UNIT - 2

**Big Data Technologies: Hadoop’s Parallel World – Data discovery – Open source technology for Big Data Analytics – cloud and Big Data –Predictive Analytics – Mobile Business Intelligence and Big Data.**

**Explain 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.

**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.

The original creators of Hadoop are Doug Cutting and Mike Cafarella.

* The name “Hadoop” itself comes from Doug ’s son, he just made the 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.
* Hadoop gives organizations the flexibility to ask questions across their structured and unstructured data that were previously impossible to ask or solve:

■ The scale and variety of data have permanently overwhelmed the ability to cost-effectively extract value using traditional platforms.

■ 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.

■ 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.

■ 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.

■ 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.

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.

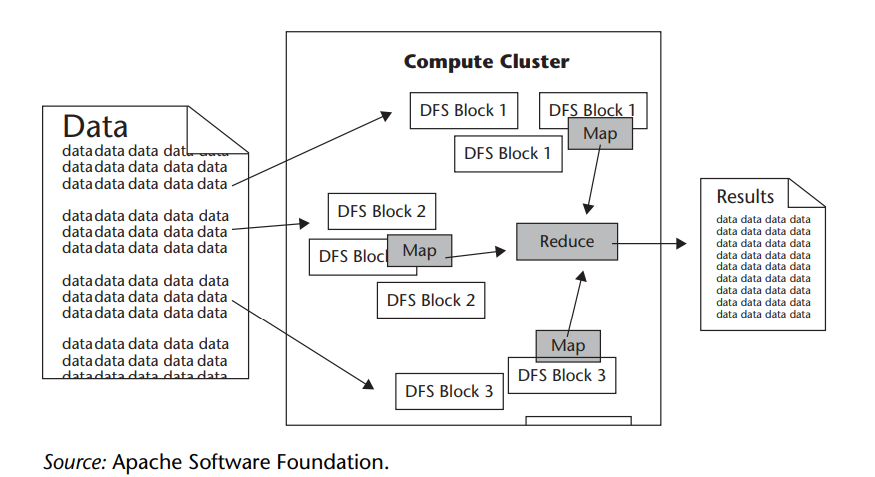
Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce is the agent that distributes the work and collects the results.

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.

**Old vs. New Approaches:** 

We interviewed data guru Abhishek Mehta to get his perceptions of the differences between the “old” and “new” types of big data analytics. The following are the observations…

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.

**The new approach is based on two foundational concepts:**

1. 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.
2. 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.

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.

**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.

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.

**Data Discovery: Work the Way People’s Minds Work:**

* 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.

**State the drawbacks of BI over the use of Qliktech?**

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”.

**Tableau** provides a company to use an interactive dashboard to track the critical metrics driving their business. Interactivity is key which Tableau provides: a click on any filter lets the executive look into specific markets or products.

“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.”

**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 confidence 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 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.

**Differences between Tableau and Qliktech?**

**Open-Source Technology for Big Data Analytics:**

What is open-source technology/software/Stack? Explain it’s advantages.

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”

“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.”

“One of the great benefits of open source lies in the flexibility of the adoption model: you download and deploy it when you need it.“

You don’t need to prove to a vendor that you have a million dollars in your budget. With open source, you can try it and adopt it at your own pace.”

The pace of software development has accelerated dramatically because of open-source software.

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

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.

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.

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, In a recent blog, explained **the significance of his firm ’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-first, solves all those problems.

**Explain** **the significance of big-data analytics for firm’s integration with open-source technology?**

**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. it’s going to the cloud, it’s just a matter of time! It’s not value add enough to collect, process and store data.

It is believed that cloud models are inevitable for every industry and it’s just a matter of when an industry will shift to the cloud model.

**Clients** are saying they “ don’t have unlimited capital to invest in infrastructure. their data is exploding both structured and unstructured. The models that they use to price products or manage risks are broken. they under immense pressure to streamline their operations and reduce headcount. How can they 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 the option to keep adding new products and services for zero additional cost—is giving rise to business models that weren’t possible before. It is “the next industrial revolution, where the raw material is data and data factories replace manufacturing factories.”

**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 in delivering software, data, and/or analytics in an “as a service” model. Whether that is in a private hosted model or a publicly shared one does not matter. The delivery, pricing, and consumption model matters.
2. **Acknowledge the business issues**. There is no point to make light of matters around information privacy, security, access, and delivery. These issues are real, more often than not heavily regulated by multiple government agencies, and unless dealt with in 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.

**Explain the emergence of cloud with growth of big data (to store it)?**

**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:**

■ Recommendation engines similar to those used in Netflix and Amazon that use past purchases and buying behaviour to recommend new purchases.

■ Risk engines for a wide variety of business areas, including market and credit risk, catastrophic risk, and portfolio risk.

■ Innovation engines for new product innovation, drug discovery, and consumer and fashion trends to predict potential new product formulations and discoveries.

■ 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.

■ 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].

**Mobile Business Intelligence Is Going Mainstream:**

Analytics on mobile devices is what some refer to as putting BI in your pocket. Mobile drives straight to the heart of simplicity and ease of use that has been a major barrier to BI adoption since day one. Mobile devices are a great levelling field where making complicated actions easy is the name of the game. For example, a young child can use an iPad but not a laptop. As a result, this will drive broad-based adoption as much for the ease of use as for the mobility these devices offer.

This will have an immense impact on the business intelligence sector. We interviewed Dan Kerzner, SVP Mobile at MicroStrategy, a leading provider of business intelligence software. He has been in the BI space for quite a while. People have been talking about mobile BI for quite some time, especially since the 1999 release of the good-old BlackBerry. However, it seems as though we have fi nally hit an inflection point. Kerzner explains his view on this topic:

We have been working on Mobile BI for a while but the iPad was the infl ection point where I think it started to become mainstream. I have seen customers over the past decade who focused on the mobile space generally and mobile applications in particular. One client in particular told me that he felt like he was pushing a boulder up a hill until he introduced mobility to enhance productivity. Once the new smart phones and tablets arrived, his phone was ringing off the hook and he was trying to fi gure out which project to say yes to, because he couldn ’t say yes to everyone who suddenly wanted mobile analytics in the enterprise.

That experience of folks who have been trying to use mobility for a long time to drive productivity and having really only pockets of success and then suddenly fl ipping over and becoming very pervasive is starting to be well understood now. In terms of why that’s the case, Dan ’s perspective on that is that with the advent of touch-driven devices, you get a set of phones that are really much more about software than they are about being a phone:

You turn off the iPhone and it’s kind of a brick, nothing to it. It doesn’t look like a phone. But you turn it on and the animating experience of it is the screen and the software that flows through that screen and the usability you get from having that touch-driven device. What ’s happened is suddenly you get a world where you actually have a form factor which lends itself to the power and flexibility, creativity, and innovation that comes with software development. That hadn ’t really been the case before. You sort of had it with some of the Palm organizer devices that were out there and you started to have it in a light-touch way with the early Blackberries. But it was always still your phone fi rst, your messaging, you weren ’t fundamentally software driven. I think the combination of multi-touch and having a software oriented device is what has unlocked the potential of these devices to really bring mobile analytics and intelligence to a much wider audience in a productive way.

**Ease of Mobile Application Deployment:**

Another inflection point for the industry is the development and deployment of mobile applications. In the past, that was controlled by the relationship with the carrier. It used to be that if you wanted to push out a mobile application, the only way you could get that application on the phone for the most part was to go through the carriers. That meant there were development environments that were sometimes very proprietary or you had to develop one set of applications for one carrier and another set of applications for a different one, maybe a third for the RIM BlackBerry environment. It didn’t lend itself to very fast detonation because there was a real channel control now. Kerzner elaborated:

One of the things that’s happened recently is that with the advent of these app stores and the maturing of the browsers on the devices into something much more powerful, now as a software provider, you can go directly to the end user. I can go to a corporation and say I ’m going to roll out a powerful global reporting application that’s also going to do deal approvals and it ’s going to totally change a whole business process. I think something that was previously locked in a desk will now give people insights into the purchasing patterns, as well as the ability to take that action. I can roll out that whole application— I never have to talk to anybody but that customer because the devices that everybody ’s lugging around are really little computers and of course you can put any software you want on a little computer and that really wasn’t the case historically in the mobile space.

**Three elements that have impacted the viability of mobile BI:**

1. Location—the GPS component and location . . . know where you are in time as well as the movement.

2. It’s not just about pushing data; you can transact with your smart phone based on information you get.

3. Multimedia functionality allows the visualization pieces to really come into play.

**Three challenges with mobile BI include:**

1. Managing standards for rolling out these devices.

2. Managing security (always a big challenge).

3. Managing “bring your own device,” where you have devices both owned by the company and devices owned by the individual, both contributing to productivity.

**Data Size 101**

Data is measured by basic units of measure that work up from a bit. A bit is represented by either a 1 (electricity fl owing) or a 0 (no electricity fl owing). This is called binary code. The code converts images, text, and sounds into numbers in order to send information from one digital device such as a computer to another. Computers use binary numbers because they are easier to handle. In binary, the digits (read and write) are worth 1, 2, 4, 8, and so on—not units, tens, and hundreds. A byte is a unit of measure and it is 8 bits put together. In ordinary numbers, “1,001” is one unit. But in binary, “1001” is one 1, no 2, no 4, and one 8, which equals 9.

