Analytics at Amazon Speed: The New Normal

Christopher Bergh and Gil Benghiat

Christopher Bergh is a founder and head chef at DataKitchen with more than 25 years of research, engineering, analytics, and executive management experience.
cbergh@datakitchen.io

Gil Benghiat is a founder and vice president of products at DataKitchen.
gil@datakitchen.io

ABSTRACT

Many companies, including Amazon and Google, have turned instant fulfillment into competitive advantage and are being rewarded in the marketplace. As consumers adapt to this "new normal," the expectation of instant delivery is crossing into other domains. For example, data analytics users can't or won't wait weeks or months for new analytics.

Data analytics teams that can successfully meet requirements for rapid delivery of new analytics will play a high-visibility role in helping their organizations compete in the on-demand economy. Enterprises can improve the speed and strength of analytics using a process and tools approach called DataOps, which draws from process innovations in software development and lean manufacturing. Organizations that correctly implement DataOps experience significant improvements in the ability to produce robust and adaptive analytics. DataOps may be implemented in seven simple steps without discarding an organization's existing analytics tools.

ANALYTICS IN THE ON-DEMAND ECONOMY

The world changed in February 2005 when Amazon Prime brought flat-rate, unlimited, two-day shipping into a world where people expected to pay extra to receive packages in four to six business days. Since its launch, Amazon Prime has completely transformed the retail market, making low-cost, predictable shipping an integral part of consumer expectations (Mangalindan, 2015). This business model, which some have called the "on-demand economy," is pop-

ping up in many industries and markets across the globe (Jaconi, 2014).

For example, some may remember video stores where movies were rented for later viewing. Today, 65 percent of global respondents to a recent Nielsen survey watch video on demand (VOD), many of them daily (Nielsen, 2016). With VOD, a person's desire to watch a movie is fulfilled within seconds. Although it's not the leading vendor, Amazon participates in the VOD market with their Amazon Video service.

Instant fulfillment of customer orders seems to be part of Amazon's business model. They have even brought that capability to IT. About 10 years ago, Amazon Web Services (AWS) began offering computing, storage, and other IT infrastructure on an as-needed basis. Whether the need is for one server or thousands and whether for hours, days, or months, you only pay for what you use and the resources are available in just a few minutes.

To successfully compete in today's on-demand economy, companies need to deliver their products and services just as Amazon has done—in other words, at Amazon speed. What might be surprising to many is how the expectations of instant fulfillment are crossing over into data analytics, which, along with everything else in the digital economy, is now expected to happen at Amazon speed with Amazon predictability.

A typical example: the VP of sales enters the office of the chief data officer (CDO). She'd like to cross-reference the customer database with some third-party consumer data. The CDO asks for time to study the problem and, days later, has planned the project. Resources will be allocated

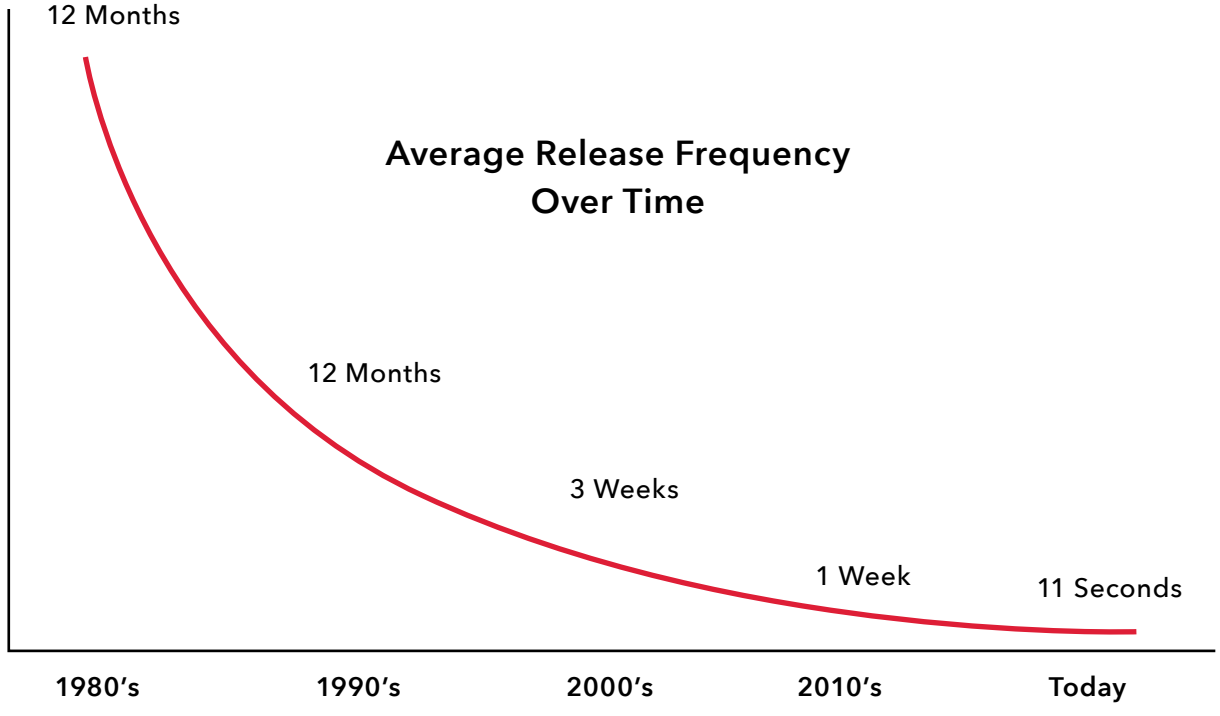

Figure 1: Average software development release frequency over time.
Source: DataKitchen, 2017.

and configured, schemas will be updated, reports will be elegantly designed, and the delivery pipeline will be thoroughly tested. The changes will take several weeks. "Not acceptable," the VP of sales fires back. The new analytics are needed for a meeting with the board later in the week. "The competition is ahead of us; we can't wait weeks." This scenario is playing out in one form or another in corporations around the globe.

When did the requirement of fulfilling data analytics at Amazon speed cross over into corporate boardrooms? Consumer services such as Amazon Prime may have had something to do with it, but this thinking is also widespread in the software industry.

SOFTWARE DEVELOPMENT AT AMAZON SPEED

Over recent decades, the average frequency of software releases has dropped from years to months to weeks to seconds. This incredible improvement stems from evolutionary changes in software development tools and methodologies. In the early days of computing, most large development organizations used the waterfall model of project planning, adapted from the manufacturing and construction industries, which required detailed planning and highly structured project management.

Projects progressed through the phases of conception, initiation, analysis, design, construction, testing, production/implementation, and maintenance sequentially with little ability to change course midway. Using the waterfall model, the average frequency of software releases in the 1980s was about 12 months (see Figure 1). The waterfall model is best suited to projects with well-understood requirements and is less effective in fast-changing industries where requirements rapidly evolve.

For much of the 1990s, most companies were still using the waterfall model and release frequency saw modest improvement. A sea change, however, occurred in the late '90s as software development organizations began to abandon the waterfall method in favor of a new methodology known as agile development.

Agile throws away the old method with its detailed up-front plans and bureaucratic sequential phases. Companies began to organize code development in short iterations, producing valuable incremental changes in short periods of time. Customers or users provided immediate feedback on each release, which was then used to influence future priorities. Agile has revolutionized software development, increasing the speed and productivity of programmers significantly (Rigby, Sutherland, and Takeuchi, 2016). Using agile methodology and principles, the average release frequency dropped to about three weeks in the 2000s.

More recently, development took another gigantic leap forward in productivity. Agile development evolved into DevOps, which includes additional process and tool optimizations that have improved the average delivery time of software projects from weeks to days. With on-demand cloud services from vendors such as AWS and the inexpensive provisioning of compute resources, organizations have been able to break down the traditional silos that separated planning, design, development, testing, maintenance, and, perhaps most important, IT. The unification of these functions on a common platform (the cloud) has enabled companies to implement *continuous delivery*. They deploy new releases to customers many times a day, sometimes every few seconds.

Amazon and Google, for example, have become extremely large yet remain nimble by adhering to agile methods, DevOps, and continuous delivery of software. They have shown that the way to grow in the on-demand economy is to apply a philosophy of continuous improvement to products and services. Companies that can follow suit will thrive in the new economy.

DATA ANALYTICS IS JUST CODE

The process of preprocessing, cleaning, checking, transforming, combining, analyzing, and reporting data—all the work that comprises data analytics—is encoded in a set of files that control the data analytics pipeline. These files include scripts, source code, algorithms, HTML, configuration files, parameter files, and containers. All of these items are essentially just code, which is written, maintained, and managed by technical personnel. As such, analytics is subject to the same challenges, the same market forces, and the same potential optimizations as software development. The problem is that most analytics teams are back in the waterfall world of painstaking bureaucracy: writing detailed specifications, planning sequential development schedules, and taking months to implement changes.

THRIVING IN THE ON-DEMAND ECONOMY

Continuous delivery has become one of the core competencies that will determine which companies thrive and which will be left behind. The burdens of the on-demand economy will fall squarely on the shoulders of data analytics teams who, in this new environment, are expected to deliver analytics at Amazon speed. In a recent survey, the technology research firm Gartner found that only about half of all CDOs in large organizations are considered successful in their roles. These CDOs work for companies that face a rocky future, because without a

AMAZON EMBRACES CONTINUOUS DELIVERY

After launching Amazon Web Services (AWS; i.e., cloud computing) in 2006, Amazon then opened up those services to be shared between development teams. This allowed them to quickly move to continuous delivery for new features and capacity. They release improvements to the AWS platform approximately every 11 seconds. In the decade-plus since its inception, AWS has become a foundation upon which the services and subscription-based economy is being built, and it is Amazon's most profitable division.

AWS developers perform thousands of software changes per day. The software is tested internally and then deployed to customers on a rapid and continuous basis. As you read this, thousands of servers are being updated to deliver new functionalities to be deployed across the range of AWS products and services. This efficiency, adaptability, and customer focus has allowed AWS to offer consistently lower prices, reducing margins for competitors and increasing AWS market share. According to Gartner, AWS is 10 times bigger than its next 14 competitors combined. More important, by providing compute resources that can be quickly provisioned to match demand, AWS has enabled numerous organizations to become much more nimble.

RELEASING IN EIGHT MINUTES AT GOOGLE

At Google, over 23,000 R&D employees build over 5,000 different services (software components) such as login, storage, and indexing. These services are shared among Google's wide array of products, which are continuously evolving. A large number of Google services are released to users multiple times in a single week. One group, the Google Consumer Surveys group, deploys code to customers eight minutes after a developer finishes writing and testing it.

Google maintains two instances of its system: one for production and one for testing. Over 100 million automated test scripts are run per day to ensure the new features released by developers work cohesively with the rest of the services. Product managers at Google deploy a feature to a small percentage of users before rolling it out to everyone. This enables them to receive feedback from users before going fully live.

The continuous releases update the feature until product managers are sure that it is robust and enhances the user experience. This allows them to keep improving the software services that form the components of their product. As a feature matures, they release it to larger and larger segments of the customer base, ensuring the new service integrates with the existing user experience. This reduces risk and keeps the product teams focused on customers' needs.

responsive analytics function, companies can't expect to compete and win in rapidly evolving markets. Since 2000, at least 52 percent of *Fortune* 500 companies have gone bankrupt, been acquired, or ceased operations (Constellation Research, 2014).

To compete in the new economy, a company needs information about customers, trends, and markets that only the data analytics team can provide. Analytics can be the core competency that enables companies to successfully navigate the rocky waters of the on-demand economy. In this new world, the CDO and the data analytics team have unique visibility. Successful analytics teams will help lead their companies to bright futures. Those who ignore the changes afoot will fade into oblivion like those former *Fortune* 500 companies.

CDOs and data analytics teams can improve their performance and quality by instituting process and tools changes that have been shown to work in software development and lean manufacturing. These changes, called *DataOps*, can help a team deliver analytics at Amazon speed while ensuring that new analytics do not disrupt operations.

DATAOPS INCORPORATES AGILE AND DEVOPS

DataOps incorporates the speed of agile software development and the responsiveness of DevOps and continuous delivery into data analytics. Similar to the continuous delivery process at Google, DataOps places a great deal of emphasis on automated testing at each stage of the data analytics pipeline in order to ensure quality. This testing supports process controls that are important to quality improvement.

DATAOPS INCORPORATES LEAN MANUFACTURING

Lean manufacturing is a key part of the intellectual heritage of DataOps. Data analytics progresses through a series of steps to produce a desired output in the form of reports, models, and views. At an abstract level, the data analytics pipeline is analogous to a manufacturing process. Statistical process control (SPC) is a well-known method used to improve manufacturing quality. If key measures are within specific limits, the process is considered to be functioning within its expected bounds. When SPC is applied to the data analytics pipeline, it can help ensure quality as well as warn the analytics team of unexpected patterns in the data, enabling them to update the analytics or develop more robust tests. The emphasis on testing in DataOps reflects the importance of SPC in achieving adaptive, robust analytics.

DATA ANALYTICS IN A DATAOPS WORLD

DataOps requires changes in both processes and tools that deliver analytics. With DataOps, CDOs and data analytics teams are able to respond to requests for changes at previously unfathomable speed. When a C-level executive asks for a new view of the data, the data analytics team responds that same day. This unlocks the productivity and creativity of decision makers, and allows key contributors to experiment with analytics, seeking new patterns and trends. Good companies that master their analytics will become outstanding companies.

The data analytics pipeline in DataOps is automated so changes can flow through to the users rapidly and continuously. Testing ensures that any updates are implemented without errors. No more waking up on Saturday morning after a long week, wondering if Friday's change broke something. No more enterprise-critical IT alerts after business hours.

Utilizing DataOps, data analytics engineers, analysts, and scientists work on changes without getting in each other's way. All changes to the analytics are captured, managed, and backed up so that they are easily reproducible. The diffuse bits and pieces of analytics that are otherwise spread across the many hard drives of individual employees are collected into one coherent repository. Team members can work on their own private copies of the data, eliminating clashes with production or interference with live, business-critical systems.

DataOps allows the data analytics team to share code and methods with each other. Changes are shared and adopted across the whole team. Complex processing is encapsulated and isolated so that modularity is improved across the data analytics code base.

DataOps enables the data analytics pipeline to be flexible enough to adapt to runtime conditions that frequently recur. You can filter a database according to any required criteria and include or exclude specific steps in the workflow. If a new model is released, the data analytics team can make both the new and old models simultaneously available. If an analysis step is interrupted, it can be restarted without losing hours of batch processing time.

DataOps stores raw data in data lakes with purpose-built data marts and data warehouses serving the organization's day-to-day analytics needs. When changes are needed, the code that generated the data warehouses and data marts is modified to accommodate the new requirements. Cloud resources allow the new modified data

warehouse to be spun up in a matter of minutes without impacting operations.

IMPLEMENTING DATAOPS

To implement DataOps, an analytics team does not need to throw away any of their beloved tools. There are tools that can help optimize the data analytics pipeline, but the methodology and philosophy of DataOps is just as important as the tools. An organization can migrate to DataOps in seven simple steps.

STEP 1: ADD DATA AND LOGIC TESTS

To be sure that the data analytics pipeline is functioning properly, it has to be tested. Testing of inputs, outputs, and business logic must be applied to each stage of the data analytics pipeline. Tests catch potential errors before they are released so quality remains high. Manual testing is time-consuming and laborious. As practiced at Google, a robust automated test suite is a key element in achieving continuous delivery, which is essential for companies in the on-demand economy.

STEP 2: USE A VERSION CONTROL SYSTEM

As previously discussed, all of the processing steps that turn raw data into useful information are source code, which can control the entire data analytics pipeline from end to end in an automated and reproducible fashion. In many cases, the files associated with analytics are distributed in various places within an organization without any governing control. A revision control tool, such as Git, helps store and manage all of the code changes. It also keeps code organized in a known repository and provides for disaster recovery. Revision control also helps software teams parallelize their efforts by allowing them to *branch and merge*.

STEP 3: BRANCH AND MERGE

To make updates, an analytics professional checks a copy of all the relevant code out of the revision control system and then makes changes to a local, private copy. These local changes are called a *branch*. Revision control systems boost team productivity by allowing many developers to work on branches concurrently. After changes to the branch are complete, tested, and known to be working, the code can be checked back into revision control, where it is then *merged* back into the trunk or main code base.

Branching and merging allows the data analytics team to run their own tests, make changes, take risks, and experiment. If a set of changes proves to be unfruitful, the branch can be discarded and the analytics team member can start over.

STEP 4: USE MULTIPLE ENVIRONMENTS

In addition to having a local copy of the code, data analytics professionals need a private copy of the relevant data. In many organizations, team members work on the production database, which often leads to conflicts and inefficiencies. With on-demand storage from cloud services, even a terabyte-sized data set can be quickly and inexpensively copied to reduce conflicts and dependencies.

STEP 5: REUSE AND CONTAINERIZE

Data analytics team members typically have a difficult time leveraging each other's work. Code reuse is a vast topic, but the basic idea is to split functions into components that can be shared. Complex functions, with lots of individual parts, can be containerized using a container technology, such as Docker. Containers are ideal for highly customized functions that require a skill set that isn't widely shared among the team.

STEP 6: CHOOSE PARAMETERS FOR YOUR PROCESSING

The data analytics pipeline should be designed with runtime flexibility. Which data set should be used? Is a new data warehouse used for production or testing? Should data be filtered? Should specific workflow steps be included? These types of conditions are coded in different phases of the data analytics pipeline using *parameters*. In software development, a parameter is some information (e.g., a name, a number, an option) that is passed to a program that affects the way it operates. With the right parameters in place, accommodating the day-to-day needs of the users and data analytics professionals becomes a routine matter.

STEP 7: USE SIMPLE STORAGE

Cloud storage vendors offer the convenience of simple, high-reliability storage at a relatively low cost. This *simple storage* is an ideal way to create data repositories, data lakes, data warehouses, and data marts for analytics. With cost-effective storage, it is possible to maintain data in its original raw form in data lakes. With the original data handy, it is easy to create new data marts and data warehouses upon request.

DATAOPS CASE STUDY

Companies that are implementing DataOps are seeing tremendous improvements to their data analytics cycle time and ability to adapt to new analytics requirements. The benefits of rapid and robust analytics flow through the entire organization.

At one pharmaceutical company, DataOps enables data analytics to be a self-service function for many individuals in the organization. They have kept the analytics tools that they rely upon but have woven them together into a cohesive pipeline. They can easily make changes to data marts and data warehouses, and a robust test suite verifies that none of the changes interrupt the flow of analytics. Enhancements are implemented quickly and released confidently, satisfying the many requests that flow in from users.

Salespeople have dashboards with forecasts, opportunities, bookings, shipments, and all the other basic information needed to do their jobs. This frees up the analytics team to focus on higher-value analytics, which help business leaders understand their growing and fast-changing marketplace. Analytics processes are updated with internal users "shoulder to shoulder," facilitating immediate feedback and greatly shortening the time it takes to provide users with applicable analytical tools.

CONCLUSION

Amazon, Google, and other companies driving the on-demand economy are transforming markets and, by association, expectations for how quickly data analytics should be implemented. CDOs and data analytics teams that can adapt to this changing environment will be successful, leading their organizations' efforts to innovate. Those who do not update their methods won't be able to keep up. DataOps is a methodology that enables data analytics teams to thrive in the on-demand economy. It allows data analytics to be updated nimbly while maintaining a high level of quality. Using seven simple steps, companies that have embraced DataOps have seen tremendous improvement in user satisfaction and development of analytics as a key competitive advantage.

REFERENCES

Constellation Research [2014]. "Research Summary: Sneak Peeks from Constellation's Futurist Framework and 2014 Outlook on Digital Disruption," February.

Jaconi, Mike [2014]. "The 'On-Demand Economy' Is Revolutionizing Consumer Behavior—Here's How," *Business Insider*, July 13. http://www.businessinsider.com/the-on-demand-economy-2014-7

Mangalindan, J. P. [2015]. "Inside Amazon Prime," *Fortune*, February 3. http://fortune.com/2015/02/03/inside-amazon-prime/

Nielsen [2016]. "On-Demand Demographics: VOD Viewing Across Generations." http://www.nielsen.com/us/en/insights/news/2016/on-demand-demographics-vod-viewing-across-generations.html

Rigby, Darrell K., Jeff Sutherland, and Hirotaka Takeuchi [2016]. "Embracing Agile," *Harvard Business Review*, May.