UNIT 2

# BIG DATA TECHNOLOGY

## 2.1 BIG DATA TECHNOLOGY

Technology is radically changing the way data is produced, processed, analyzed, and consumed. On one hand, technology helps evolve new and more effective data sources. On the other, as more and more data gets captured, technology steps in to help process this data quickly, efficiently, and visualize it to drive informed decisions. Now, more than any other time in the short history of analytics, technology plays an increasingly pivotal role in the entire process of how we gather and use data.

### 2.1.1 The Elephant in the Room: Hadoop's Parallel World

There are many Big Data technologies that have been making an impact on the new technology stacks for handling Big Data, but Apache Hadoop is one technology that has been the darling of Big Data talk. Hadoop is an open-source platform for storage and processing of diverse data types that enables data-driven enterprises to rapidly derive the complete value from all their data.

We spoke with Amr Awadallah, the cofounder and chief technology offi cer (CTO) of Cloudera, a leading provider of Apache Hadoop-based software and services, since it was formed in October 2008.

**He explained the history and overview of Hadoop to us:**

Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models. The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers. Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage.

The name "Hadoop" itself comes from Doug 's son, he just madethe word up for a yellow plush elephant toy that he has. Yahoo! hired Doug and invested significant resources

into growing the Hadoop project, initially to store and index the Web for the purpose of Yahoo! Search. That said, the technology quickly mushroomed throughout the whole company as it proved to be a big hammer that can solve many problems.

In 2008, recognizing the huge potential of Hadoop to transform data management across multiple industries, Amr left Yahoo! to co-found Cloudera with Mike Olson and Jeff Hammerbacher. Doug Cutting followed in 2009.

Moving beyond rigid legacy frameworks, Hadoop gives organizations the flexibility to ask questions across their structured and unstructured data that were previously impossible to ask or solve:

1. The scale and variety of data have permanently overwhelmed the ability to cost-effectively extract value using traditional platforms.
2. The scalability and elasticity of free, open-source Hadoop running on standard hardware allow organizations to hold onto more data than ever before, at a transformationally lower TCO than proprietary solutions and thereby take advantage of all their data to increase operational efficiency and gain a competitive edge. At one-tenth the cost of traditional solutions, Hadoop excels at supporting complex analyses— including detailed, special-purpose computation—across large collections of data.
3. Hadoop handles a variety of workloads, including search, log processing, recommendation systems, data warehousing, and video/image analysis. Today 's explosion of data types and volumes means that Big Data equals big opportunities and Apache Hadoop empowers organizations to work on the most modern scale-out architectures using a clean-sheet design data framework, without vendor lock-in.
4. Apache Hadoop is an open-source project administered by the Apache Software Foundation. The software was originally developed by the world 's largest Internet companies to capture and analyze the data that they generate. Unlike traditional, structured platforms, Hadoop is able to store any kind of data in its native format and to perform a wide variety of analyses and transformations on that data. Hadoop stores terabytes, and even petabytes, of data inexpensively. It is robust and reliable and handles hardware and system failures automatically, without losing data or interrupting data analyses.
5. Hadoop runs on clusters of commodity servers and each of those servers has local CPUs and disk storage that can be leveraged by the system.

### Components of Hadoop

The two critical components of Hadoop are:

1. **The Hadoop Distributed File System (HDFS)** : HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and

distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.

2. **MapReduce:** Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce is the agent that distributes the work and collects the results.

Both HDFS and MapReduce are designed to continue to work in the face of system failures. HDFS continually monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails, or data is damaged, whether due to hardware or software problems, HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster.

Likewise, when an analysis job is running, MapReduce monitors progress of each of the servers participating in the job. If one of them is slow in returning an answer or fails before completing its work, MapReduce automatically starts another instance of that task on another server that has a copy of the data. Because of the way that HDFS and MapReduce work, Hadoop provides scalable, reliable, and fault-tolerant services for data storage and analysis at very low cost.

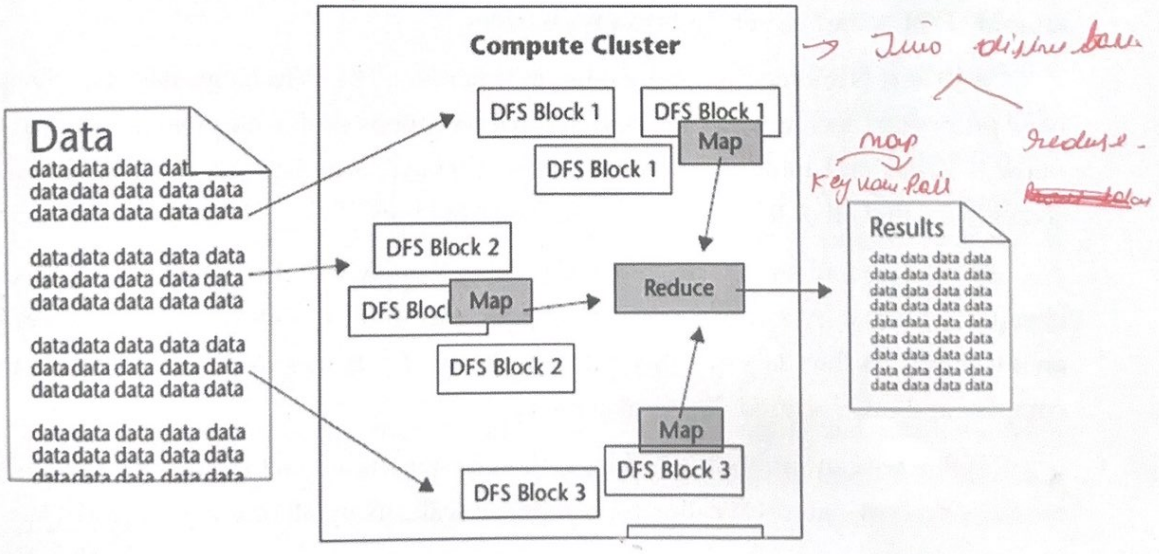


Fig: Apache Software Foundation.

## 2.1.2 Old vs. New Approaches

When we interviewed data guru Abhishek Mehta to get his perceptions of the differences between the "old" and "new" types of big data analytics. Mehta is a former Bank of America executive and MIT Media Lab executive-in-residence. He recently launched Tresata, a company that is developing the first Hadoop-powered Big Data analytics platform focused on financial industry data. Here is a summary of what Mehta told us:

The old way is a data and analytics technology stack with different layers "cross-communicating data" and working on "scale-up" expensive hardware. The new way is a data and analytics platform that does all the data processing and analytics in one "layer," without moving data back and forth on cheap but scalable ("scale out") commodity hardware. This is a mega shift and a complete game changer!

The new approach is based on two foundational concepts. Number one, data needs to be stored in a system in which the hardware is infinitely scalable. In other words, you cannot allow hardware (storage and network) to become the bottleneck. Number two, data must be processed, and converted into usable business intelligence where it sits. Put simply, you must move the code to the data and not the other way around. That is a fundamental departure and the primary difference between the old way and the new way. In the old ways, you had the multiple tiers of the stack and in the new way we have what is essentially a horizontal platform for data. The data sits in one place, you never move it around. That 's the "secret" to big data analytics.

And here 's another important point to remember: The technology stack has changed. New proprietary technologies and open-source inventions enable different approaches that make it easier and more affordable to store, manage, and analyze data. So it 's not a coincidence that all of this change is occurring right now.

Hardware and storage are more affordable than ever before, and continuing to get cheaper which allows for increasingly larger and more ambitious massively parallel architectures. As the sheer quantity and complexity of data increases, our ability to handle complex and unstructured data is also rising.

Today we can run the algorithm, look at the results, extract the results, and feed the business process—automatically and at massive scale, using all of the data available.

We continue our conversation with Mehta later in the book. For the moment, let 's boil his observations down to three main points:

1. The technology stack has changed. New proprietary technologies and open-source inventions enable different approaches that make it easier and more affordable to store, manage, and analyze data.
2. Hardware and storage is affordable and continuing to get cheaper to enable massive parallel processing.
3. The variety of data is on the rise and the ability to handle unstructured data is on the rise.

## 2.2 DATA DISCOVERY: WORK THE WAY PEOPLE'S MINDS WORK

There is a lot of buzz in the industry about data discovery , the term used to describe the new wave of business intelligence that enables users to explore data, make discoveries, and uncover insights in a dynamic and intuitive way versus predefined queries and preconfigured drill-down dashboards. This approach has resonated with many business users who are looking for the freedom and flexibility to view Big Data. In fact, there are two software companies that stand out in the crowd by growing their businesses at unprecedented rates in this space: Tableau Software and QlikTech International.

Both companies' approach to the market is much different than the traditional BI software vendor. They grew through a sales model that many refer to as "land and expand." It basically works by getting intuitive software in the hands of some business users to get in the door and grow upward. In the past, BI players typically went for the big IT sale to be the preferred tool for IT to build reports for the business users to then come and use.

In order to succeed at the BI game of the "land and expand model," you need a product that is easy to use with lots of sexy output. One of the most interesting facts about Tableau Software is that the company 's chief scientist and cofounder, Pat Hanrahan, is not a BI software veteran—he 's actually an Academy Award-winning professor and founding member of Pixar! He invented the technology that helped change the world of animated fi lm. Harahan 's invention made it possible to bring some of the world 's most beloved characters to the big screen, such as Buzz Lightyear and Woody the cowboy. Imagine the new creative lens that Pat brought to the BI software market!

When you have a product that is "easy to use," it also means that you have what Harahan and his colleagues call the "self-service approach," versus the traditional approach with heavy reliance on IT. Pat, co-founder Chris Stolte, and colleague Dan Jewett stated in a recent whitepaper.

Analytics and reporting are produced by the people using the results. IT provides the infrastructure, but business people create their own reports and dashboards.

The most important characteristic of rapid-fi re BI is that business users, not specialized developers, drive the applications. The result is that everyone wins. The IT team can stop the backlog of change requests and instead spend time on strategic IT issues. Users can serve themselves data and reports when needed.

The traditional practice of trying to anticipate the analytic needs of each employee is impossible—can an IT department really read the minds of business users? Business users are more productive when answering questions with their own tools.

There is a simple example of powerful visualization that the Tableau team is referring to. A company uses an interactive dashboard to track the critical metrics driving their business. Every day, the CEO and other executives are plugged in real-time to see how their markets are performing in terms of sales and profit, what the service quality scores look like against advertising investments, and how products are performing in terms of revenue and profit.

Interactivity is key: a click on any filter lets the executive look into specific markets or products. She can click on any data point in any one view to show the related data in the other views. Hovering over a data point lets her winnow into any unusual pattern or outlier by showing details on demand. Or she can click through the underlying information in a split-second.

We also spoke with Qliktech 's CTO, Anthony Deighton, to get his view on the world of data discovery. Deighton is an ex-Seibel executive who has been with Qliktech since 2005. He is responsible for guiding product strategy and leads all aspects of the company 's R&D efforts for its product suite, named QlikView. Deighton started off the interview with a very simple message: "Business intelligence needs to work the way people 's minds work. Users need to navigate and interact with data any way they want to—asking and answering questions on their own and in big groups or teams."

One capability that we have all become accustomed to is search, what many people refer to as "Googling." This is a prime example of the way people 's minds work. Qliktech has designed a way for users to leverage direct— and indirect—search. With QlikView search, users type relevant words or phrases in any order and get instant, associative results. With a global search bar, users can search across the entire data set. With search boxes on individual list boxes, users can confine the search to just that field.

Users can conduct both direct and indirect searches. For example, if a user wanted to identify a sales rep but couldn't remember the sales rep 's name—just details about the person, such as that he sells fish to customers in the Nordic region—the user could search on the sales rep list box for "Nordic" and "fish" to narrow the search results to just the people who meet those criteria.

## 2.3 OPEN-SOURCE TECHNOLOGY FOR BIG DATA ANALYTICS

Open-source software is computer software that is available in source code form under an open-source license that permits users to study, change, and improve and at times also to distribute the software. The open-source name came out of a 1998 meeting in Palo Alto in reaction to Netscape 's announcement of a source code release for Navigator (as Mozilla).

Although the source code is released, there are still governing bodies and agreements in place. The most prominent and popular example is the GNU General Public License (GPL), which "allows free distribution under the condition that further developments and applications are put under the same license." This ensures that the products keep improving over time for the greater population of users.

Some other open-source projects are managed and supported by commercial companies, such as Cloudera, that provide extra capabilities, training, and professional services that support open-source projects such as Hadoop. This is similar to what Red Hat has done for the open-source project Linux.

"One of the key attributes of the open-source analytics stack is that it 's not constrained by someone else 's predetermined ideas or vision," says David Champagne, chief technology officer at Revolution Analytics, a provider of advanced analytics. "The open-source stack doesn 't put you into a straitjacket. You can make it into what you want and what you need. If you come up with an idea, you can put it to work immediately. That 's the advantage of the open source stack—flexibility, extensibility, and lower cost."

The old model 's end state was a monolithic stack of proprietary tools and systems that could not be swapped out, modified, or upgraded without the original vendor 's support. This model was largely unchallenged for decades. The status quo rested on several assumptions, including:

1. The amounts of data generated would be manageable

2. Programming resources would remain scarce
3. Faster data processing would require bigger, more expensive Hardware

Many of those underlying assumptions have now disappeared, David writes:

The sudden increase in demand for software capable of handling Significantly larger data sets, coupled with the existence of a worldwide community of open-source programmers, has upended the status quo. The traditional analytics stack is among the first "victims" of this revolution. David explains how it has changed the game of enterprise software:

The old model was top-down, slow, inflexible and expensive. The new software development model is bottom-up, fast, flexible, and considerably less costly.

A traditional proprietary stack is defined and controlled by a single vendor, or by a small group of vendors. It reflects the old command-and-control mentality of the traditional corporate world and the old economic order.

David then makes the case for an open-source analytics stack. For David, who is a leading proponent of open-source analytics, it 's a logical leap: An open-source stack is defined by its community of users and contributors.

No one "controls" an open-source stack, and no one can predict exactly how it will evolve. The open-source stack reflects the new realities of the networked global economy, which is increasingly dependent on big data.

It 's certainly fair to argue whether the new analytics stack should be open, proprietary, or a blend of the two. From our perspective, it seems unlikely that large companies will abandon their investments in existing proprietary technologies overnight.

Our hunch is that open-source and proprietary solutions will coexist for a long time, and for many good reasons. In fact, most proprietary vendors have been designing their solutions to plug and play with technology such as Hadoop.

For example, Teradata Aster designed SQL-H, which is a seamless way to execute SQL and SQL-MapReduce on Apache Hadoop data. Tasso Argyros is copresident of Teradata Aster, leading the Aster Center of Innovation. In a recent blog, Argyros explained the signifi cance of his fi rm 's integration with open-source Hadoop:

This is a significant step forward from what was state-of-the-art until yesterday. This means that [in the past] getting data from Hadoop to a database required a Hadoop

expert in the middle to do the data cleansing and the data type translation. If the data was not 100% clean (which is the case in most circumstances) a developer was needed to get it to a consistent, proper form. Besides wasting the valuable time of that expert, this process meant that business analysts couldn 't directly access and analyze data in Hadoop clusters. SQL-H, an industry-fi rst, solves all those problems

## 2.4 THE CLOUD AND BIG DATA

It is important to remember that for all kinds of reasons—technical, political, social, regulatory, and cultural—cloud computing has not been a successful business model that has been widely adopted for enterprises to store their Big Data assets. However, there are many who believe that some obvious industry verticals will soon realize that there is a huge ROI opportunity if they do embrace the cloud.

There will be Big Data platforms that companies will build, especially for the core operational systems of the world. Where we continue to have an explosive amount of data come in and because the data is so proprietary that building out an infrastructure in-house seems logical. I actually think it 's going to go to the cloud, it 's just a matter of time! It 's not value add enough to collect, process and store data.

Abhishek Mehta is one of those individuals who believes that cloud models are inevitable for every industry and it 's just a matter of when an industry will shift to the cloud model. He explains that his clients are saying, "I don 't have unlimited capital to invest in infrastructure. My data is exploding—both structured and unstructured. The models that I use to price products or manage risks are broken. I 'm under immense pressure to streamline my operations and reduce headcount. How am I going to solve these problems?"

Market economics are demanding that capital-intensive infrastructure costs disappear and business challenges are forcing clients to consider newer models. At the crossroads of high capital costs and rapidly changing business needs is a sea change that is driving the need for a new, compelling value proposition that is being manifested in a cloud-deployment model.

With a cloud model, you pay on a subscription basis with no upfront capital expense. You don 't incur the typical 30 percent maintenance fees—and all the updates on the platform are automatically available. The traditional cost of value chains is being completely disintermediated by platforms—massively scalable platforms where the marginal cost to deliver an incremental product or service is zero.

The ability to build massively scalable platforms—platforms where you have option to keep adding new products and services for zero additional cost—is giving rise business models that weren't possible before. Mehta calls it "the next industrial revolutio where the raw material is data and data factories replace manufacturing factories." pointed out a few guiding principles that his firm stands by:

1. **Stop Saying "Cloud."** It's not about the fact that it is virtual, but the true value lies delivering software, data, and/or analytics in an "as a service" model. Whether that in a private hosted model or a publicly shared one does not matter. The deliver pricing, and consumption model matters.
2. **Acknowledge the Business Issues.** There is no point to make light of matters aroun information privacy, security, access, and delivery. These issues are real, more ofte than not heavily regulated by multiple government agencies, and unless dealt with i a solution, will kill any platform sell.
3. **Fix Some Core Technical Gaps.** Everything from the ability to run analytics at scale in a virtual environment to ensuring information processing and analytics authenticity are issues that need solutions and have to be fixed.

## 2.5 PREDICTIVE ANALYTICS MOVES INTO THE LIMELIGHT

To master analytics, enterprises will move from being in reactive positions (business intelligence) to forward leaning positions (predictive analytics). Using all the data available– traditional internal data sources combined with new rich external data sources—will make the predictions more accurate and meaningful.

Because the analytics are contextual, enterprises can build confidence in the analytics and the trust will result in using analytic insights to trigger business events. By automatically triggering events, the friction in business will be greatly reduced. Algorithmic trading and supply chain optimization are just two typical examples where predictive analytics have greatly reduced the friction in business. Look for predictive analytics to proliferate in every facet of our lives, both personal and business. Here are some leading trends that are making their way to the forefront of businesses today:

1. Recommendation engines similar to those used in Netflix and Amazon that use past purchases and buying behaviour to recommend new purchases.
2. Risk engines for a wide variety of business areas, including market and credit risk, catastrophic risk, and portfolio risk.

3. Innovation engines for new product innovation, drug discovery, and consumer and fashion trends to predict potential new product formulations and discoveries.
4. Customer insight engines that integrate a wide variety of customer related info, including sentiment, behaviour, and even emotions. Customer insight engines will be the backbone in online and set-top box advertisement targeting, customer loyalty programs to maximize customer lifetime value, optimizing marketing campaigns for revenue lift, and targeting individuals or companies at the right time to maximize their spend.
5. Optimization engines that optimize complex interrelated operations and decisions that are too overwhelming for people to systematically handle at scales, such as when, where, and how to seek natural resources to maximize output while reducing operational costs – or what potential competitive strategies should be used in a global business that takes into account the various political, economic, and competitive pressures along with both internal and external operational capabilities.

Today we are at the tip of the iceberg in terms of applying predictive analytics to real-world problems. With predictive analytics you can realize the uncontested market space [competitive free] that Kim and Mauborgne described in Blue Ocean Strategy.

## IMPORTANT QUESTIONS

1. What is big data technology ?
2. Explain about the elephant in the room: hadoop's parallel world.
3. Discuss briefly old vs. new approaches of big data.
4. What is data discover? Explain about work the way people's minds work.
5. What is open-source technology for big data analytics ?
6. Explain about the cloud and big data and predictive analytics moves into the limelight.