DataBeat 2013
Redwood City, California
December 4-5, 2013

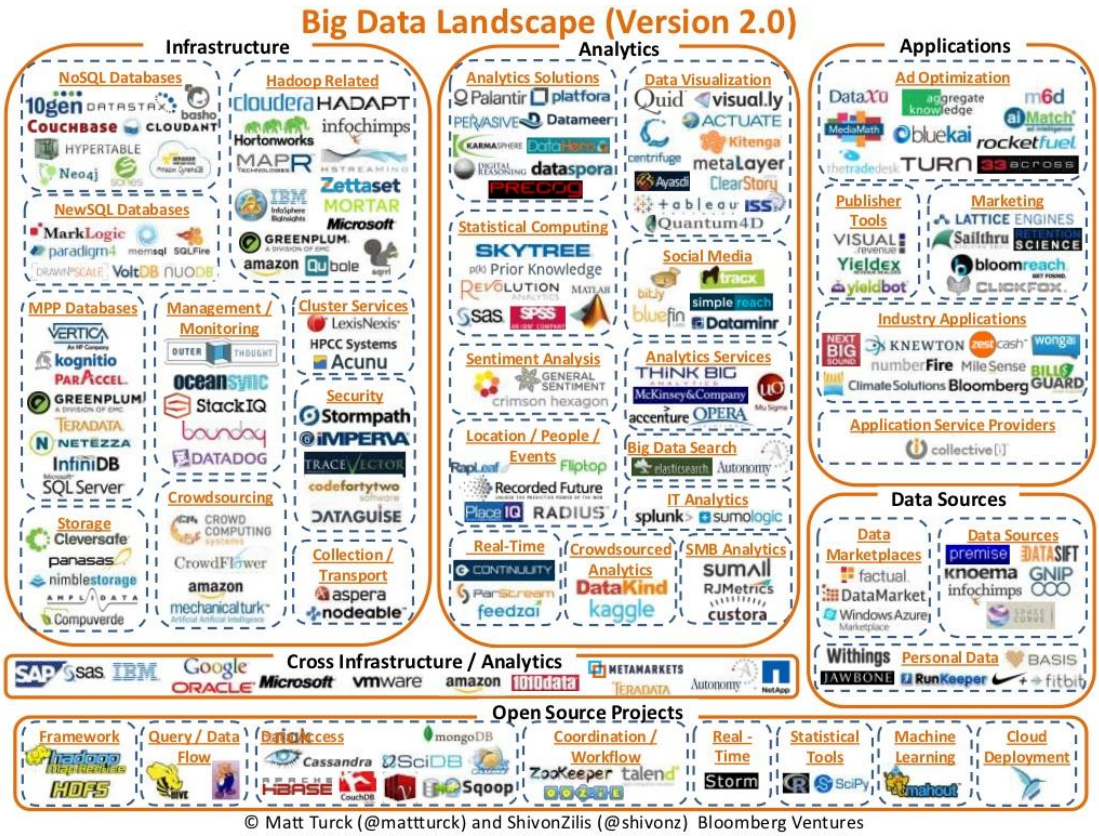
M. R. Pamidi, Ph. D.
Editor-in-Chief
IT Newswire

Executive Summary

Big Data and Data Science are rapidly encroaching every aspect of our lives—healthcare, sports, financial services, retail, manufacturing. While there is a great deal of pioneering work being done, many companies are blindly purchasing big data tools and hiring expensive old-fashioned statisticians calling themselves *Data Scientists* without first defining their data science opportunities. At best, the results can be nonsensical data. At worst, distraction takes over and companies lose focus on their *core* value. Organizations should thoroughly evaluate the real business value Big Data can add to their bottom line. Truly effective data science work is about practical service to the business first and the trappings of big data second. Don't fall prey to Big Data vendors with their sexy marketing presentations and Consultants with their exotic titles if they can't help solve your problems.

Big Data Landscape

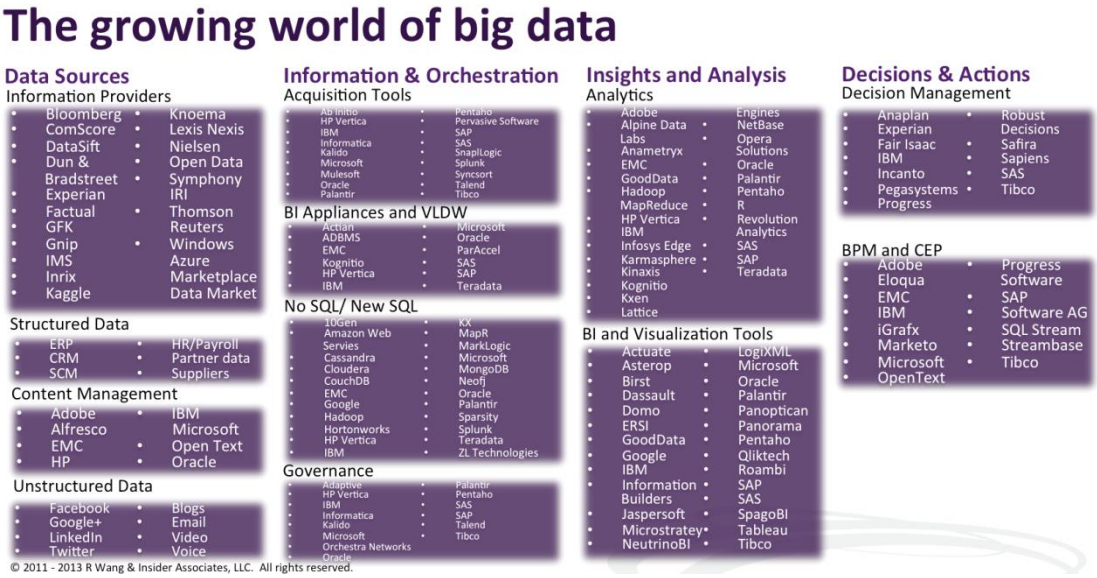
To say this industry is complex and still emerging is an understatement. Below is the vendor landscape, as of October 2012:



© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

Source: Slideshare

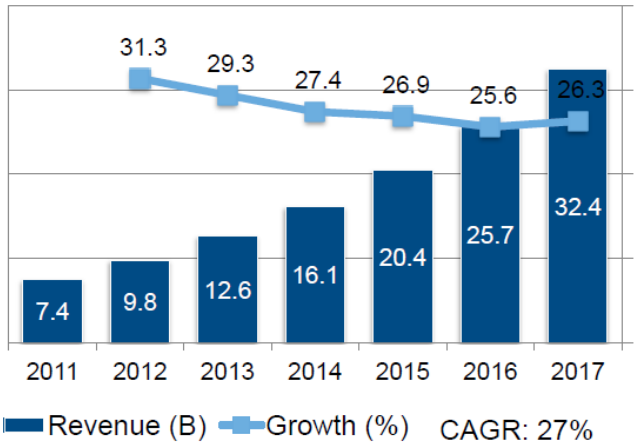
Here is another version as of April 2013:



Source: Software Insider

The State of Data Science

Data science is potentially driving the most accelerated transformation in the human experience in recent times. Although still emerging as a discipline, it's already making an impact on our ability to solve some of the world's biggest problems and deliver on some of the world's biggest opportunities. We have more data ever before. Big Data tools market is exploding. There are more opportunities springing up as US$12.6 billion will be spent on Big Data this year, expecting to grow to $32.4 billion by 2017, according to IDC.



Big Data Market Forecast

Source: Worldwide Big Data Technology and Services 2013 – 2017 Forecast, December 2013, IDC

This should turn into more visibility into the enterprise and customer insights, more Big Data jobs, and more universities offering Big Data programs. Beware many institutions and individuals are putting a wrapper over traditional specialties, such as Statistics, and labeling them *Data Scientist*; in actuality they may be just a *Data Practitioner*. The concept of Big Data is nothing new; Roger Magoulas, Research Director at O'Reilly Media, talked about in 2005.

Big Data in Financial Services: The New York Stock Exchange

Steve Hirsch, Chief Data Officer, Intercontinental Exchange Group (The New York Stock Exchange, ICE-NYSE) gave a fabulous presentation on how Big Data is playing a major role at the NYSE.

The NYSE was one of the early, fearless adopters of streaming data and Big Data. The financial industry has always been one of the key drivers of data and analytics innovation as well as an

early adopter of data technology. Near-real-time streaming data was invented in 1867. Edison's patented ticker tape machine of 1869 is a priceless museum piece. Of course, not everybody had a ticker tape printer at home during those days.

Edison Gold and Stock Telegraph Ticker (1869)
Source: Wikipedia

Real-time Quotron machines followed in the 1980s. Again, not everyone had a Quotron machine at home.

Quotron Terminals
Source: Computer History Museum

The Regulation of Exchanges and Alternative Trading Systems (Regulation ATS, sounds better than RATS!) of 1998 resulted in the Rise of the Machines and Archipelago connected islands of trading platforms and democratized data, akin to what Sabre did to the airlines reservation systems in the 1960s.

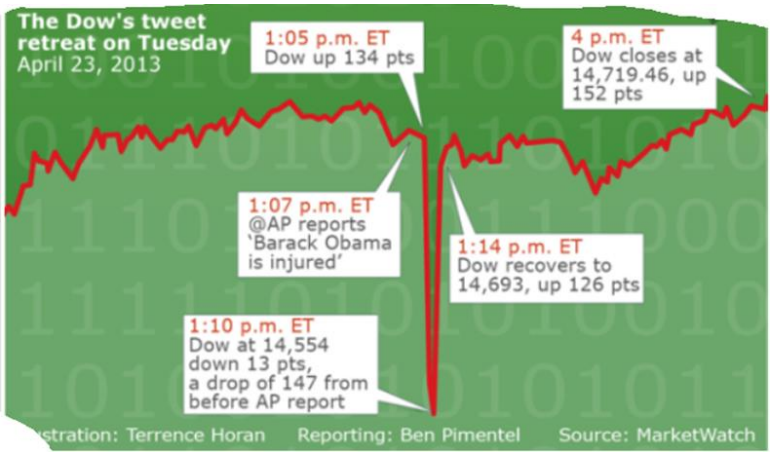
The Regulation National Market System (Regulation NMS), passed in 2007, connecting the web of exchanges, pushed the responsibility down to both the brokers and the exchanges to ensure that they get the best price for their customers, across all markets. This increased the complexity of the markets dramatically and created another data deluge on the compliance side of the financial industry.

Today, the NYSE is a showpiece of Big Data:

- Has 10 PB of data
- Handles over 14 billion quotes in a single day
- Requires 48 network pipes to subscribe
- Peaks at 10+ million messages per second and generates over 30 billion messages in a single day
- Generates 7 TB of fresh data each day for surveillance, compliance, and analytics
- Latency is measured in microseconds
- Completes trading transactions in 120 microseconds

The most sophisticated traders have built data-aware trading "machines" that start the day driven by predictive models based on recent and historical data, but "react" to real-time data inputs from sources such as the Associated Press (AP), Facebook, and Twitter feeds. This is absolutely essential since fake and false tweets and retweets can easily cause havoc in the market as it happened in April 2013. Hackers took over the AP Twitter account and falsely claimed there had been explosions in the White House and that President Obama was hurt. The

AP immediately confirmed the news was not true, but the tweet had been retweeted 3,000 times, and the damage had been done (see graphic below), before the market recovered. Today, to rebuild confidence in the markets, the NYSE is using Data analysis at the heart of the current evaluation of the U. S. structure.



"Flash Crash" of April 2013

Domo

Josh James, CEO (founder of Omniture that he sold to Adobe in 2009 for $1.8 billion) founded the company after he was frustrated that he couldn't have adequate access to data about his own business. The data he wanted was trapped in multiple systems, databases, spreadsheets, and presentations. A Domo 2013 CEO survey revealed:

- 70% of CEOs lack real-time access to the data they need
- 92.5% don't trust the data
- 93.3 % are regularly concerned they can't make sense of data

Domo has 350 customers, raised $63 million in VC, and claims customer data on a Domo platform is secure, real-time, historical, combined, drillable, shared, published, actionable, trended, mobile, collaborative, alertable, exportable, monitored, interactive, assignable, organized, visualized, uniform, backed up, snapshotted, managed by exception…did I leave out any adjectives or verbs?

Big Data in Healthcare: Kaiser Permanente Southern California

Medicine has been seen as a scientific field for centuries. Yet, newly available troves of individual data, from genetic sequences to past clinical history, are redefining the levels of exactitude doctors can achieve—and shifting expectations. We may soon expect our human doctors to be aided by powerful algorithms that help them quickly consider millions of variables.

Dr. John Mattison, Chief Medical Information Officer, Kaiser Permanente Southern California (KPSC), emphasized one of the 'V's—Variety—of Big Data. (For inquiring minds who want to know, the other 'V's are Volume, Velocity, Variability, Value, and Vendor.) He underscored the need to realize the real importance of Big Data and not be carried away by vendor marketing and hype. To repeat his quotes:

- *The future is already here, it is just unevenly distributed*. – William Gibson
- *The half-life of facts: Why everything we know has an expiration date*. – Samuel Arbesman
- *41% of all medical literature is subsequently refuted*. – John Ioannidis

The real applications of Big Data are:

- Compression of transformation time, reducing cycle time for discovery and implementation from decades to days
- Decision Support for care processes and care practices
- Clinical Practice and Data Capture
- Analytics, Modelling, and Simulation
- Inpatient opportunities in early detection/intervention and level of care/transfers
  - From risk adjustment to earlier detection/prediction to care
  - Co-occurrence trends, especially infections
  - Moving from 'offline periodic analytics' to real-time in-memory analytics

- A family of Multi-omics:
  - Panarome, comprising genome, transcriptome, proteome, metabolome, lipidome, epigenome
  - Metagenome/microbiome for autism, multiple sclerosis, obesity (Of the two identical twins, one is lean; the other is obese, why?)
  - Phenome for EHR + PHR
  - Socialome
  - Exposome for fixed and mobile sensors, physical, biologic markers, strep throat, RSV (Respiratory syncytial virus), H. flu
  - Personal Sensoromes

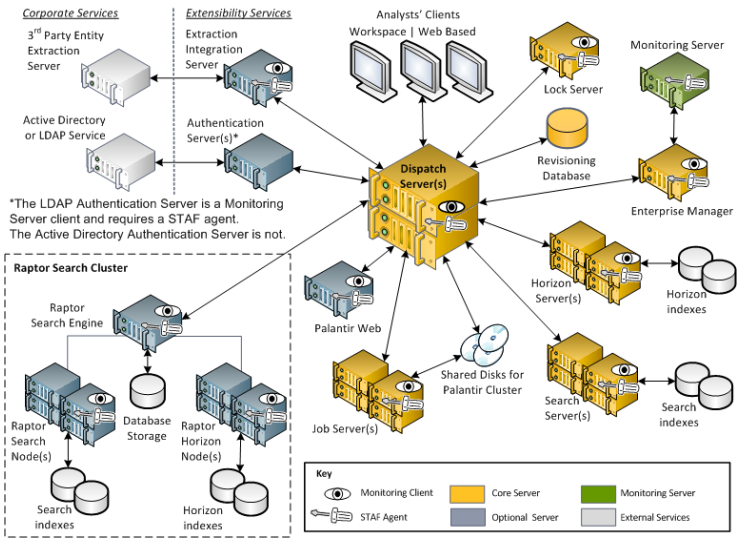
Big Data demands cross-disciplinary collaboration from different perspectives and multiple platforms. On example is the 2103 Nobel Prize in Chemistry which comprised three contributors from three countries, three universities, two different technologies—Newtonian and Quantum; and one new combined method for understanding how protein structure drives function.

Big Data has barely begun to scratch the complex surface of healthcare, genomics, and bioinformatics. And, to quote H. L. Menken, "For every large complex problem, there is almost always at least one very simple solution…and it's usually wrong."

Big Data in Intelligence, Cybersecurity, and Public Safety: Palantir

Ari Gesher of Palantir gave an interesting presentation on what his company is doing in these areas. Palo Alto, CA.-based Palantir is well-funded and has strong presence in the three-letter acronym intelligence community (CIA, FBI, NSA…) in D. C. and the Beltway around Washington, D. C., and now boasts of a market cap of $9 billion which, we believe, is highly inflated. Palantir's slogan seems to be ***Connect*** *the people who care to the data that* ***matters without friction***.

Regardless, Ari talked not about AI (Artificial Intelligence), but about IA (Intelligence Augmentation). Palantir core server architecture is shown below; more details can be found here.



Palantir Server Architecture
Source: Palantir

Some of the application areas the speaker discussed (obviously, he couldn't talk about the work being done for the spooky guys; otherwise, he had to kill the audience!) are public safety (as in Hurricane Sandy of 2012), Rogue Traders, Counter-Terrorism, and Credit Card Bust-Out. The last one was an intriguing discussion of how crooks commit credit card fraud, which is beyond just stealing credit card information and involves the following steps:

1. Open accounts under fake identities
2. Use accounts, drive up credit lines
3. Hit magic number, max out all cards
4. Write fake checks to zero balances

5. Max cards a second time
6. Disappear
7. Open accounts under fake identities…

Big Data and Analytics in Sports

Probably the most interesting presentations were by Jason Fass, CEO, Zepp Labs and Mike Crowley, CEO, InfoMotion Sports Technologies.

Zepp Labs uses wearable 3D motion technology to help baseball, golf, and tennis players (an ice hockey version is in the works) improve their swing, accuracy, speed, etc., by capturing, measuring, and analyzing your swing in three dimensions and recording 1,000 data points per second.

InfoMotion sells an intelligent, instrumented 94Fifty Smart Sensor Basketball that measures muscle memory that the human eye can't see, learns the strengths and weaknesses of players at any level, adapts as the player improves, and provides basic, intermediate, and advanced level training to build better shooting and ball-handling skills – fast. It's like having the best coaches in the world with you every day of the year. The December 9, 2013 issue of *Sports Illustrated* has a brief but interesting write-up (may require paid subscription to access) on it. The basketball sells for US$295.

"Isn't it too pricey for many inner-city kids where many of the top athletes in America come from?"
"Yes, but remember many of these kids save their pocket money on food and sundries to buy expensive athletic shoes."
"So, you expect them to be hungry, unclean, and play basketball? And, why are you focusing on just basketball?"
"There are an estimated 300 million basketball players in the world. And, we'll be introducing a soccer product to address an estimated 800 million players around the world."
Good points!

Additional Tidbits

- Verizon has 100 million customers and collects 5 PB to 8 PB of data per day
- GE, despite often being labelled as stodgy, is transforming itself in software, especially in its Internet of Things and the Industrial Internet initiatives. It sells aircraft jet engines at cost, but makes money on 30-year services and maintenance contracts. Remember, for instance, the Boeing 747 was first introduced in February 1969, some of them are still running on GE engines. GE believes in a centralized data model which is very crucial for the complete lifecycle of an aircraft engine, including configuration management.
- One academic claimed in 10 years Data Scientists will rule the world. It'll be what MBAs are now; the U. S. graduates about 160,000 every year. Sure, that's all we need: Number-crunching, algorithm-driven quants running Wall Street and hedge funds—moving money around, creating no real wealth, believing in Destructive Creation vis-à-vis Silicon Valley's Creative Destruction!

The Future of Data Science

Some commonly asked questions are:

- Where do the data end and the application begin?
- Or, are they fused forever?
- Is there a need for cross-functional organizational collaboration and the tightly coupled practices of data science and software development?
- What new areas of study are going to matter next?
- What is going to drive success, and how should we be thinking about our future?

Conclusion

Although, this was a well-organized conference with Pivotal gaining a lot of visibility, we feel many of the sessions could have delved deeper into technical aspects of Big Data and Science. There was very little discussion on some of the pitfalls, 'gotchas', and lessons learned from early implementers.

