CapTech® Others Talk,
We Listen.

WHITE PAPER

AGILE AND BIG DATA: SPRINTING TO BUSINESS INSIGHT

Introduction

Few businesses today can afford to use the traditional waterfall methodology to build and deliver data products. Across industries, the business environment is moving faster than ever, causing data product requirements to change at record speed as well. Data that was important only a month ago may be irrelevant today.

Every two days, we create as much data as we did from the beginning of time to 2003, according to Eric Schmidt, chief executive officer of Google.[1] In this environment, waterfall can't reasonably be relied upon to help businesses turn data into insights as quickly as is necessary. Organizations that can't keep pace are at risk of being left behind.

A common thread in the failures of Blockbuster, Borders, and Circuit City involved the inability to innovate. Many companies today use a combination of agile methodologies and big data tools to quickly convert data into business insights, speeding innovation. Spotify and Netflix, which leverage agile as well as big data tools, exemplify successful innovation.

Although some organizations have balked at using agile in a big data environment, our experience suggests that it is not only workable—it is essential. This paper discusses how the combination of agile and big data tools accelerates time to insight, driving business innovation.

Agile versus waterfall in the big data world

CapTech has argued for some time that agile methodologies can deliver powerful results in a big data environment. In a 2013 blog, for example, I wrote: "Delivering a [data] warehouse project using agile will feel uncomfortable at first. It probably won't work perfectly and you may have to adapt the processes to work with your organizational policies and procedures. Stick with it, listen to your team and don't be afraid to make adjustments as you go (it is agile after all). If you go in with an open mind and a clear roadmap, agile will deliver business value faster than you ever imagined possible."[2]

That reasoning still applies today, although the need for agile transformation has become more urgent.

Continued resistance to agile reflects problems that some organizations have had in attempting to use agile in traditional enterprise data warehouse (EDW) projects; i.e., projects not involving big data technologies. Three major problems make agile largely unworkable in that environment:

- **Data modeling isn't easy to do iteratively.** In agile, work is accomplished in sprints, or iterative time boxes, with the team delivering usable components of a larger project in a matter of days or weeks instead of waiting to deliver the entire project at one time. The process of creating a data model for a traditional data warehouse is too rigid and structured to support iterative work. Generally, developers have to create the entire model in order to determine if it will work; i.e., deliver business value. If there are issues with the model, the team might need several weeks to revise it. This isn't necessary in a big data environment, where the model can be adjusted while still in progress.

- **Excellent source metadata is required up-front**. To map and build a data model in the traditional environment, the team needs a full understanding of the data it will be working with. Access to excellent metadata is key. With the organization focused on creating solid metadata, months might pass before data model development can get under way. In the meantime, the business evolves, customer needs change and data comes and goes at steadily increasing velocity. The combination of agile and big data tools can help organizations resolve this problem. Teams can leverage known data in order to begin delivering value even before all metadata is discovered and documented.
- **Accommodating requirements changes is difficult when these changes affect the data model.** If the business recognizes that its needs are changing in the midst of a traditional project, requirements are revised. By the time the data model can be modified accordingly, however, months might pass. This leaves both IT and business teams frustrated; meanwhile, the business struggles to keep pace with the competition, despite significant investments in data products. Agile and big data address this problem by ensuring that business input is obtained throughout the project, enabling the team to revise data models immediately, and not solely after the project is completed.

> If agile isn't the standard in your big data ecosystem, you're probably heading for serious trouble.

Now that many organizations have moved beyond traditional EDWs and embraced big data technology and modern data architectures, it is no longer as challenging to deliver insight quickly using agile such as scrum or Kanban. (Scrum is a framework for structuring and delivering work iteratively. Kanban is another type of agile methodology; it limits work in process and creates continuous delivery through incremental improvement.) Not only is agile faster, but it also improves flexibility and quality. If agile isn't the standard in your big data ecosystem, you're probably heading for serious trouble.

Data governance in the big data world

Through the use of big data technologies, businesses can capture all the data they want, even if they aren't sure it's valuable. They can decide later whether to build data models and put the data to use. This wasn't feasible prior to the advent of big data tools, when businesses needed to know what they were going to do with data before they could even capture it.

Although this change brings new opportunities to businesses, it also presents issues. Where is data coming from? Who has access? How reliable is it? And how is it cataloged? Unless such questions are addressed, businesses soon find themselves operating data swamps instead of data lakes.

Data governance is the key, addressing such questions as the origins, or lineage, of data, who can access data and what they can do with it, how data is catalogued (i.e., metadata), and the quality and completeness of data.

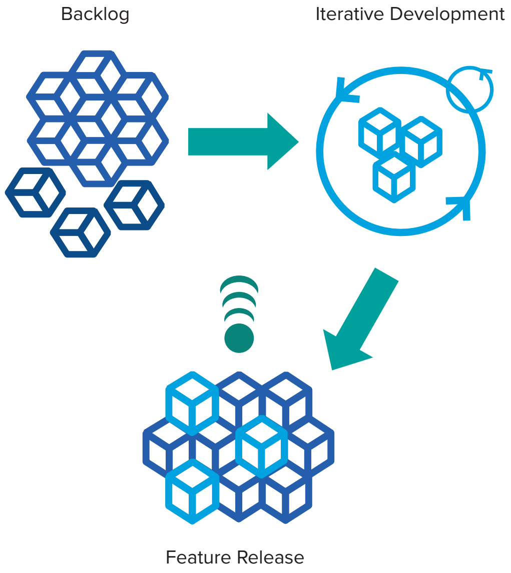
By addressing these issues clearly, data governance keeps data scientists doing what they do best – finding insight. It also keeps the organization safe and compliant by providing full documentation of how data is used and by whom.

For more information about data governance, please view our white paper, "Five ways modern data governance will make your organization more productive."

Why agile and big data deliver quicker insights

Schema on read supports agile data modeling. A schema defines how data is to be read; for example, it may specify that the first nine digits in a data file are Social Security numbers. Data files typically can be viewed in many different ways through many different schemas.

In traditional data warehouses, schemas have to be developed before data can be ingested. This approach is referred to as schema on write. In a big data environment, schema on read becomes possible, with data being applied to the schema as the data is brought out of storage, not as it is brought in. In other words, data isn't organized until it's used.



This eliminates time-consuming bottlenecks that traditionally have developed as organizations have built schemas. With data storage now relatively inexpensive, businesses can quickly load as much data as they want into a data warehouse and retain it in raw form for later use. Placing all the data in common location, typically a data lake, makes it conveniently accessible to users across the enterprise.

> "This approach enables the business to solve given problems immediately, instead of waiting many months…"

Using agile, the business gains even greater advantages. Data engineers can slice up the important data in the data lake, develop rudimentary schemas and conduct analysis and discovery work in two- to four-week sprints, delivering chunks of business value rapidly. As the agile team gets feedback from business users, the team can perfect the rudimentary schemas, ensuring these fully meet business needs. This approach enables the business to solve given problems immediately, instead of waiting many months to receive an overarching (and probably out-of-date or off-target) solution designed to solve a wide range of problems.

Real questions are answered with every sprint. Agile puts business and IT teams together in a co-located space, giving the development team direct access to the business when questions arise. Developers inevitably have many questions for the business, and having the business team readily available to respond is a huge time-saver.

More importantly, the business can see and evaluate data much more quickly than in a traditional warehouse environment. That's because big data technologies enable the development team to quickly profile and share vast amounts of data. The business team then can determine whether the data will answer real questions. These teams are no longer hindered by rigid data models or the lengthy timelines that typically accompany building these models.

If the agile team is mining data, it's essential that people with business understanding as well as data exploration and data preparation skills be readily available. The role of the data scientist encompasses all three of these areas of expertise; however, the role is often filled by multiple people. In an agile environment, people with all of these skill sets are typically located in the same room, so they can quickly interact, share information, and provide feedback. In the waterfall methodology, with resources/skill sets located in different places and, in many cases, different time zones, data science activities, such as data mining, become far more complex and time-consuming, delaying the possibility of answering business questions.

The data lake serves multiple needs, including production workload and data discovery. Organizations that have adopted a modern data architecture and a data lake have set themselves up for success. A data lake brings all of the organization's data together in a single repository (the lake) and makes it available to data consumers throughout the organization. (See sidebar on data governance.)

With a data lake, project teams can quickly find the data they need in order to develop business insights. If the data isn't available, new data sets can easily be brought into the lake.

A well-established lake provides an area for discovery as well as an area for production. In the discovery area, the team can use data exploration and mining tools as needed. If new tools are introduced, these will need to be moved to the production environment, along with the associated data product, once value has been proven.

Co-location brings the business and IT together. A key reason that agile can deliver business insights faster is that these methodologies promote frequent and direct communication among members of the project team.

The Agile Manifesto states that the "most efficient and effective method of conveying information to and within a development team is face-to-face conversation."[3]

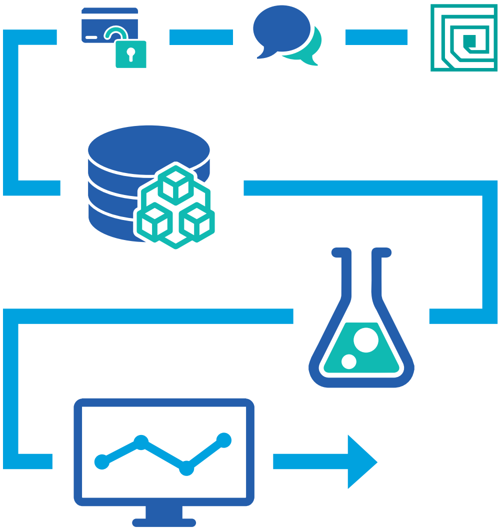
A recent Forrester Consulting Study commissioned by CapTech stresses that "frequent, clear communication" is "very important to project success." It also notes the advantages of using agile methodologies as well as onshore/nearshore resources in digital transformation projects.[4]

Businesses clearly recognize the relationship between communication and project success. The Forrester study, based on a survey of 300 IT & business professionals of U.S. enterprises responsible for decisions related to third party service providers indicated that:

- 50% of respondents said that "lack of communication can lead to a host of issues, including the issues with the deliverable project and project delays."
- 37% said that "lack of communication" was among the most common causes of delays of third-party projects.
- 65% of respondents told Forrester they had brought work previously performed offshore back onshore due to poor project experience. Of these firms, 36% cited "poor communication with offshore staff" as a reason for bringing work back onshore.

Agile support timely, direct, and clear communication, which can only enhance results.

Open source tools increase efficiency, speed, and cost-effectiveness. Modern data architecture is primarily made up of open source tools such as Hadoop, Spark, Cassandra, Caravel, R, Nifi, and Kafka. These allow agile teams to try new capabilities without having to deal with costly licensing concerns. At many businesses, obtaining a license means dealing with a time-consuming internal procurement process and negotiating with the vendor. That can take weeks, not to mention substantial money.



In contrast, open source tools give a team the autonomy to innovate, using freely available and rapidly advancing open source technology. These tools also save time, enabling the team to deliver business insights more rapidly. If and when value is delivered, the procurement staff can be engaged to negotiate a contract with a vendor that distributes the open source software.

The use of open source technology also gives teams the flexibility to modify or contribute to the open source project as needed. If they are needed, proprietary tools always remain an option, and they work well with many open source products.

The cloud further increases speed

As agile teams build data products, they likely will need access to computing resources and storage that traditionally would take weeks or months to procure, configure, and install in the data center.

The cloud eliminates this constraint, enabling teams to procure computing resources and storage in seconds – yet another way to get to the business of data discovery quickly. Although the cloud isn't an agile methodology or, for that matter, a big data technology, many companies that use agile and big data tools leverage the cloud in order to gain additional speed to insight.

For organizations that don't want to risk putting production information in the cloud, considerable innovation work can still be done, leveraging the speed of the cloud infrastructure, the power of big data, and agile. Using these tools and capabilities with only test data to build prototype data products in the cloud will still help organizations improve speed to insight. Once the agile team is comfortable with the prototype, the prototype can be moved to an on-premises data center for final acceptance and production deployment.

DevOps identifies problems before they enter production, adding speed

Bringing operations teams into the development process can solve potential problems before they become real problems, further increasing speed to value. A DevOps approach enables the development team to get feedback from the operations team during the development process. This not only helps identify issues before they may crop up in production but it also gives the teams an opportunity to determine how issues will be addressed if they do arise in production.

The velocity at which businesses are moving today makes it critical that data products be as effective as possible in production. At Amazon, for instance, more than $80,000 in sales are generated every minute.[5] In 2014, customers in those countries where Amazon's Prime service was available ordered 34.4 million items, or 398 items per second, on "Prime Day," surpassing Amazon's Black Friday sales.[6]

In that environment, even a slight change in how data and analytics are used can have an enormous impact. If a data product can make a positive difference, getting it into production quickly is key. If a data product is deficient, resolving the problem or pulling the product out of production quickly is equally critical. In either case, DevOps can make a tremendous difference.

Conclusion

Few large businesses today can afford to build and deliver data products using traditional waterfall methodologies. Businesses and the data they rely on for decision-making are moving too quickly.

According to ScienceDaily,[7] "A full 90 percent of all the data in the world has been generated over the last two years." With this much data moving this rapidly, businesses have to find insight quickly or be left behind.

Agile enables teams to build small slices of functionality that deliver business value each and every sprint. The big data environment enhances efficiency by removing traditional data modeling constraints. The use of open source technologies, common in the big data environment, removes software licensing constraints.

With this powerful combination of agile and big data tools, it is no wonder that a small co-located cross-functional agile team can deliver business value each and every sprint, helping businesses improve decision-making, innovate, and compete more effectively.

CapTech: Helping you sprint to insight

At CapTech, we have deep experience in using big data and agile to help clients bring new data capabilities to end users quickly and efficiently. We also help clients adopt the right agile framework within their own organizations. Whether you're looking for highly targeted solutions such as building a big data product, are considering changes in how you execute your data and analytics projects, or need help getting started with the cloud, our Agile Transformation and Big Data teams can help you deliver business insight faster.

End Notes

[1] "People Aren't Ready for The Technology Revolution." Aug. 5, 2008. The Huffington Post. Available at http://www.huffingtonpost.com/2010/08/05/google-ceo-eric-schmidt-p_n_671513.html.

[2] "Delivering Data Warehousing and BI Projects using Agile." June 28, 2013. Ben Harden, CapTech. Available at https://www.captechconsulting.com/blogs/delivering-data-warehousing-and-bi-projects-using-agile\.

[3] "Manifesto for Agile Software Development." February 2001. Available at www.agilemanifesto.org.

[4] Onshore/Nearshore Services Thrive In The Age Of The Customer: Selecting The Right Partner For Your Company's Digital Transformation, May 2016. A commissioned study conducted by Forrester Consulting on behalf of CapTech. Available at http://www.captechconsulting.com/site%20assets/white-papers/onshore-nearshore-services-thrive-in-the-age-of-the-customer.

[5] "The Data Explosion in 2014 Minute by Minute." Susan Gunelius, ACI Information Group. July 12, 2014. Available at http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/.

[6] "Amazon Prime Day shattered global sales records." CNN Money. July 16, 2015. Available at http://money.cnn.com/2015/07/15/news/amazon-walmart-sales/.

[7] "Big Data, for better or worse: 90% of world's data generated over last two years." ScienceDaily. May 22, 2013. Available at https://www.sciencedaily.com/releases/2013/05/130522085217.htm

AUTHOR BIO

Ben Harden | bharden@captechconsulting.com

Ben Harden leads the Data and Analytics practice at CapTech and has over 18 years of enterprise software development experience in the areas of data warehousing, metadata management, data governance, analytics, engineering, and enterprise scale Hadoop data ingestion and refinement. He is also an Agile Scrum coach who specializes in making data delivery teams successful using the Agile methodology.