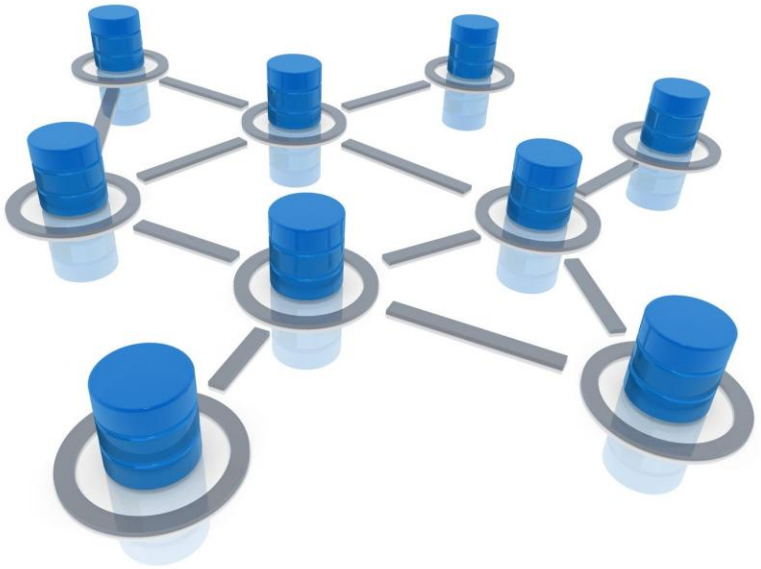
BIG DATA BASICS

An Introduction to Big Data and How It Is Changing Business



James Higginbotham
LAUNCHANY.COM

INTRODUCTION

Amazingly, 90% of the data in the world today has been created only in the last two years. With the increase of mobile devices, social media networks, and the sharing of digital photos and videos, we are continuing to grow the world's data at an astounding pace.

However, big data is more than just the data itself. It is a combination of factors that require a new way of collecting, analyzing, visualizing, and sharing data. These factors are forcing software companies to re-think the ways that they manage and offer their data, from new insights to completely new revenue streams.

ABOUT THIS BOOK

This book addresses the confusion around big data and what it means for your software business. Whether you are an executive, a product manager for an existing product, or a startup founder, this book will help you:

- Define big data and the factors involved when moving to the scale of big data
- Review case studies of recent big data projects
- Learn how the big data process works
- Understand big data architectures and the technology landscape
- Identify new business opportunities and revenue streams

ABOUT THE AUTHOR

James Higginbotham is the founder of LaunchAny, an Austin-based firm that specializes in launching software products. James has launched over 50 products in his career. He enjoys working with companies on their technology and product strategy. James has experience in architecting, developing, and deploying SaaS and big data applications to the cloud.

James serves on the advisory board of a non-profit organization that focuses on helping people become advocates and activists for their charities and causes. When not writing or consulting, he enjoys landscape photography.

James can be reached at james@launchany.com

WHAT IS TODAY'S DATA VOLUME?

In what units do we express the size of the world's information today?

A) Megabyte (10^6)

B) Petabyte (10^15)

C) Yottabyte (10^24)

TODAY'S DATA VOLUME

The answer: C) Yottabyte.

A Yottabyte is the equivalent of ~250 trillion DVDs worth of information, 90% of which was generated in only the last 2 years. In the future, we will likely see the amount of available data double every 2 years.

TODAY'S DATA GROWTH

The amount of data growth today is astounding.

Every 60 seconds:

- 98,000 tweets are created on Twitter
- 695,000 status updates are generated on Facebook
- 11 million instant messages are sent
- 698,445 Google searches are conducted
- 168 million emails are sent
- 1,820 TB of data is created
- 217 new mobile web users added

THE CHALLENGE OF TODAY'S DATA

Today's volume and growth of data creates a huge challenge when it comes to collecting, storing, and analyzing the data to find important metrics and trends.

And this challenge is extending beyond Facebook, Twitter, and Google:

"…through 2018, big data requirements will gradually evolve from differentiation to 'table stakes' in information management practices and technology. By 2020, big data features and functionality will be non-differentiating and routinely expected…"

– Gartner on Big Data (2012)

THE DRIVERS OF BIG DATA

HOW WE GOT HERE

TODAY'S BUSINESS ANALITICS

Today's business analytics explores past business performance (data) to gain insight and drive business planning (action). The data sets used have historically been limited to what we can capture within standard business processes:

1. Transactions
2. Surveys/sentiment
3. Click-logs

While data sets were large, they were often predictable and allowed for manageable data analytics.

THE NEW FACE OF BUSINESS ANALITICS

We now have the Internet of Things:

1. Multiple devices (e.g. mobile phones, tables, GPS, RFID)
2. Streams of data from each device
3. Consumer-driven content (e.g. social networks and social collaboration)
4. Online marketing and detailed click tracking
5. Advertising

The amount of data being generated from these new sources exceeds our previous analytic capabilities. A new strategy is required to accommodate these new requirements.

HYPE OR A GAME-CHANGER?

Big data is disruptive to everything businesses must manage:

1. More data sources (web, mobile, third-party)
2. More generated data
3. Batch or real-time processing
4. Right-time business insights

This is placing a heavier burden on technology solutions to deal with these new processes. IT can no longer solve these upcoming issues with a database and reporting tools. Big data solutions are now required.

GARTNER ON BIG DATA

Big data will drive $28 billion of worldwide IT spending in 2012, according to Gartner, Inc. In 2013, big data is forecast to drive $34 billion of IT spending.

"Despite the hype, big data is not a distinct, stand-alone market, it but represents an industrywide market force which must be addressed in products, practices and solution delivery," said Mark Beyer, research vice president at Gartner. "In 2011, big data formed a new driver in almost every category of IT spending. However, through 2018, big data requirements will gradually evolve from differentiation to 'table stakes' in information management practices and technology. By 2020, big data features and functionality will be non-differentiating and routinely expected from traditional enterprise vendors and part of their product offerings."

DEFINING BIG DATA

WHAT MAKES DATA BIG?

GARTNER SAYS:

"Big data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making."

Source: Gartner

BIG DATA DEFINED

We prefer to simplify the definition, focusing on the key differentiators of big data:

"Big Data Is Right-Time Business Insight and Decision Making At Extreme Scale"

THE 5 V'S OF BIG DATA: VOLUME

Big data may involve large amounts of data, from Terabytes to Petabytes and beyond.

It exceeds the vertical scaling capabilities of a single instance or storage cluster.

Examples include:

- Device measurements across millions of devices every 15 minutes (e.g. smart grid)
- User-generated content (e.g. Twitter/Facebook)
- Historical market transaction data (e.g. Netflix/Amazon)

THE 5 V'S OF BIG DATA: VELOCITY

Big data may also include the speed of incoming data. Even if the data itself is small (e.g. temperature from a sensor), the speed in which it arrives combined with the quantity of devices can exceed current capabilities. Velocity may also involve the speed in which analysis and insight is needed.

Examples include:

- Fraud detection
- Predicting Trends
- Real-time analytics

THE 5 V'S OF BIG DATA: VARIETY

Big data often involves combining multiple data types together in the processing, generating new insights previously not available. This may require multiple stages of data conversion and aggregation, prior to incorporation with other data sets. The larger the data set, the more involved this process may become.

Examples of sources that may be combined to provide greater insight:

- Transactions, click streams, and content for trend analysis
- Multiple real-time medical devices for patient assessment

THE 5 V'S OF BIG DATA: VERACITY

Not all data is considered equal. Determining the trust level of data by validating it against multiple data sources is often necessary. This may require additional levels of processing or filtering prior to data inclusion, adding complexity.

Examples include:

- Filtering spam
- Normalizing data/Master Data Management
- Audit trails/data lineage

THE 5 V'S OF BIG DATA: VALUE

The end result of applying big data solutions to your data is to drive business value. This value can create new stories or even new opportunities previously hidden in the data.

Examples include:

- Identifying new markets or demographics
- Netflix determining that the series House of Cards would succeed based on the viewing habits of their viewers
- Political campaigns using data to steer conversation with voters in near real-time

THE 5 V'S OF BIG DATA: PUTTING IT TOGETHER

"Big Data Is **Right-Time Business Insight** and **Decision Making** At **Extreme Scale**"

Velocity **Variety** **Veracity** **Volume**

Value

TECHNOLOGY CHALLENGES OF BIG DATA

WHY IS BIG DATA A BIG DEAL?

HARDWARE LIMITATIONS

Today's hardware is amazingly fast. A consumer desktop may have up to 8 CPU cores, each clocked at approx 3.4 Ghz. Compare that to only 10 years ago, when desktop CPUs had only one core and clocked at a peak of 1.8 Ghz.

Even with these advances, today's hardware can't keep up with the speed of data generation. Plus, CPU speed isn't the only limitation.

MODERN HARDWARE FACTS

Memory (RAM) - Fast & Expensive

- Max per computer – 256 GB

Hard Disk (HDD) - Slow & Cheap

- Read speed - 30MB/sec (typical) to 100MB/sec (burst)
- Max local storage per computer - 24 TB

Network - Slow

- Max speed 1.5Gb/s (typical) up to 10 Gb/s (max)

Processor - Speeds Flattening

- Individual core speeds remain at 3.4 – 4.0Ghz
- Number of cores per processor are increasing

OVERCOMING HARDWARE LIMITS

To overcome these modern limitations, a method of fast, cost-effective storage is needed. Currently, the approach is to utilize multiple machines to spread data out, increasing the maximum speed of reading data. This is often called “sharding”:

1 Hard drive = maximum of 100MB/Sec

10000 Hard drives x 100MB/Sec = maximum of 100 GB/sec (102,400 MB/Sec)

Each machine, or “node”, stores part of the data. Nodes may have extra copies of data to ensure data is not lost in the case of node failure.

DATA STORAGE LIMITATIONS

Beyond the hardware limitations, there are also storage limitations. Today, most structured data storage is managed in a relational database. Most relational databases enforce a set of rules to ensure that data is consistent (CAP Theorem). They also ensure transactions are atomic – they either succeeded or failed (ACID).

With these rules in place, it becomes much harder to spread data out to multiple nodes to increase retrieval speed and therefore processing speed.

Newer databases are offering different approaches to the CAP Theorem and ACID compliance to overcome these limitations.

CAP THEOREM

CAP stands for Consistency, Availability, and Partition Tolerance. CAP determines how your data storage will operate when data is written and the availability of the data when you go to retrieve it later (even under failure). You cannot have everything from all three categories, so each database must choose what they will implement and what they will sacrifice. In general:

1. Relational databases choose Consistency and Availability (CA), ensuring writes are consistent and immediately available across all instances
2. Many new database vendors are opting for Availability and Partition Tolerance (AP), accepting new/updated records without immediate confirmation ("eventually consistent")
3. Other database vendors are opting for Consistency and Partition Tolerance (CP), allowing arbitrary loss of messages to some instances, while the system continues to be available

Many vendors are experimenting with various combinations to satisfy specific use cases, and are also choosing to require specific infrastructure/architecture to support their implementation.

TECHNOLOGY LIMITATIONS REDUCE SCALABILITY

By understanding the CAP theorem, we can see that traditional relational databases generally require slower performance to ensure transaction consistency across one or more database servers. This is due to the requirement that the storage of data must occur on each database server, limiting vertical scaling to the speed of the slowest server's speed of storage.

While transaction consistency may be critical for some systems, when datasets reach extreme scale, traditional databases often cannot keep up and require alternative approaches to data storage and retrieval.

The result: big data architectures are required to overcome these limitations as our data grows beyond the reach of a single server or database cluster.

BIG DATA CASE STUDIES

WHO IS USING BIG DATA?

CASE STUDY: PROGRESSIVE INSURANCE

Progressive Casualty Insurance Company, uses big data as part of its “pay as you drive” program, offering drivers the chance to lower their insurance premiums based on real-time analysis of their driving habits. Data is collected from a “black box” installed into the car, then pushed to their servers for analysis. The results are then used to analyze a driver’s habits, make recommendations, and adjust their insurance rates.

Source: McKinsey

CASE STUDY: FRAUD DETECTION

Amazon and PayPal use a big data architecture for fraud detection for their secure e-commerce and payment platforms. They identify and flag potential fraud quickly by performing real-time analytics of transaction data. Prior to big data architectures, the identification of fraud was only available after days or months of analyzing transaction patterns.

Source: McKinsey

CASE STUDY: HEALTHCARE

Seton Healthcare needed a way to reduce the occurrence of high cost Congestive Heart Failure (CHF) readmissions. By proactively identifying patients likely to be readmitted on an emergent basis, they applied predictive models and examined analytics through which providers can intuitively navigate, interpret and take action.

The benefit: For Seton, a reduction in costs and risks associated with complying with Federal readmission targets. For Seton's patients, fewer visits to the hospital and overall improved patient care.

CASE STUDY: IMPROVING CUSTOMER SATISFACTION

Hertz needed a way to determine customer satisfaction by collating information from surveys. By applying advanced analytics solutions, the company was able to process the information much faster. The results are now half the time it previously took, while at the same time providing a level of insight previously unavailable to the company.

CASE STUDY: US PRESIDENTIAL ELECTION

The Obama campaign set out to create, as Chris Wegrzyn (director of data architecture for the DNC) described it, "an analyst-driven organization by providing an environment for smart people to freely pursue their ideas."

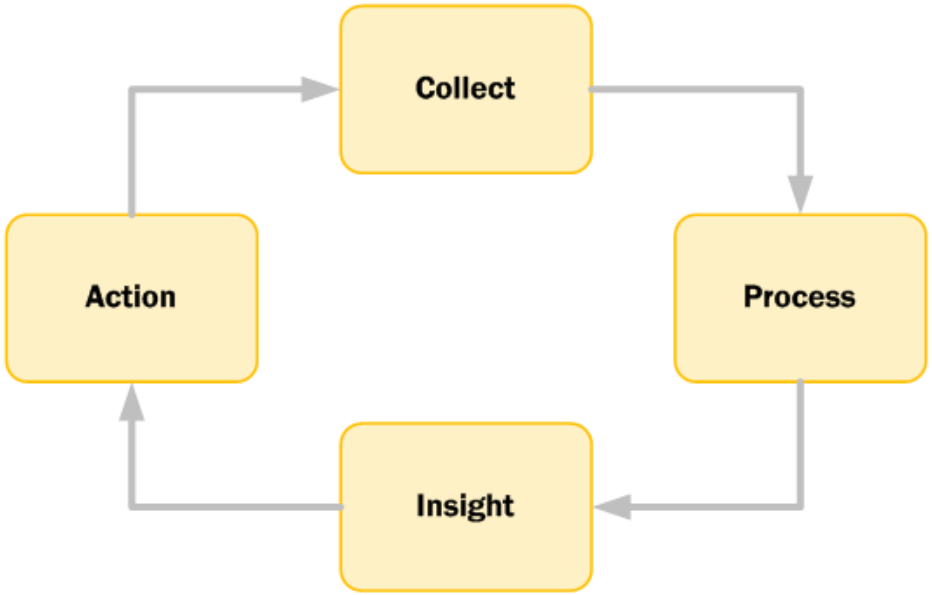
The project combined a SQL-friendly platform for analytics, rather than requiring a knowledge of Java or statistical analysis. In addition, the platform provided enough horsepower to enable analysis "at the speed of thought." The result was a friction-free analytic environment that combined data from a variety of sources and provided just-in-time results that fed directly to the campaign.

THE PROCESS OF BIG DATA

HOW DOES IT WORK?

THE BIG DATA CYCLE: OVERVIEW

The big data cycle is comprised of four steps that result in insight and action. The results of each cycle are often used for the next cycle, creating a continuous insight loop.



THE BIG DATA CYCLE: COLLECT

Data is first collected from devices, log files, click-tracking, third-party data sources. This may include traditional ETL processing but typically on a much larger scale. If data streams generate large bursts of data, it is quickly stored and processed later.

Storage strategies are critical, as the processing cycle (step 2) often needs fast access to the data. Therefore, data storage should be in close to the processing nodes to reduce network I/O issues.

THE BIG DATA CYCLE: PROCESS

The process step includes all computational tasks required to process the data. MapReduce is the most widely known pattern to processing big data. It works by processing subsets of the data across machines and then combining the results. It is often used in batch-based processing, meaning that operations occur on a known snapshot of data.

Demands for real-time processing are emerging, requiring the evaluation of data as it becomes available rather than when it all has been received. However, most big data problems can be managed in batch, as real-time processing is often difficult to maintain without falling behind when data velocity is high.

THE BIG DATA CYCLE: INSIGHT

The third step is to perform queries and analytics over processed data to find metrics and KPIs. The insight stage may require multiple passes of processing and analytics to obtain the desired metrics.

Analytics may be pre-calculated or adhoc. Adhoc query support over big data may yield results in seconds to hours (or even days) instead of sub-second as we may see today with data sets stored in relational systems.

Real-time analytics are complex but are beginning to emerge in big data. This allows for more immediate actions when patterns emerge prior to all data becoming available. This is also being applied to healthcare, particularly in hospitals and emergency rooms where multiple devices are capturing data about a patient all in real-time.

THE BIG DATA CYCLE: ACTION

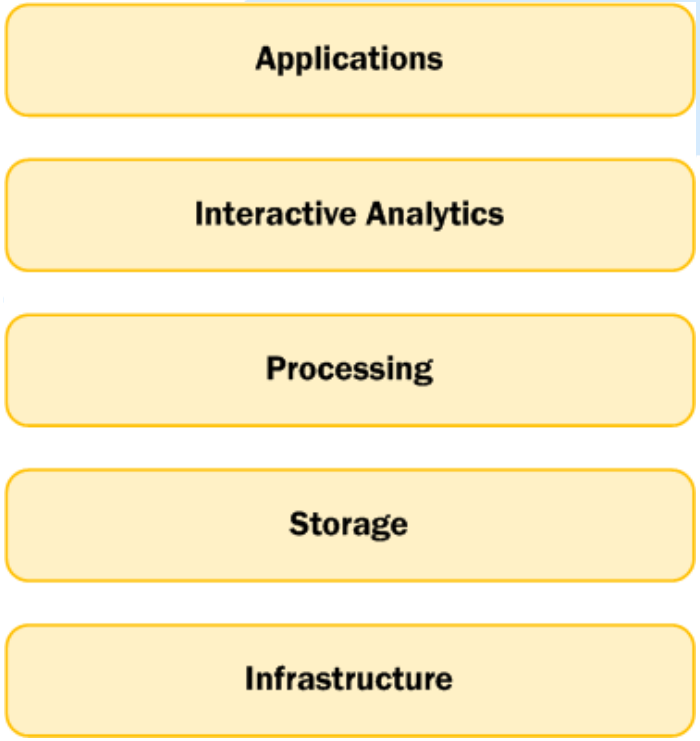
Once insight has been gained, action may be taken. This may take the shape of manual or software-driven adjustments to systems and processes. Action taken often starts the cycle over again, allowing for analyzing trends as a result of the changes or the availability of new data points.

THE BIG DATA LANDSCAPE

TECHNOLOGIES, PLATFORMS, AND SOLUTIONS

BIG DATA ARCHITECTURE

A big data architecture is generally composed of five layers, with each layer requiring one or more technologies:



BIG DATA ARCHITECTURE: INFRASTRUCTURE

Infrastructure refers to the compute resources necessary to perform calculations and analytics across big data sets. It may be comprised of:

- "Bare metal" hardware (i.e. racked servers)
- A virtualized, private cloud
- A virtualized, public cloud

While virtualization isn't a requirement to support a big data architecture, it is common as it allows hardware to be more fully utilized. It also allows for the elastic growth or reduction of assigned resources to running big data analytics, based on internal priorities and overall hardware availability.

BIG DATA ARCHITECTURE: STORAGE

As discussed previously, data storage and performance is often considered the biggest bottleneck in big data processing. Big data architectures often combine a number of storage options, based on processing and reporting requirements. This may include:

- Clustered and/or distributed file storage for unstructured data
- Semi-structure data stores such as NoSQL databases
- Structured data stores, such as relational databases and data warehouses

The goal of the storage layer is to ensure that data is as close as possible to compute nodes, reducing the network impact of data transfer and node wait time.

BIG DATA ARCHITECTURE: PROCESSING

The processing layer provides query, analytics, and workflow support on top of the infrastructure and storage. Often, the processing layer will consist of one or more of the following solutions:

- Custom query/analytics code (e.g. MapReduce)
- Data pipeline processing frameworks that offer SQL or alternative scripting of queries/analytics
- Workflow frameworks that coordinate multiple jobs and data source integrations into a single workflow solution for one-time or scheduled execution

Each of these solutions require varying degrees of programming experience. For example, MapReduce often requires a more low-level understand of programming , while other tools may only require an understanding of the available data and basic SQL skills.

The processing layer is often where most of the exploration and analysis occurs, commonly generating data sets that are used by other processing jobs to perform further analysis.

BIG DATA ARCHITECTURE: INTERACTIVE ANALYTICS

While the processing layer is commonly managed by developers and data scientists, the interactive analytics layer exposes analytics data to end-users. While not a complete solution in itself, this layer offers one or more advantages:

- Adhoc queries of analytics data
- Visualization of analyzed data sets
- Export of data sets into alternative data stores (e.g. relational databases, data warehouses, etc)

This layer is optional and commonly made available to information systems users that are comfortable with manipulating raw data sets and familiar with query/analysis tools.

BIG DATA ARCHITECTURE: APPLICATIONS

For users that are not comfortable with interactive analytics, applications may be provided to visualize and view big data processing results. Commonly, applications contain:

- Information dashboards
- High-level exploration tools
- Visualizations, such as charts, graphs, and tables
- Summarized and detailed reports

Applications may be custom built for in-house consumption, purchased from commercial vendors, or offered in a multi-tenant, SaaS-based product.

UNDERSTANDING HADOOP

WHY HADOOP IS EVERYWHERE

WHAT IS HADOOP?

The Hadoop is a software framework that allows for the processing of large data sets across clusters of computers using simple programming models. It offers local computation and storage on each machine, detects and handles failures, and can scale to thousands of machines.

Hadoop is licensed as open source using a Apache v2 license. As a result, there is an ecosystem of additional projects that can be used to customize and extend Hadoop for a variety of situations.

Several commercial vendors have licensed Hadoop and offer a custom distribution, tailored to specific uses and/or hardware.

HOW HADOOP WORKS

Hadoop is composed of several components, including:

- HDFS, which stores the data, both pre and post-processed
- MapReduce, the engine that processes the data across multiple nodes

HDFS allows very large files to be split across server nodes, allowing the data to be both available and a manageable size for MapReduce processing.

MapReduce works by taking a big job and splitting it into smaller jobs. A good example of this is counting the words in a 600 page book. If it takes someone 10 hours to count the number of words in a book, it might take only 1 hour if the book were split evenly between 10 people. The “map” step is responsible for counting the words for each portion of the data, while the “reduce” step combines the results from each map step to determine the final result. Hadoop allows this process to be performed on data sets too large to be managed on a single machine, and often much faster as well.

BEYOND HADOOP

BIG DATA ALTERNATIVES

IN-MEMORY PROCESSING

Some big data processing can be built to operate strictly in memory, without the need for slower storage devices such as hard drives. This has the advantage of very fast data access and processing, but often requires more advanced programmer expertise than a simple MapReduce job. It also has hardware limitations due to the maximum amount of memory that a single server may have installed, as well as the cost limitations of the additional server hardware.

This approach is often used a first step to work out processing algorithms on smaller data sets, before the need to manage the hundreds or thousands of Hadoop nodes commonly required for extremely large data sets. It may also be used with larger compute clusters, when the speed of analysis outweighs the additional hardware cost.

ELASTIC MAP/REDUCE

EMR is a solution from Amazon that offers Hadoop-as-a-Service. Data is first loaded into their Simple Storage Service (S3). Then a job is defined and executed, with the results being stored back into S3.

Amazon EMR reduces the setup, configuration, and management of a custom Hadoop cluster and is a great alternative for quickly getting up-and-running with Hadoop. In addition, Amazon's Data Pipeline can be used to define, schedule, and execute workflows that involve EMR.

However, customization options are limited and many Hadoop extensions are not available to Amazon EMR. It is a great alternative when the jobs or the data set is limited but MapReduce is the preferred method for processing.

UNDERSTANDING NOSQL

IS SQL GOING AWAY?

WHAT IS NOSQL?

NoSQL is a relatively new term in software architecture. It generally means the use of databases that do not require or offer SQL data structures and queries.

While NoSQL is a recent term, the concept of using databases other than those adhering to the SQL standard isn't new. In fact, there have been a variety of database options available for some time, include LDAP directory services, object databases, and XML databases. These alternative data stores provide different performance capabilities and storage options to traditional SQL-based data stores.

NOSQL + BIG DATA

While many NoSQL databases offer different methods of data storage and retrieval, some were designed to address the issues commonly found with big data: large data sets, high read/write demands, and distributed processing support.

Vendors targeting big data architectures often integrate with frameworks such as Hadoop, while others offer drop-in replacements to the HDFS storage component of Hadoop.

Before you select a NoSQL vendor, determine how you plan to use it, including how it integrates with your big data architecture. Otherwise, the solution may just become another ETL import/export process with little-to-no added value.

THE API OBSESSION

HOW DO BIG DATA AND API GO TOGETHER?

THE RECENT API OBSESSION

There has been a good deal of discussion, both positive and negative, about the recent obsession of APIs. Some say that it is ushering the next wave of product companies, while others are asking, “What’s the big deal?”

To the non-technical, APIs might be a bit of mystery and magic – build an API, sell your company for lots of money. To the technical, it offers a way to build and share your work with other developers, making money along the way.

APIs themselves are nothing special. That’s right, I said it…and I’m not taking it back.

SO, WHAT IS AN API, ANYWAY?

An application programming interface (API) specifies how software components should interact with one another.

Traditionally APIs were used to build a standalone application. Now web APIs enable applications to talk to other applications across the Internet using the standards of the web.

If you have done any of the following, you have experienced an application that used APIs without you even knowing it:

- Published a tweet using a Twitter web or mobile interface
- Viewed the weather on your mobile phone or tablet
- Shared files between your desktop and a mobile device (e.g. Box, Dropbox, etc)

A LITTLE API HISTORY

You may not have realized it, but APIs have been around for a long time.

Before the current incarnation, the most popular way to distribute APIs were through SDKs. They were commonly used to integrate new functionality into applications, or to construct an entire application (e.g. the Windows Mobile or Android SDKs).

SDKs define APIs that developers build their code against for a specific programming language (e.g. Java, Ruby, or C/C++). They do this by providing a contract with the developer on what they will do, how to call the code, and what the result will be. The developer then builds code on top of the SDK to get the functionality they desire.

After the SDK came the initial attempt at web services. Many of these early web services used technologies such as SOAP, WSDL, and a number of other acronyms. These technologies required a deep understanding of their specifications before the web services could be used. Additionally, the high cost to develop and deploy these services prevented many companies from offering them in the early stages of the company.

WHAT CHANGED?

Simplicity

The newest web service APIs speak a simple language: HTTP, the language of the web. Combined with other common usage patterns such as REST and HATEOAS, web APIs promote easy integration from any programming language – no SDK required!

Lower Cost

In addition, the cost of infrastructure has decreased with the introduction of cloud computing, allowing for rapid provisioning of a data center with only a credit card. Compare that to just 10 years ago when hardware had to be purchased and installed in a co-location facility before the API was ever deployed.

New Business Models

Finally, the market that was previously accustomed to the SDK business model started to accept new ways of paying for access to web APIs. The pay-per-use or subscription models are the most common that we see today. This market shift has even opened an amazing 20 new API business models available to product companies today.

APIS, EVERYWHERE

The API market has been hot for several years and doesn't show signs of slowing down. In fact, more than 85% of Enterprises will have an API program by 2018 (Layer 7 survey). Here are just a few companies using APIs to their advantage:

Best Buy is using APIs to make their product data and reviews available to mobile applications. This helps their customers price compare and make more informed decisions at purchase time.

Amazon Web Services allows anyone with a credit card to design and provision their own data center. AWS APIs offer servers, databases, and data analytics engines to anyone that needs them.

Netflix gained significant market adoption by enabling any device that has Internet access to stream movies and shows. Now, anything that connects to a TV likely offers Netflix streaming.

API + BIG DATA = NEW OPPORTUNITY

Combining what we just learned about APIs with our new understanding of Big Data, we can now move our data from inside knowledge to revenue-generating services. How? There are several ways to accomplish this:

1. Creating private APIs to extend additional insight to your existing partners
2. Marketing your services to developers through public APIs
3. Generating new product revenue through commercial APIs
4. Increasing internal product value and intelligence with Big Data analytics services wrapped in APIs
5. Offering an API to your raw or analyzed datasets (Data-as-a-Service)

IN SUMMARY

FIVE BIG DATA FACTS

REVIEW: FIVE BIG DATA FACTS

1. Big Data Is Right-Time Business Insight and Decision Making At Extreme Scale
2. Big Data is here to stay and will be commonplace by 2018 (Gartner)
3. Big Data offers new business model and business transformations
4. Big Data architectures require a different approach than traditional IT
5. Big Data moves organizations from “Data is an Asset” to “Data is Revenue”

THE BIG DATA CHALLENGE

HOW DO I GET STARTED?

FIVE STEPS TO LAUNCHING A BIG DATA ARCHITECTURE

1. **Size the opportunities and threats.** Your opportunities may range from improving core operations to creating new lines of business
2. **Identify your internal data sources and gaps.** Knowing what you have, what you need, and how you plan to combine them is important to create insight
3. **Define insights and goals.** By determining your goals and the insights required, it helps provide a single-mindedness that will avoid the "analysis paralysis" common with big data projects
4. **Align organizational priorities.** Without organizational alignment, big data projects risk being pushed to the side or reprioritized by other concerns
5. **Create a hiring and execution plan.** The skills and execution required for big data is different than traditional information systems. Partner with a service provider to find the right skills and internal talent necessary to execute your big data project

LET'S TALK BIG DATA AND YOUR BUSINESS.

LaunchAny has a passion for leveraging proven and emerging technologies to launch the next generation of business software and services. We can help you to develop and implement your big data strategy.

Our services include:

- Big data architecture and infrastructure
- Cloud strategy and implementation
- Custom API and application development

Contact us to find out more about how we can help you launch your next project.

Telephone: +1.512.537.8493
Email: info@launchany.com
Web: launchany.com
Twitter: @launchany

ABOUT THE AUTHOR

James Higginbotham is the founder of LaunchAny, an Austin-based firm that specializes in launching software products. James has launched over 50 products in his career. He enjoys working with companies on their technology and product strategy. James has experience in architecting, developing, and deploying SaaS and APIs to the cloud.

James serves on the advisory board of a non-profit organization that focuses on helping people become advocates and activists for their charities and causes. When not writing or consulting, he enjoys landscape photography.

James can be reached at james@launchany.com