Table of Contents

Preface

CHAPTER 1: BI, Analytics, and Big Data: The “A-ha!” Moment

Without tools, data is nothing but an expense
What you’ll learn
Taming—and exploiting—the data beast
Old solutions won’t work for the new normal
BI and analytics at-a-glance
Defining terms

CHAPTER 2: New Platforms Bring New Approaches

Setting the scene
Really big, big data
Brave new world of big data analytics
How we got here
RDBMS benefits package
The fox in the hen house
IT evolution, not revolution
Which platform to use?
  Data warehouses
  Massively parallel processing systems (MPPs)
  Hadoop
Data lakes vs. data warehouses
Choices
Criteria for selecting the right big data analytics platform
Defining terms

CHAPTER 3: Navigating the SQL-on-Hadoop landscape
Where SQL fits
Batch SQL
Interactive SQL
In-memory SQL
Operational SQL

CHAPTER 4: The Analytics Revolution
System overload at Kobo
Analytics
Big data has opened some new options
Preparing for analytics

CHAPTER 5: Analytics at the Speed of Light
The move to in-memory
What’s the deal with streaming?
The SAP HANA platform
SAP HANA Vora
Use cases

CHAPTER 6: The Promise of Predictive Analytics and Machine Learning - Seeing the Future
Making educated guesses
Predictive analytics
Machine learning
Tools and terms

CHAPTER 7: Your Turn - How to Make Analytics Work for You
8 Steps

The open source equation

CHAPTER 8: Big Data Analytics: How It’s Being Done in the Real World

Winning characteristics

UnitedHealthcare

TransUnion

Harte Hank

comScore

CHAPTER 9: Conclusions and Recommendations
Preface

Much of the history of computing and information technology has been less than glamorous, often comprised of “business plumbing” work like payroll, HR, data storage, and the like. Then along came analytics and business intelligence (BI), precipitating a collective “A-ha!” moment that reverberated from IT right through the C-suite. Here was a set of truly business-first technologies and solutions capable of optimizing decision-making to lower costs, speed time-to-market, and drive real competitive advantage, among other vital benefits.

Now BI and analytics are being paired with the latest IT phenomena, big data. The impact of this union promises to be a disruptive game-changer, both for business and IT. The reason is simple. Somewhere in the area of 75% of data available to the enterprise today goes unused. This data holds the keys to a deeper, richer understanding of customers, competitors, markets, and trends.

The distinction of use cases

This book is intended to discuss BI and analytics in a capacity where the business is not yet leveraging a big data platform to build new applications to serve business needs. It is rather common for people to build brand new applications directly atop a Converged Data Platform. There are many benefits from the application development side of the house.

As new applications are built on this platform, new analytics can be applied. This brings an inherent benefit of not having to move the data between the system of record and the platform for which analytics are performed as they are one in the same. While legacy systems are still in place, coexistence is a very happy place to be, and that is the major focus of this book.
Without tools, data is nothing but an expense

Without the tools to organize, analyze, and properly mine that data, it remains just that—an ever-growing pile of data, doubling in volume every two years, with little or no business value. Most infrastructures and solutions in place today cannot efficiently scale to process and analyze this data.

Plus most of this new data is unstructured or semi-structured (from here on only referred to as unstructured), hailing from sources like email, social media, video, and others. Traditional systems tend to choke when fed large volumes of data that is not completely structured.

For these and other reasons, all eyes today are upon solutions built from the ground up to gain business value from big data. These BI and analytics solutions for big data have one strong point of commonality, namely the Hadoop framework, and in particular Hadoop distributions that include Apache Spark. Forrester calls Hadoop “the new core of the analytical enterprise,” adding that Hadoop “is mandatory for firms that wish to double-down on advanced analytics and create insights-driven applications to help them succeed in the age of the customer.”

What you’ll learn

This book closely examines the promise, potential, and significant dynamics of analytics and BI on the big data platform.

You will see that while the disruptive potential of big data is vast, the acceptance and adoption of BI and analytics on big data is absolutely essential for success in virtually every organization.

A good analogy is today’s energy companies; had they not adopted advanced technologies for extracting oil and gas from previously unreachable sources, they would be out of business today. This is the same problem with BI and analytics on big data.

Adapt and flourish—or perish.
Taming—and exploiting—the data beast

Data is in the DNA of Experian.

This nearly $5 billion information services leader relies fully on the timely analysis of growing mountains of data to support its four lines of business.

The fast-track growth of two businesses—credit services and marketing services—spawned a steep upswing in sheer volumes of consumer and business data Experian had to process. Upgrading its Microsoft Windows/SAN infrastructure loomed as overly expensive and complex. Plus, even upgrading would have left Experian without enough processing power, hampered further by storage and data-access limitations.

Experian management decided to meet brave new challenges with a bold new solution. They chose the MapR Converged Data Platform to bust out of the restraints of its in-house database and move forward with renewed processing power to handle the onslaught of big data.

The results offer a glimpse into a future where IT and the business look at data differently, as organizations evolve into data-centric businesses. At Experian, storage expenses have dropped, while storage capacities and capabilities have grown. The company can now process more financial data faster than ever, and provide clients with enhanced insights from a deeper pool of data. For each business unit at Experian, the switch to Hadoop for big data has meant faster time-to-market for services.

The future is now when it comes to big data and analytics.

And it couldn’t arrive too soon, given the tidal wave of change washing over organizations as data volumes grow at breakneck speeds.

Old solutions won’t work for the new normal

Traditional infrastructures offering real-time analysis of normalized data in warehouses or data marts worked swimmingly for 25 years. Big data is anything but normal, both in its hockey-stick growth and in its basic format, namely unstructured.

Traditional systems cannot scale to handle the volumes in anything approaching a cost-effective way. Nor can they efficiently—if at all—process and provide a platform for analysis of data from social media, videos, emails, and other emergent sources.

Data is evolving. So it is not surprising that the worlds of BI and analytics are also in a state of metamorphosis.

BI and analytics at-a-glance

We can break down the evolution of BI and analytics into three easy pieces, or chapters.

1980s and 1990s – The era of IT-driven analytics. The CFO wants a report? Go to IT and ask for it. The CEO needs a report? Ditto. The reports and spreadsheets created back then remain a mainstay of BI today. But as senior managers gained a deeper appreciation for
the value of data, so did their demand for reports. The notorious result was called report backlog.

2000s – Several self-service BI and analytics tools emerged and the backlogs diminished. Still, these tools often required a measure of technical expertise. But surely they whet appetites for more and better self-service tools as productivity within the business analyst community soared.

The present – Big data is synonymous with a growing variety of data formats, like JSON and complex flat schemas, often stored in a data lake. In an era of schema-free data analysis and exploration, IT support can be a thing of the past. And this translates into near-instant data analysis by non-IT stakeholders. Take marketing VPs, for example. Their ability to improve customer lifetime value and conversion rates depends on immediate access to data flowing in from campaigns and from external sources. The faster they get it and analyze it, the quicker they can adjust on-the-fly to shifting market conditions.

You can learn more from this video that describes the evolution of BI over the last three decades from being IT-driven to analyst-driven with self-service tools. [Here’s a link to the blog post with video embedded: mapr.com/bi-evolution]

Data from various recent surveys confirms the growing enthusiasm for big data analytics.

One such survey from Wikibon polled 300 organizations that either deployed or were evaluating big data analytics projects. One key finding: A majority believe big data analytics represents an entirely new source of competitive advantage, not just a complement to data warehouses and existing BI workloads.

Other significant findings include:

- Companies are moving steadily from pilots to actual deployments of big data analytics.
- As organizations gain more IT expertise in evaluating and deploying these solutions, increasing numbers of them report actual “success” in the deployments.
- Initial primary use cases for big data analytics include IT operations support emphasizing cost savings, which help defray deployment costs, and ETL (extraction, transformation, and loading) of data from both heterogeneous and homogeneous sources.
- As users grew the number of Hadoop clusters deployed from 1 to 2 or more, the number of administrators assigned to each cluster dropped dramatically from 3.5 for one cluster to less than 1.5 when 3 or more clusters are deployed. This dynamic reflects how rapidly IT staff acquires new skills for this environment.
- Challenges early on include ensuring that basic integration and operational performance work smoothly together. There are also challenges to maintaining application performance at very high data volumes.
Subsequent chapters in this ebook will demonstrate the tremendous store of benefits that await both IT and the business as they more fully embrace BI and analytics on big data.

**Defining terms**

**Business Intelligence** – Stripped of all else, BI describes decision making driven by data rather than, say, an executive’s gut instinct. BI not only encompasses the generation of data, but also the analysis and eventually visualization of data so that business analysts and business leaders make the most informed decisions about products, strategies, market timing, and other mission-critical factors. Without question, the major trend in BI today is in moving away from IT-generated or IT-assisted reports to a BI world steeped in user self-service. The other trend, of course, is in the utilization of non-traditional, unstructured data. Over time, this will mean less dependence upon structured or normalized data often residing in a warehouse or data mart. This latter trend is essential for organizations to grasp, given that most data growth going forward will be of the unstructured type. **Gartner pegs** the global BI and analytics market at a shade under $17 billion this year, growing at a steady 5% annual clip.

**Analytics** – Data from different sources often exhibit valuable patterns that convey market intelligence. Think of analytics as unearthing and then interpreting those patterns into clear business language, using data visualization as a key tool. Organizations leverage analytics to reduce costs, make faster and more informed decisions, and to test the waters a priori before developing new products and services. When it comes to analytics and big data, customer analytics far and away is the **three drivers of interest** in big data analytics are finding correlations across disparate data sources; predicting customer behavior; and predicting product or services sales.

**Predictive Analytics** – A subset of analytics, this describes the use of data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. **Gartner says** there are several attributes that separate this analytical approach from several others, including:

- A predilection toward prediction of events or trends
- Very rapid analysis often measured in hours, not days as with traditional data mining techniques
- Stresses ease of use for non-IT professionals to “roll their own”
- An emphasis on highly relevant business insights rather than longer-term “blue sky” analyses—think very targeted marketing
- The use of “scores” to rank predictive models, with the higher scores signaling a greater likelihood that a predicted outcome will in fact occur. Credit bureaus are big users of predictive analytics, using loan histories, bank balances, income, and other factors to predict debt repayment ability.
**Big Data** – Best thought of as all the data both internal and external to your company, the gathering, storage, and analysis of which would prove immensely valuable to your organization. One other important fact: traditional infrastructures cannot handle the sheer volumes and complexity (unstructured) of big data. The market is immense. **Big data will revolutionize operations** the same way the Internet did.

**Data Lake** - An area in which raw data can be housed which was previously too expensive to store and process. Data is managed and governed here as well as staged for moving to a data warehouse, if so required. The data lake can also act as an online archive for infrequently accessed data.
Setting the scene

Venerable businesses with global brands didn’t get that way by standing still and ignoring the mandate for continuous change.

Thus when Wells Fargo Capital Markets mulled the changing nature of data and how the changes could impact core businesses, IT leaders decided it needed a big data platform to support analytics in a way traditional systems could not. Founded in 1852, parent company Wells Fargo is a global giant with 265,000 employees and $1.8 trillion in assets.

As the company’s data mix was changing to include vast volumes of unstructured data, IT focused attention on platform solutions supporting NoSQL and Hadoop.

Really big, big data

What’s the meaning of big data and analytics at Wells Fargo Capital Markets? Consider this. The group regularly processes market tick data, which includes all the price points for thousands of equities and the movements of those stocks. Up to three million ticks per second.

Being in a highly regulated industry, Wells Fargo listed security considerations as a top criterion for platform selection. Also on the company’s wish list were superior scalability to handle spikes in big data volumes, ultra-high performance for customer-facing apps, and multi-tenancy to efficiently share IT resources. And the platform had to support the most robust analytics tools available.

As told by Paul Cao, Director of Data Services at Wells Fargo Capital Markets,

“The MapR solution provides powerful features to logically partition a physical cluster to provide separate administrative control, data placement and network access.”
Brave new world of big data analytics

Wells Fargo is one of many forward-leaning organizations that have recognized the need for new platforms for BI and analytics in a big data world.

**New platforms.**

For many CIOs and IT leaders, these two simple words can cause panic, especially for those leaders seasoned enough to recall the many disastrous ERP ventures of a generation ago. So, now they are asking what a new platform or platforms for big data mean.

- Are we talking about a clear break with the past?
- Do we now put aside the enormous investments in so-called “traditional” data processing?
- Will CIOs be asked to shutter warehouses, abandon data marts, establish data lakes, and oversee a complete retooling and training on everything “new” for the IT staff?
- Are we talking forklift-style upgrades?

The short answers to these four questions are “No,” “No,” “No,” and “No.”

**How we got here**

It will be helpful to take a quick look back at where we’ve come from in order to frame the discussion in a more IT-evolution style, because that is what the new platform discussion is really all about.

Before there was big data, there was just data, which was processed by sophisticated databases, and excellent tools that were developed in the 1970s. The most popular were (and still are) relational database management systems (RDBMS), which are transactionally based. The structured query language (SQL) is the decoding ring for managing data and simplifying processing within RDBMS.

Other iterations of DBMS include columnar, key/value, and graph. For the most part, they worked with structured if not highly structured or normalized data, often residing in a warehouse or special purpose data mart.

Another form—object databases—was IT’s first foray into working with less structured if not unstructured data, like videos and images. They are placed in specialized data repositories and usually require specialized skill sets and specialized infrastructure to make them work. In other words, they are expensive to run.

**RDBMS benefits package**

Billions and billions of dollars globally have been invested in the infrastructure to run these databases, and the people to operate and refine them for various vertical market ap-
Applications. For real-time transaction processing, they remain the undisputed king of the hill.

**Other RDBMS benefits include:**

- Recoverability from failure is very good, right up to the most recent state in most instances
- A RDBMS can be distributed easily in more than one physical location
- RDBMS virtually guarantee a high degree of data consistency
- SQL is easy to learn
- There is an enormous installed base of IT talent familiar with RDBMS
- Users can carry out reasonably complex data queries

What’s the downside? The truth is, as long as the data being managed is structured and relational in nature, there are few downsides. Scalability is a problem, as most of these systems are proprietary, and core storage is very expensive, especially as the database grows. But these venerable databases and their entourage of tools and applications are highly visible in every Fortune 1000 company for a good reason: they deliver value.

**The fox in the hen house**

But then came big data, a lot of it coming from the unstructured hinterlands. It encompassed data from clickstreams, website logs, photos, videos, audio clips, XML docs, email, Tweets, etc.

Initially to IT, most of this data resembled the background noise emanating from deep in the universe—just a lot of noise. But remember this: A man named Arno Penzias deciphered that deep space background noise in 1964, eventually interpreting it as proof of the since-validated Big Bang theory of the universe. He won a Nobel Prize.

And so it is with big data. As it turns out, locked in all those disparate big data sources are invaluable insights into customer behavior, market trends, services demand, and many other nuggets. It is the Big Bang of information technology.

With big data far and away the biggest component of the overall growth in data volumes, and with the relative inability of traditional analytics platforms and solutions to efficiently handle unstructured data, the analytics landscape is undergoing profound changes.

**IT evolution, not revolution**

But here is the important thing to bear in mind. Big data analytics is not going to replace traditional structured data analytics, certainly not in the foreseeable future.

Quite to the contrary. As stated in *The Executive’s Guide to Big Data & Apache Hadoop,* “Things get really intriguing when you blend big data with traditional sources of information to come up with innovative solutions that produce significant business value.”
So you might see a manufacturer tying its inventory system (in an RDBMS) with images and video instructions from a document store-based product catalog. This would help customers help themselves to immediately select and order the right part.

Or a hotel chain could join web-based property search results with its own historical occupancy metrics in an RDBMS to optimize nightly pricing and boost revenues via better yield management.

Coexistence, not replacement. That is the correct way to view the relationship between Hadoop-based big data analytics and the RDBMS and MPP world. Thus organizations are wise to focus on Hadoop distributions that optimize the flow of data between Hadoop-based data lakes and traditional systems. In other words, keep the old, and innovate with the new.

**Which platform to use?**

There are three basic data architectures in common use: data warehouses, massively parallel processing systems (MPP), and Hadoop. Each accommodates SQL in different ways.

**Data warehouses**

Data warehouses are essentially large database management systems that are optimized for read-only queries across structured data. They are relational databases and, as such, are very SQL-friendly. They provide fast performance and relatively easy administration, in large part because their symmetrical multiprocessing (SMP) architecture shares resources like memory and the operating system, and routes all operations through a single processing node.

The biggest negatives are cost and flexibility. Most data warehouses are built upon proprietary hardware and are many orders of magnitude more expensive than other approaches. In one financial comparison conducted by Wikibon, the break-even period for traditional data warehouse was found to be more than six times as long as that of a data lake implementation.

Traditional data warehouses can also only operate on data they know about. They have fixed schemas and aren’t very flexible at handling unstructured data. They are good for transactional analytics, in which decisions must be made quickly based upon a defined set of data elements, but are less effective in applications in which relationships aren’t well-defined, such as recommendation engines.

**Massively parallel processing systems (MPPs)**

MPP data warehouses are an evolution of traditional warehouses that make use of multiple processors lashed together via a common interconnect. Whereas SMP architectures share everything between processors, MPP architectures share nothing. Each server has its own operating system, processors, memory and storage. The activities of multiple process-
ors are coordinated by a master processor that distributes data across the nodes and coordinates actions and results.

MPP data warehouses are highly scalable, because the addition of a processor results in a nearly linear increase in performance, typically at a lower cost than would be required for a single-node data warehouse. MPP architectures are also well suited to working on multiple databases simultaneously. This makes them somewhat more flexible than traditional data warehouses. However, like data warehouses, they can only work on structured data organized in a schema.

However, MPP architectures have some of the same limitations as SMP data warehouses. Because they require sophisticated engineering, most are proprietary to individual vendors, which makes them costly and relatively inflexible. They are also subject to the same ETL requirements as traditional data warehouses.

**MPPs & SQL**

From a SQL perspective, MPP data warehouses have one major architectural difference: in order to realize maximum performance gains, rows are spread sequentially across processors. This means that queries must take into account the existence of multiple tables. Fortunately, most MPP vendors hide this detail in their SQL instances.

**Hadoop**

Hadoop is similar in architecture to MPP data warehouses, but with some significant differences. Instead of rigidly defined by a parallel architecture, processors are loosely coupled across a Hadoop cluster and each can work on different data sources. The data manipulation engine, data catalog, and storage engine can work independently of each other with Hadoop serving as a collection point.

Also critical is that Hadoop can easily accommodate both structured and unstructured data. This makes it an ideal environment for iterative inquiry. Instead of having to define analytics outputs according to narrow constructs defined by the schema, business users can experiment to find what queries matter to them most. Relevant data can then be extracted and loaded into a data warehouse for fast queries.

**Data lakes vs. data warehouses**

---

Let’s take a look at the main differences between a data lake and a data warehouse (summarized from KDNuggets):

**Data:** While data is structured in a data warehouse, data lakes support all data types: structured, semi-structured, or unstructured.

**Processing:** Data is schema-on-write in a data warehouse, while it’s schema-on-read in a data lake.
Storage: It can be expensive to store large volumes of data in a data warehouse, while data lakes are designed for low-cost storage.

Agility: In a data warehouse, data is in a fixed configuration and is much less agile, while data in a data lake is easy to configure as needed.

Users: The data lake approach supports all users (data scientists, business professionals), while a data warehouse is used primarily by business professionals.

The foremost use case for Hadoop continues to be the “data lake” because it stores a lot of unstructured data for refinement and extraction into relational “data marts” or data warehouses. In fact, Gartner said that they have seen a big uptick in inquiries from customers about data lakes as evidenced below:

> Just looked at the numbers. @Gartner_inc inquiries on data lakes increased 72% from 2014 to 2015.
> — Nick Heudecker (@nheudecker) April 12, 2016

David Menninger of Ventana Research says that data lakes can provide unique opportunities to take advantage of big data and create new revenue opportunities. His article on “data lakes being a safe way to swim in big data” provides a good view on how adoption of data lakes is on the rise.

As noted above, there are many parallel efforts to bring the power of SQL to Hadoop, but these projects all face the same structural barriers, namely, that Hadoop is schema-less and the data is unstructured.

Applying a “structured” query language to unstructured data is a bit of an unnatural act, but these projects are maturing rapidly. Below is an architecture diagram that shows how some of these different approaches fit together in a modern data architecture.

**Data architecture:**
Choices

In summary, there are valuable use cases for each platform; in fact, data warehouses, MPP data warehouses, and Hadoop are complementary in many ways. Many organizations use Hadoop for data discovery across large pools of unstructured information called data lakes, and then load the most useful data into a relational warehouse for rapid and repetitive queries.

Criteria for selecting the right big data analytics platform

First let’s state there is no one right way to select a big data analytics platform, but there are many wrong ways to do so. The best advice is to consider a checklist like the following. As you will see, selecting the right platform is more of a self-examination of business needs and IT features that will fulfill them.

- **Know the business requirements.** This may seem like simple table stakes more than sound advice, but it is the single most mission-critical criteria when selecting a big data analytics platform. Doing this right means having keen insight into your organization’s requirements in 3-5 years, not just in the next 12 months. What new revenue streams might the business explore? How might compliance and regulatory requirements impact data use going forward? What previously untapped data sources hold the most potential value to your business if it could be accessed and analyzed? How are data volumes from those sources expected to grow? These are questions
that bestride the IT-business edge, answers to which are vital to your platform selection.

- **Security as always is job #1.** Real-time cyber attack detection and then mitigation have never been of more critical importance. Your big data platform must be highly capable of analyzing data from an ever-growing myriad of sources and devices, then using the most advanced security analytics to prevent attacks from ever happening. Be certain in your platform selection that vital functions such as app log monitoring, fraud detection, event management, intrusion detection, and other security chores are handled better on the platform you choose than on any other, and in real time.

- **Look for open and interoperability.** Always look for solutions that support open source and for components that support Hadoop’s APIs. And look for solutions that interoperate with existing apps so that all apps will work with data you store in Hadoop.

- **Big data and scalability are a power couple.** If the platform cannot scale to meet the ultra aggressive requirements of big data volumes, then “that dog won’t hunt,” as the saying goes. comScore watched as their initial Hadoop cluster grew to process more than 1.7 trillion events per month globally. You surely do not want to be facing such a processing load without first checking to be sure the platform can handle it without performance degradation.

- **Warehouse integration is often needed.** Typical organizations store lots of data in warehouses and data marts. To cut costs, they will want to move some of that archived data to less expensive platforms such as Hadoop-based data lakes. And then at some point, the data may well need to migrate back the other way. Be sure to closely examine the ease with which this data integration can take place on big data analytics platforms you are considering. Remember, those distributed and relational database management systems are here to stay for many years to come. The name of the game when it comes to platforms for big data analytics is coexistence and cooperation, not replacement of one for the other.

- **The future is now, and it’s all about self-service.** A basic question to ask is just how data will be used and accessed. In the world of big data analytics, IT needs to cast itself in the role of a bystander. That means the platform you choose must support a rich array of self-service tools that are truly user friendly—for any user, not just by quants and data scientists.

- **Is multi-tenancy important to you?** Multi-tenancy allows you to share IT resources while letting different business units, as well as data from partners and customers to co-exist on the same cluster. But the highest levels of customizable security must ac-
company this feature. If the platforms you are considering don’t measure up, consider others instead.

Defining terms

MapReduce – This is a programming model—some say “paradigm”—and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Thus, it is tailor-made for the big data era where deriving speedy insights into massive, often disparate data volumes is essential.

Hadoop – A specific approach for implementing the MapReduce architecture, including a foundational platform and a related ecosystem. This Java-based programming framework supports the processing of large data sets in a distributed computing environment. So again, we have a perfect fit for the big data era of enormous often-disparate data volumes. **Wikibon estimates** Hadoop will comprise a $22 billion US market by 2018.

Hadoop ecosystem – The Hadoop ecosystem includes additional tools to address specific needs. These include Hive, Pig, Zookeeper, and Oozie, to name a few.

Open source – This is software, including Apache Hadoop and Linux, whose source code is available for modification or enhancement by anyone. Source code is typically the domain of heads-down coders, and seldom is seen or touched by users. Source code is changed or manipulated to enable discrete applications to run on it.

SQL – The broadly accepted language of choice for data query, manipulation, and extraction from an RDBMS.
Where SQL fits

SQL is one of the most mature and ubiquitous languages used in data processing. Designed in the early 1970s and standardized in 1986, it is used by millions of people and is the lingua franca of relational database management systems.

SQL was designed specifically for relational platforms with well-defined schemas, or data structures. When the type and location of data elements are clearly known, SQL is a fast and flexible way to extract information and combine it in flexible ways. It’s also pretty simple to use. A typical simple SQL query that retrieves the records of customers between 30 and 45 years old might look like this:

```
Select * from CustomerName
Where Age Between 30 and 45;
```

Business intelligence and visualization tools rely on SQL to load data for what-if analysis, standard reporting, and dashboarding.

Within the big data landscape there are multiple approaches to accessing, analyzing, and manipulating data in Hadoop. Each depends on key considerations such as latency, ANSI SQL completeness, ability to tolerate machine-generated SQL, developer and analyst skillsets, and architecture tradeoffs.

A good way to think about various SQL approaches is to segment them by latency characteristics. Below is an illustration that can serve as a framework:
There are broadly four classes or categories of SQL workloads:

**Batch SQL**

Apache Hive is designed for batch queries on Hadoop by providing a declarative abstraction layer (HiveQL - an incomplete subset of SQL), which uses the MapReduce processing framework in the background. Hive is used primarily for queries on very large data sets and large ETL jobs. The queries can take anywhere between a few minutes to several hours depending on the complexity of the job.

**Interactive SQL**

Technologies such as Impala and Apache Drill provide interactive query capabilities to enable traditional business intelligence and analytics on Hadoop-scale datasets. The response times vary between milliseconds to minutes, depending on the query complexity. Users expect SQL-on-Hadoop technologies to support common BI tools such as Tableau and MicroStrategy (to name a couple) for reporting and ad-hoc queries.

**In-memory SQL**

In-memory computing has enabled new ecosystem projects such as Apache Spark to further accelerate query processing. SparkSQL uses in-memory computing while retaining full Hive compatibility to provide 100x faster queries than Hive.
**Operational SQL**

Unlike batch and interactive queries that are used by business teams for decision making and operate as read-only operations on large datasets (OLAP), point queries are typically done by OLTP and web applications, operating over smaller datasets and typically include insert, update, and deletes. The expected latency is usually very low (e.g., milliseconds) due to the high volume of requests from these applications.

Most big data platforms such as Hadoop are designed to manage large amounts of unstructured data. There is no schema and no consistent format, so SQL queries can’t return reliable results without a tedious extract/transform/load procedure. This has caused some people to question whether SQL is too long in the tooth to be applied to big data. Maybe it’s best to move on to something else.

That thinking is changing, however. Because SQL is so well-known, the open source community has developed a variety of solutions that make SQL useful in both batch and interactive analytics applications.

The best-known of these is Apache Hive. First released in 2010, Hive can be used to process long-running batch queries on Hadoop that convert them to the MapReduce model. Hive has its own version of SQL called HiveQL that is well-tuned for use on very large data sets and large ETL jobs.

Newer technologies such as Impala and Apache Drill provide interactive query capabilities, which are a key feature for exploratory business intelligence (BI) applications. Impala and Drill also enable users to leverage popular BI tools such as Tableau and MicroStrategy on data stored in Hadoop. Drill is an especially interesting engine because, unlike Hive, it can query schema-less files, as well as JSON, Hive, HBase, MongoDB, and even cloud storage systems from Amazon, Microsoft, and Google. Whereas Hive builds a relational layer on top of Hadoop, Drill accesses the underlying data directly and can discover schemas on-the-fly, making it faster and better-equipped for exploration and discovery.

For fast in-memory processing, Apache Spark has an SQL module that provides a common way to access a variety of data sources using unmodified Hive queries. For use cases that involve embedding SQL within machine learning programs, SparkSQL is a better choice.

Mitsutoshi Kiuchi, a Senior Consultant with Creationline, has a great blog post that specifically compares Apache Drill and SparkSQL and the use cases each is best suited for.

To summarize, Apache Drill and Apache Spark support semi-structured data such as CSV and JSON, as well as relational databases through JDBC. Both are able to analyze a mix of structured data sources and semi-structured data sources. Both engines can execute queries on and transparently capture data from a wide variety of data sources, with Spark being able to conduct more varied processing than SQL queries.

If you need to perform complex math, statistics, or machine learning, then Apache Spark is a good place to start, but if you are considering Spark only for SparkSQL, then you should consider starting with Apache Drill.
In addition, most relational database and data warehouse vendors have developed their own engines for querying Hadoop data from within their own environments. So as you can see, the community has done a good job of covering all the major uses of SQL in the unstructured big data context. The following table compares and contrasts capabilities of some of the more common SQL-on-Hadoop technologies.

<table>
<thead>
<tr>
<th>Key Use Cases</th>
<th>Drill</th>
<th>Hive</th>
<th>Impala</th>
<th>Spark SQL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self-service Data Exploration / Ad-hoc queries</td>
<td>Batch / ETL / Long-running jobs</td>
<td>Interactive BI / Ad-hoc queries</td>
<td>SQL as part of Spark pipelines / Advanced analytic workflows</td>
</tr>
<tr>
<td>Data Sources</td>
<td>Files Support</td>
<td>Files Support</td>
<td>Files Support</td>
<td>Files Support</td>
</tr>
<tr>
<td></td>
<td>Parquet, JSON, Text, all Hive file formats</td>
<td>Yes (All Hive file formats)</td>
<td>Yes (Parquet, Sequence, RC, Text, AVRO ...)</td>
<td>Parquet, JSON, Text, all Hive file formats</td>
</tr>
<tr>
<td>HBase / MapR-DB</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Beyond Hadoop</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Data Types</td>
<td>Relational</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Complex / Nested</td>
<td>Yes</td>
<td>Limited</td>
<td>No</td>
<td>Limited</td>
</tr>
<tr>
<td>Schema-less / Dynamic Schema</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Metadata</td>
<td>Hive Meta Store</td>
<td>Yes</td>
<td>Limited</td>
<td>No</td>
</tr>
</tbody>
</table>

**CHAPTER 3: Navigating the SQL-on-Hadoop landscape**
Migrating basic SQL queries across traditional, MPP, and Hadoop warehouses is pretty straightforward. As with any programming language, using the standard constructs and avoiding vendor and platform-specific extensions make migration easier. However, consider how queries will be applied in the long-term when selecting which query language to use. For example, Hive has a number of Hadoop-specific extensions that don’t translate to the ANSI-standard SQL. That means that if you plan to load data from your Hadoop data lake into your data warehouse, you’ll have to modify the queries to take into account the different processing back-end.

Apache Drill has the advantage of being fully ANSI:SQL 2003 compliant and has the ability to discover schema on-the-fly. This reduces the amount of time needed for data modeling and schema development. Below are a few examples of SQL queries that show Apache Drill in usage as well as a high level architecture that shows how Drill works with various visualization tools such as Tableau, Qlik, and Spotfire, among others.
SELECT b.location.state AS state, b.location.city AS city, count(*) AS businesses from dfs.`/data/business.json` b
GROUP BY b.location.state, b.location.city
ORDER BY businesses DESC LIMIT 10;

// We’re in Vegas and looking for a good place for lunch
>SELECT name, category, rating
FROM dfs.`/data/business.json` b
WHERE b.attributes.`good for`.lunch = 'true'
AND rating > 3
AND b.location.city = 'LAS VEGAS'
ORDER BY rating DESC LIMIT 2;

Query data with any levels of nesting
SELECT b.name from dfs.`/data/business.json` b
WHERE b.business_id IN
(SELECT r.business_id FROM dfs.`/data/review.json` r
GROUP BY r.business_id HAVING SUM(r.votes.cool) > 10000
ORDER BY SUM(r.votes.cool) DESC);

CREATE OR REPLACE VIEW dfs.tmp.BusinessReviews AS
SELECT b.name, b.rating, r.votes.helpful,
  r.votes.cool, r.`date`
FROM dfs.`/data/business.json` b, dfs.`/data/review.json` r
WHERE r.business_id = b.business_id;

Get started with Apache Drill here — download it here.

Drill is designed for a wide set of use cases
The good news is that there is no shortage of choices when it comes to using SQL for data warehousing. The trick is to think of the use case and what makes the most sense, but also to keep in mind the long-term implications and growth of use cases.
System overload at Kobo

Kobo

Toronto–based Kobo is in the business of knowing what people want to read. The developer of E-Ink Readers, customized tablets and e-reader applications for the most mobile platforms, has an online bookstore that features more than four million titles for customers in 190 countries and 71 languages. Kobo collects exhaustive statistics about what and how its customers read and how they interact with its website. It even analyzes the texts of its titles to determine how to promote them. Kobo’s data store has quickly grown to more than 150 terabytes.

That was more than its SQL Server relational database could handle. Kobo needed to manage more data at a lower cost, but without disrupting the wide range of business intelligence tools it had in place.

Kobo can now load an unlimited amount of unstructured data from sources that include e-commerce sites, mobile apps, and raw text and metadata from books. It now performs clickstream analysis to track customer activity and generate recommendations in real time.

For business analytics, Kobo uses Apache Hive to extract data to tables that load into Tableau for business intelligence. Data scientists are also running MapReduce–based machine-learning algorithms to “teach” computers how to understand customers’ preferences. Thanks to the use of NFS by the MapR Platform, Kobo was able to do all of this without disrupting their existing IT infrastructure or even changing existing queries.

Kobo’s story is just one example in which big data is redefining the way companies think about data. No longer is analytics limited to cleaned and scrubbed records that have been
pre-selected for analysis. The process is more iterative, exploratory and, ultimately, more fruitful.

Data warehouses have been around a long time, but they have been limited to working with normalized records in a well-understood schema. They have also historically been proprietary and expensive, with costs of up to $30,000 per managed terabyte. That made data warehouses the domain of only the world’s largest companies and it made analytics a specialized and arcane discipline.

The cost barriers have now fallen thanks to a combination of open source technology, the Hadoop ecosystem and the power of commodity processors. Research firm Wikibon has estimated that processing big data with Hadoop is about 30 times cheaper per terabyte than it was using a traditional warehouse, and those costs continue to fall.

**Analytics**

But the bigger revolution is arguably in analytics. Analytics has become the domain of everyday business users instead of being solely for data scientists working with complex statistical languages. Visualization tools like Tableau are offering unprecedented ease of use, while sophisticated modeling engines like SAS are expanding to accommodate the exploding variety of data types that can be modeled. Machine learning and predictive modeling are also making analytics more forward-looking to enable businesses, for example, to anticipate customer behavior based upon previous patterns or changes in the environment.

One of the biggest changes in the analytics world sparked by big data is that outputs have become more prescriptive and process-oriented. In the past, analytics involved asking specific questions of historical data to prove or invalidate theories. For example, do you usually sell more ice cream in July than in November? Because the data domain was well known, there was little opportunity for discovery or exploration.

Introducing a Hadoop-based “data lake,” which contains a combination of structured and unstructured data, has changed the kinds of questions we ask.

For example, when analyzing ice cream sales, we can introduce public data like weather records, along with completely unstructured variables like new advertising campaigns or news events that might affect interest in ice cream or a particular brand of ice cream.

As we introduce new variables, the analytics engine produces not just calculations, but possible avenues of further inquiry. Machine learning is a recent development that applies algorithms that continuously churn through results to refine and improve queries.

**Until recently, there were basically two kinds of analytics:**

1. **Decision-support analytics**, which enables decision making by humans primarily through the use of visual tools such as charts and graphs

2. **Descriptive analytics**, which combines historical data along multiple dimensions to uncover insights
Big data has opened some new options

Streaming analytics processes information arriving continuously in real time to quickly identify patterns that may require action. For example, data from a flow meter in an oil pipeline can be monitored to detect changes in volume that would indicate that a rupture is likely to occur, or a data stream from a point-of-sale system could identify discounts or up-sells to present to the customer before checkout is completed.

Predictive analytics is the use of data, statistical algorithms and machine-learning techniques to identify the likelihood of future outcomes based on historical data. Predictive models use known results to predict values for different or new data. Results are typically represented with a probability rating based on confidence in input variables. One particularly interesting use of predictive analytics is correlation analysis, which looks for independent variables that correspond to similar behaviors.

Experimenting with shipping products that it thinks customers will want, even if they haven’t ordered them.

Machine learning is an evolution of the decades-old concept of artificial intelligence in which analytic models are built automatically by the computer itself. Iterative algorithms “learn” from the results of previous queries to continuously refine questions and identify hidden insights. For example, American Express built a machine learning mobile phone application that provides customized restaurant recommendations based upon recent spending histories and card member profile data. These recommendations are then given proactively to consenting customers.

Another notable example of machine learning is IBM’s Watson computer, which can sort through millions of medical case records and find correlations that reveal courses of inquiry that researchers should follow. Facebook, Google, and Microsoft have all contributed their machine learning engines to the open source community, so you can expect rapid advances in this area.

American Express uses machine learning to stop fraud and delight customers

Financial services giant American Express is using machine learning to improve decisions and better leverage data. It’s in the unusual position of being able to see data from both the customer and the seller sides of the business. The question was how to use that insight to improve decision-making.

About five years ago, American Express recognized that traditional databases would not be enough to effectively handle the level of data and analytics the company needed. It became an early adopter of Hadoop as a platform for machine learning.
American Express is privy to millions of decisions its card-holders make every day. If it could become just a little bit smarter about understanding those decisions, it could create a huge competitive advantage.

American Express chose three areas of focus: fraud detection, new customer acquisition and customer experience. Fraud detection is a huge area of concern for all credit card companies. Modeling methods must consult a variety of data sources including card membership information, spending details, and merchant information in order to stop fraudulent transactions before a loss is incurred while still allowing normal business transactions to proceed quickly. Fraud detection systems must flag suspicious events early and make decisions in a few milliseconds against a vast dataset. Machine learning has provided an improvement over traditional linear regression methods, taking the precision of predictions to a new level.

For new customer acquisition, American Express applied machine learning to personalizing interactions on its website. There are many different credit card plans from which to choose on the company’s website. When 90% of new customers came via direct mail campaigns, there was no chance to customize offers, but in recent years the amount of new customer acquisition via online interactions has risen to 40%. American Express uses machine learning to match visitors to targeted marketing campaigns and deliver offers that are most likely to appeal to individual visitors.

Machine learning at American Express

In the area of customer experience, American Express applied machine learning to build a mobile phone application that provides customized restaurant recommendations. The machine learning algorithm uses recent spending histories and card member profile data
to predict which restaurants a customer might enjoy based upon the customer’s history and preferences of others with similar spending patterns. Customers get a better experience and merchants get feedback on the quality of their offers.

The recommendation system uses large amounts of user historical and behavior data to train a machine learning model to identify recommendation indicators based on historical co-occurrence of users and actions. The computational heavy lifting to train the model can be done ahead of time, making the actual recommendations nearly instantaneous.

The successful use of machine learning at American Express required an infrastructure with the following characteristics.

**Scalability** – Models learn from motifs observed across a large number of historical actions, so scalable storage is required to support the database as the volume of transactions grows.

**Speed** – Fraud detection requires nearly instantaneous performance at scale. This is challenging to achieve when large numbers of transactions hit the network at the same time. The ability to predict these anomalies so that appropriate computing power and bandwidth can be allocated is essential to detecting fraud without disrupting customer experience. American Express is machine learning algorithms can query millions of records and return a response in less than one second.

**Compatible non-Hadoop File Access** – Big data systems usually need to bridge to legacy data stored in relational databases and flat files. American Express took advantage of MapR’s real-time read/write filesystem with NFS access to run legacy code alongside Hadoop applications.

**Data Versioning** – Building big data applications at scale requires version control so that machine learning algorithms can compare historical and current data. MapR’s transactionally consistent snapshots for data versioning made it possible for the systems to store training data from an exact point-in-time, so that old and new data could be compared to see what is really causing changes in behavior.

**Federation Across Multiple Data Centers** – American Express deployed its machine learning projects across multiple data centers either to share results or to improve the quality of source data. MapR mirroring was used to creates a consistent remote replica of data as well as to provide a disaster recovery platform.

Adapted from “*Machine Learning at American Express: Benefits and Requirements*” by Ellen Friedman

Self-service analytics is all about enabling faster data exploration. But unlike BI, which basically put an easy-to-use layer on top of a query engine, self-service analytics permits users to interact with the data directly. Software providers like Tableau, Qlik, and Datameer have written connectors and filters that enable non-technical users to integrate unstructured and public data sources into their analysis via Hadoop.
Visual analytics uses graphical imagery to display large and complex data sets in ways that make relationships and variables easier to identify. A simple example would be arrest records or home prices overlaid on a map. Basic charts and graphs have been available on BI dashboards for years, but visual analytics built on a big data layer makes it possible to incorporate a much wider variety of unstructured data. One common example is tag clouds derived from analyzing tweets. Tag clouds show quickly which topics are sparking the most discussion without requiring detailed examination of individual posts.

Sentiment analysis attempts to figure out how people feel about a topic from the words they use. This has long been considered one of the most challenging forms of analytics because of the many language and cultural nuances involved, but machine learning is making rapid progress. Among the many applications of sentiment analysis is measuring consumer attitudes toward brands and products and determining popular opinions on topical issues.

Preparing for analytics

Being successful in data analytics requires taking a disciplined approach to identify questions and relevant data.

1. **Identify the problem** - Involve business users at the front end to nail down specifics about what they want to accomplish. The more detail you can gather about need, scope, market conditions and goals the better. Defining specific data needs leads to the selection of tools and modeling techniques.

2. **Choose and prepare data** - Depending on the scope of the question and the type of data needed (structured vs. unstructured), locate, access, clean, and prepare the data. This could involve combining transactional data in a relational format, unstructured data in a format like JSON and existing analytical data in a warehouse. Platforms like Hadoop and NoSQL databases are ideal for combining multiple data types.

3. **Choose platforms** - Examining the types of data that will be required helps you choose the tools to manage it. Is the data structured or unstructured? Streaming or static? High- or low-volume? Will analysis be done mostly by data scientists or business users? Is there a need for an ETL process? A schema? There are open-source options for all of these scenarios.

4. **Explore data** - Use interactive analytics to take a first pass through the data in a data lake to identify relevant variables, trends and relationships. Visual tools are especially effective at this stage to pinpoint clusters and find outliers. Look for irrelevant variables that can be deleted as well as gaps in information. This is also the time to determine what transformation will be needed.
5. **Transform data and create models** - Bring a skilled analyst or modeler into the process to use statistical, data mining or text mining software to select and transform key variables. Use an iterative approach to find the models that produce the best results.

6. **Test and validate models** - Test and validate models with the business user. Document approved models with scoring criteria, code and metadata for storage in a model repository. This enables models to be reused and reviewed for quality control or auditing purposes.

7. **Deployment models** - Let users enjoy the fruits of your careful preparation.

8. **Review and assess models** - Evaluate the effectiveness of the model against real-world results. Remember that no model is perfect, and that model performance may degrade over time as variables change. Be prepared to revise or retire models that no longer deliver quality results.

These core practices will serve you well in every analytics scenario, whether batch or interactive. The most exciting new technology to come online is streaming or high-speed in-memory analytics. Let’s look at that next.
The move to in-memory

The big data industry has been abuzz over the past year about in-memory and streaming analytics. Interest was kicked off by the arrival of Apache Spark in early 2014, and has continued with other open-source innovations like Apache Flink, a real-time processing engine, and Apache Apex, both of which are top-level projects. On the commercial side, SAP HANA is a powerful alternative that has unique features for enterprise customers.

In-memory processing is up to 100 times faster than batch processing in some applications. One obvious reason is that all calculations are done in memory rather than written to disk, as with Hadoop’s MapReduce framework. Another factor is the use of a message queuing system like Apache Kafka, which can manage streams of data coming in from multiple sources at very high speed. Kafka is essentially the file management component of in-memory analytics processing. Hadoop still plays a role as the storage layer.

This kind of high-speed analytics is driving much of the innovation in the market right now. Research firm Wikibon has forecast that nearly 60 percent of all big data spending will be tied to in-memory or streaming analytics by 2026, up from just 6 percent today.

Some of the appeal of in-memory processing is apparent. Answers are quicker and data is fresher. Spark also resolves some of Hadoop’s complexity issues by providing a single administrative interface rather than Hadoop’s somewhat complex constellation of independent technologies.

Perhaps more importantly, in-memory processing opens up new applications of analytics. Data processing in real time enables decisions to be made rapidly in a way that may influence a process, such as delivering customized offers to customers at the checkout counter rather than afterwards in the mail. In effect, analytics become more prescriptive rather than descriptive. That’s one reason why predictive analytics has caught fire. By combining historical data with real-time process data, companies can predict how different scenarios will unfold and adapt accordingly.

The concept of online analytical processing (OLAP) has been around for years. OLAP enables end-users to perform ad hoc analysis of data in multiple dimensions to make better decisions. Data is usually extracted from production data sources and loaded into an analytical database using a multi-dimensional or “cube” structure. The cube lends itself well to
drill-down exploration using SQL statements to look at different slices of data. OLAP engines are built to handle a high volume of read queries, in contrast to online transaction processing (OLTP) engines, which are optimized for fast and repetitive write operations.

OLAP is a mature technology that is typically used with large volumes of structured data. However, big data has introduced new elements to this process by incorporating unstructured data, such as text messages and documents, typically in a data lake. This has created a need to build upon the OLAP concept. Real-time data adds yet another new dimension.

Real-time analytics is new. It provides a means for analytical operations that were traditionally done with OLAP to incorporate current data. When using Spark, real-time analytics can also merge structured and unstructured data, although some of the benefits of full SQL compatibility may be lost.

Traditional data warehousing and real-time analytics both have their value. For example, a traditional data warehouse can dig into years of data to determine that certain products sell best on certain days of the week. Real-time analytics can be used to monitor weather data or a stream of tweets to forecast that sales will spike on a particular day or in the next two hours.

What’s the deal with streaming?

Streaming analytics is stateless, which means there is no persistent memory. While data may be saved for later analysis, the stream processor is generally limited to running simple computations across a stream of arriving events, comparing streamed data to a baseline or triggering a program if an anomaly is detected. Below is an example of an architecture that illustrates how companies are incorporating newer event streaming technologies such as MapR Streams.

Stream processing solution template:
OLAP uses stored data to make calculations that involve much larger amounts of information. Data may be stored in memory or on storage media. OLAP is usually used on structured data, and is typically associated with data warehousing.

To use a simpler example, by using streaming analytics one can identify that Citigroup stock is up two dollars today and can trigger a sell order. OLAP can be used to correlate Citigroup stock to exchange rates, unemployment statistics, oil prices and other macroeconomic factors to see if they correlate with today’s stock price.

By combining OLTP, OLAP, and streaming capabilities using unstructured or semi-structured data, users can realize new capabilities. For example, operational business data can be combined with external data sources for more powerful analytics. Data management is simpler because all data can be contained on a single Hadoop cluster, thereby alleviating integration challenges. OLAP models also have access to larger volumes of data distributed across thousands of nodes.

The combination of these processing models can yield insights that were impossible with legacy OLTP and OLAP technologies. For example, retailers can correlate Twitter messages with business metrics like sales and profitability to determine whether customer sentiment impacts sales. This can then be used to adjust inventory and even prices. Tweets are a combination of semi-structured and unstructured formats and are disorganized, but when combined with sentiment analysis and behavioral data they can yield fascinating relationships that would be otherwise invisible.

Real-time analysis isn’t right for every application. It requires fast processors, ample memory and instantaneous access to data, all of which consume resources. When crunching through years of historical data to detect long-term patterns, traditional data warehouses and data mining with Hadoop clusters can usually do the job just fine. When you need to make decisions on the spot, OLAP and streaming analytics come into play.

When considering what tools to use for your own big data projects, you need to be aware of the concept of the time value of your data. If decisions can be made based upon structured transactional data, and if split-second decision-making isn’t critical, then OLAP is sufficient. If analysis of unstructured data like email promotions, news reports or social media data is involved, then consider a data lake powered by an engine like Spark on Hadoop. If split-second decision-making is required based upon data streaming in from a cash register, motion detector or temperature sensor, then Spark or Apache Flink are options to consider.

If the application is self-contained, repetitious and uses a combination of transactional and unstructured data, then an integrated HANA analytics engine is a good choice to provide excellent performance, flexibility, and high scalability with a minimum of data scrubbing and preparation.

Big data was traditionally backward looking, focused on historical data. Streaming analytics opens up the potential of current data. Now there’s a new class of analytics that looks forward and predicts the future. We’ll look at that next.
The SAP HANA platform

A new breed of products combine OLTP functionality with in-memory analytics to provide the best of both worlds. SAP’s HANA is one example. It’s an in-memory data platform for real-time analytics and applications, but it sits atop an OLTP engine so that ETL is minimized and analytic results are much fresher.

The SAP HANA platform is actually much more than just a database. HANA is a collection of predictive analysis and business function libraries, text search, application services, calculation engines and information modeling tools layered on top of an in-memory database. This provides the ability to perform predictive and text analysis on large volumes of data in real-time. HANA is also tightly integrated with R, a programming environment for statistical computing and graphics. It includes text search and analysis capabilities that are ideal for analyzing unstructured data in a data lake, such as text documents, Twitter messages, and email.

The fact that HANA is both a transactional and an analytical database has significant benefits in both timeliness and ease-of-use. Traditional data warehouses require organizations to extract data from their production databases, transform it into a consistent format, and load it into an analytical database, a process called ETL. This effort can be highly manual and take hours or even days for large data sets. A recent study by Xplenty Ltd. found that one-third of business intelligence professionals spend over half their time cleaning up raw data to load into analytics platforms.

HANA enables an organization to capture transactions and report on them in real time, which skips the entire ETL process. While some other commercial products provide OLTP and OLAP functions in the same box, none can accommodate these two to the degree that HANA can.

SAP HANA Vora

SAP also provides a product called SAP HANA Vora. It is a distributed computing engine that integrates with Apache Spark and makes OLAP modeling on Hadoop data accessible via the well-known SQL language. It also tightly integrates with HANA to give users the capability to perform real-time analytics on the combination of HANA and Hadoop data. For example, users can expose HANA data to the Hadoop environment and use Vora’s computing engine and OLAP Modeler for real-time drill-down exploration. Or they can expose Hadoop data to the HANA environment and build applications leveraging the HANA platform, for use cases where big data analysis needs to be incorporated into transactional workloads in real-time. (Try out a HANA + Vora use case lab experience for free.)

The architecture below illustrates how MapR and SAP HANA Vora come together to bridge the gap between Hadoop, HANA, and enterprise data.

Contextual analysis across SAP HANA, Hadoop, and enterprise data:
Use cases

There are many unique applications of OLAP that extend beyond the traditional data warehousing model.

- Retailers can offer discounts, coupons, cross-sell, and upsell opportunities to customers while standing at the checkout counter based upon analysis of the contents of their shopping baskets.
- Credit card processors can spot and block potentially fraudulent charges based upon recognized patterns before the transaction is committed and money is lost.
- Systems administrators can run analytics on system logs in real time to look for anomalies that might indicate potential bottlenecks or equipment failure.
- Algorithms can detect changes in buyer behavior — such as stocking up before a snowstorm or a local football game — and adjust inventory levels accordingly, avoiding shortages, and increasing sales.

One relevant example is Hong-Kong-based furniture maker Man Wah. It boosted sales 13% and improved the productivity of its financial staff by one-third by replacing a time-consuming financial reporting process with real-time analytics based on SAP HANA. Instead of printing out reports at the end of each month — a process that by itself consumed a half-day of time — the company gave individual business units the ability to customize reports to their own requirements based on up-to-the-minute data.

**Man Wah** sales people can now predict changes in customer demand, and the procurement department can forecast market hotspots in advance to adjust production levels and minimize shortages. Instead of waiting four months to determine the profitability of an or-
der, Man Wah now has that information before the order is even signed. That allows it to adjust sales and pricing strategies promptly.

Luxury auto maker Mercedes-AMG tied HANA to its engine sensors as part of its rigorous testing program. A typical engine employs 300 sensors and can generate up to 30,000 data elements per second. Engineers used to have to wait until the nearly one-hour test was completed to analyze the data and spot anomalies. With HANA, they can correlate live sensor data with historical data to pinpoint potential performance problems instantly. The process has cut one full day from the weekly testing cycle and increased the number of processes that show improvement by 76%.
Making educated guesses

Machine learning and predictive analytics are two manifestations of the same concept: using algorithms to make educated guesses. Both disciplines use historical data to underlie their assumptions, but they express the outcomes differently. The popularity of both disciplines has surged with the arrival of big data for a couple of important reasons.

One is that predictive analytics and machine learning both require large amounts of data and processing power. Low-cost, scale-out file systems like Hadoop have brought the price of data management way down, and open source software running on commodity servers has done the same for CPU cycles.

Another reason is that these tools require iterative analysis, often involving unstructured data. For example, recommendation engines, which have elements of both predictive analytics and machine learning, dig through preferences and behavioral information in a data lake to find relationships between data elements that aren’t obvious on the surface. That means constantly creating new relationships on the fly to test and score them. It also means that different data types need to be combined and measured, such as Tweets and transactions.

Relational databases aren’t good at dealing with any of this, but graph analytics handles it without a hiccup. Machine learning and predictive analytics are transforming the role of computers from repetitive records management engines to intelligent advisors. Here’s a quick overview of these disciplines.

Predictive analytics

Data mining is good at telling us what has happened in the past. Predictive analytics uses historical and current data to make predictions about the future, using a combination
of statistical, data mining, machine learning, and analytic techniques to assess risk within a particular set of conditions.

The simplest form of predictive analytics, is one every driver is familiar with: the “check engine” light in your car. Sensors in the engine compartment continually sample data about operating conditions and compare it to preset thresholds that indicate something is amiss. If the data from sensors falls outside the set parameters, a warning is triggered. The automotive sector is undergoing significant changes with the Internet of Things and data from sensors and devices. Below is an example architecture that illustrates how data from cameras and sensors on vehicles gets processed, stored, and analyzed.

**MapR automotive IoT architecture:**

Predictive analytics are a lot more sophisticated than that, but the concept is similar. There are many applications, ranging from predictive maintenance to marketing. For example, predictive analytics can be used to predict the success of email promotions by identifying patterns common to previous successful campaigns. Retailers use predictive analytics to determine the likelihood that customers who buy certain products will also buy other products. Not surprisingly, the insurance industry thrives on predictive analytics.

In healthcare, predictive analytics can determine the likelihood that patients will develop certain conditions based upon existing conditions, family history, and even DNA. There are a dizzying number of variables to consider, and huge amounts of unstructured information to process, but today’s engines take it all in stride. Below is an architecture that paints the picture of what a modern healthcare system looks like.

**MapR healthcare architecture:**
This enables companies like the Next-Generation Sequencing, a new approach to DNA analysis that sequences genomes much faster than was previously possible. Faster sequencing means more rapid diagnosis.

Genomic research involves massive databases, so Hadoop has been a godsend to researchers who need to manage them cost effectively. The real payoff comes when data from multiple sources can be correlated to find relationships that aren’t obvious to human researchers.

Thanks to Hadoop and Apache Spark, Novartis Institutes is able to represent diverse data as a vast knowledge graph with trillions of edges that is stored in their distributed file system and manipulated with Spark. Predictive analytics let researchers easily model complex and changing biological data sets via graph manipulations. In this way, they can bring together research from many sources to identify factors that predict the development of medical conditions and identify drugs that can treat them early.

One particularly promising new application of this technology made possible by the Internet of Things (IoT) is in the equipment maintenance field. Companies ranging from manufacturers to bus lines have huge investments in machinery, and machines break. Predictive maintenance in the oil and gas industry helps increase efficiency and reduce operational costs, and this is a fast growing solution for MapR.

The historical approach to maintenance has been to do what we do with our cars: schedule service every 5,000 miles whether we need to or not. That’s the price we pay to not end up stranded on the highway.

If machinery could be continually monitored and analyzed to look for signs of impending failure, then repairs could be scheduled more efficiently, less money wasted on needless replacement, and downtime would be minimized. The combination of IoT-enable sensors and predictive analytics is making that real, and revolutionizing equipment maintenance in the process.
You might ask what distinguishes predictive analytics from ordinary analytics? After all, the purpose of mining data has always been to turn up insights that lead to better decisions.

The principal difference is in output. Unlike traditional analytics, predictive models yield prescriptive results that suggest specific courses of action—“check engine” lights, if you will—or forecast the likelihood of a specific outcome. It’s part of the transition of enterprise computers from “systems of record” to “systems of intelligence.” The output of data analysis in the future will increasingly be advice or even decisions. In one notable example, Amazon is testing a service that uses predictive analytics to deliver certain types of products—like books—to people before they have even ordered them based upon preferences shown by past behavior.

**Machine learning**

Machine learning is subset of predictive analytics that is a modern take on artificial intelligence, a concept which dates to the earliest days of computing. Machine learning is a better term, though. Rather than broadly mimicking human intelligence, computers are best at doing specific tasks well. Machine learning uses self-teaching algorithms that continually get “smarter” as they repeatedly cycle through data and look for patterns that lead to desired outcomes.

A simple example of machine learning is a spam filter. When a message hits your mail stream, the filter initially matches things like the sender’s domain and IP address to a list of known spammers. It then matches the contents to a set of rules about words, phrases, punctuation, salutations, use of images, and other patterns that are common to spam messages. It uses a scoring algorithm to determine the likelihood that the message qualifies. Based on thresholds set by the user, it either lets the message pass or dumps it into the spam folder.

Over time, the machine learning engine continuously combs through messages in the spam folder and looks for new patterns that are common to offending items. It tests these observations continually to figure out which correlations matter. It may even compare its notes to other spam filters on a network to improve the intelligence of the whole group. This constant iterative process is why spam filters are so sophisticated and keep getting better.

**There are three basic types of machine learning:**

**Classification** – This is the spam technique we just described. Classification takes a set of data with known characteristics specified by a supervisor and learns how to label new data with similar characteristics. The criteria are typically pre-defined so that the machine works quickly. This speed is one reason classification is effective in cases like credit card fraud detection, where the patterns are well known and split-second decisions are critical.

**Clustering** – Clustering algorithms discover groupings that occur in collections of data and assigns records to categories based upon similar traits that the machine discovers without supervision. An example of clustering is customer profiling. By matching sales data
to known demographics, a clustering algorithm can classify customers into groups such as early adopters, value-seekers, and bargain hunters. These clusters may be invisible to humans. In fact, clusters often turn out to form around factors nobody anticipated.

**Collaborative Filtering** – This technique is based upon the assumption that people who have similar tastes will tend to like the same things. A well-known example of collaborative filtering is Netflix’s recommendation engine. It determines which programs you’re likely to watch based upon the choices of others like you. Amazon’s recommendation engine works the same way.

All machine learning algorithms use a testing mechanism that can be described most simply as a fancy decision tree. The learning process begins with a set of test data and assumptions provided by a user or another program. The algorithm then runs through the sample data repeatedly to test the assumptions and determine if they match the expected outcome. If they don’t, different assumptions may be used. Data that has a high correlation value to the desired result is assigned a higher ranking, and the model is run repeatedly as additional data is added. Over time, patterns become clearer and more variables are tested.

For example, a machine learning analysis of the NCAA basketball teams most likely to reach the final four in the March Madness tournament might sift through 20 years of data about successful teams and look at statistics like free-throw percentage, field-goal percentage, turnovers, average points per game, the seniority of the players, and records against certain other teams. As it compares these statistics to success on the court, it keeps the data points with the highest correlative value and discards others. Then it delivers a model that can be used to size up the teams in this year’s tournament.

The technique of modeling relationships between objects is called graph processing. Graph databases are used to attack big modeling projects like recommendation engines or customer profiling. Users of Apache Spark can apply the GraphX API to quickly create and change relationships to be tested by a machine learning library.

Below is a matrix to help understand the different technologies available for predictive analytics and machine learning. This includes tools from the Hadoop ecosystem as well as more proprietary tools used by enterprises.
Big data platforms enable analysts to store and process unstructured data at unprecedented scale, which makes them ideal for machine learning. Managed service provider Solutionary has a security and compliance platform that analyzes network traffic for its clients on a massive scale, looking at patterns of behavior, anomalous activities, and attack indicators. It then enriches and correlates data across global threats and trends to provide context and actionable alerts to clients. As the number of clients grows, so does the quality of its security detection technology.

Using Hadoop and Spark, Solutionary can now globally detect and analyze activity across all clients within milliseconds, whereas before adopting these technologies the
same task would take thirty minutes. In computer security, a half hour is enough time for a distributed denial of service attack to immobilize a company, so time is precious. Solutionary is now using machine learning, predictive modeling, and analytics to improve detection of Advanced Persistent Threats, which are one of the most damaging forms of attack and also one of the most difficult to sniff out. As Solutionary continually feeds new data into its machine learning algorithms, the system steadily improves its detection capabilities.

**Tools and terms**

There are many tools to choose from in this category. A few notable open source options include:

- **Apache Mahout** is a popular library of highly scalable algorithms that offers many options for building and deploying machine learning projects.
- **GNU Octave** is an interpreted language primarily used for numerical computations and linear and nonlinear problem solving.
- **Konstanz Information Miner (KNIME)** provides a data analytics platform that can be used to analyze trends and predict potential results using a visual workbench that combines data access, data transformation, initial investigation, predictive analytics, and visualization.
- **MLlib** is a scalable machine learning library for Apache Spark. It provides common learning algorithms and utilities for processes like classification, regression, clustering, collaborative filtering, and dimensionality reduction. While not a programming language in itself, it automates much of the coding work required to create machine learning solutions.
- **OpenNN** is an open source class library for building neural networks, oriented toward advanced users who have good C++ skills.
- **PredictionIO** is a machine learning platform that comes with a large library of built-in routines and is said to be easier to learn than alternatives. Salesforce.com acquired PredictionIO in early 2016 but said it will continue to maintain the code as open source.
- **R** is a programming language that originally was created to enable academic statisticians and people with extensive programming skills to perform complex data statistical analysis and display the results in publication-quality graphics. R has grown significantly in popularity over the last few years for use in big data analytics by businesses. Instead of having to write complex R code, users can now create sophisticated data models using point-and-click interfaces. R is now widely supported by analytics vendors.
- **TensorFlow** is a software library for numerical computation using data flow graphs. Originally developed by Google and released to open source in late 2015, it is extensively used by the search giant in speech and image recognition.
- **Torch** is a scientific computing framework with wide support for machine learning algorithms using graphical processors. Torch is said to be easy to use and efficient, thanks to an easy and fast scripting language.
Now that we’ve covered the major categories and uses of predictive analytics at a high level, let’s look at how to put them into action in your organization.
Your Turn – How to Make Analytics Work for You

8 Steps

Whether you’re tackling your first analytics project or building on a solid big data foundation, there are certain tried-and-true steps you can follow to keep yourself focused on goals and minimize risk and scope creep. Here are eight steps to this process.

1. **Determine the problem you’re trying to solve or the opportunity to uncover.** This process should be driven by the business, not by the IT organization. Don’t bite off more than you can chew, particularly if you’re new to analytics. Look for a manageable task with few variables and clearly defined deliverables. When possible, seek tasks that lend themselves to incremental results.

2. **Understand the value of the outcome.** This helps in budgeting and calculating ROI. Is the outcome of this project going to uncover a $100 million revenue opportunity or cut your equipment maintenance expenses by $10 million? If so, then you can afford to be pretty generous with the resources you throw at it. But if it’s going to save you $2 million and cost $3 million to implement, either redefine the scope or adjust your budget.

   If the outcome is speculative or uncertain, ask the business stakeholders to agree on a percentage of likelihood that the objective will be achieved. Also factor in profit margin. So if the $100 million revenue opportunity has a 70% likelihood of being achieved, it has payback potential of $70 million. If the profit margin on that new business is 50%, then the return for ROI purposes is $35 million.

3. **Determine criteria for success.** The problem with big data projects is that they can go on forever because there is always more data to consider. Projects also can suffer from scope creep, as stakeholders continually add questions that can take you down new rat holes. Successful projects are defined as much by what they don’t tackle as what they do. Get agreement from the business side on what success looks like, being aware that perfection is the enemy of good enough.

4. **Define your data needs.** This critical stage will impact your project on several dimensions, ranging from budget to time to tool selection. Among the questions to ask:
• Do you already have the data or do you need to acquire it?
• If you need to acquire the data, what cost and time commitments will be needed to get it?
• Is the data you need publicly available or do you have to pay for it?
• How much scrubbing or pre-processing will you need to do?
• How timely must it be?
• How much do you need to store, and how quickly must it be retrieved?
• What are your needs for retention and archiving?
• Are there any regulatory or privacy issues to consider?
• Do you need to take any special security measures? If so, at what cost?

5. **Identify tools.** Determine first if the desired outcomes can be achieved with tools you already have. There’s no reason to bring in Hadoop and Spark if your RDBMS or data warehouse will do the job just fine. Start with the functionality you need, then work backwards to choose tools. If you’re experimenting with new tools, start with a project that you know you can deliver and bring in shiny new objects on the periphery to test for additional value.

Don’t forget scalability. Will the tools you choose grow to accommodate the raging success of your project? Consider both open source and proprietary commercial platforms. As explained below, there are merits to each. Most importantly, never bet the farm on an unfamiliar technology. The needs of the business should guide your technology choices. There is no faster way to fall flat on your face than to put your faith in a technology based on sales pitches or promises.

6. **Determine resource gaps and shortfalls.** Figure out whether you have the in-house resources to complete the project or if you need outside help. Determine whether you have the necessary equipment and software or if you have to acquire it. Using a cloud platform is an excellent way to minimize the cost of acquiring new equipment. Equally important, consider what skills your staff will need to acquire to tackle the project. Training takes time and money, so factor those costs into the equation.

7. **Think agile.** Projects that deliver incremental results have a better chance to succeed than big all-in-one solutions. They also adapt better to changes in personnel, budgets, strategies, market conditions, and other factors that invariably occur. Take a page from agile programming and look for ways you can divide the task into pieces that deliver value continuously and keep the momentum going.

8. **Start small.** Use a sandbox approach with a sample of data and a subset of deliverables. Test results to see if they meet expectations. Then layer in additional data and technology, while testing continuously. Data lakes are an excellent exploratory vehicle.
The open source equation

Open source licensing is revolutionizing software by democratizing access to technology that was once available only to large enterprises. There is a myth that open source is cheaper than commercial software. After all, there’s no license cost. But the reality is that open source software comes with built-in expenses that can actually make it more expensive in a number of cases.

There are several compelling benefits of open source:

- It’s free. Software that may carry six-figure licensing fees in the commercial world costs nothing to download and use. This can save significantly on capital expenditure costs.
- If the open source package you choose lacks features you need, you’re free to customize it.
- Enhancements and bug fixes are made quickly because a large community of developers is constantly improving the code.
- Training and support resources are inexpensive or free in keeping with the open source culture.

Be careful, though.

These benefits generally apply only to the most popular open source projects. A lot of open source software never builds a substantial user base and can languish for lack of support. Switching costs are also low, which means there’s a greater possibility you can be “marooned” with an open source package if users move en masse to something else.

Commercial software still has some appealing benefits:

- “One throat to choke.” A commercial vendor is under pressure to make sure its software performs as advertised, and the vendor is accountable for any fixes that need to be made.
- Support resources are generally better with commercial packages, particularly telephone or on-site support personnel, who can address problems that individual customers are experiencing.
- Functionality may be superior, particularly in vertical markets or special use cases that don’t have a critical mass of community interest, but represent an attractive commercial opportunity.
- Usability is often better. User interfaces are a notoriously weak feature of open source software because the techies who develop the tools don’t need it. Witness the struggles Linux on the desktop has had against less functional but easier-to-use alternatives on PCs and Macs.
It’s possible to have the best of both worlds. Many popular open source solutions today are available in both free “community” and commercial editions from vendors. The community versions are feature-rich but may lack enterprise-class support or advanced features that are available as closed-source options.

Users can thus freely test solutions before making a commitment. They may never need to upgrade to get 24X7 support or special options, but they have the freedom to do so if they need to. In the meantime, they can continue to enjoy all the benefits of the open source model outlined above.

One example of this hybrid licensing model is the MapR Converged Data Platform. This powerful big data suite is available in a Community Edition for free usage, as well as a Converged Enterprise Edition for critical deployments requiring high availability and other enterprise-class features.

The code base for both products is the same. Both include MapR-FS (POSIX file system), MapR-DB (NoSQL database management system), MapR Streams (global publish-subscribe event streaming system), Hadoop, and related open source projects. The Converged Enterprise Edition has some advanced features that the Community Edition lacks, such as advanced multi-tenancy, global replication, 24X7 support, and integrated disaster recovery. Users of the Community Edition are free to add these features on their own, but many enterprises appreciate the convenience of a packaged solution and the highest quality of technical support.

The hybrid model makes it economically feasible for software development companies to offer free editions of their products while still generating the revenue that goes back into development. Although some MapR code is proprietary, the company participates in and contributes to the open source community and all of the APIs for the platform are open and standard.
Winning characteristics

This section of the ebook features snapshots of leading-edge organizations parlaying strategies for big data analytics to drive tangible business value. Their experience is where the rubber meets the road, where claims and hype meet reality.

First, here are some characteristics that most of these organizations shared in common as they ventured into big data exploration and exploitation.

• Big data is new data, so discover new value in it. Hadoop-based big data analytics is not simply a new way of processing and analyzing the same data you’ve been using. It’s all about exploiting untapped data. This means discovering new facts, new correlations, and new business opportunities.

• Stress ease of use. The most elegant and intelligent data solutions are those that conceal the immense complexity underlying unstructured data analysis, and present non-IT analysts with self-service options and tools that are truly easy to use. This is much simpler said than done. But ease of use accelerates productivity as well as overall ROI of big data analytics as more workers exploit it.

• Follow the data. Data today is everywhere and anywhere, and the tools to explore and exploit it have to go where the data is. That means developing an IT strategy that supports multiple platforms from RDBMS to Hadoop, and distributed queries across them. As outlined in detail in Chapter 2, it’s all about coexistence between Hadoop and traditional processing systems, not replacement or even competition.

• Different data, different best practices, different skill sets. Unfortunately many of the time-honored best practices for data management just don’t translate to the world of big data analytics, as least not during the discovery process, early users have learned. When working for example with highly unstructured data like video and social media, some early adopters have turned to data serialization technologies that
facilitate capture and storage of such data. All of this means the IT staff needs to acquire some new skills suited to the brave new world of big data. But isn’t that what IT staffs have been doing for the last 50 years?

**UnitedHealthcare**

UnitedHealthcare had a 10-terabytes-big problem. That is the volume of claims data its Payment Integrity group had to manage every day. That data was ad-hoc, heavily rules-based, and siloed randomly hither and yon. That was particularly problematic for a group tasked with the responsibility of paying claims on time and correctly.

IT leaders at UnitedHealthcare believed the path out of this data bog involved transitioning to a predictive modeling environment grounded on a Hadoop big data platform. But they wanted that platform to integrate any new tools or technology; it had to have multi-tenancy capabilities, and the platform had to have Direct Access NFS for direct ingestion of complex data sets from a variety of sources.

Working with MapR, the group set about building a predictive analytics ‘factory’ to identify inaccurate claims, but in a repeatable way. Hadoop is UnitedHealthcare’s framework for a single platform to handle structured and unstructured data from a wide variety of internal and external sources.

The end results are truly eye-popping, including a 2,200% return on UnitedHealthcare’s big data project. In fact this project was so successful for big data analytics that IT leaders are campaigning to have other parts of the company embrace Hadoop and big data. Areas of the business they have identified cover their clinical business and care provider networks, and in improved overall customer experience.

**TransUnion**

TransUnion never had a data problem, per se. As a credit reporting and information management leader, the company routinely gathered a mountain of data from clients across a broad swath of industries.

The challenge was finding a way to enable TransUnion’s analysts to experiment within and across data domains in order to innovate, outmaneuver, and outsmart the competition. A legacy technology stack was just not going to cut it.

TransUnion wasn’t seeking to replace its legacy systems, which did a fine job handling mission-critical customer facing applications. Instead, IT leaders opted for a hybrid architecture that kept traditional systems in place, but surrounded them with a Hadoop-based platform for emerging big data assets.

Hadoop and the tools that comprise the Hadoop ecosystem have proven to be a great fit for the new generation of data scientists. TransUnion has been hiring people proficient in tools like R and Python. Data visualization tools like Tableau are further enabling some analysts to dive deeply into data lakes containing information they always knew they had, but really couldn’t exploit with legacy tools.
What’s more, the productivity of analysts on this Hadoop platform is inspiring others at TransUnion to polish their skills as well, further leveraging the Hadoop investments.

**Harte Hank**

_Harte Hanks_ routinely pulls tidal volumes of data from emails, social media, transaction histories, and many other sources as it provides a multidimensional view of its customer’s clients. That’s what database marketing firms do.

Over time, these data sources have become increasingly comprised of unstructured big data. To store and crunch streams of such data, Harte Hanks started out using a free version of Hadoop.

To take its big data analytics game up several notches, the company replaced that with the Hadoop distribution from MapR. The main reason for the switch was due to the multi-tenancy capabilities delivered by the MapR Converged Data Platform. This allowed each customer of Harte Hanks to easily scale up or down according to individual demand.

In addition, the MapR Platform is highly resilient with no single point of failure in the entire system. If any nodes on the Hadoop cluster goes down, data is replicated to other nodes. Ultimately, the MapR Distribution has given Harte Hanks vastly improved customer service through deeper analysis and flexibility of stored data, and increased database capacity and performance.

**comScore**

_comScore_ and digital data are practically synonymous. Since its founding nearly 17 years ago, the company has provided digital marketing intelligence and digital media analytics in online audience measurement, ecommerce, and advertising.

To care for clients that include the world’s biggest digital companies (Google, Facebook, Twitter, and Microsoft, to name a few), comScore captures nearly 2 trillion interactions per month — equal to 40% of the monthly page views of the entire internet. That’s big data on steroids.

As client data volumes grew astronomically, along with customer demands for speedier service, comScore IT leaders upgraded to a super-robust, open source platform for its data processing. That platform is the MapR Converged Data Platform, chosen for its scalability, high availability, ease of use, and data protection.

**How are things working out?**

IT leaders report that comScore can more easily manage and scale the Hadoop cluster. The company can create more files and process data far more quickly, giving comScore a 3X performance increase running the same data and code as before. Looking ahead, comScore will leverage Hadoop’s preeminent scalability, having already grown the Hadoop cluster to more than 450 servers with 17,000 cores. Meanwhile, the cluster contains more than 10 petabytes of disk.
We noted in the very beginning of this ebook that “the disruptive potential of big data is vast and the acceptance and adoption of BI and advanced analytics on big data is absolutely essential for success in virtually every organization.” That’s a bold statement, but one for which the intervening chapters have offered substantial proof. Organizations and businesses across every vertical market are already leveraging Apache Hadoop and big data analytics to uncover the substantial treasures buried in mountains of unstructured data, which today comprise 75% of all new data being gathered. The deep insights gained almost immediately create advantages over competitors that are slow on the uptake of big data analytics on Hadoop.

How quickly is this brave new world evolving? Consider that just ten years ago, virtually none of the now vast array of today’s open source analytics tools were available. Today, there is a rapidly growing vendor ecosystem of Apache Hadoop and big data analytics solutions providers. The potential of these tools to completely change the way we think about data and its application has galvanized this market.

Below are a few key recommendations to help guide your efforts in using data to make your organization all that it can be.

Just because Hadoop is 100% open source, doesn’t mean you should roll your own. Would you build a car from scratch just because all the parts are available in the public domain? Unless you are a Formula One racing team, no! In theory, you could assemble a team of crack in-house IT staff, add in some consulting help and try to build out your own Hadoop platform. Not many organizations do that, however, and there is a good reason why.

There is no match for the expertise, skill and financial commitment of well-established Apache Hadoop vendors already selling Hadoop distributions that are ready to rock your big data world. The better ones have well-established, ongoing technical relationships with the world’s leading analytics, networking, hosting, and compute engine vendors, such as: SAS, Cisco, SAP, Amazon, Google, HP, Teradata, and many others.

Just consider that the overwhelming number of companies out there whose core business is not software development. Don’t be lured by the open source siren and go it on
your own unless you’re the adventurous type. Much of the work has already been done for you.

Don’t hastily part with what has worked. As we pointed out in chapter 2, by far the preferable strategy to follow now and in the foreseeable future is one of productive coexistence between traditional RDBMS now in place and the Apache Hadoop clusters being deployed. They are not the same, by any means. The traditional systems still do certain vital tasks better and faster than emerging systems, especially when the data is truly relational. Plus, vendors have had some 30 years of experience developing the tools to exploit that data efficiently. In maybe 10 years, possibly less, we could be talking about RDBMS as ‘sunset’ technology. But not now.

Get familiar with all analytics options for big data. In chapter 4, we showed that we have swiftly moved from basically two types of analytics in the RDBMS world to at least a half dozen in the Hadoop big data analytics sphere. Every organization has specialized needs. Now you have that many more options to select the most highly specialized, targeted analytics solution. A trusted partner can help you to decipher the nuances among these various options, ever mindful of emerging developments that can impact a decision you make today.

Go with the power of 2. Just as we advocate coexistence with RDBMS, we also recommend you get the best of both worlds when it comes to combining OLAP functionality with the newer in-memory analytics. The SAP HANA iteration we reviewed in chapter 5 is one proven solution, with scores of verifiable use cases already available. New technologies are doing away with the arduous extract/transform/load process and making it possible for you to detect trends and predict the future using live data. Consider the possibilities of transforming your business and your customer relationships with that kind of power at your fingertips. Which goes to our final recommendation.

Think different.

We borrow from Apple’s famous — if slightly ungrammatical — slogan for this item. Human beings are predisposed to incremental change, but the opportunities with big data are much greater than that. We’ve given you just a taste of some of the examples in this ebook. Imagine what you could do if you knew details about each customer’s preferences, priorities, and engagement history and were able to put that knowledge into action right at the point of sale.

Think of the possibilities of teaching machines to comb through years of operational data and find efficiencies that no human analysts would detect. Imagine the value of being able to respond instantly to a change in the market driven by a news event or popular new fashion because you were tapped into conversations in social media. These are just a few of the possibilities that big data analytics presents.

Thanks to the new economics of software, they are now within reach of anyone.