Big Data Essentials

Many of the big data tools are open-source and Linux-based. Explore the fundamentals of big data, including positioning it in an historical IT context, available tools for working with big data, and the big data stack.

Table of Contents

1. Introducing Big Data
2. The Biggest Wave Yet
3. Emerging Technologies
4. Defining Big Data
5. Key Terms for Data
6. Sizing Big Data
7. The Original Key Contributors
8. The Distro Companies
9. Apache Software Foundation
10. Apache Projects
11. Other Apache Projects
12. Other Open Source Projects
13. The Big Data Stack
14. Big Data Components
15. NoSQL Databases

Introducing Big Data

Big Data is the technology defining the first half of the 21st century. It is the merging of scientific supercomputing with commercial application. In this video, we'll put Big Data into perspective, by understanding the range and the extent of its impact on the modern world. Let's get started. Big Data is, and will affect, everything and everyone. Not just those in technology, but everyone in the world. Let's look at some of the most immediate impacts to IT jobs, to the data center, to how we create knowledge, its impact to business, and perhaps most importantly the impact to global culture. Regarding IT jobs, I believe it will completely redefine our technology job descriptions in IT. Many of the current jobs will become outmoded, but at the same time hundreds of thousands, maybe even millions of new careers are going to become available. They will all be great, high-paying jobs that will be rewarding and challenging. Regarding the enterprise data center, it's already being dramatically changed, and is undergoing the kind of revolutionary change we only see in a generation.
Heading: Impact of Big Data.

Big Data will create new careers and skills sets in IT, it will necessitate numerous projects and changes in data centers, bring about significant changes to analytics, can be highly disruptive in many directions of business, and will enable greater data level connectivity in global culture.

Big Data's impact in terms of technology in use, in operations, in a relationship to the business, affects every single data center on the planet. And the impact to the way we gather and decide knowledge is incredible. It is absolutely incredibly powerful the way we can create new knowledge with Big Data. It's completely changing how we collect, analyze, and gain insights. It is directly responsible for an acceleration of knowledge in all aspects of our lives; from research, to government, to business. And in regards to business, I can only say it's having unpredictable disruptions to our industry, government, and research. Because data touches every industry and everywhere there is data flowing, Big Data will be there. And I believe a number of long-standing industries will undergo rapid and broad-scale changes, as Big Data changes the competitive landscape. I also particularly believe that it's going to completely change the way data flows throughout the globe. I'm going to advance a concept called Global Data Fabric and I believe that it's conceivable, we will be able to create planetary level file systems that map all the world's data for use by advanced computing systems.

Now let's take a look at what we really do mean by Big Data, and here's the real answer. Big Data is supercomputing. First the primary technology of Big Data has been in use for a few decades. We call it supercomputing. And initially it was reserved for very select use, such as scientific forecasting of weather. The use of clustering, parallel processing, and distributed file systems is well understood, but was out of reach for commercial use. But today this supercomputing technology is available to all of us, and it is what is powering Big Data in the 21st century. Today a Big Data cluster will have 20 racks, 400 nodes, 4800 physical disks, and will be easily managing 3200 terabytes. That's 3200 terabytes of data. The future may well bring clusters with millions of nodes, managing zettabytes of data. We are entering a new era of hypercomputing. The other really important thing about Big Data is it is affordable supercomputing. Previous supercomputing technology was expensive, extremely expensive, and out of reach for most commercial use, and now it is not.
Heading: Supercomputing for All of Us.

Big Data is possible because the capability to parallel process computing jobs across large amounts of distributed data was made possible.

What makes big data commercially viable is it affordable at scale.

A Big Data supercomputing platform can be built for a fraction, and I do mean a fraction, of the cost of earlier supercomputing. And this is what's driving the explosive growth of Big Data. In summary Big Data is supercomputing for all of us. Now the term Big Data has a number of different meanings and I think it's very important, to kind of understand the context in which the term Big Data is being used. I group them into one of four categories – Technology, Analytics, Business, and Marketplace. In regards to the Technology, Big Data is a collection of a wide range of technologies, almost all of it open source used in a distributed computing environment to create low-cost, but high-yield supercomputing platforms. And when the term Big Data is used in regards to Analytics, we are referring to a dramatic change in the practice of data analytics; the hunt for value in the data.
Heading: Big Data in Context.

The different meanings associated with Big Data include reference to open source supercomputing technology, an advance in data analytics for fusing data into knowledge, porting over to data dependent industries in the business world, and a high value, explosive opportunity in the market place.

The core concept that's driving this change is that the sample size is now the entire data set. And that the answer can be found in minutes or hours, not weeks and months. This is leading to a significant change in how we map, analyze, and find new knowledge from data. And then when the term Big Data is used in regards to business, we are really talking about a very significant business disrupter – certainly I believe, larger than the Internet and is causing major changes in business, government, and research. And finally perhaps the most exciting, when people use the term Big Data in regards to the marketplace. It's because Big Data is this generation's high-value, explosive growth technology with estimated multiple billions of dollars of associated revenue.

The Biggest Wave Yet

Let me advance the argument that Big Data is the biggest technology wave yet. For just a brief moment, follow me through a history of computing. Our aim is to put Big Data in the perspective of technology changes from the past 70 years. Going back to the 1950s when John Bardeen, William Shockley, and Walter Brattain working at Bell labs created the modern transistor, and from this we got Silicon valley and the age of information. In the 1960s, it was the creation of the mainframe and the impact it made to our entire economy because all of a sudden all the electronic processing that we could do for the exchange of goods, services, and even money. Just take an example of the credit card industry, which was created really by IBM and Bank of America. Where would we be today without that foundational financial institution of our modern economy? And then in the 1980s, one of the most revolutionary changes in all of human history was the introduction of the PC. *Heading: Digital Paradigm Shifts.*

A timeline displays the progress of computing over the decades, beginning with 1940 and ending with 2020. The 1950s shows the transistor, the 1960s shows mainframes, the 1970s shows mini computers, the 1980s shows the PC, the 1990s shows the World Wide Web, the year 2000 shows smart phones, 2010s shows social media, and the 2020s shows Big Data.

In 1982, Time Magazine named the PC "The Machine of the Year." It had that big of an impact on our entire culture. Think – within just a few short years of it being introduced, every office worker had a PC on their desk, many of us purchased them for home and school use. Entire new industries in entertainment, education, and business were all created as a result of the introduction of the PC in the 1980s. And then came the really big bang for technology, the introduction of the Internet, which provided access to information on a global and real-time scale. Today there are over two billion people on the Internet; over one-third of the global population. And the changes to our economy, our commerce, our culture, our politics, and even our perception of each other have been changed dramatically and permanently. And then we have the introduction of smartphones. It puts a computer in your pocket. Think about it. Today there are over five billion mobile phones in use across

the planet and this connects us faster and wider than we've ever been connected in all of history. And then came the introduction of Big Data, and the first real introduction of Big Data was social media. Social media is the online social interaction among people to create and share information.

One of the most popular being photos. But empowering the social media are the Big Data recommendation engines. Those engines analyze and suggest connections to other members, they make suggestions on goods to purchase or places to travel, they find topics that peek our interest. Social media is now one of the largest segments in technology and still growing, and the end result of all this is larger and more significant impact to everyone on the planet. Now there are many of us involved in Big Data who strongly believe this will be the biggest technology wave yet. We see this technology wave completely and radically changing all other competing structures of the past 70 years.
Heading: The Biggest Wave Yet.

The scope of big data computing and the possibilities it presents means that it will be a significant wave of change in the Information Technology sector.

Tracking data with our mobile phone is being poured into Big Data solutions by the phone company, the government, market research companies, and other businesses. Each organization is finding value in analyzing this data and each of these organizations will interact with us differently because of this data. Cell phone tracking data, in of itself, is becoming a multi-billion dollar industry. These changes will dwarf, and I do mean dwarf, the impact of the Internet. So as you learn about Big Data, my goal is to convince you to agree with this statement; the changes are going to be nearly inconceivable. In the next few lectures I will present facts and comments and ideas of why I believe these bold statements are true.

Emerging Technologies

There are several emerging technologies all coming together at the same time, which are foundation of Big Data and the changes to global computing. Let's look at them individually and then look at the result. These technologies are flash memory, mobile devices, cloud computing, distributed computing, advanced analytics, and in-memory applications. The use of flash memory in solid-state drives allow computers to become universal. Additionally flash memory delivers random-access speeds of less than .1 milliseconds. Now compare this to disk access of 3 to 12 milliseconds. I believe future Big Data solutions will use a lot of flash memory to improve access time to data. Mobile devices represent computers everywhere. Mobile devices serve the dual role of being the creator of much of our Big Data inputs, and the receiver of the outputs from Big Data solutions. And we all know the impact of Cloud Computing, how it's impacted storage, databases, and services. And these...of many of these have moved onto the cloud, creating an entire new economy of computing. Cloud computing offers great access for rapidly deploying Big Data solutions. And then the heart of Big Data is the technology for distributed computing.
Heading: Technologies Driving Big Data.

The technologies that are driving big data are flash memory, mobile devices, cloud computing, distributed computing, advanced analytics, and in-memory analytics.

Big Data's large-scale distributed computing systems, based on open-source technology, are providing direct access and long-term storage for petabytes of data, while powering extreme performance. And then right alongside this are the advancements to analytics. Big Data's advancement to analytics, called data analytics, is transforming business intelligence from slow and obtuse to real-time and immediate. And then one of the most amazing recent developments is in-memory databases. These are increasing database performance by fifty to a hundred-fold. And Big Data is the platform for many of these state-of-the-art applications. Now I'm going to introduce the idea of the Global Data Fabric. Big Data really is the new global biosphere for computing. Today the modern world is connected at a bit level. As you use the Internet and ask for data, it's streamed through you bit by bit. But as Big Data scales up and out, the world will become connected at a data level. Let me give you an example, let's use health monitoring. Now consider a technology device, perhaps the size of a watch, that could be strapped to the wrist of an at-risk of heart attack end-user. *Heading: Global Data Fabric.*

The medical field example of how data is transported through different channels. In the example, the presenter explains how data moves from an end user mobile device, through big data repositories, alert and dispatch systems, incident management, medical care management, until it find its way into medical research.

The device continuously records a number of health data points related to the heart. And this data is then streamed through your mobile device into a Big Data repository. Here the inputs are continuously monitored and a pattern recognition applied to determine if there's an oncoming heart event of concern. And when the system matches to an alert, obviously a warning will be immediately sent out. Now think about this. If this were to be an incident, the information could be shared with the ambulance service and the hospital, perhaps even before the ambulance arrives on scene. And at the hospital, the use of RFI tags on medical instruments and medications can be used to create the patient's medical chart. While back at the Big Data repository, all of this data can then be mapped to other data, such as environmental factors or dietary decisions. And then it can be analyzed again to determine a comprehensive understanding of the health factors affecting heart health. From this new insurance risks could be actualized, or new health advisories could be developed, or new target drugs could be developed. Now this is Global Data Fabric in action and this is the way you have to understand Big Data. It is the centerpiece for the entire biosphere of modern computing.

Defining Big Data

Big Data is about data but it's about data on a global scale. To help us grasp the size and the immensity of Big Data, in this video, we'll learn some of the key terms used in characterizing and sizing data. Let's begin. As you engage in the subject of Big Data, and in particular as you reach a broader audience, it is a good practice to have a very precise

definition of Big Data memorized. Now I encourage my students to memorize the following: "Big Data is a supercomputing environment engineered to parallel process compute jobs across massive amounts of distributed data for the purpose of analysis."

There are other organizations that have provided definitions to Big Data and one of the most popular is the Gartner's vector model. Gartner is an American Information Technology Research and Advisory company. Gartner early on recognized the emergence and the importance of Big Data. Now according to Gartner, the technical term "Big Data" is used to describe data overload. They further define three vectors to describe Big Data – increasing variety, increasing velocity, and increasing volume. Increasing variety implies a wide variance in the types of data. Increasing velocity is the creation and the volumes of creation of data and the speed at which it's being delivered. Increasing volume means data in volumes greater than terabytes, all the way up to zettabytes. Now I strongly encourage you to memorize both the statement and the Gartner vector model.
Heading: Gartner's Vector Model.

A diagram of Gartner's vector model is displayed. It comprises a circle with Big Data at its center. The circle is divided into three segments: Variety, Velocity, and Volume. The Variety segment shows an arrow leading from Structured to Structured & unstructured. The Velocity segment shows an arrow leading from Batch to Streaming data. The Volume segment shows an arrow leading from Terabytes to Zettabytes.

Key Terms for Data

Now let's take a moment and learn some of the key terms used for Big Data. There are a lot of terminology that's created on the fly to describe Big Data. But here are a few terms I use frequently and the definition I intend for them. Now the differences are subtle but they really do help your conversation when talking to people about Big Data. The four terms I recommend are the Big Data ecosystem; the distributed computing environment; one of my most favorite, supercomputing platform; and the Hadoop cluster. The Big Data ecosystem. This is everything – people, operational practices, software, computers, everything. And I use this term when talking to the business side to ensure everyone understands that there's much more than just hardware and software. Distributed computing environment – this is the term used to describe the actual supercomputing platform, plus all the associated infrastructure and systems; those systems that provide the network, those systems that provide the Edge services, and in particularly those systems that are feeding input or taking data out of the actual cluster. I use this term a lot when talking to IT, and particularly when discussing the need to identify and map out all of the various data sources for the Supercomputing Platform.
Heading: Terms for Big Data Computing.

The big data ecosystem includes people, operations, computers, and all of the things which support them. The distributed computing environment comprises the supercomputing platform and all its supporting subsystems. The supercomputing platform is made up of the hardware and software that builds up a cluster, and a Hadoop cluster is a cluster that was built using Hadoop technology.

And now my favorite term, the supercomputing platform. This is all the hardware and software components used to build a supercomputing cluster. This is the most common term I use when talking about Big Data. But I particularly mean the actual cluster used to produce the Big Data solution. And the Hadoop cluster is the term I use when I want to talk specifically about Hadoop. Now let's jump over and talk about the different categories of data. And it really falls into two broad groupings. There is structured data and there is unstructured data. Structured data is tightly predefined in its type and format, while unstructured Data is everything else. Structured data, I think of as being tightly pre-defined prior to creation and we primarily associate it with SQL and RDBMSs. Now nearly all transactional data is structured. And by this we mean a programmer must have created a schema ahead of time. Then the schema defines the tables, consisting of column definitions and rows of records. A schema tightly defines the data that must be entered into a record.
Heading: Categorizing Types of Data.

Data can either be structured or unstructured.

Heading: Structured Data.

Almost all transactional data is structured to be used in relational databases. It's up to the programmer to build a schema before the data is created. Structured data can be defined and bounded by its format and data type and other criteria.

It's not unusual for a programmer to further define the data as being unique, or not unique, or not known. It could be defined as a particular data type or even set to a particular string, think of a phone number or a ZIP code. Now let's compare this to unstructured data. Well over, easily well over, 95% of the world's digital data is unstructured. And it consist of things such as text files, logs, e-mail, books, photos, audio. Now we all know there is tremendous value in dealing with unstructured data. Obviously there is value in the document contents itself, but there is value in discovering patterns. Even in something as simple as counting the nulls or determining the percentage of dirty data. There's a lot of data that is structured to a point. Now this is an important comment. I personally like to refer to it as semi-structured or loosely structured data. By this I mean any data element that comes with metadata is an example of semi-structured. Think about an MPEG file. How much structured data, how much metadata, how much direction on the content comes in a bit string? A lot.
Heading: Unstructured Data.

Because unstructured data has metadata to describe it, it is better called semi-structured data. For example, in stream data for sequence video or audio files and header information for e-mail messages.

Now let me introduce key-value pairs. Key-value pairs are very simple model for representing data. And most unstructured data can be expressed and managed as key-value pairs. It has very broad applicability in Big Data. It is really very simple. A user provides a key and all of the values related to the key are returned. A key may or may not be unique and there are great advantages to programmers in key-value pairs. The primary is that it's

an open-ended data structure. This means that the data structure can be extended without modification to existing code, to existing data, and most importantly to an existing schema. Let me give you a few examples. An address book; an address book relates a name – the key – to contact information – the value. A bank account; a bank account using the account number – the key – to associate with account details – the value. A movie database; the movie database contains titles of movies that relate to a key. And when you select that key, the video streaming begins on which you can view your movie – the value.
Heading: Examples of Key-Value Pairs.

Some of the examples of key-value pairs include an address book, a bank account, and a movie database. In an address book, the key of a last name is entered and the program returns an address or phone number. With a bank account, an account number is entered and the program returns a list of statements. Similarly in a movie database, the key of a movie title and a movie which can be streamed is returned.

So let me now answer this question. Why are key-value pairs so important to Big Data? I would say that Big Data was purposely intended to take on unstructured Data. And I would say that key-value pairs are the most critical strategy, for managing that unstructured data. Additionally key-value pairs are the cornerstone of MapReduce and NoSQL databases. They were intentionally, purposely built for key-value pairs. And as both MapReduce and NoSQL databases is scaled affordably, this is what gives the power to Big Data. And in regards to programming, it's important to understand the key-value pairs can be associated with any value. And they provide a very fast access method and it's simple, fast, and scalable. And most importantly key-value pairs really have a schema that is loose, flexible and can be modified on the fly.

Sizing Big Data

Big Data requires us to extend our language for sizing of data. This table contains a list of key terminology that you should use correctly. Most of these sizes are unfamiliar to most of us so if you're going to work in around Big Data, it is very important to learn them and to use them correctly. This table is another one of the things I encourage students to take the time to memorize. So a terabyte is 10 to the 12th bytes of data, a petabyte is 10 to the15th bytes of data and an exabyte is 10 to the 18th. These are commonly used and soon we will be using zettabytes and yottabytes. Now always remember a zettabyte is smaller than a yottabyte. There are a lot of ways to think and talk about data sizes, but let us do something fun to give it some concept. One way is to consider a gigabyte of data as being the size of a golf ball. After all a golf ball is nearly the same size as the flash drive you dangle around your neck. Now a gigabyte is 10 to the 9th. Or a 1 with 9 zeros behind it. While the size of the petabyte is 10 to the 15th. A 1 with 15 zeros behind it. If you wanted to process a petabyte of golf ball data, you will need 10 to the 6th number of golf balls.
Heading: Terms of Data Sizing.

A table containing three columns and ten rows is displayed. The columns are: Title, Abbreviation, and Number of Bytes. Each row contains the title of the data size, its abbreviation, and then the number of bytes.

Kilobyte: KB: 10 to the power 3.
Megabyte: MB: 10 to the power 6.
Gigabyte: GB: 10 to the power 9.
Terabyte: TB: 10 to the power 12.
Petabyte: PB: 10 to the power 15.
Exabyte: EB: 10 to the power 18.
Zettabyte: ZB: 10 to the power 21.
Yottabyte: YB: 10 to the power 24.

Heading: A Football Stadium Filled with Golfballs.

To calculate how many golfballs make up a petabyte, you would divide 10 to the 15th by 10 to the 9th. This give you 10 to the 6th. Or one million golfballs - a football stadium of golfballs.

I can reach this number quickly by subtracting 9 from 15. Now 10 to the 6 is one million golf balls, that's enough golf balls to fill a football stadium. In Big Data we commonly deal in these football stadium size collections of data. Now let's look at some real world data and compare it to our football stadiums. I want to look at some data that can commonly be found on the Internet. Walmart handles more than 1 million customer transactions every hour and their customer database is over 2.5 petabytes of data. This is 2.5 football stadiums of data. While the Library of Congress receives over 235 TB of data every single year and the total collection of the Library of Congress now houses over 60 petabytes of electronic data, this is 60 football stadiums of data. Now let's think about mobile phones. I said earlier that we have 5 billion phone users in the world today. Each one of them can generate up to a terabyte of data records every single year. This is 5 million football stadiums of data every single year.
Heading: Imagining Really Large Data Sets.

Walmart processes more than a million customer transactions hourly and stores 2.5 petabytes of customer data. The Library of Congress collects 235 terabytes of new data per year and stores 60 petabytes of data. Over 5.5 billion mobile phones were used in 2014; each phone creates one terabyte of call record data yearly.

The Original Key Contributors

Big Data originated with a few tech giants of Silicon Valley. In this video we'll learn about the original contributors, as well as the primary distributors of Hadoop software. Let's take a look. Big Data's real ancestry is with high-performance computing for science. This is where the size of the processing and the datasets exceeded any monolithic computer's capability. And from this came parallel processing, distributed data, and clustering. The first really significant supercomputers were used for weather forecasting. This was done as a project between NOAA and the US military with partners such as IBM. And even in the 1990s these supercomputers could run for seven days to produce a fourteen-day forecast. The same work we now solve in just a few hours. The primary difference between scientific

computing and Big Data is that scientific computing is normally focused on crunching large, complex equations, while Big Data is more focused on discovering patterns. So let's take a moment and look at the original use case that created Big Data. The original use case was to monetize click streams into advertising dollars. To accomplish this the entire Web had to be indexed, which in turn meant a number of new technology problems had to solved. Problems such as, how do I set up a crawler that crawls through the entire Internet and tracks down every website?
Heading: Origins of Supercomputing.

Supercomputing started as high performance supercomputing for science. The first supercomputers were used for WEAX weather forecasting by the US military in collaboration with NOAA and IBM.

Even with supercomputers, it took seven days of processing to produce a 14 day WEAX forecast.

Heading: The Original Use Case for Big Data.

To monetize world wide web usage meant that the world wide web needed to be indexed. This presented problems such as how to crawl the web and collect searches on every website, how to complete the collection quickly enough, how each website would be indexed and when it would be reindexed, how to organize and prioritize page rankings, and how to match search queries to the most relevant pages.

How do I identify what has changed since the last reindexing? How do I even index all these different websites and capture all that data? How do I organize and prioritize the page rankings? And another common use case was how to match lookup queries to the most relevant pages. Now many people, many universities, and many companies all participated in creating the foundations of Big Data; each building on the work of each other. But the following are considered be the original key contributors, to solving the architectural challenge of distributed computing for commercial use. And that's Google and Yahoo. I particularly also would like to mention Doug Cutting. Now we're going to look at each of these individual organization's contribution. Let's begin with Google. Google has led this technology and has led the creation of supercomputing computing environments. Perhaps one of the most important things in their leadership is that they made the decision to publish two academic papers describing Google's Big Data technology. The first paper was the Google File System and the second was the Google BigTable Architecture. Now you can quickly find these papers on Google itself. They are seminal to Big Data and I encourage you to download and read both of them. Let's talk about Yahoo's contribution. They are a pioneer in building low-cost supercomputing platforms.
Heading: The Original Key Contributors.

Google, Yahoo, and Doug Cutting solved the problem of commercial distributed computing and parallel processing.

Heading: Google's Contribution.

Google was committed to creating a new paradigm in computing and invested heavily in the idea. In 2003/2004 Google published the Google File System (GFS) and Google BigTable Archetecture white papers, which described Google's technologies.

Heading: Yahoo's Contribution.

Yahoo is a pioneer at building low cost supercomputing platforms and making significant contributions in the platforms' architecture, software, and operations. Yahoo created Hadoop, the open-source software and Yahoo continues to make significant contributions to Big Data development.

Their engineers have made many significant contributions in share nothing. They have designed, architected, and deployed some of the first massively-sized clusters. And perhaps very important to us is that in 2006, Yahoo hired Doug Cutting, who then became the father of Hadoop. Most importantly to us, Yahoo has always encouraged the open sourcing of Hadoop. And they supported not only the Hadoop project, but many other Apache projects. Even to this day they continue as a major contributor and test-bed for Hadoop. Doug Cutting is considered the father of Hadoop. He also has developed Nutch and Lucene. As I said earlier, in 2006 he was hired by Yahoo to develop Hadoop. He started by working from the papers released by Google for the Google File System and Big Table. He promoted it inside Yahoo as an open-source project. Doug currently serves as the Chief Architectural Officer at Cloudera. Now as a funnier side, everyone should know how he came up with the name Hadoop. Doug's son had a yellow stuffed elephant that he dragged around and he called it Hadoop. And that's how we got the name, Hadoop.
Heading: Doug Cutting's Contribution.

Doug Cutting also developed Nutch and Lucene.

The Distro Companies

Hadoop software is open source but similar to Linux and MySQL. There are companies which support unique distributions of the open source software. These distribution companies are commonly referred to as distros. Each of these distro companies has unique go-to market strategy. The four companies we're going to talk about are Cloudera, Hortonworks, IBM, and MapR. Cloudera uses an open source stack of Hadoop, but they have built their own management console which they release with their distribution. Hortonworks is pure open source and their strategy is to include only open source software. IBM uses many open source software components, but they have used their years of experience in the enterprise to build additional value-add tools. MapR address some of the early issues with Hadoop by making their own improvements and readying it for commercial use. The distros all have common approaches to releasing the Hadoop software. All make contributions to the Apache projects related to Big Data and all the distro companies provide both free and paid software bundles. The intent of their release is to ensure a working environment. They're all committed to preventing software version mismatch, which can be a very tough challenge to an engineer new to Big Data and one of the primary reasons for the popularity of the distros.

Heading: Comparing Distro Companies.

The various distro companies have their own advantages. Some of the features of major distro companies:

Cloudera is open-source but has its own management console, Hortonworks provides pure open-source play, IBM has positioned itself as a big corporate partner, and MapR has managed to customize and commercialize the open source.

Currently the standard core is Hadoop, Hbase, Hive, Pig, Sqoop and Flume. But each distro then tailors their own selection of additional software components. Generally all distros release their software as tar, RPM, Yum and APT. Cloudera offers a Hadoop solution to solve data intensive business problems. They are a major leader in Hadoop and they are recognized as one of the first independent distros. Cloudera offers its enterprise customers a family of products and services that well complement the open source Hadoop platform. Cloudera serves a wide range of customers including financial services, healthcare, digital media, advertising, networking, and even telecommunications. One small comment I'd like to add: they are strongly backed by Intel. Now let me talk about the CDH Bundle. That is the Cloudera Distribution of Hadoop. You can find it at `cloudera.com/hadoop`. It is the baseline product set that we're familiar with Hadoop, which includes Hadoop, Hive, Pig, Hbase. Also tools like Sqoop, Flume, Mahout, and Whirr.
Heading: Introducing Cloudera.

Cloudera is the first distro company in history and a major Hadoop leader. They offer comprehensive training, architectural consulting, and technical support.

A diagram depicts the different components of Big Data. For example, unstructured data, such as movies, audio, and images, and databases such as Forecast, Financial, Legacy, and CRM. Also tools such as Decision extracting tools, Data mining tools, and Predictive modeling tools. These tools are used by business users, for statistical analysis, and by predictive modeling experts respectively.

Heading: The CDH Bundle.

The Cloudera Distribution of Hadoop (CDH) includes Hue, Impala, Mahout, Whirr, and Cloudera Enterprise Manager.

Add-ons include Cloudera's Enterprise Manager. Now this is a very important offering from Cloudera as it is a tool that's intentionally built to install and manage a supercomputing platform. Hortonworks is a major competitor to Cloudera and they are a second large player in the expanding Hadoop marketplace. In 2011 the Yahoo division responsible for much of the development of Hadoop was spun off to form its own company, Hortonworks. They are a fully open source solution, and they do not offer any commercial software. Hortonworks offers its enterprise customers services for the open source Hadoop platform, and they call that the HDP Bundle. The Hortonworks Bundle can be found at `hortonworks.com/hdp/downloads`. It obviously also includes the standard baseline tools such as Hadoop, Hive, Pig, and Hbase, but they also include a whole range of

extra tools like Sqoop, Flume, Mahout and Whirr. They continue to add additional tools all the time like the current offering includes Storm, Falcon, and Knox. HDP is a great choice for Hadoop.
Heading: Introducing Hortonworks.

Hortonworks offer pure open source comprehensive training, architectural consulting, and technical support.

A diagram depicts how a Big data system might work with Hortonworks bundles. The diagram shows Image, Email, HTML, and DB content attached to a box that contains the Edge node, Name node, and Data nodes sections. The Edge node section contains Tomcat and Oozie, the Name node section contains Tresata libraries, Oozie, Hive, MapReduce, and Hbase, and the Data nodes section contains the MapReduce jobs, Hive jobs, Hadoo pHDFS, and Apache Hbase.

There is a two sided arrow which connects the box with a Total view of customer graphic.

Heading: The HDP Bundle.

The Hortonworks Data Platform (HDP) also includes these additional tools: Accumulo, Storm, Mahout, Soir, and Falcon.

Well let me introduce IBM. Obviously we all know that they are a major leader in all levels of computing, but they've taken a very strong lead in Big Data. They've made many commercial improvements on Hadoop, many of them tailored from their years of experience in the enterprise. IBM packages their Big Data solution as a product called Infosphere Big Insights. It comes in two editions. There's a free and a commercial. The commercial tool comes with all kinds of additional application tools, data analyst tools, and data visualization tools. It's a great product. And finally let me introduce MapR. They also are one of the early distro companies and they built an early reputation for being enterprise ready. They have an intense focus on performance and availability. Their products are called M3 and M5. M3 is the free version and M5 is a paid enterprise version. It does come with a customized file system, and they've spent a tremendous amount of effort doing distributed namenode to remove the single point of failure for a Hadoop file system. It also is a tremendously high value product.
Heading: Introducing IBM.

IBM focused on enterprise level applications.

Heading: The Infosphere Bundle.

The IBM Infosphere bundle includes an enterprise edition of IBM Infosphere Big Insights with application tools, data analysis tools, and data visualization tools.

Heading: Introducing MapR.

A diagram shows how MapR systems work. In a box called Supercomputing platform, there

is a Data access server, which is connected via a LAN to a Job node and a Data node. The Job node contains the MapReduce and Oozie servers, which are connected. The Data nodes section contains Hadoop and Hbase. The Job node is connected via VPN to two Oozie client user access computers.

Heading: MapR Bundle.

MapR is a customized file system which has a distributed namenode arch to remove SPOF for HDFS.

Apache Software Foundation

The Apache Software Foundation plays a very important role in Big Data. They are directly responsible for managing many of the software development projects, such as Hadoop used in Big Data. In this video you will learn about their contribution. Let's get started. The majority of Big Data software is open source – this is why it's important to have a good understanding of the power and the limitations of open source software. What was once considered the realm of high end but low money high tech companies, is now becoming the standard across industry. Open source operating systems and databases are now widely accepted. Open source languages, such as R, are becoming the standards and new open source applications are rapidly gaining acceptance and customers. Now let me talk about the values and the risks of open source software development. It has some distinct advantages over commercially developed products but also comes with some very distinct risks. One of the advantages is you can quickly gain a global community of contributors who will focus hundreds of hours of high talented developers on product creation. Another great advantage is you can get a leading edge product rapidly into market and obviously one of the most distinct advantages is most open source software comes with a free edition. Additionally for the programmers you get access to the source code so that you can heavily modify it for your own custom needs.
Heading: Open Source Software.

Open source is a software usage philosophy of open collaboration and free access to software. Some of the popular open source programs are Linux, MySQL/Postgres, R statistical programming language, and Hadoop.

Heading: Values and Risks of Open Source.

Some of the values of open source software are having a global community of developers, rapid development and rapid to market, low cost point, and access to source code.

The risks which come with open source software include the development community fading away and having numerous overlapping products.

Now let me discuss some of the risks. The first is that your development community fades away. By this we mean that your product is no longer considered cutting edge and leading in its market space and the development community moves on to other products. Another

really significant problem that you need to understand in development of open source projects is that you can have numerous overlapping products causing fierce competition for a fairly niche market. Now I want to introduce one of the key players in Big Data and that is the Apache Software Foundation. I have their mission statement here and it really does a great job of summarizing their value added. They are a group of software developers who are committed to developing free and open source software and perhaps one of the most important statements is that they are collaborative and consensus based in their development. I would also like to discuss their software solution in terms of a license; they have a very liberal distribution policy for their licensing. For example, reuse allows for reuse of the source and the bundles without reservation. Their extensions are allowed for all programmers, who can then easily modify source for local customization. And for the commercialization, they do allow you to build on their products. However, if you are going to use an Apache product for commercial purposes have legal review of the license.
Heading: Apache Software Foundation.

"Apache Software Foundation is a non-profit organization dedicated to Open Source software projects. It is a decentralized community of software developers who are committed to developing free and open source software. Apache projects are collaborative, consensus-based development." Source: apache.org.

Heading: Apache Licenses.

The Apache license allows the licensee to reuse, extend, and commercialize their use of the software.

Apache Projects

Apache organizes software development into top tier projects and subordinate projects. Each project has a community of committers – these are software developers who contribute peer review code to the primary projects. These projects are all defined by consensus and by pragmatic software licensing. In other videos, we're going to learn about Hadoop. It is a software framework that allows for the distribution of processing of large data sets using a fairly simple programming model. The initial code was originally developed by Doug Cutting, who by the way is committed to a number of Apache projects. You can go to apache.org web site and look up all these various projects. I suggest you start with Hadoop. It is interesting to look up the list of committers and the companies they work for.
Heading: Apache Software Foundation.

"Apache projects are defined by collaborative consensus based on processes, an open, pragmatic software license and a desire to create high quality software that leads the way in each field." Source: apache.org.

Heading: Apache Hadoop.

"Hadoop is an Open Source project lead by Apache. It underpins the Big Data movement

and is a rapidly advancing major technology. Original design intent was to handle massive amounts of data, quickly, efficiently and inexpensively. It handles all data types, structure and unstructured. If the data consists of bits and bytes Hadoop will store it on commodity hardware." Source: apache.org.

Other Apache Projects

There are hundreds of Apache projects. We will learn about a number of them in other videos but let me discuss a few interesting examples. Accumulo is a very interesting example of an Apache open source project. Accumulo is a key-value NoSQL database. It's directly built on Google's Big Table architecture. It is written in Java and one of its most important features is it has a cell-level security. And one of the most interesting things about Accumulo is it originated at the US National Security Agency in 2008. And the US government made the decision to submit it as an open source project in 2011. Today Accumulo is ranked as the third most popular NoSQL database. Now let me discuss Avro. Avro is a data serialization system for data input and out of Hadoop. It has a rich data structure and I have to tell you it has much more than just Hadoop, and is very commonly used in other projects outside of Hadoop. It has a fast, compact, binary data format. Very easy to program and very easy to translate. And another one of the interesting things about Avro is it's another Apache project that Doug Cutting was involved with. He created the original spec and some of the original code for Avro.
Heading: Introducing Accumulo.

Accumulo is built as a column family-store database on top of HDFS. It is classified as NoSQL, uses key-value pairs, has no schema, and has column-oriented views of data.

Heading: Introducing Avro.

Avro provides simple integration with dynamic languages.

And finally let me introduce Oozie as another Apache project commonly used in a Hadoop cluster. Oozie is a workflow and coordination tool. It runs...across the supercomputing platform. It allows jobs to run in parallel while waiting for input from other jobs. One of the interesting advantages to Oozie is it comes with a very complex scheduling tool. This allows for coordination of jobs waiting for other dependencies within the supercomputing platform. Oozie is a great tool to master if you are going to run Hadoop supercomputing.
Heading: Introducing Oozie.

Oozie workflow definitions are written in XML. It is better than cron.

Other Open Source Projects

I want to ensure you understand that there are more open source communities than just Apache. Two good examples are Cascading and MongoDB. Cascading is a proven application development platform. It is purposely designed for building data applications on a Hadoop cluster. The goal is to provide a higher-level interface to MapReduce.

MapReduce is built in Java and does require advanced programming skills. Cascading uses a data flow model, which consists of pipes and multiple joiners, taps, and other constructs. This makes it easier for the programmer to manage the translation, deployment, and execution of the workflow on the Hadoop cluster. Again Cascading is an alternative open source project, it is not an Apache project.

A very popular open source project is MongoDB. MongoDB began in 2007. It is a NoSQL database using key value pairs, though it is more frequently classified as a document-oriented database. MongoDB is used extensively by many of the early adopters of Big Data. It does come as a free download and it has full access to the source code. It is released under both the GNU General Public license and the Apache license. So when you require a NoSQL database MongoDB is a great alternative open source solution for consideration.

The Big Data Stack

A supercomputing platform is anything but an elegant solution. It is built on a complex stack of entwined technology and like all complex things, we need mappings to assist us in understanding and managing it. In this video I will introduce you to a map for understanding your Big Data architecture. Let's jump in. As we discussed in a previous video, Big Data primarily uses open source software. One of the problems with an open source strategy is it generates a large number of computing software projects. A Hadoop cluster is constructed by combining any number of these software products. Software selection is driven by use case, product reputation, team expertise, and some amount of what we know works now. But such a wide selection can make it difficult to understand what does what and why is it there. We suggest one of the first architectural drawings everyone should use is a simple drawing showing a functional view of the Big Data architecture. A functional view is a simple model. It defines the Big Data solution into functional areas and then maps the selected software into the correct functional layer. We recommend starting with a simple model we call the Big Data stack. Let us define the layers for our functional view of the Big Data stack.
Heading: Introducing the Big Data Stack.

"Big Data is made up of many functional elements and these are built up into a number of different systems for flowing and analyzing data. These systems can be seen as a layered model. This is not a technical definition per se but a way of understanding all of the many components and how they fit together to provide a supercomputing platform." Source: Will Dailey, Instructor.

The first layer is Infrastructure. This includes all the hardware and software used to support and to operate a Hadoop cluster. Obviously this will include software versions of the operating system but it should include all the commonly used tools for monitoring and reporting on the Hadoop cluster as well. The second layer we call the Data Repository layer. It deals with the movement of data within a distributed computing environment. It is centered on the Hadoop distributed file system, the primary repository, but should include data transport tools such as Sqoop and Flume. Additionally we commonly assign all the no

SQL databases, such as Accumulo and HBase as a form of data repository. The Data Refinery layer deals with the manipulation and processing of data using the parallel processing framework. The primary technology is Yarn and MapReduce. We will learn more about these two technologies in a later video. The Data Factory layer is an ever expanding number of tools designed to interface into Hadoop. All of them make it easier to access the full power of Hadoop. Many of them actually allow the users to create compute jobs in an easily understood language, such as SQL, and then they translate these inputs into MapReduce jobs. This is why we refer to this class of software as data workers and why we call this layer the data factory.
Heading: The Infrastructure Layer.

A diagram shows the various levels of a supercomputing platform. The foundation layer is Infrastructure. The layers above it are: Data repository, Data refinery, Data factory, Data fusion, and ending with Business value.

Infrastructure comprises utilities, network, and hardware which build or support the supercomputing platform.

Heading: The Data Repository Layer.

The Data Repository layer is in charge of controlling the movement and storage of data within the supercomputing platform.

Heading: The Data Refinery Layer.

The Data Refinery layer is responsible for parallel processing compute jobs in order to manipulate and process data. This is done using primarily Yarn and MapReduce, both products of the Hadoop project.

Heading: The Data Factory Layer.

The Data Factory layer comprises tools which are designed to make it easier to access the full offer of Hadoop. The tools include Hive, Pig, Spark, and Oozie.

The Data Fusion layer is the application layer and the true business end of the Big Data solution. These are applications many purposely built to take advantage of Hadoop's power which create new information by combining multiple data sources; we frequently refer to this action as data fusion. They are combining the different data to gain new knowledge. This is where all the data analytics and the data visualization tools reside in our model. Examples are a machine learning technology, such as Mahout, or Data Visualization tools, such as Datameter or Pentahoe. And no model would be complete without understanding why we're building the Hadoop cluster. What software tools do we use to interface into the business? What tools do we use to generate reports, which will clearly explain our value to the business? We use a range of tools to express the requirements, the service levels, the request for changes, the cost and the expenditures, all are important parts of our Big Data stack.
Heading: The Data Fusion Layer.

The Data Fusion layer comprises tools used to create applications, implement algorithms, and visualize data. The tools include Mahout, Datameter, Pentahoe, and Tableau.

Heading: The Business Value Layer.

The Business Value layer is the product of the supercomputing platform and uses a number of tools to express the requirements, service levels, and costs and expenditures for the business.

Big Data Components

Big Data technology consists of a large number of software components. Many of these are Apache open source projects. But there are certainly many other great tools outside of Apache projects. Our intent here is to give you just the most common open source tools available to use in constructing a Big Data platform. These tools are continuously in update and many new tools are being introduced on a regular basis. Now I would like to add a comment regarding software names. There is always some meaning to the names given to the software projects, but there is no criteria defined for the namespace. This means there is no connection between their names. And even though we have a software product we use in Hadoop called Zookeeper, not all software is named after animals. Flume is named after the water race used in some saw mills to bring logs to the mill. It was purposely designed for extracting large amounts of data into and out of Hadoop. Flume is well suited for gathering web logs from multiple sources by the use of agents. Each agent is freestanding and the agents are easily connected one to the other. Flume comes with many connectors, making it fast and easy to build reliable and robust agents. And best of all Flume is highly scalable across many machines.
Heading: Overview Big Data Components.

A diagram shows the various tools used on a Big Data platform including Flume, MapReduce, Oozie, and Ambari.

The name Sqoop is a combination of SQL with the word Hadoop. It is a great tool for exporting and importing data between any RDBMS and Hadoop Distributed File System. Sqoop uses both a JDBC and a command line interface. It supports parallelization across the cluster and has the ability to deploy as a MapReduce job to manage the export or import. I use this tool frequently and have always found it to be reliable and a stalwart ally in the battle for moving data into a Hadoop cluster. Hive was originated at Facebook. This factory worker is a strongback in any Hadoop cluster. It is one of the first tools learned by many new Big Data users. Hive is a SQL-like interface into Hadoop. It allows those of us who are comfortable with SQL to use common SQL commands and common relational table structures to create MapReduce jobs without having to know MapReduce. Hive treats all the data like it belongs in tables and allows us to create table definitions over the top of the data files.
Heading: Introducing Hive.

Hive convert inputs into MapReduce jobs and organize unstructured data metadata into tables.

The other strongback in a Hadoop Cluster is Pig. Pig was originally created at Yahoo. It was named on the fact that pigs eat anything. Pig is a script language for data flow. It also converts the scripts into MapReduce jobs. Pig also provides its own vernacular. Obviously the scripting language is called Pig Latin and the Pig Shell is called grunt. And let us not forget that there is a common storage for Pig scripts we call the Piggy Bank. These are just a few of the most common tools used in a Hadoop cluster. To learn more about the more popular tools, I suggest you go to one of the distro companies such as Cloudera, Hortonworks, IBM, or MapR and review their distribution for a supercomputing platform.
Heading: Introducing Pig.

The Pig schema is optional at runtime.

NoSQL Databases

Let us begin our conversation about NoSQL with a few common facts. First NoSQL means Not Only SQL. This implies they can use SQL as well as other low-level query languages. NoSQL is a broad range of databases. The broadest definition is they are non-relational distributed store databases. Nearly all NoSQL databases use a key value pairing as their central architecture. All are designed to work in a distributed computing environment. NoSQL databases are powerful because of the simplicity of their design for the fact that they massively scale out, and that they are significantly faster when used for the right use cases. Most NoSQL databases are open source. On our last count there was over 75 offerings, many highly customized specific use cases. Selecting and deploying a NoSQL database truly does require platform engineering. There's a large amount of complexity, overlap, and constant change in this technology arena. We strongly suggest you seek expert advice when making this decision. NoSQL databases are in the forefront of new engineering to take full of advantage of emerging technologies such as solid-state drives and all-memory applications.
Heading: NoSQL Databases.

NoSQL experience is required when selecting a database.

NoSQL can roughly be categorized into one of four groupings – Columnar, Document, Scientific, and Graph. The columnar databases are purposely built databases that implement key-value stores on what is called a wide column. These are some of the most popular and most frequently chosen NoSQL databases. The document store databases are used to store documents. These are also widely used. The scientific-oriented NoSQL databases are used for large-scale algorithm crunching. And the graph NoSQL databases are used for graph structures. These define the datasets in terms of nodes, edges, and properties. Some of the most frequently-used columnar databases in Big Data include Accumulo, Cassandra, Druid, and HBase. This is not at all an inclusive list, but all these are open source projects. And two of these are Apache projects – Accumulo and HBase. HBase is a highly popular NoSQL database. It is commonly included in Hadoop distributions. HBase is a purposely

built columnar data store. It was intended to be deployed on top of Hadoop Distributed File System, but it can run standalone. HBase is based on Google's Big Table model of data storage. It is a NoSQL database using the key-value pair for data storage. It has remarkably low latency views on large data sets.
Heading: Categories of NoSQL.

Graph databases use graph structures that define datasets by nodes, edges, and properties.

Heading: Columnar Databases.

Some of the more widely used databases are Accumulo, Cassandra, Druid, and Hbase. Hbase is an extension of Hadoop.

Heading: Introducing HBase.

Hbase has no schema and provides a column oriented view of data.

A diagram shows data being placed in a database, going through the key-value store and the output showing the saved data. For example, the first data input is ABC 3356. For example, data is put into the key-value store, which has the format "Name":"ABC Co.". The data is saved as "Web":"abc.com".

Comparing RDBMSs to NoSQL is a complicated subject. But let me make a few key points everyone should understand when dealing with NoSQL databases. RDBMSs are superior for transactional processing. They are uniquely designed to store and retrieve rows of information. While NoSQL databases are superior for finding subsets of data from massive datasets. RDBMSs are highly engineered to guarantee data integrity, while NoSQL databases take a very different approach. But this does not mean they lose or corrupt data. If you are interested in learning more about NoSQL databases, there's a lot information on the Internet. There's some great Skillsoft online classes and a growing number of subject matter experts. Because of its power and speed, NoSQL databases play, and will continue to play, an important role in our supercomputing platform.