big team of data scientists needed to create predictive analytic models and deploy them. Even then, it often took months or even years to create a model and then deploy, and doing that came at great expense. Sometimes, due to the wait, the model would be outdated by the time it hit production.

With Predixion, predictive models are created through a simple wizard-driven interface that any BI professional can understand. Predixion enables collaboration and sharing of predictive models among business analysts, data scientists, and information consumers to expedite the predictive process. It integrates into existing enterprise software applications, so predictive insights can be more easily consumed at the decision point. Predixion also enables existing predictive applications to be quickly modified, shared, and redeployed into disparate environments and across multiple data sets.

Predixion argues that their key differentiators are: 1) that it does not require a data scientist, since predictive models can be created by a business analyst with minimal training; 2) their solution can be deployed in a much shorter period of time than competing solutions, and 3) predictive results are delivered directly to the business users who need to take action on the information.

Clients include GE, Chevron, United States Armed Forces, Kaiser Permanente, Carolinas Healthcare Systems, ThriveHD, and Maritz.

Competitive Landscape: Predixion will compete against IBM, SAS, and Alpine Data Labs.

SiSense

What they do: Provide a Big Data analytics solution for business users.

Headquarters: New York, NY with an R&D Center in Tel Aviv, Israel

CEO: Amit Bendov. He was formerly CMO of Panaya and SVP of Worldwide Marketing at ClickSoftware.

Founded: 2010 (they were technically founded in 2004, but were really just a side project for the five founders until 2010, and their official launch was 2012).

Funding: Last week, SiSense closed a $30 million C round of new financing led by DFJ Growth, with participation from existing investors Battery Ventures, Genesis Partners and Opus Capital. This follows a $10 million Series B round of funding led by Battery Ventures with participation from Opus Capital and Genesis Partners, and a $4 million Series A round that was secured in 2010.

Why they’re one of the 50: SiSense argues (and I tend to agree) that many Big Data analytics solutions are like battleships: they're expensive, complicated to operate, and are actually overkill for most businesses, which just don't need that much processing. The typical business does not need to analyze petabytes of data. Rather, they'd be happy to glean insights from terabytes of data, but that's either too expensive or forces them to rely on in-memory solutions, which cannot later scale to handle massive amounts of data, if and when the need arises.

The startup’s eponymous flagship product is built to offer Big Data analytics technology to businesses of all sizes. With no coding or scripting required, business analysts can analyze data themselves, without having to draw IT or data scientists into the process. SiSense claims that its software allows non-technical users to analyze 100 times more data than current in-memory analytics solutions, and it does so 10 times faster. There's no need to set up complex data warehouse systems or OLAP cubes.

SiSense software is powered by SiSense's Elasticube technology, which features a columnar data store, strong data compression, parallel processing, and advanced query optimization to offer analytical processing power previously available only with high-end solutions.

The recently introduced, SiSense 5, is an analytics solution that enables non-technical users to join multiple data sources, analyze billions of records, and share interactive dashboards from any device, including mobile. Its back-end is powered by a proprietary In-Chip technology architecture that SiSense claims yields 100x more data scalability and 10x faster performance than in-memory architectures.

And be sure to check out Crunch Analytics, which takes CrunchBase’s data, plugs it into SiSense Prism, and cranks out all sorts of interesting information about the startups attracting VC funding.

Customers include NASA, ESPN, Samsung, Target, eBay, fiverr, Wix, and bookings.com.

Competitive Landscape: SiSense’s competitors include Tableau, QlikView, and SAP HANA.

Focusing on Performance Management

From data center operations to virtualization to applications, these startups use Big Data to analyze the performance of these complicated systems.

CloudPhysics

What they do: Provide intelligent operations management for virtualized workloads.

Headquarters: Mountain View, CA

CEO: Jeffrey Hausman, who was previously the senior VP at Symantec responsible for the $1 billion Information Availability and Intelligence group.

Founded: 2011

Funding: In late June 2014, CloudPhysics secured a $15 million Series C round of funding led by Jafco Ventures with participation from the company's existing investors Kleiner Perkins Caufield & Byers and Mayfield. This brings total funding to date to $27.5 million.

Why they're one of the 50: Virtualization and cloud management platforms lack actionable information that admins can use to better design, configure, operate, and troubleshoot their systems. However, having lots of data points is not enough. Not all data is equally valuable. In order to go beyond basic data capture, decision makers need to be able to validate, evaluate, and assess information from a variety of perspectives in order to make real, impactful decisions.

CloudPhysics goal is to analyze the world's IT data knowledge and use the learnings to transform computing, driving out machine and human costs in ways never before possible. Today, their servers receive a daily stream of 100+ billion samples of configuration, performance, failure, and event data from their global user base.

CloudPhysics SaaS product combines Big Data analytics with data center simulation and resource management techniques. CloudPhysics argues that this approach uncovers hidden complexities in the infrastructure, discovers inefficiencies and risks that drain and endanger resources, and enables what-if analyses that can inform every data center decision.

In June, CloudPhysics added new capabilities to its product, rolling out a Storage Analytics tool and Smart Alerts. Using its global dataset, CloudPhysics examines metadata and trends from thousands of datacenters and bakes these learnings into the Storage Analytics. These high-resolution performance and capacity analytics run across a customer's entire virtual infrastructure, evaluating the configuration and behavior of storage resources at the datastore, VM and guest level; highlighting potential trouble spots; and pushing recommendations for preventive actions.

Smart Alerts is a feature intended to make alerts more predictive than reactive. According to CloudPhysics, other virtualization management products trigger alerts based on a “dumb,” often arbitrary number or static threshold, creating significant noise about events that have already taken place. In contrast, CloudPhysics Smart Alerts are analytics-driven and change the focus from reactive to preemptive troubleshooting.

Customers include Sanofi, Thiel Capital, and Zettagrid.

Competitive Landscape: This space is a land grab at the moment. Competitors include Splunk and Sumo Logic.

Concurrent

What they do: Provide Big Data application infrastructure.

Headquarters: San Francisco, CA

CEO: Gary Nakamura, who previously served as SVP and GM of Terracotta.

Founded: 2008

Funding: Concurrent received $10 million in Series B financing in June 2014. The round was led by new investor Bain Capital Ventures, with the participation of existing investors Rembrandt Ventures and True Ventures. Previously, Concurrent received $4 million in Series A financing in March 2013, and a $900,000 seed investment in August 2011. Both investments were led by led by True Ventures and Rembrandt Venture Partners.

Why they're one of the 50: While MapReduce is a key programming tool for Big Data applications, there is a growing realization that it is rather difficult to use. Meanwhile, Pig and Hive arguably fall short of enabling applications to do more than just simple ad-hoc analysis.

Concurrent's products are intended to significantly accelerate developer productivity and provide visibility to developers creating Big Data applications .Concurrent makes it possible for enterprises to leverage their existing skillsets to create, deploy, run, and manage data applications at scale.

The startup's flagship product, Cascading, is an alternative API to MapReduce for processing large data sets on technologies like Hadoop. Its Java-based application open-source framework enables developers to build robust data processing and data management applications for deployment on clusters running in the cloud or within private data centers.

Concurrent's application performance management product for Big Data applications, Driven, enables developers, data analysts, data scientists, and operations personnel to visualize their data applications metrics in real-time, via the seamless collection of internal runtime and execution metadata.

Users can isolate and resolve data application failures and performance problems immediately, as well as quickly develop and debug Cascading applications. Driven enables enterprises to make the most of their Big Data by addressing the biggest pain points around enterprise application development and application performance management.

Customers include Twitter, eBay, The Climate Corp, Etsy, Razorfish and DataSong (formerly UpStream).

Competitive Landscape: Concurrent is rather uniquely positioned, with most of its competition coming from roll-your-own projects relying on MapReduce, Pig, Hive, and other open-source tools in Hadoop ecosystem.

Search, Warehousing, and "People Analytics"

These startups are pushing the boundaries of analytics, focusing on such features as search, low-cost warehousing, and even "people analytics."

LucidWorks

What they do: Provide enterprise search tools to help navigate Big Data.

Headquarters: San Francisco, CA

CEO: Will Hayes, who recently served as the head of technical business development for Splunk.

Founded: 2008

Funding: Total venture funding stands at $16 million (from Granite Ventures, Walden International, In-Q-Tel and Shasta Ventures).

Why they're one of the 50: IT organizations are beginning to collect orders of magnitude more data than they gathered even a few years ago. Collecting data is one thing; however, making actual use of it is another. Enterprise search clearly has a role to play in terms of making Big Data accessible. The challenge is doing it in a way that other applications can utilize.

LucidWorks Search is designed to help developers build highly secure, scalable and cost-effective search applications, while providing a simple and comprehensive way to access open-source search technologies.

LucidWorks Big Data is an application development platform that integrates search capabilities into the foundational layer of Big Data implementations. The product is built on a foundation of key Apache open-source projects and enables organizations to quickly discover, access and evaluate large volumes of structured and unstructured data. LucidWorks Big Data and LucidWorks Search work hand-in-hand to accelerate and simplify the building of highly secure, scalable and cost-effective search applications.

ADP is a **named customer.**

Competitive Landscape: Competitors include Endeca, Autonomy and Elasticsearch.

MammothDB

What they do: Provide a low-cost alternative for enterprise analytics and data warehousing.

Headquarters: Sofia, Bulgaria

CEO: Steve Keil. Prior to MammothDB, he was CEO of Sciant, an outsourcing company that was acquired by VMWare.

Founded: 2012

Funding: The startup is mostly self-funded, but it has also received an EU innovation grant; total funding is ~400,000€.

Why they're one of the 50: The majority of Big Data solutions (data warehouses + analytics tools) on the market today are priced high enough to be out of reach for most of the mid-market. Mid-sized companies that want enterprise-scale analytic tools balk at paying $750k+ for a scalable database, and as data size and complexity grows, so does the price.

In fact, according to Gartner, approximately 90 percent of the enterprise market has yet to deploy data warehouses and BI/Big Data analytics solutions, with the sky-high price tag being a major reason why.

MammothDB argues that it can provide enterprise-scale Big Data analytics for a fraction of the price, opening up the market to smaller companies.

MammothDB leverages open-source technologies, such as Hadoop and MySQL, and places a columnar database on each Hadoop node. This hybrid approach utilizes key parts of the Hadoop framework, but is otherwise a fully SQL-compliant data engine. CEO Steve Keil says that MammothDB's performance results on large data sets are multiple times better than Oracle, Impala, or MySQL.

Customers include DHL, Publicis, Savvy, Austria Telekom, Transmetrics, and Piano Media.

Competitive Landscape: Teradata is closest competitor from an architectural standpoint, but MammothDB will also compete with Oracle (Exadata), Vertica, Netezza, Hana, Hadapt, and Greenplum.

VoloMetrix

What they do: Develop cloud-based analytic software to measure a company's "people analytics," or how their employees spend their time and whom they spend it with.

Headquarters: Seattle, WA

CEO: Ryan Fuller, who was formerly a manager at Bain & Company and a solutions architect at Cognos.

Founded: 2011

Funding: In April, the startup closed a $3.3 million Series A round led by Shasta Ventures, the VC firm which also provided $1.6 million in seed funding.

Why they're one of the 50: Many businesses are complex and geographically dispersed. In addition, most teams and individual employees are faced with a multitude of competing demands on their time. It is difficult to determine if people are really spending their time on the most important priorities for the company, or if they are distracted by other competing demands within the organization.

IDC predicts that the worldwide market for enterprise social networking will exceed $4.5 billion by 2016.

To address the lack of visibility into how employees spend their time, VoloMetrix applies Big Data analytics to help businesses better understand employee productivity and behavior and how to map those to business outcomes.

VoloMetrix' cloud-based Social Enterprise Intelligence software works by applying privacy filters and then analyzes anonymous, real-time information from corporate email, calendar, instant messaging, and social platforms to provide deep actionable insights into the ways teams are spending their time on the most important business priorities.

VoloMetrix argues that while there are some tools on the market that provide analytics for individual social networks, VoloMetrix' software is the only one that looks at activities that are happening across the entire organization. In fact, no other solution looks at email which is still the "largest social network" nor do these other solutions look at all collaboration applications holistically.

Only by looking across the entire organization can managers get an accurate picture of how teams collaborate and the way work really gets done.

Competitive Landscape: Competitors include RelateIQ, Relationship Science, DataHug, and Activate Networks.

Data Connections and Preparation

These startups help connect various types of data and clean it up for analysis.

Datameer

What they do: Provide an end-to-end and self-service data analytics application that is purpose built for Hadoop.

Headquarters: San Francisco, CA

CEO: Stefan Groschupf. Prior to Datameer, he spent several years architecting and implementing distributed big data analytic systems for such companies as Apple, EMI Music, Hoffmann La Roche, AT&T, the European Union, and others. He was one of the early contributors to Nutch, the open-source project that spun out Hadoop.

Founded: 2009

Funding: To date, Datameer has raised $36.8M in funding, most recently (in December 2013) raising $19M in Series D funding from Next World Capital, Redpoint Ventures, Kleiner Perkins Caufield & Byers, Software AG, Workday, and Citi Ventures. Previous funding rounds have been from KPCB and Redpoint Ventures. It is worth noting that Workday is an OEM customer turned investor, and Citi is a customer also turned investor.

Why they're one of the 50: It's no secret that Hadoop isn't the easiest technology in the world to use. Hadoop has no end-user interface, so you have to know MapReduce to use it as a platform for big data analytics. In other words, using Hadoop for Big Data is typically an IT-intensive project.

Datameer removes the need to code a custom solution, enabling any business user to integrate, analyze and visualize their data.

Sitting natively on top of Hadoop (instead of on top of a traditional RDBMS database, which business intelligence tools typically sit on top of), Datameer handles data integration, analytics, and visualizing the results, whereas other solutions typically only solve one or two parts of that equation.

Customers include Visa, Workday, Citi, British Telecom, CDW, Newegg, MachineZone, Cardinal Health, Sears, and the U.S. Women's Olympic Cycling team, Vivent, and Trustev.

Competitive Landscape: As with many of the startups on this list, open-source projects remain one of the strongest sources of competition, since many engineers prefer to build rather than buy. Datameer's most direct commercial competitors include Platfora, Trifacta, ClearStory, and Tableau.

Paxata

What they do: Provide tools that make data preparation more accessible to business analysts, helping companies shorten data preparation time from days to minutes.

Headquarters: Redwood City, CA

CEO: Prakash Nanduri, who was previously co-founder and VP of Velosel Corporation (acquired by TIBCO in 2005). More recently, he worked as Head of Product and Technology Strategy at SAP.

Founded: 2012

Funding: $10 million from Accel Partners.

Why they're one of the 50: Business analysts often have to spend weeks preparing data for analysis. Data often comes from disparate sources, and it must be scrubbed of inaccuracies (or just noise). In real-time environments, such as e-commerce, that delay translates into lost revenue.

To solve this problem Paxata has developed a native multi-tenant cloud solution that provides a set of pre-built data preparation services that are automated via proprietary machine learning, latent semantic indexing, statistical pattern recognition, and text analytics techniques.

An elastic in-memory data preparation engine operates over a large variety and volumes of structured and unstructured data in real-time, enabled by a vector query processor with columnar data storage. A governance backbone time-stamps and versions all data modifications at a tenant-, user- and cell-level.

After data is collected and “cleaned,” users get a Paxata “Answer Set,” which they can then plug into whichever BI tool they are already using.

Customers include Dannon, UBS, Pabst Brewing Company, and Box.

Competitive Landscape: Paxata will compete with ClearStory, Datameer, Informatica, Tamr, Trifacta, and others.

Tamr

What they do: Provide a data connection platform that reduces the time and effort of connecting and enriching data sources.

Headquarters: Cambridge, MA

CEO: Andy Palmer. Prior to Tamr, Palmer, along with co-founder & CTO Dr. Michael Stonebraker, co-founded the early Big Data analytics company Vertica Systems, which was acquired by HP.

Founded: 2012

Funding: $16.1 million in financing led by Google Ventures and New Enterprise Associates (NEA).

Why they’re one of the 50: Enterprises want to use all the data at their disposal today, including internal data sources and external public data sources, as well as feeds that will soon come from the Internet of Things. However, for most organizations, this is little more than a pipe dream. Most organizations can’t take advantage of all of this disparate data because it takes too long and costs too much to integrate, curate, and prepare data for analytics using traditional methods. According to Tamr, it's not uncommon for companies to have dozens or even hundreds of people manually working on curation, depending on what business they are in.

Tamr intends to help enterprises integrate and curate data at scale, while abstracting the complexity of Big Data's "variety" problem. Tamr provides discovery, linking, cleaning, preparation, and curation for internal and external data sources.

Tamr is an open data connection platform. It uses a combination of advanced algorithms, machine learning, and guidance from data experts to fuse structured and semi-structured data, with the goal of making all enterprise data ready for analysis within existing business intelligence and analytical tools. The Tamr system can identify sources, understand relationships, and curate the massive variety of siloed data in the enterprise. In early customer tests, Tamr claims that it was able to connect, curate, and prepare enterprise data in days or weeks instead of months or quarters.

Customers include Thomson Reuters and Gloria Jeans.

Competitive Landscape: Trifacta, Paxata, and ClearStory Data are the most direct competitors, although Tamr argues that those systems can benefit from Tamr's curation capabilities.

Trifacta

What they do: Provide a platform that enables users to transform raw, complex data into clean and structured formats for analysis.

Headquarters: San Francisco, CA

CEO: Joe Hellerstein, who in addition to serving as Trifacta's CEO is also a Professor of Computer Science at Berkeley. In 2010, *Fortune* included him in their list of 50 smartest people in technology, and *MIT Technology Review* included his Bloom language for cloud computing on their TR10 list of the 10 technologies "most likely to change our world."

Founded: 2012

Funding: Trifacta is backed by $41.3 million in funding raised in three rounds from Accel Partners, Greylock Partners, Ignition Partners, and individual investors.

Why they're one of the 50: According to Trifacta, there is a bottleneck in the data chain between the technology platforms for Big Data and the tools used to analyze data. Business analysts, data scientists, and IT programmers spend an inordinate amount of time transforming data. Data scientists, for example, spend as much as 60 to 80 percent of their time transforming data. At the same time, business data analysts don't have the technical ability to work with new data sets on their own.

Trifacta argues that the problem of data transformation requires a radically new interaction model – one that couples human business insight with machine intelligence. Trifacta's platform combines visual interaction with intelligent inference and "Predictive Interaction" technology to close the gap between people and data.

Trifacta's Predictive Interaction technology elevates data manipulation into a visual experience, allowing users to quickly and easily identify features of interest or concern. As analysts highlight visual features, Trifacta's predictive algorithms observe both user behavior and properties of the data to anticipate the user's intent and make suggestions without the need for user specification. As a result, the cumbersome task of data transformation becomes a lightweight experience that is far more agile and efficient than traditional approaches.

Lockheed Martin and Accretive Health are **early customers.**

Competitive Landscape: Trifacta commonly competes with manual, hand-coded approaches to data migration. Other companies in the data transformation niche include Cambridge Semantics, Paxata, Informatica, and Tamr.