Modern **Business Intelligence:**

Leading the Way for Big Data Success

Sushil Thomas & Steve Wooledge

ARCADIA DATA

Modern **Business Intelligence:**

Leading the Way to Big Data Success

Sushil Thomas & Steve Wooledge

ARCADIA DATA

Arcadia Data Inc.
999 Baker Way, Suite 120
San Mateo, CA 94404

Visit us on the web: www.arcadiadata.com

The text, graphics and examples included herein are for the purpose of illustration and reference only. The specifications on which they are based are subject to change without notice. No legal or accounting advice is provided hereunder.

Arcadia Data reserves the right to revise or withdraw this document or any part thereof at any time.

ISBN: 978-0-692-94721-0

Contents

Introduction

What is the value of big data if the ability to analyze it and extract business insights is largely confined to data scientists? The answer is "very limited," as data scientists today are as scarce as they are expensive. But that is the reality facing most enterprises today. They are struggling to enable business users and analysts without coding skills to analyze big data, for many reasons.

IT leaders are rapidly arriving at the conclusion that they need a new, distributed approach to big data analytics. Leading industry analysts agree that data volume is growing at a phenomenal rate with "unstructured data" leading the way. It is this data where companies believe they will be able to differentiate from their rivals and gain a competitive edge. Legacy data repositories are ill-suited for this challenging task, which easily explains why big data has been embraced so strongly.

Other terms related to "unstructured data" include "semi-structured data" and "multi-structured" data. These terms have nuanced distinctions, but for the purposes of this paper, the term "unstructured data" is used as the umbrella term for all three types, as the key point is that big data largely deals with data that is not structured per the relational model.

Due to the escalating growth in unstructured data creation, many enterprises are realizing traditional approaches to data management are not enough. As a result, these organizations are exploring options such as looking to data scientists to complement the tasks traditionally assigned to business analysts.

This book will look closely at the emerging trends in big data and the pressing need for analytics solutions that emphasize more user-friendly approaches, such as more sophisticated visualization techniques. There are significant changes brewing that can potentially and irreversibly disrupt the traditional analytics landscape, delivering heretofore unprecedented business insights.

Chapter 1: A Brief Overview of the Big Data Ecosystem (Hadoop, Spark, and Beyond)

As mentioned in the introduction, big data offers the greatest opportunity for organizations of all sizes to truly distinguish themselves and forge real competitive advantage. For example, the relative success of one company's marketing program versus that of a competitor may boil down to which of the two does the best job of leveraging big data analytics on sentiment data scraped from social media feeds on Facebook or Twitter. That is very different from analyzing data warehouse data related to marketing campaigns conducted in the past. In short, big data can be used to glimpse the future of what consumers are likely to do, not just what they have already done.

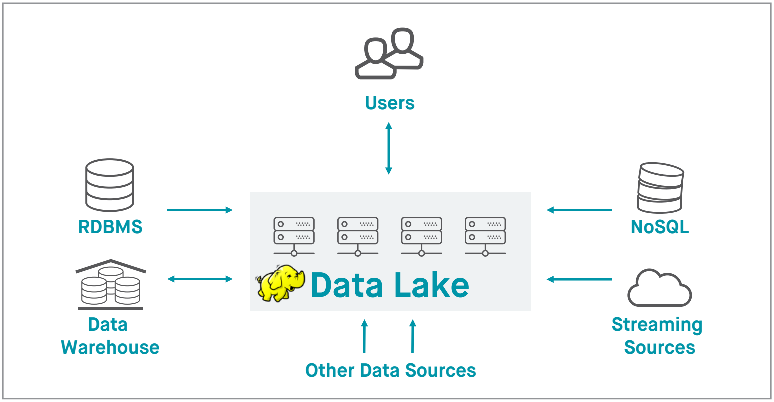
The key, as will be shown later in this book, is to deploy the analytics tools that enable the enterprise to explore and exploit big data.

The Big Data Ecosystem Starts with Apache Hadoop

According to Alexa Internet, a leading commercial web traffic and analytics company, as of March 2017, three of the most commonly visited websites in the United States are Amazon, Facebook, and LinkedIn. Despite their different missions, one thing that each of these organizations have in common besides their phenomenal success is that they operate and maintain some of the biggest Hadoop clusters in the world. Since publicly launching in 2006, this open source, Java-based framework has extended far beyond its open source / search engine roots to become the premier platform for aggregating, storing, and processing extremely large data sets in distributed environments using commodity hardware.

Ten years later, Hadoop adoption and expansion has continued at a dramatic pace. In fact, leading experts believe that Hadoop is predicted to grow[1] at a compound annual growth rate of 59% through 2020. This more than doubling every two years mirrors IDC's predictions[2] of growth in overall data volume. In fact, Forrester Research[3] predicted that eventually all enterprises will adopt and deploy Hadoop somewhere in their organization. A client survey by Gartner in September 2016[4] shows that 73% of organizations either have or plan to invest in big data, and that number increases to 86% for large enterprises.

Hadoop's core strength is capturing and storing multiple data types (unstructured, semi-structured, and structured) in almost limitless amounts, while offering a comprehensive framework for high-level big data analytics.



Apache Hadoop Has Emerged as the De Facto Standard for Data Lakes

Hadoop is an ultra-scalable platform that was designed to exploit the collective power of hundreds if not thousands of clustered computing nodes. Like other successful open source projects such as Linux, Hadoop has an active community and has attracted the attention of open source vendors as well as legacy players offering their monetization strategies.

The source of Hadoop's capability to store and process enormous data sets is a robust programming model that controls the commodity hardware-based computing nodes. Hadoop runs multiple data nodes and distributes workloads across a typical deployment using the MapReduce engine as one compute option. Hadoop and its ecosystem are highly fault-tolerant because of the built-in redundancy capabilities to prevent data loss should a node fail. Many other processing engines such as Apache Spark have increased in popularity for in-memory performance considerations, and can be resource managed using Apache YARN, adding tremendous flexibility to the big data ecosystem.

In with the New — and the Old, Too

The Hadoop ecosystem doesn't behave like a rogue set of technologies. It has embraced both SQL, the global standard for communicating with and querying traditional relational database management systems (RDBMS), as well as online analytical processing (OLAP), commonly used in multidimensional data analysis. In doing so, Hadoop attracted thousands of IT professionals trained in these traditional query languages.

As a result, the demand for these approaches spawned their own sub-ecosystem of open source projects and startups, as well as support by industry-leading software vendors such as Oracle and Microsoft. At last count, there are over 20 different SQL-on-Hadoop offerings while OLAP-on-Hadoop is quickly catching up.

Four of the more well-known SQL-on-Hadoop offerings include:

Apache Hive: Apache Hive is considered by the big data community as the first native SQL-on-Hadoop engine. It is mostly used in conjunction with traditional BI tools for batch data preparation as well as ETL (extract, transform, load) processes. Hive is supported by every major Hadoop distribution and has a very active open source community sponsored by the Apache Software Foundation (ASF).

Apache Impala: Originally conceived by Cloudera and then donated to the Apache Software Foundation community, Apache Impala is a SQL-on-Hadoop engine that runs directly on top of a Hadoop installation. While Hive conducts its operations in batches, Impala works in real time and shines in multi-user interactive BI and analytics operations.

Apache Drill: Apache Drill is recognized as the first distributed SQL query engine that incorporated a schema-free JSON (JavaScript Object Notation) object model. Drill defines its schema dynamically ("schema on read") as opposed to others which require a predefined one. It also operates with its own execution engine which includes in-memory processing for fast, interactive ad-hoc querying.

Spark SQL: Spark SQL is a key component of Apache Spark. Spark SQL introduced a data abstraction called DataFrames which offers support for both structured as well as semi-structured data. It provides a domain-specific language (DSL) to manipulate these DataFrames in Scala, Java, or Python.

By providing support for these traditional approaches, Hadoop truly offers both scalability and flexibility as well as the potential to empower business analysts and other non-IT users to reap the benefits commonly associated with big data analytics.

Make Way for Spark

The big data ecosystem is certainly not limited to Hadoop. As of 2016, the most popular open source project globally is Apache Spark.[5] Spark has captured the attention and imagination of data scientists, and increasingly, business analysts. It is an ultra-high speed general processing engine that is compatible with Hadoop and can access data sources including the Hadoop Distributed File System (HDFS) and Apache HBase.

Many industry analysts believe that long term, Spark may be a leading candidate if not the leader in providing the most user-friendly, fastest processing solution for advanced big data analytics.

According to the official Spark project page,[6] more than 1000 organizations use Spark in production environments most notably Amazon, eBay, and Pinterest. Many of these organizations run Spark on massive clusters covering thousands of nodes, to perform both ETL as well as data analyses on multi-petabyte data stores with ease.

While traditional MapReduce (and YARN) incorporates a disk-based data processing approach, Spark uses an in-memory application framework. This is one key reason why Spark advocates are able to claim it is 100 times faster than MapReduce.

Other reasons for its wide-scale adoption include:

Strong language support: Spark has various APIs for three of the most popular programming languages used by open source and commercial software developers (Java, Python, Scala), as well as for common data management languages such as SAS, SQL, and R, each of which are heavily used by data scientists and business analysts.

Multiple deployment options: Spark can be deployed on-premises with a storage engine such as HDFS, as well as via cloud (public, private, hybrid). Spark provides an interactive shell and can be operated in batch mode, so it can be used in almost any setting.

Advanced data operations: Spark not only offers the "map" and "reduce" functions that are inherent in the MapReduce framework, it also added support for operations commonly available with databases such as "filter," "join," and "group by." With such operations, a word count function can be written in Spark in only 4 lines of code versus 100 in MapReduce.

Graph data support: Spark natively provides the capability to handle graph data (i.e., sets of nodes/vertices/points connected by edges) via its GraphX service. When used in conjunction with data stored as rows and columns, this offers the benefit of quickly analyzing relationships between entities.

Whether the Hadoop ecosystem can continue to evolve to enable greater adoption and use by line-of-business managers and other non-IT staff remains to be seen, but the traction so far is very promising. For now, Spark's greatest appeal is to data scientists for whom the framework can offer considerable productivity gains in the development of big data analytics solutions.

Other Platforms: NoSQL, NewSQL, Object Stores

Other technologies in the big data ecosystem that work well alongside Hadoop, SQL-on-Hadoop, and Spark are worth mentioning here. These classes of technologies include NoSQL, NewSQL, and object stores. Each of these are intended to solve big data challenges with different approaches as noted below. These technologies are typically used with Hadoop/Spark to handle big data analytics, as none of these technologies were designed specifically for analytics to the level Hadoop and Spark were. Still, their complementary use with Hadoop/Spark make them an important part of any modern data architecture.

Arguably, the most significant technology is the NoSQL databases, which were originally created to overcome the limitations of RDBMSs when using big data. Instead of a well-defined, tabular data model as specified by the relational model in RDBMSs, NoSQL databases "denormalized" data that allowed you to put as much data as you wanted into a single record. This meant that if you wanted to save or retrieve information about a specific entity, such as a person or an invoice, you only needed to access a single database record. This modeling of data allowed other advantages such as a scale-out architecture that leveraged low-cost, commodity hardware, just like Hadoop. In conjunction, the performance was superior and much more cost-effective than RDBMSs because of the simplicity of the database reads and writes.

And while NoSQL databases often displaced RDBMSs for certain workloads, they are certainly not a drop-in replacement. The term NoSQL was originally interpreted as "'No' to SQL" for a short while until the industry acknowledged that NoSQL should not be viewed as a direct replacement to RDBMSs. The acronym morphed into "not only SQL" to suggest that NoSQL and relational databases could act as complements in a data center.

Some of the early technologies that helped to give rise to the NoSQL movement include MongoDB, CouchDB, Apache Cassandra, and Apache HBase, though many other NoSQL databases are available today. And the early origin of the term "NoSQL" is actually a little more complicated than described above, but one thing to note here is the irony of the name. While the pioneering NoSQL databases had no SQL interface, more and more NoSQL databases today are adopting SQL as the query language, making NoSQL an increasingly obsolete name. Just as it had helped with Hadoop, SQL promises to make NoSQL databases easier to adopt and deploy in production environments.

NewSQL represents the emerging class of technologies that leverage the key characteristics of RDBMSs but were architected to solve the scale limitations. Since traditional RDBMSs were designed to run on a single hardware server, handling scale for more data and more users typically meant upgrading your hardware. With NewSQL, you get all the benefits of RDBMSs while also gaining the benefit of scaling out on commodity hardware (again, like Hadoop), necessary for growing big data volumes. NewSQL databases are typically run in-memory to overcome the latency of disk accesses and of coordinating data sets across nodes in a cluster. Since the NewSQL databases are still in their early stages, they are less certain to impact the BI world in the near future as other emerging technologies. As such, much of the discussion in this book will not specifically call them out as data platforms for BI. Examples of NewSQL databases include MemSQL, VoltDB, and ClustrixDB.

Finally, object stores represent yet another approach to addressing big data challenges. These are simply low-cost places to store large volumes of data. Object stores are ideal for "objects" which in this context are large files. Each object/file is typically accessed via a standard URL. The advantages are purely about cost and convenience, as there's no compute layer that processes the data. These advantages have made object stores very popular for big data environments. Also, many websites take advantage of object stores because object stores are one of the most inexpensive ways to store data that needs to be accessed across the cloud. Many popular technologies including Spark can leverage object stores such as Amazon S3 as the means for storing big data.

References

1. http://www.businesswire.com/news/home/20160531005571/en/Global-Hadoop-Market-Growth-59.37-CAGR-2020

2. https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

3. http://www.cio.com/article/3024879/big-data/the-top-5-hadoop-distributions-according-to-forrester.html

4. http://www.gartner.com/newsroom/id/3466117

5. http://www.techrepublic.com/article/apache-spark-rises-to-become-most-active-open-source-project-in-big-data/

6. https://spark.apache.org/

Chapter 2: BI and Analytics Meet Business Transformation

One question frequently asked by BI vendors as well as industry experts is "What is the difference between BI and analytics?" In fact, when doing a Google search for "difference between BI and analytics,"[1] you will receive over 3 million results (as of the time of the writing of this book).

BI and analytics continue to be important capabilities in organizations around the world. In fact, BI is enjoying a rebirth, courtesy of the data explosion brought to the enterprise by big data, social media, the Internet of Things (IoT), and other sources. Because of this fact, it is definitely time well spent to review: 1) What the difference is between BI and analytics, and 2) How BI has changed over the years.

What is the Difference between BI and Analytics?

From a high level, BI and analytics both have the same purpose: both help organizations tap into their data to improve decision making. However, how they actually achieve that goal is quite different.

Traditionally, BI leverages historical data to learn from past decisions, while analytics draw on different sources to predict future results. In other words, BI answers questions like, "what happened with…," "when…," "who…," and even "how many…," while analytics answers questions such as "what if…" and "what's next…" Both apply to static data, while analytics often includes time-series data, which is a series of data points captured at specified periods over time. Examples of time-series data includes up-to-the-second stock prices, measurements from equipment sensors, and GPS readings in a navigation device.

	Business Intelligence	**Analytics**
Focus	past, present	present, future
Typical users	line of business (LOB), IT	LOB, IT, data science, business analysts
Questions answered	"What happened?" "When?" "Who?" "How many?"	"What if?" "What will happen?" "What's next?"
Deliverables	ad-hoc query alert canned report custom report data visualization dashboard balanced scorecard data application	ad-hoc query classification algorithm regression model segmentation model dashboard canned report custom report
Methods	roll up slice and dice filter sort	descriptive modeling data mining text mining multimedia mining statistical / quantitative analysis

Comparison of BI and Analytics

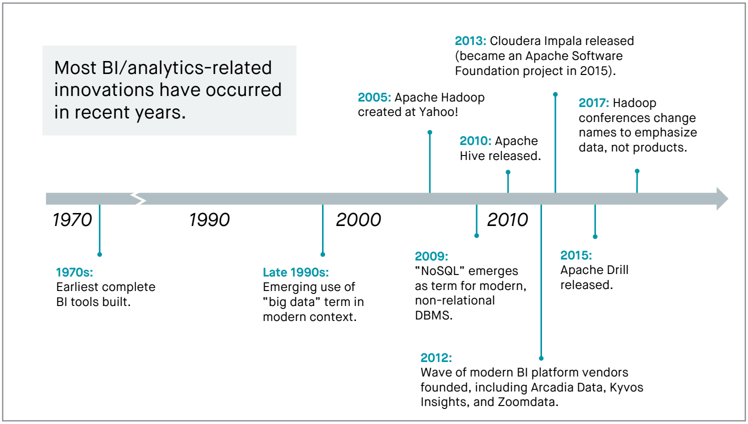
Further, the methods each take are quite different. BI products provide users with tools to perform ad-hoc queries as well as create reports, dashboards, and scorecards as well as offer APIs to embed these objects into or even build applications. Some also provide monitoring and alert capabilities, notifying essential personnel if a threshold has been reached. On the other hand, analytic applications incorporate various mining and modeling frameworks as well as provide support for statistical and other quantitative analyses.

Since organizations often were interested in gaining the benefits associated with BI and analytics, they often had to buy multiple applications, often from different vendors, requiring time, money, and resources to maintain.

With the increased amount of unstructured data being created from various sources, organizations have begun to rethink about how to store, analyze, and exploit this information, whether it's stored on-premise or even on a private, public, or even hybrid cloud.

A Brief History of BI

It may come as a surprise, but the term "business intelligence" was first used in 1865 by Richard Miller Devens in his book Cyclopedia of Commercial and Business Anecdotes. However, it was nearly a century later when it was first adopted by the technical community when IBM researcher Peter Luhn published "A Business Intelligence System" in 1958. Even then, it took until the 1980s for BI to gain visibility. This is when Bill Inmon and Ralph Kimball conceived the first "data warehouse" which brought data from multiple sources, storing the results in a central "system of record" (in their case, a mainframe computer).


Most BI/analytics-related innovations have occurred in recent years.
1970
1990
2000
2010
1970s: Earliest complete BI tools built.
Late 1990s: Emerging use of "big data" term in modern context.
2005: Apache Hadoop created at Yahoo!
2009: "NoSQL" emerges as term for modern, non-relational DBMS.
2010: Apache Hive released.
2012: Wave of modern BI platform vendors founded, including Arcadia Data, Kyvos Insights, and Zoomdata.
2013: Cloudera Impala released (became an Apache Software Foundation project in 2015).
2015: Apache Drill released.
2017: Hadoop conferences change names to emphasize data, not products.

Timeline of Recent BI/Analytics Events

Initial data warehouses were tightly managed by corporate management information systems (MIS) departments. They controlled the entire process from end to end—creating, scheduling, and distributing "canned" reports to company's executives and other line-of-business counterparts. When an individual or group required information that was not provided in those prepackaged documents, they submitted formal requests which took between days, weeks, and in some cases months to assemble. As time progressed, demand for these ad-hoc reports multiplied in number, resulting in increased frustration and backlogs.

By the early 1990s, companies interested in cutting costs started to migrate their data management operations from mainframes to “client-server” environments using relational databases to house their online transactional processing (OLTP) while introducing the world to online analytical processing (OLAP). As part of this platform shift, the first generation of independent BI companies emerged to simplify report development and distribution. Enterprises often purchased large quantities of these products, but the vast majority of reports were still developed and maintained by MIS, with pockets of non-technical “power users” scattered across their ranks for good measure.

At the same time, use of computer-aided manufacturing (CAM) applications grew, including material resource planning (MRP), supply chain management (SCM), and product data management (PDM) systems. Computer-aided design/computer-aided drafting (CAD) products forever changed how people managed development of new “things.” These “things” eventually would gain “smart” capabilities as well as the ability to connect with each other, which would be referred to as the “Internet of Things.”

While enterprise resource planning (ERP) products were initially developed in the 1980s, they became more popular around this time due to the desire for companies to cut costs through using less expensive UNIX servers instead of mainframes and minicomputers. Further, traditionally non-technical departments entered the information age through the introduction of sales force automation (SFA) and customer relationship management (CRM) applications, transforming operations for their respective organizations as well.

Even though the target audiences for each of these systems were very different, one key detail all of them had in common was rudimentary (or in some cases, non-existent) reporting capabilities. This spawned the beginnings of the "analytical applications" market with both incumbent BI players as well as startups racing to reverse engineer these systems to address this glaring oversight.

In the mid to late 1990s, companies began to transform the Internet, which previously was mostly limited to government, academia, and increasingly, high tech companies, into a communications channel as well as a source for generating revenue. BI embraced this new approach, offering "zero footprint" BI clients that were accessible by a desktop web browser. When used in conjunction with "analytical applications," this helped set the stage for the advent of the enterprise performance management (EPM) market.

In the early 2000s, companies that were unable to invest in the human, hardware, and software capital associated with deploying traditional enterprise resource planning (ERP) and CRM systems on their own, finally were given a way to do so with the introduction of application service providers (ASP). Companies like Asera, Corio, and others offered their customers the ability to share a common installation. BI once again followed, which laid the seeds for the "Self-Service BI" movement.

Today there is an increasing number of tools claiming to offer close to real-time analytics for business analysts of all stripes—not just data scientists. But as we will see in later chapters, not all these offerings are created equal.

WITHER THE RDBMS? NOT SO FAST...

Where do these latest developments leave the traditional BI world of structured data-driven data warehousing? After all, most of these traditional approaches are relatively expensive. Their ability to discover new patterns and business insights is highly limited to the subset of "normalized" data that has been painstakingly extracted and placed into a data warehouse or mart. Yet there is a rich and broad-based set of tools to very quickly analyze that information— and a great deal of legacy IT expertise to support that environment.

It turns out that most enterprises today are not debating if data warehouses, enterprise data management systems, or even (big) "data lakes" will emerge as their ultimate data repository. They will all co-exist, and will continue doing so for several years to come. As technologies like Hadoop/Spark, NoSQL, NewSQL, and object stores are all complementary with the traditional technologies, organizations should expect to use some combination of these technologies together.

Since the vast majority of new data being created is recognized as "unstructured," it is highly probable that use of traditional RDBMS systems will wane over time as the use of Hadoop/Spark and the other big data technologies will increase. After all, these big data technologies were designed to elastically accommodate data growth.

THE PRESENT AND FUTURE OF ENTERPRISE REPORTING

Traditional data warehouses and enterprise data hubs have little (if any) support for streaming or unstructured data in their native format unless IT endures the pain and expense of preparing that data for the structured warehouse. Organizations are also finding it desirable to move some little-used warehouse data over to Hadoop where it can be stored more economically.

So the bottom line is that both environments—Hadoop and RDBMS/enterprise data warehouses—each have unique, prized attributes. RDBMs are the past and present. Big data analytics on Hadoop represent the present and future, especially due to the massive growth of valuable data that businesses collect. As Forrester VP and Principal Analyst Boris Evelson summed it up,[2] "Most BI applications are smart and only request aggregate, not detailed, result sets, minimizing network traffic between the app and database servers. But as big data volumes increase and enterprises mature their data mining and exploration applications, there'll be an increasing requirement to analyze data at a detailed level, putting strain on network bandwidth." Big data technologies like Hadoop were designed to handle this growing requirement.

SELF-SERVICE BI

Another important characteristic of this new BI era is the concept of self-service BI—essentially a take on analytics allowing business users to analyze mission-critical enterprise data with minimal or no IT intervention. According to Gartner,[3] by 2017, most business users will actually have access to the self-service tools they need to do this. Gartner also[4] asserts that by 2020, self-service BI will comprise a full 80% of all enterprise reporting.

It is important to note that self-service BI does not take humans out of the business decision equation. The specialized knowledge and insights business users have developed will be integral in solving business problems for years to come. Also, self-service BI should not be confused with self-sufficient BI. IT or some other entity still has to provide trusted data to be analyzed, so data quality remains pre-eminent. For organizations that did not purchase a cloud-based BI tool, the BI systems themselves must be maintained and updated when necessary. Still, the importance of self-service BI cannot be overstated, given the predictions of very substantial growth in this critical area of enterprise reporting.

HADOOP ANALYTICS CASE STUDY

Consider the case of a company known as a leader in information technology. This large vendor has a division that produces and sells efficient application-integrated data storage solutions. A large volume of data gets created by their solutions in customer deployments, representing a wealth of information that could be mined to assist customer support, product design, and even marketing activities. This environment is a classic IoT analytics environment where large streams of data from many remote sources are continually analyzed to benefit both the company and the customer base. For example, quick identification of a failed drive can enable rapid support, providing a positive customer experience, not to mention the avoidance of customer data loss. Also, an analysis of which product features are most used and most underused can help the company provide customers with guidance on how to get the most value out of their deployment.

The company was collecting hundreds of millions of data points in half a million files per day from tens of thousands of servers. The data collection, however, was not the real problem. The company used Hadoop as the data platform as a means to cost-effectively scale for the growing volumes of data. While this division was new to Hadoop, the challenge was to offer the data to end users in a way that was granular, comprehensive, consistent, and quickly accessible across business teams. When they implemented an in-cluster analytics solution that was architected for Hadoop, they were able to identify potential sales opportunities, underutilized and poorly provisioned systems, historical and real-time details on equipment reliability, customer usage patterns, and more. Users were able to create interactive data applications with no coding required, allowing a powerful self-service BI environment that enabled agility and collaboration across teams.

References

1. https://www.google.com/search?q=difference+between+bi+and+analytics

2. https://www.arcadiadata.com/blog/forrester-gives-bi-a-hadoop-shot-in-the-arm-for-big-data-scalability/

3. http://www.informationweek.com/big-data/big-data-analytics/11-tips-for-successful-self-service-bi-and-analytics/d/d-id/1324728

4. https://upside.tdwi.org/Articles/2016/02/17/Age-of-Self-Service.aspx?Page=3

Chapter 3: Rise of the Citizen Data Scientist

The concept of "citizen journalism" refers to common citizens "playing an role in collecting, reporting, analyzing, and disseminating news and information."[1] While citizen journalism in the United States has been around nearly as long as the country itself, it has become more commonplace substantially since the late 1980s. Citizen journalism is not unique to the United States; individuals and groups around the world have embraced the concept as well, such as during the 2010 Haiti earthquake, the Arab Spring, the 2013 protests in Turkey, and more recent events such as the Euromaidan events in Ukraine and the Syrian Civil War.

According to Terry Flew, Professor of Media and Communication at the Queensland University of Technology in Brisbane, Australia, there were three key elements which led to the rise in this viral information sharing approach:

- Open publishing
- Collaborative editing
- Distributed (online) content

All three elements are possible due to technological advances that simplified the business of journalism to all, even those who are not trained in journalistic practices. The use of the term "citizen" as an adjective to describe individuals empowered by technology also applies in "citizen data scientist."

In 2015, Gartner coined the aforementioned term, characterizing such an individual as "a person who creates or generates models that leverage predictive or prescriptive analytics but whose primary job function is outside of the field of statistics and analytics." While the challenges of these armchair data analysts are different than their commentator counterparts, their primary objective is virtually the same. Some individuals discourage the use of the term "citizen data scientist"[2] as in many cases it simply describes the work a business analyst or "power user" of a BI tool. However, we'll use it for purposes of this book as a good way to describe analytics beyond what a casual business user might want to do, namely leverage advanced analytical processing with a simple visual interface.

The broader story is about making data more valuable to more users. In the emerging world of big data analytics, "if the right BI-user applications can be built, this will empower a new generation of business data consumers, much broader than just the technical specialist pool of data scientists, DBAs, and analysts," contends Nik Rouda, senior analyst at ESG.[3] "Opening up access to insights in Hadoop would trigger a virtuous cycle of data utilization. As more users draw more value, that additional value would draw more innovation in the ways that Hadoop is leveraged across the business."

The Imperative of User-Friendly Analytics

The bottom line is that big data analytics solutions can create significant business value by delivering relevant insights to the people best equipped to act on them for the good of the enterprise—and those people are business people. The success of digital natives such as Uber and Lyft transforming transportation, Amazon and eBay redefining retail, and YouTube and Hulu enabling "cable cutters" is closely linked to their ability to analyze data better than their competitors. Even companies in traditional industries definitely see how big data is truly accelerating business transformation.

The challenge, of course, is putting in place the right platform and the right tools to allow increasingly more business analysts to undertake big data analytics projects. Gartner recommends starting by “facilitating ingestion, preparation, and analysis of complex data currently beyond the reach of business information analysts.” Next, organizations need to “increase the range of analytics capabilities available to users by deploying tools” for data discovery, self-service data preparation, and behavioral analytics.

While data warehouses have been around since the 1980s, mass adoption has been primarily limited to large enterprises for the most part, due to their costs. Traditionally, data warehouses were centrally managed “on premises,” with storage supplied by a storage area network (SAN) or network-attached storage (NAS) devices. As the number of data warehouse consumers grew, the amount of system resources including storage, memory, and network bandwidth increased proportionally. Since data warehouses were originally intended to be repositories for structured data, maintaining and scaling such an environment proved to be costly from a Capex, Opex, or even a labor perspective.

Hadoop as the Platform Game-Changer

Hadoop has radically altered these platform economics by leveraging inexpensive commodity hardware. That is, storage no longer needs to be centralized—it can be allocated to the low-cost nodes that also handle processing.

A key element of Hadoop is its distributed processing manager. By allowing the different nodes on a Hadoop cluster to handle processing of its own stored data, Hadoop greatly minimizes the movement of data, which minimizes latency associated with data movements. This setup is known as "data locality," in which the processing work is done where the data resides, versus moving the data to designated processing nodes. More importantly, the technique of distributing work across many nodes for parallel processing leads to significant throughput gains. The end result is performance that approaches that of traditional data warehouses, but at a fraction of the cost.

Finally, Hadoop addresses the relative unreliability of distributed clusters of commodity hardware by storing (replicating) three copies of all data, with each of these copies being distributed across nodes. Should one node fail, its data is still available on two other nodes. Despite the replication of data, the use of commodity hardware still allows lower overall costs as compared to traditional configurations that rely on high-end servers.

Thus the platform piece of the big data analytics equation rests on a solid foundation of Hadoop clusters. But the primary tools for analysts on Hadoop remain in the hands primarily of highly technical specialists such as data scientists who are comfortable with procedural languages, R, SAS, and Spark. Declarative approaches with SQL and SQL-like processing engines are possible, but are not yet mature enough for complex, machine-generated SQL from mature BI tools. This means that data analysts must be willing to get their hands dirty with writing SQL. The more casual end user who prefers an intuitive graphical user interface (GUI) and data visualization cannot rely on traditional tools to get true self-service access and analysis directly against big data platforms.

Data Visualization Comes to the Fore

The relationship between clear data visualization and subsequently communicating analyses and insights about that data is obvious. It’s simply easier for most people to intrinsically understand complex relationships visually as opposed to when they are presented as rows and columns and tables filled with text and numbers.

How important is data visualization when it comes to expanding big data analytics use beyond statisticians and data scientists to business analysts? In 2016 Gartner made significant changes[4] to its vaunted Magic Quadrant for Business Intelligence and Analytics Platforms. Gartner predicated its changes on the belief that enterprise analytics has evolved today to the point of being both more business-centric and more user friendly. Most organizations have incorporated a bimodal IT approach—simultaneously emphasizing safety and accuracy (via "traditional and sequential" approaches) as well as agility and speed (through more "exploratory and nonlinear" models).

However, as time progresses, companies will replace legacy BI products with more sophisticated yet more user-friendly tools. These "modern BI platforms," such as those providing advanced visualization capabilities, will not only support sophisticated big data analytics, they also will not require the intervention or oversight from IT. This level of self-service provides organizations with the agility to discover new insights from data in a faster, more iterative approach that allows hypothesis testing. Compare this to the traditional BI data flows where data is carefully prepared and organized to answer specific, known business questions based on requirements of the business in a much more centralized and governed approach.

In essence, data visualization tools allow business analysts to literally “see” the reasoning behind the big data analyses and discover new insights more quickly. As a result, it should not be a surprise that Gartner and other industry analyst firms maintain that data visualization is becoming a “must-have” for quickly communicating insights gained from big data analytics and converting those insights into actionable business decisions.

References

1. Bowman, S. and Willis C. "We Media: How Audiences are Shaping the Future of News and Information" 2003, The Media Center at the American Press Institute

2. http://www.kdnuggets.com/2016/03/mirage-citizen-data-scientist.html

3. http://go.arcadiadata.com/Accelerating-Business-Insights-DZ.html?utm_source=Website&utm_medium=Referral&utm_campaign=ebook

4. http://www.informationweek.com/big-data/software-platforms/gartner-bi-magic-quadrant-inflection-point-has-arrived/d/d-id/1324233

Chapter 4: Democratizing Big Data

As mentioned in the introduction of this book, Gartner and industry experts maintain there are far too few data scientists to meet the current demand, and those that are available are expensive. This sets the stage for the emergence of "citizen data scientists" leveraging powerful big data analytics tools. This trend is part of the larger movement many call "democratizing big data"—the enabling of more and more business users to quickly access the data they need to perform analyses and enterprise reporting of their own making, independent of IT.

However, for organizations to move closer towards true "self service BI" their data must be freed from their traditional silos. Decision makers, influencers, and even observers need a unified view of such data, and getting that has traditionally involved a number of highly labor- and therefore cost-intensive processes. One potential solution that organizations have explored is dumping data into a single, large "data lake" that holds tidal volumes of raw big data in its native format until requested.

Leave the Data Where It Is: A Case Study of Data Democratization in Action

To truly "democratize" data, it is essential to leave it where it naturally resides, such as in the Hadoop platform. Then the only thing what would be needed is an intelligent layer above these many and varied data types that can transparently integrate all the data. Ideally a visualization tool, this top-down "smart" layer will provide everyone what they need—namely, a unified view of all data regardless of its source.

The democratization of big data and the ensuing benefits is illustrated in a fast-growing marketing analytics technology provider to major brands. Acquired by Neustar in late 2015, MarketShare[1] helps marketers make better decisions more rapidly, offering both decision analytics and prescriptive recommendations to help clients optimize marketing spending. At the heart of MarketShare's value proposition are data-driven recommendations leveraging big data analysis.

After moving from MySQL to Hadoop, Neustar soon realized it wanted to provide dynamic rather than predefined reporting capabilities supplied by tools that more of their business analysts could leverage. For MarketShare, the set of "small data" tools it was using just didn't work on big data. It would take MarketShare a full day and a half to develop customer-specific data sets then transform and load them into an Oracle database. Analysts then had to manually produce one-off reports, and then embed them into a cloud application for analysis—another day and a half process. The result was static, predefined reports instead of the highly dynamic type of reporting MarketShare wanted and needed.

The solution was a native visual analytics and BI platform for big data, which gave business analysts the ability to drill down into the raw data details on individual customer interactions. Overall, this solution eliminated labor- and time-consuming data extraction and data movement, allowing analysts to point directly to data stored in highly elastic cloud platforms like Amazon S3 for very fast ad-hoc visualizations. Business analysts now could create sophisticated reports on the fly by simply selecting client-specific parameters. This is vastly different from what MarketShare did previously when analysts waited and waited for data to be moved into the relational DBMS. As a result, reporting time and effort has been slashed from two full-time equivalents for three days all the way down to one full-time equivalent for a half-day.

Democratization To-Do List

The MarketShare experience also outlines the value proposition of an analytics-as-a-service solution, which can deliver very timely, relevant and insightful big data analytics but without the heavy costs of investing up front in infrastructure. Essentially all that is needed is big data— something most all organizations have plenty of.

Thus, in the interest of fostering this democratization of data and realizing the fuller potential of big data as a competitive tool used by non-IT business professionals, organizations should consider the following:

- **Deploy** solutions proven to enable access to new sources of data and new kinds of data as well, most notably big data in all its many forms and from its various sources of origin.
- **Expand** and leverage new analytics capabilities, in particular ones that can uncover new insights such as those derived from predictive and prescriptive analytics. Traditional BI tools are good at deriving insights from events that have already occurred, such as transaction data. Predictive analytics deliver insights into events and behaviors that haven't

even occurred yet, giving organizations an opportunity to respond a priori.

- **Push** relentlessly to expand the pool of business analysts exploiting and leveraging big data through advanced analytics tools, such as data visualization. A sort of "multiplier effect" in this regard can result in newfound business value from the one thing organizations have plenty of, and that is data.

References

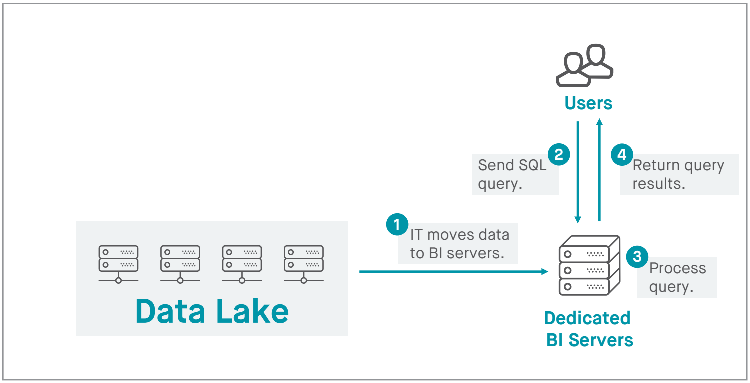
1. https://www.arcadiadata.com/lp/faster-big-data-insights-case-study-constellation-research/

Chapter 5: Common Approaches to Big Data Analytics

Let's look at four popular approaches to business intelligence architecture incorporating big data and the strengths and weaknesses of each.

The Dedicated BI Server (a.k.a. "Traditional BI")

This architecture is common to legacy BI platforms such as SAP BusinessObjects, IBM Cognos, OBIEE, MicroStrategy, and more recently, Qlikview, and Tableau. Traditional BI employs a dedicated middle tier BI server with connectors to back-end data sources. Users then access data via a local desktop application or web browser that is primarily fueled with data from the BI server. In terms of the data architecture, at the end of a long and complex ETL process, the most granular data is typically stored in the data warehouse, then aggregated (usually for performance reasons) and stored in data marts. The BI server will then query data from the data marts or the data warehouse directly, and cache results locally (often in memory) for consumption by the end-user clients as needed.



Dedicated BI Server ("Traditional BI") Architecture and Process Flow

PROS

- **Incumbent platform.** Most organizations already have made BI investments and have the resources in-house to support and maintain them. Barriers to adoption are typically low as a result. Often, BI tools are used to simply extract or embed results into other desktop tools like Excel or PowerPoint for convenience and easy sharing with more across the enterprise which again leverages existing investments and reduces user training needs.

- **Predictable environment.** Since ETL systems will produce clean (albeit limited) data sets, BI systems are designed to efficiently answer questions that their user communities have defined in advance.
- **Semantics makes "sense."** Since traditional BI clients (client server or web-based) provide access via a predefined semantic layer that puts in place formal business rules and metadata definitions, it facilitates common understanding across the enterprise.
- **Solid performance.** Performance is usually good on the desktop client and/or BI server, assuming the query results fit nicely within the physical resources of the desktop or BI server hardware.

CONS

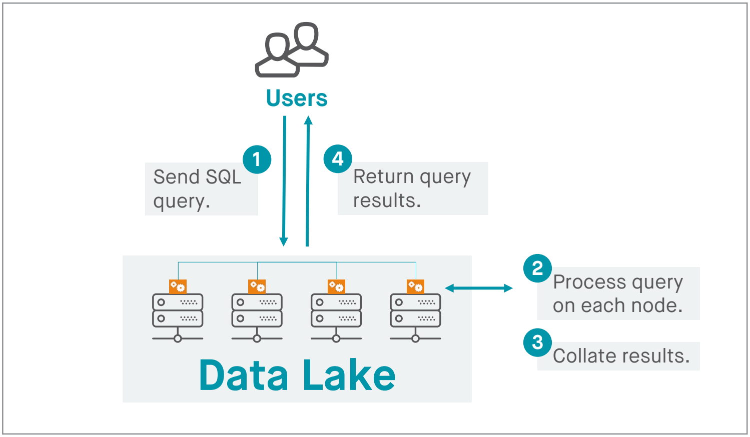
- **Significant scaling costs.** An increased number of users means a significant increase in costs from a hardware, software, and administrative perspective. Since traditional BI architectures required dedicated resources, these updated architectures often complicate management as well as introduce new potential security and administration weaknesses because they are additive to the security within the modern big data platform itself.

- **Tradeoff between data granularity and performance.** Traditional BI architectures cause organizations to make frequent tradeoffs between the high fidelity data and end-user performance. Why? Because they are scale-up, SMP servers which can be clustered, at best, but are unable to handle the volume and complexity (i.e., semi-structured, schema-less data) available in massively parallel, distributed data platforms. As discussed above, these data volumes cause SQL queries to run in minutes or hours, so organizations typically create extracts and roll-ups (i.e., a cube or data mart) on top of the original data to enable fast BI performance and keep business users happy. This aggregation causes valuable granular detail to be lost along the way, which limits the value of the analytics and the ability to innovate. For example, a retail organization wanting to report on sales transactions, for performance reasons might decide to aggregate the underlying data to exclude the individual customer detail. The impact of this aggregation would remove the ability of an analyst to later join in demographic or social media data collected on individual customers from other sources.
- **Architectural complexity.** Since the typical BI configuration is to add separate BI servers for extracts, more distinct data silos are added to the overall data architecture. This

leads to administrative, maintenance, and security complexity with more chance for error.

SQL-on-Hadoop Engines with BI Tools

The emergence of SQL-on-Hadoop offerings such as the ones mentioned in chapter 1 are often used to enable existing BI tools. These tools reduce (or in some cases eliminate) the need to extract the data into a data warehouse, mart, or cube before it can be analyzed. Running SQL queries directly on a data lake using Hadoop or cloud platforms such as Amazon S3 provides easier access to the most fine-grained data available.



SQL-on-Hadoop Architecture and Process Flow

PROS

- **No dedicated servers required.** Analyzing data directly in the big data platform leads to significant advantages. First, with no data movement from the platform to dedicated BI servers, the overall architecture is simpler and easier to maintain. Second, no external BI servers or external ETL software means lower hardware/administrative costs. Third, all raw data is immediately available to allow details on fine-grained data, unlike the summaries and aggregations used in dedicated BI servers. Finally, governance and compliance frameworks are easier to support by not creating separate copies of data in external repositories.
- **Unified security.** Related to the above, since there is no data movement, data can be secured by the platform's security controls, thus simplifying data protection and lowering the costs of securing your data. In the case of Hadoop, integration with security technologies like Apache Sentry and Apache Ranger enable the unified security model.
- **Lower learning curve.** Both "citizen data scientists" and RDBMS power users take advantage of tools and skills they already have. This leads to easier adoption since users don't have to learn unfamiliar new tools.
- **Maximize current investments.** Companies can leverage existing BI tools, skills, and training, thus avoiding additional expenditures on new BI tools.

- **Self-service.** With minimal IT intervention to run news types of queries, SQL-on-Hadoop can effectively provide self-service analytics that give much more flexibility to business users.
- **Reduced dependence on ETL.** Since many SQL-on-Hadoop tools offer ETL functionality natively, the need for more sophisticated (and costly) ETL tools can be reduced as well.
- **Cost-effective scalability.** By deploying analytics that leverages a big data architecture, you can easily cost-effective scale-out by incrementally adding more commodity nodes to a cluster.

CONS

- **Less mature SQL support.** While these SQL-on-Hadoop engines are intended to be used with popular BI tools, they tend not to support the SQL syntax as extensively as other veteran technologies. This limits the types of queries that can be run from your BI tool.
- **Limited track record.** While some tools suggest massive performance gains, these results are often performed under ideal circumstances. It is important to validate their claims in your environment first.
- **Concurrency limits.** Since consumers have grown to expect real-time or near-real-time results while performing

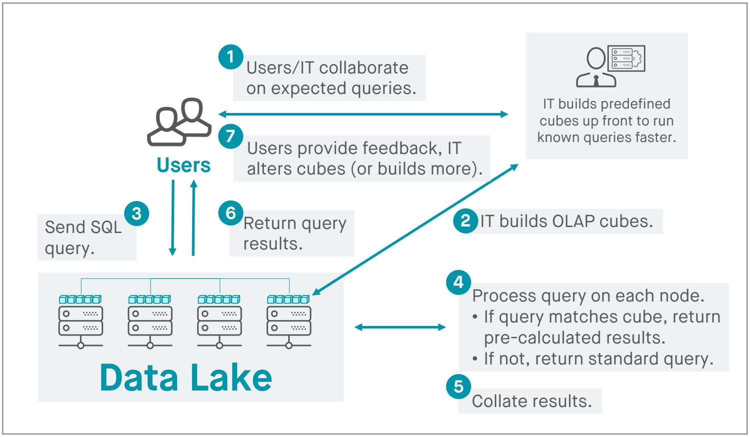
their analyses or accessing data via their dashboards, these products may not provide that due to their need to process SQL commands across a widely distributed cluster. This is especially true when a typical BI deployment has hundreds, if not thousands of concurrent users.

- **New skills and unfamiliarity with Hadoop.** The concept of accessing large amounts of structured and unstructured data for analysis is new, and will require some ramp-up time for users and IT organizations alike. Basic familiarity with Hadoop or cloud system skills are needed set up and maintain the data store.

OLAP on Big Data

An alternate approach embracing a modern data platform, OLAP (online analytical processing) concepts started to creep their way into the world of big data analytics in the early 2010s because of the scale and performance constraints of traditional BI approaches. The concept of OLAP involves predefining materialized views or virtual data "cubes" (which actually are pre-summarized aggregations of the underlying data) and directing queries to the appropriate level in the cube to return results much faster than the underlying granular data.

Many organizations experiencing frustration with the less than interactive response times of SQL-on-Hadoop engines (see previous section) are looking to close the performance gap by deploying such big data OLAP solutions in between their BI tools and their data in Hadoop. Modern big data OLAP solutions avoid the movement of data and achieve scale by deploying their cubes directly into the Hadoop environment, right next to the granular data from which they are summarized. Example big data OLAP technologies include Apache Kylin, AtScale, Kyvos Insights, and Zoomdata.


1 Users/IT collaborate on expected queries.
IT builds predefined cubes up front to run known queries faster.
Users
7 Users provide feedback, IT alters cubes (or builds more).
3 Send SQL query.
6 Return query results.
2 IT builds OLAP cubes.
Data Lake
4 Process query on each node.
• If query matches cube, return pre-calculated results.
• If not, return standard query.
5 Collate results.

OLAP on Big Data Architecture and Process Flow

PROS

- **No dedicated servers required.** Analyzing data directly in the big data platform leads to significant advantages. First, with no data movement from the platform to dedicated BI servers, the overall architecture is simpler and easier to maintain. Second, no external BI servers or external ETL means lower hardware/administrative costs. Third, all raw data is available to allow details on fine-grained data, unlike the summaries and aggregations used in dedicated BI servers. Finally, governance and compliance frameworks are easier to support by not creating separate copies of data in external repositories.

- **Maximize current investments.** As with SQL-on-Hadoop offerings, companies can leverage existing BI tools which can visualize the output from cubes, providing a faster path to big data adoption with a reduced learning curve.
- **Fast queries and high user concurrency.** Because aggregates are predefined, stored and accessed, sub-second response times are achieved for predefined queries.
- **Cost-effective scalability.** By deploying analytics that leverages a big data architecture, you can easily achieve cost-effective scale-out by incrementally adding more commodity nodes to a cluster.

CONS

- **Requires up-front modeling.** Cubes require significant investment from IT and other technical staff in terms of time to design, deploy, and administer, which makes them costly to build and maintain before queries and exploration can begin. This requirement often adds significant latency, inhibiting an immediate and real-time analytical environment.

- **Ongoing assembly required.** Since big data OLAP products lack the ability to automatically develop new cubes when new data are added, time and effort is required to develop, deploy, and test these enhancements prior to deployment. These updates can take weeks or months to complete, potentially losing opportunities for revenue in the process.
- **Not real-time.** Batch data updates are required for cubes, which typically happen once per day and can take hours to updated based on their size.
- **Lacks ad-hoc freedom.** The tradeoff for scale and performance is limited flexibility. Users are confined to OLAP views and data cubes that are prepared by IT in advance, so they have little latitude to experiment. If the data is not in the cube, the BI tool needs to bypass the cube so all the benefits of pre-aggregated performance are lost.
- **Increased administration.** Due to the need to develop multiple microcubes, many of these providers require separate security for these structures in addition to what they have currently. While this may be manageable initially,

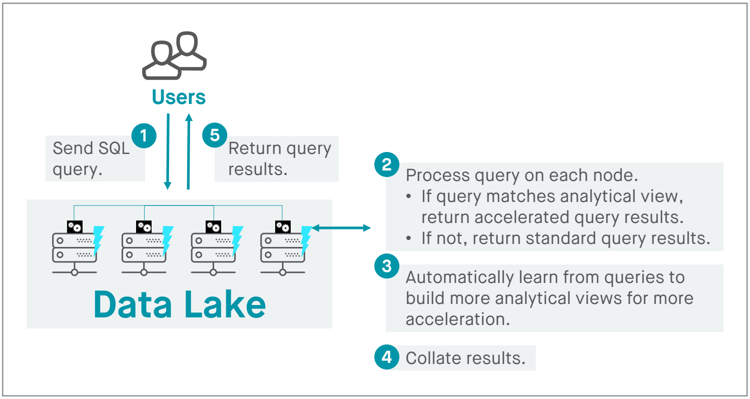
as the number of cubes and users grow, this initially simple task can become very unwieldy extremely quickly.

- **New skills and unfamiliarity with Hadoop.** The concept of accessing large amounts of structured and unstructured data for analysis is new, and will require some ramp-up for users and IT organizations alike. Basic familiarity with Hadoop or cloud system skills are needed set up and maintain the data store.

At the end of the day, big data OLAP cubes offer a stop-gap solution that exists only to enable legacy BI technology, which was never designed to work natively with big data, to achieve fast query performance for certain predictable questions. They do not remove the cost or the complexity of the legacy BI architecture. In fact, they increase it by adding yet another middle layer into the equation.

Native Visual Analytics and BI for Big Data

This is a relatively new approach to BI that is made possible by embracing distributed scale-out architectures such as Hadoop, Spark, and cloud-based object storage such as Amazon S3. As with SQL-on-Hadoop and big data OLAP, native visual analytics keeps data in the big data platform. It also leverages the power and scalability of the platform to run large-scale analytics. The BI and visualization UI is written from the ground up to take advantage of the analytics platform, but the system can optionally leverage existing BI tools as well. This approach gets the benefit of having learned from prior BI approaches, and improves upon known challenges to boost performance and scale while reducing complexity and latency.


Users
1 Send SQL query.
5 Return query results.
Data Lake
2 Process query on each node.
• If query matches analytical view, return accelerated query results.
• If not, return standard query results.
3 Automatically learn from queries to build more analytical views for more acceleration.
4 Collate results.

Native Visual Analytics Architecture and Process Flow

PROS:

- **No dedicated servers required.** Analyzing data directly in the big data platform leads to significant advantages. First, with no data movement from the platform to dedicated BI servers, the overall architecture is simpler and easier to maintain. Second, no external BI servers or external

ETL means lower hardware/administrative costs. Third, all raw data is immediately available to allow details on fine-grained data, unlike the summaries and aggregations used in dedicated BI servers. Finally, governance and compliance frameworks are easier to support by not creating separate copies of data in external repositories.

- **Unified security.** Related to the above, since there is no data movement, data can be secured by the platform's security controls, thus simplifying data protection and lowering the costs of securing your data. In the case of Hadoop, integration with security technologies like Apache Sentry and Apache Ranger enable the unified security model.
- **Lower learning curve.** As an option, standard BI tools can be used with native visual analytics. That way, both "citizen data scientists" and RDBMS power users take advantage of tools and skills they already have. This leads to easier adoption since users don't have to learn unfamiliar new tools.
- **Maximize current investments.** As an option, companies can leverage existing BI tools, skills, and training, thus avoiding additional expenditures on new BI tools.
- **Self-service.** With minimal IT intervention to run news types of queries, and no time-consuming and resource-intensive cube building, native visualization analytics can effectively provide self-service analytics that give much more flexibility to business users.

- **Query acceleration for performance and high user concurrency.** Native analytics entails optimizations such as caching and pre-computing subsets of queries that accelerate a wide range of end user queries. This also enables hundreds and even thousands of concurrent users on the system.
- **Cost-effective scalability.** By deploying analytics that leverages a big data architecture, you can easily cost-effective scale-out by incrementally adding more commodity nodes to a cluster.

CONS

- **Fear of the unknown.** Since this approach is relatively new, organizations may encounter resistance due to a lack of understanding why this approach is superior to others.
- **New skills and unfamiliarity with Hadoop.** The concept of accessing large amounts of structured and unstructured data for analysis is new, and will require some ramp-up for users and IT organizations alike. Basic familiarity with Hadoop, NoSQL, or cloud system skills are needed set up and maintain the data store.

	Traditional BI	SQL-on-Hadoop	OLAP on Big Data	Native Visual Analytics
Market maturity	●	◔	◔	◔
Use of existing/ common skills	●	◑	◑	◑
Architectural simplicity	◔	◕	◑	●
Enables self-service BI with access to granular detail	○	◕	◔	●
Works with legacy BI software	●	●	●	●
Ad-hoc query performance at scale	◔	◑	◑	◑
Prepared/ dashboard query performance at scale	◑	◕	●	●
Cost-effective scale	○	◑	◑	●
User concurrency	◔	◑	●	●
Data granularity support	◔	◔	◔	●
Unified security support	○	●	◔	●

Chapter 6: Making Big Data Actionable

Today's "Brave New World" of big data BI offers the potential of visualization to create new perspectives and insights. However, many of these initial offerings are still limited to data that has been extracted and prepared, often by costly data scientists. To realize the full potential of big data, business intelligence applications must evolve to support visualization directly to all data stores. They must eliminate the dependence on data cubes and predefined views while also integrating real-time and streaming data from sources like intelligent devices in the Internet of things.

Forrester Research has said that "customer obsession and big data call for more BI muscle," observing that nearly one-third of organizations now store and process more than 100 TB of structured data. “To process and analyze all this data efficiently and effectively, application development and delivery pros working on BI initiatives need highly scalable and distributed BI platforms and flexible and agile technologies,” wrote Principal Analyst Boris Evelson.[1]

The Next Generation of Business Intelligence

The change begins with the client. Since many legacy desktop BI tools failed to offer capabilities to support real-time collaboration, IT organizations had to ensure there were enough server resources to share analytics. A small implementation could carry a hefty price tag as it becomes deployed enterprise-wide. Since desktop BI clients leverage local copies of data downloaded from a server, they tax network bandwidth and introduce version-control problems, not to mention additional security risks as well. Further, many BI tool users simply use them to extract data, while relying on spreadsheets and other tools for more advanced analysis and formatting. Users then typically exchange these reports by emailing them to each other, exacerbating the data duplication and version control nightmare.

This is not to say that these traditional BI tools are not used for analysis or building visualizations. In many instances, a company's IT organization is tasked with building complex analytical applications such as balanced scorecards and dashboards using native functionality as well as APIs offered by these vendors. However, these pre-canned views are often designed to provide little or no ability to conduct ad-hoc queries or even "what if" questions that were not anticipated when the cube or view was created.

The need for BI servers should also be revisited in light of the limitations they place on working with unstructured data. Extracting and transforming data is a time-consuming process that is inconsistent with today's need for rapid decision-making. ETL also requires skills that are in short supply. And the fact that data scientists are often responsible for data wrangling only worsens the situation. Accenture found that 80% percent of new data scientist jobs created between 2010 and 2011 had still not been filled two years later.[2] Numerous studies have reported that the average salary for data scientists in major markets exceeds $200,000. That doesn't include the tax that BI servers place on IT infrastructure and budgets. Hadoop and the cloud has revolutionized the economics of data. Why not extend those benefits to business intelligence?

The Browser Is the New PC

It is highly unlikely that the minds that developed what is now known as today's Internet would have imagined how it has impacted our lives. With the increased proliferation of smartphones and other devices simplifying access to this "information superhighway" coupled with the demand for a more uniform "user experience," various companies have attempted to find ways to leverage these advances.

It is highly recommended that BI clients should leverage the native functionality provided by one of these modern browsers which support advanced programming languages like JavaScript and HTML5. Among the powerful new features HTML5 provides are the ability to accept non-HTML data, store and operate upon data locally, play video and audio without the need for plug-ins, display two- and three-dimensional graphics, and optimize for local processing resources, all of which would prove useful for an individual looking for more than text.

By incorporating these native capabilities, browser-based BI tools reduce the need for application installation and maintenance. This in turn significantly reduces licensing fees and IT support costs in both the short as well as long term.

When Machine Learning Meets Data Preparation — Introducing Smart Acceleration

Machine learning can now substitute for much of the data preparation work that once required human interaction. By analyzing data usage patterns over time, the analytics platform can "understand" underlying data and accelerate queries for improved performance and higher concurrency. Simply put, the engine essentially identifies and prepares the data based upon demonstrated user preferences.

The Arcadia Data Smart Acceleration™ is an example of this. Smart Acceleration is a framework that includes a recommendation engine for derived views (called "Analytical Views") of raw data based upon dynamic data usage analysis within the data lake, whether it is a Hadoop cluster or a cloud platform. Arcadia then transparently reroutes data queries to the Analytical Views, providing automated acceleration when needed for production and high concurrency uses. Queries are routed to these views in a cost-based manner. Views are stored in HDFS, Amazon S3, or other distributed data platforms and cached when in use for optimal performance.

In-cluster on in-cloud processing enables the analytics engine to scale linearly with the data for greater speed and easier management. Data is automatically modeled and maintained within the Hadoop cluster or cloud environment using simple logical data models that aren't tied to specific data cube structures. Users work with a dashboard or application that presents consolidated views of data, which they then point and click to drill through or across to the raw data source on the data platform. Intuitive visualization enables instant micro-segmentation, network graph analysis, event and time-series analytics, and dimension/measure correlations.

Multiple Data Sources

Arcadia Enterprise is designed to work directly with a wide variety of relational, real-time, and NoSQL data sources including HDFS, Amazon S, Apache Spark, Apache Kudu, Solr, MapR-FS, and more. These can be used in any combination, enabling structured and unstructured sources to be combined in a single view. For example, a single visual can combine data from relational, NoSQL, and Hadoop sources.

Users can also create views that combine real-time/streaming and historical data. This addresses an important shortcoming of legacy BI systems, which is that they only work on historical data. Folding in real-time data streams opens whole new applications of BI. For example, equipment managers can overlay streaming data from sensors on historical lifecycle data to see if there are signs of imminent equipment failure, or marketers can monitor real-time click data on a new advertising campaign to see how performance compares to previous efforts and can make adjustments on the fly.

By deploying directly on the Hadoop cluster or cloud environment, Arcadia Enterprise takes advantage of distributed scale-out architectures to accommodate hundreds or even thousands of users with virtually no degradation in performance. Instead of needing to purchase expensive new BI servers to handle increased demand, IT organizations can simply add low-cost servers to the cluster.

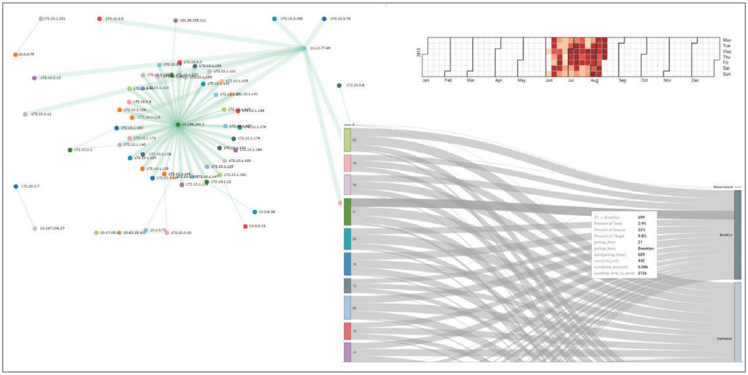
Advanced Visualization Goes Mainstream

Even though nearly two-thirds of people are visual learners,[3] most IT reports are still composed of rows and columns. It's difficult for the average end user to read rivers of numbers, much less derive patterns from them, which is why visualization is a must-have feature for any new BI platform.

Fast processors, increasingly high-resolution displays, and powerful analytics engines are enabling a wide variety of new visualization options. For example:

- **Network graphs ("network maps")** are used to identify relationships between related items and clusters such as when visualizing a social network or displaying a market basket analysis.

- **Correlation heat maps** provide a graphical representation of data where individual values contained in a matrix are represented by different colors.
- **Path visualizations** are collections of funnel visualizations which display information across a sequence of time-stamped events, such as conversion data for website visitors or airline delays by time of day.



Advanced Visualizations Reveal Insights That Raw Data and Basic Visualizations Cannot

Visualizations can also be overlaid on maps, calendars, workflow diagrams, and still images like screen captures. A series of visualizations can be displayed over time like a movie to illustrate time-series analyses. The arrival of affordable virtual and augmented reality hardware will undeniably expand these options as well.

Comparing these powerful new ways to visualize data to the traditional bar and pie charts provided by spreadsheets and legacy BI tools is like comparing a watering can to a garden hose. The tools that work most closely with the underlying data give users the latitude to quickly explore new visualizations and combine a variety of data sets and types fluidly, and should be included in any organization's requirements list when selecting an analytics platform.

Modern Data Platforms Continue to Evolve

The commercial software and open source communities are constantly innovating in areas such as Hadoop management. For example, YARN was added to Hadoop version 2 to handle resource management for Hadoop jobs. Apache Ambari is a management environment focused on ease-of-use for Hadoop installation, system management, and operations, most commonly used in vanilla Apache Hadoop and Hortonworks HDP. Cloudera Manager and MapR Control System are proprietary but highly-functional commercial alternatives. A BI system that works directly with Hadoop data needs to take advantage of these and many other management tools for performance and capacity tuning, especially as their deployment grows.

Cloud deployments are becoming increasingly popular especially for the rapid provisioning capabilities. While the cloud has historically been revered for lower cost of ownership, it is the "elasticity," or the ability to easily expand and contract a big data deployment, that is the key benefit today. Organizations no longer have to wait days or weeks for physical hardware servers to be set up in the data center, as cloud instances on third-party vendors can typically be provisioned in minutes. This capability is vital for supporting environments where data sets and user loads continue to grow and require more hardware resources. It is also valuable for handling short-term load bursts, where the extra capacity can later be shut down to reduce expenditures on unneeded resources. Most major big data technology vendors support third-party cloud deployments, so it is a viable option for advanced BI workloads. There are even popular cloud-native options like Amazon EMR, and emerging options like Snowflake as a cloud-only data warehouse technology, that round out the cloud landscape.

Another emerging technology that looks to change the way data is analyzed is the class of technologies around real-time streaming data. This includes technologies such as Apache Kafka, MapR-ES (formerly MapR Streams), RabbitMQ, and

IBM WebSphere MQ. These products, particularly the former two, are commonly used in big data deployments, especially those based on Hadoop, to handle event streaming as a required complement to batch-oriented, historical data. In fact, Cloudera and Hortonworks both support Kafka as part of their offerings, and MapR provides MapR-ES, which is their Kafka-compatible event streaming engine. For added capabilities, specialized stream processing engines like Spark Streaming, Apache Flink, Apache Apex, Apache Storm, and StreamSets, just to name a few, help with analyzing the data as it is delivered. Visualizations on event data are becoming more mature as well, and new innovations to take advantage of high speed streaming data in a graphical way are going to pave the way for gaining faster insights from big data.

Use Cases

MARKETING

Clickstream analysis is a critical tool for understanding how users traverse a website, which paths lead to a transaction, and which cause them to hit a dead end. The most effective form of clickstream analysis combines server analytics, such as load times and data transmission volumes, with e-commerce analytics such as the amount of time shoppers spend on certain pages, which items they add to or remove from their shopping carts, coupon use, and payment preferences. This involves huge amounts of data, which is why Hadoop is commonly used as a foundation for analysis. By working directly with Hadoop, marketers can more quickly identify sales opportunities or impediments. When streaming data is added, they can conduct live A/B and multivariate tests to compare multiple offers with each other and with historical performance.

Management of complex multi-channel marketing campaigns involves many moving parts, including clickstream analysis, offer codes, audience segmentation, time-series analysis, and multivariate testing. This complexity makes it all but impossible to anticipate which views and reports will be needed. Marketing analysts need to explore raw data and create visualizations on the fly, and the faster the better, particularly when time-sensitive actions like ad-bidding decisions need to be made in seconds.

Customer 360 profiles, considered the Holy Grail of marketing, integrate demographic, psychographic, behavioral, and preferential data derived from potentially thousands of sources. This gives companies a holistic view of customers to see new opportunities as well as avoid redundant communications that causes customers to lose faith. The more data marketers can integrate, the better the customer profile, leading to a greater customer experience.

FINANCIAL SERVICES

Financial risk analysis can involve a nearly unlimited number of variables that affect the behavior of markets and investment vehicles. It requires tolerance for both volume and speed, since time is money when making trading decisions. Using analytic tools that work directly on live data gives analysts a critical timing edge.

Fraud detection is a core practice in financial services, where seconds count when determining whether to approve or deny a transaction. Big data platforms like Hadoop are being applied to this task in innovative ways, such as integrating behavioral analytics, demographic profiling, pattern recognition, and trend analysis. The result in fewer fraudulent activities, but also fewer unnecessary transaction declines and more sales for retailers.

SECURITY

Security operations centers (SOCs) are widely using big data technologies for cybersecurity. Cybersecurity is increasingly becoming an analytical discipline as traditional endpoint devices are becoming less effective. SOCs gather and comb through vast amounts of server, network, and database log data looking for patterns that indicate a breach, as well as for anomalies that might indicate a breach. Timing is critical to isolating and containing an intruder. SOC personnel don't have the luxury of waiting for extract databases to run analyses.

Systems monitoring/management also entails analysis of many log files. As data processing environments grow ever more complex, the number of variables that contribute to slowdowns and outages grows accordingly. In the same way that SOCs analyze log data to find intruders, system managers can use live analysis and comparative historical data to more quickly avoid or remedy performance and availability problems before they impact the business.

OPERATIONS (INTERNET OF THINGS)

Streaming data analysis is becoming more and more critical in today's business environment. To understand the data volumes that IoT involves, consider this one statistic: autonomous vehicles are expected to generate and consume 40 terabytes of data[4] for every eight hours on the road. With 50 billion new connected devices expected to join the Internet over the next three years, the data management challenges will be unprecedented. The only practical way to manage and make sense of this volume of data is by using an analytics platform that directly accesses the storage layer with no unnecessary data movement.

References

1. How to Scale Business Intelligence With Hadoop-Based Platforms, Forrester Research, Sept. 13, 2016.

2. The Team Solution to the Data Scientist Shortage, Accenture, 2013

3. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=587201

4. http://www.networkworld.com/article/3147892/internet/one-autonomous-car-will-use-4000-gb-of-dataday.html

Chapter 7: Selecting a Next-Generation Business Intelligence Platform

Big data has enabled new applications and created new consumers for business intelligence and analytics. "Where in the past BI was limited by rigid data definitions, highly constrained data types, and the cost of delivering speed and scale, big data is now eliminating those constraints," wrote[1] Enterprise Strategy Group Senior Analyst Nik Rouda. A new class of applications "will empower a new generation of business data consumers, much broader than just the technical specialist pool of data scientists, DBAs, and analysts."

One big question is, "What should these next-generation solutions provide?" We suggest the following:

- They will reduce and in some cases eliminate ETL steps and intermediary data stages.

- They will natively support Hadoop, cloud, NoSQL, and other modern data platforms. By doing so, this will reduce the creation of disparate task-specific data silos.
- They will offer rich visualization features that are extensible to accommodate third-party visualization engines such as maps.
- They will empower more business users to create their own BI content, curate streaming and stationary data feeds, provision applications, and foster collaboration with peers.

The reason greater scrutiny will be required is because the market is rapidly evolving, with many new entrants taking advantage of big data platforms to approach BI in new ways. Each has different strengths and weaknesses. At the same time, existing vendors such as Oracle are retrofitting their BI platforms to support native Hadoop. This a trend appears to be unstoppable; in the words[2] of Forrester Research's Boris Evelson, "It's a question of when, not if."

Here are some additional capabilities to consider:

Self-service for business users. If offering self-service BI is important to your organization, look for the ability to browse data structures and sources at a fine level of granularity, create semantic relationships across multiple sources, and set hierarchies and logical data sets.

All users should be able to assemble their own dashboards with the ability to drill down to raw data and resume and pivot through multiple data levels easily. End users should also be able to create and publish visualizations to anyone with a browser. "Citizen data scientists" should have the ability to author their own BI content, curate data, create calculated measures, provision applications, and easily collaborate with others.

Data visualization features. Most BI packages provide basic visualizations such as bar and line graphs and pie charts. Modern BI systems should provide support for sophisticated visualizations like funnel charts, network graphs, dendrograms, packed bubble charts, heat maps, geographic map displays, and interactive legends. Extensibility is important to accommodate new visualizations. For example, the D3.js JavaScript library[3] enables developers to take advantage of the full capabilities of modern browsers without being tied to a proprietary framework.

If many of your users are non-technical, pay extra attention to ease-of-use, and the strategies that were applied to maximize it. Users should be able to build a chart very quickly, and certainly without writing any code. If you want a system that incorporates user experience (UX) expertise and best practices, look for products with leading browser-based innovations. HTML5 is extremely powerful for deploying analytics across a large user base, so look for technologies that use it for rich, responsive interfaces. Also look for technologies that incorporate principles from Material Design. Initially announced in June 2014 at the Google I/O conference, Material Design is a design language developed by Google that aims to unify user experience across their products and other platforms such as iOS as well as modern web browsers. Material Design[4] improves the overall digital experience for end users to make their analytical activities more intuitive.

Advanced analytics support. Advanced analytics goes beyond looking at historical data, and often deals with real-time responses. A common use of advanced analytics is on time-series data for a variety of analytical tasks, such as calculating averages, identifying variances, and making predictions. It is used in cybersecurity, network management, and website traffic analysis where streaming and historical data can be combined to identify outliers and exceptions.

In a security scenario, real-time analytics can trigger alerts based on unusual activity from a particular node or IP address. Once discovered, analysis can be conducted across users, endpoints, and networks within a specified time window to look for correlations and patterns.

In an e-commerce setting, it can be used to create heat maps that show the most active areas on a website and predict preventable actions like cart abandonment. It can also drive recommendations to not only give customers ideas on what they might need to purchase, but also to help the seller increase sales.

Derived data is another particularly powerful form of advanced analytics. It applies the output of one set of calculations to the input of another in a single pass, thus cutting down on query overhead and enabling much richer derived visualizations to be created.

Query engine support. There are may tools that enable some degree of SQL queries to be run against Hadoop data, but consistency and maturity vary widely. At a minimum, look for support for ANSI-standard SQL to lower the learning curve. There are many SQL-on-Hadoop engines besides the ones described previously, including IBM Big SQL, Teradata QueryGrid, Apache HAWQ, Microsoft PolyBase, Presto, and Splice Machine. Be sure to also check if their query engines can handle both normalized and denormalized data sets, and whether they support time-saving features like aggregate-awareness and multipass SQL.

Data discovery and exploration. Users should have the ability to browse data sources, structure, and content with full granularity and transparency. Good data discovery capabilities enable access to data inside and outside of Hadoop and the cloud from within a web browser without proprietary drivers or extracts. Discovery should work across multiple relational databases as well, such as MySQL/MariaDB, PostgreSQL, Oracle, Microsoft SQL Server, and Amazon Redshift. Look for sampling support. This enables the system to retrieve only a small percentage of the underlying data for discovery purposes, greatly reducing query times. Finally, look for specialized processing engines that are optimized for discovery. Users should be able to write queries however they want, without worrying about the underlying processing engine.

Hadoop support. New Hadoop versions are released frequently, so ensure that the platform integrates with whatever version(s) of Hadoop you use. If your use case involves commercial versions, such as Cloudera, Hortonworks, or MapR, look for an analytics or BI platform that is certified as compatible by those vendors. Native HDFS API support is important without the need for a separate extract engine or intermediate data structures. Also ensure that the platform integrates with the same Hadoop cluster manager that you use, such as Apache Ambari or Cloudera Manager. It should also integrate with Hadoop metadata frameworks, such as the Hive Metastore and HCatalog.

Native Hadoop security. Many traditional BI and visualization tools rely on decentralized security models, which complicates the process of extracting and managing data from Hadoop. In these cases, administrators must redefine security roles and privileges redundantly at both the Hadoop tier and again in the BI environment. Integrating with native Hadoop security enables administrators to control data access at a granular level, from the platform through to the UI. Look for centralized role-based access control (RBAC) that is integrated with Hadoop-native projects like Apache Sentry and Apache Ranger. Authentication and group membership administration should integrate with underlying directory sources based on Active Directory, Kerberos, LDAP, or SAML, as well as role membership and privilege information from Apache Sentry. Data permissions should be defined in the cluster, including discrete access control down to a single row or column in the data. If users plan to publish their data applications externally, make sure that published data can be securely provisioned and controlled down to the exact dataset level.

Vertical integration. Determine whether the tools that the product provides for data preparation, data modeling, semantic modeling, and reporting/analysis are seamlessly integrated with each other. This reduces or eliminates the need for extract databases. Some platforms integrate data preparation, modeling and reporting/visualization but may have only limited compatibility with third-party tools. Check partnerships to see if the platform integrates with other visualization, preparation, and management tools you may already use. Integration with popular open source data management and analytics tools like Apache Kudu, Apache Spark, Apache Impala, and Apache Solr is desirable.

Multiple deployment options. Even if your organization hasn't made the jump to the cloud yet, it's highly likely you will move at least part of your infrastructure to a public, private, or hybrid cloud at some point. Your BI engine should support both on-premises as well as cloud-based data sources. Look for compatibility with popular cloud operating systems and storage platforms, such as Amazon S3. Be sure to ask potential vendors if their analytical and visualization engines can seamlessly accommodate data from multiple sources without requiring extracts or intermediate steps. Can all or part of your data store be moved to the cloud without breaking existing queries or reports? Can data from HDFS and S3, for example, be combined in a single query? Each organization will have different tolerance for the trade-off between flexibility and functionality.

Ease of administration. To minimize complexity, look for platforms that integrate natively with existing Hadoop administration tools like Apache Ambari and Cloudera Manager.

In most cases, you will want to avoid introducing yet another administrative tool into your environment.

Recommendations

Business Intelligence has kept pace with many advances in both hardware and software technologies. However, because they traditionally relied on data extracts and dedicated infrastructures, legacy business intelligence solutions have limited the ability to explore the full range of new data sources and types that are available to them. This is particularly true in the area of streaming or time-sensitive data because of the long lead times commonly associated with ETL.

With the arrival of Hadoop, Spark, Apache Kafka, and other open-source solutions, the cost of storing and processing data has plummeted, opening vast new opportunities for innovation and insight. Many tools have emerged to tap into these data sources, but most still require data translation, transformation, integration, ETL, and intermediary platforms between the Hadoop store and the analytic front end, providing incremental improvements over legacy BI approaches.

The concept of self-service has changed the user experience. Business users are more emboldened than ever to take care of their own data management needs, and they have little patience for long delays getting at their data. Given that the shortage of data scientists is unlikely to abate in the near future, a solution that involves throwing more people at the problem is not attractive. A far more practical solution is to pacify users by giving them self-service access to analytics and visualization with the power to drill down into the back-end data stores.

AN EXAMPLE TECHNOLOGY TO CONSIDER

Arcadia Data is one of a new breed of visual analytics companies that is taking an entirely different approach to BI by cutting out intermediate steps and enabling their users the ability to query big data sources with a full range of high-performance analytic and visualization options.

Arcadia Enterprise is a native visual analytics platform for big data that is designed from the ground up to work in concert with Hadoop, cloud environments, and other modern data platforms. This eliminates the need for a large portion of ETL required to "fit" data into legacy BI tools and the many related complications and failure points that we have outlined here, including delays caused by data preparation, data duplication, version conflicts, security vulnerabilities, and more. The Arcadia Data approach minimizes risk by never copying or moving data out of the core data lake, hub, or platform.

Arcadia Enterprise works with popular big data technologies in use today like Hadoop, S3, Hive, Cloudera Manager, Ambari, Apache Sentry, and Apache Ranger, rather than saddle customers with additional layers of tools. Arcadia Data is also committed to making its extensive library of visualizations available through web and mobile browsers. This removes barriers between users and their data, eliminates client licensing costs, and delivers major operational efficiency benefits.

One of the biggest limitations associated with conventional BI is cost-effective scalability of the analytics server. Because Arcadia Data runs natively on the data platform itself, it transparently scales linearly with Hadoop and cloud environments.

Until now, BI users did not have the ability to access real-time data, much less integrate it with batch sources. Arcadia Enterprise provides the means to do this with drag-and-drop simplicity, opening up a world of new applications in the process. For example, factory managers in an instrumented IoT environment can match live data streams from sensors on the floor to historical data illustrating averages and thresholds. Managers can see in real time when equipment is malfunctioning and replace it without downtime. Or web server administrators can monitor clickstream data and adjust resources in real time to minimize performance delays.

References

1. http://go.arcadiadata.com/Accelerating-Business-Insights-Report-Twitter-Revolution.html

2. https://www.arcadiadata.com/lp/forrester-wave-hadoop-bi-research-report/

3. https://d3js.org/

4. http://www.techrepublic.com/article/google-material-design-the-smart-persons-guide/

Conclusion

This is an exciting time for technologies that enable data-driven companies. While there have been many impactful IT-facing innovations over the last 40 years—hardware virtualization, public cloud, flash drives, just to name a few—the most promising data-focused ones have only arisen in the last several years. The proliferation and maturation of big data technologies have given organizations more power to successfully leverage data to drive the operations of their business. As data-driven companies try to accomplish more and more with their huge volumes of data, they continue to explore new technologies to stay ahead.

Hadoop has been, and likely will continue to be, a key technology for managing big data. Its ability to handle a wide variety of data types, the scale-out architecture on commodity hardware, and its widespread support make it an ideal data management platform for today and the future. But Hadoop certainly does not and should not operate in isolation. Related technologies like Spark, NoSQL, NewSQL, and object stores will also play critical roles in building out modern architectures.

And while big data innovations have largely pertained to data management, focus on end users particularly regarding BI and analytics, will continue to grow. The notion of self-service, especially important when it comes to gaining competitive advantage via data agility, will emerge further as a strong theme. More powerful tools such as those that enable the "citizen data scientist" will dominate the BI/analytics landscape. And related to that, the phenomenon of "data democratization" is a key aspiration that will let companies get more value from data with less overhead.

The advent of big data technologies resulted from the recognition that traditional technologies could not efficiently handle the volume, velocity, and variety of data that is inherent in today's business environment. This simply means that a separate class of technologies was needed to handle emerging challenges. This evolution is certainly true in the BI/analytics world. As data-driven companies seek more data, more agility, more insights, all with lower costs, a new paradigm is required.

One cannot expect to use traditional technologies on big data without making significant compromises that ultimately reduce the value of that big data. With newer technologies, you will not have to make those compromises that limit your ability to compete. The industry has evolved from a traditional BI approach to new approaches such as SQL-on-Hadoop, OLAP on big data, and native visual analytics. While these newer technologies have the obvious disadvantage of a shorter track record, the proof points among current innovating companies using these technologies show the distinct operational advantages that make the investment worthwhile.

Native visual analytics is arguably the most intriguing approach because it provides more advantages for analyzing big data than the other approaches. An integrated suite of analytics and query acceleration provides benefits to both the end users as well as the IT administrators. End users get powerful, easy-to-use visualizations with faster performance, while the IT team has significantly lower overhead around deployment and data management activities. Support for complex data types lets end users analyze a much wider range of data sources, a key tenet of big data, and reduces the ETL effort on the management side. And immediate access to granular data means end users can get details, not just summaries and aggregations, without time-consuming IT intervention.

Arcadia Data is an example of a technology provider for native visual analytics. With capabilities that cater to a wide range of end users including the power users, they make the Gartner vision of "citizen data scientists" real. Visual analytics and more advanced data discovery need no longer be limited to a few senior managers using algorithms designed by data scientists. Instead, data professionals can focus on building operational programs and models based upon discoveries enabled by the people who use the data. They can also work with larger volumes of varied data types while also retaining the ability to drill down to a granular level, all within the same application. Support for a variety of big data environments including Hadoop and the cloud means that the IT team can choose the deployment model that best suits their needs while giving end users the data access they need.

About the Authors

Sushil Thomas is a technologist with a passion for data and applications at scale. He was a part of Sun Microsystems in the first internet boom and helped design and develop the platforms that ran most of the internet at a time when Sun's mission statement — "The Network is the Computer" — seemed visionary. He helped usher in the utility storage revolution at 3PAR and moved the industry from an era where terabytes of storage were considered big data to an era where petabytes of data are routinely stored and processed. At JovianData and Aster Data, he worked on the challenge of bringing business value and business users to big data. He was acquired into Teradata where he designed the first big data appliance that Hortonworks and Teradata took to market at scale. In 2012, he co-founded Arcadia Data to take on the ongoing industry challenge of making big data platforms accessible to non-technical business users. He looks forward to a future where data lakes become data playgrounds and is doing his part to ensure that business users are allowed in.

Steve Wooledge leads marketing at Arcadia Data and has worked in the analytics and business intelligence enterprise software market for 15 years at leading-edge companies including Business Objects, Aster Data, Teradata, and MapR Technologies. An entrepreneur at heart, he has a passion for bringing innovative, disruptive technology to market to help businesses create competitive advantage and has served as an advisor to several enterprise analytics software startups. Steve has been a featured speaker at dozens of international industry trade shows including O'Reilly's Strata Data and Gartner's Data and Analytics Summit, sharing best practices of organizations who have successfully unlocked the value of data for business applications. He has appeared in Forbes and was a contributing author to dozens of industry trade publication articles and whitepapers while living the dream in San Jose, California with his wife Vicki and their five children.

This is an exciting time for data-driven companies. The proliferation and maturation of big data technologies give you more power to successfully leverage data to drive your business operations. While big data innovations have largely pertained to data management, the focus on end users, particularly regarding BI and analytics as well as data-centric applications, will continue to grow. The notion of self-service BI, especially important when it comes to gaining competitive advantage via data agility, will emerge further as a strong theme.

As you seek more data, agility, and deeper insights, all with lower costs, a new paradigm is required. Applying legacy BI technologies to the modern world of big data is often not the answer. Instead, modern approaches to analytics and data applications were designed to address challenges of growing data volumes while also reducing the dependence on IT for day-to-day data management activities.

In this book, we'll explore new self-service technologies that enable you to stay ahead of the game whether you are a business analyst, data scientist, or data/application architect.

ISBN 978-0-692-94721-0

9 780692 947210 >