Computer Aided Decision Systems - Industrial practices using Big Analytics

Professor Deepu Philip
Department of Industrial and Management Engineering
Professor Amandeep Singh
Imagineering Laboratory
Indian Institute of Technology Kanpur

Lecture 23

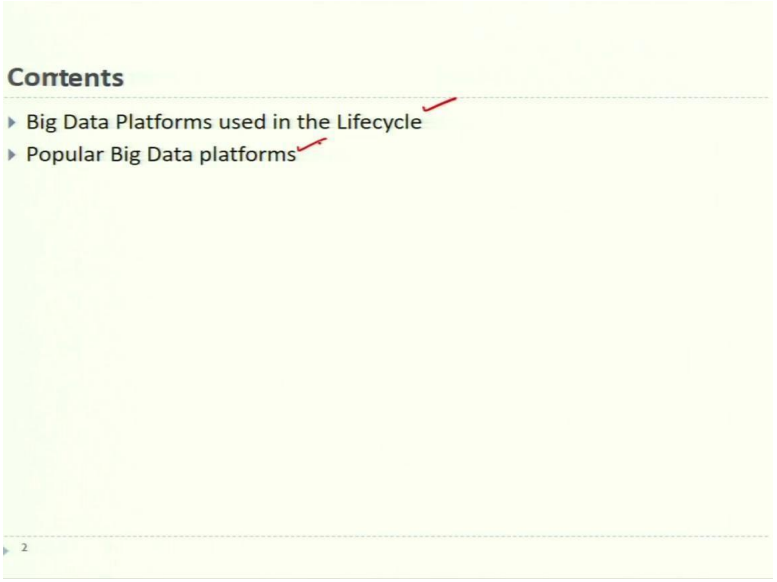
Big Data Analytics Tools and Software (Part 1 of 4)

Welcome back to the course on Big Data Analytics, where we were trying to discuss the Computer Aided Decision Support Systems using Big Data Analytics.

(Refer Slide Time: 0:25)

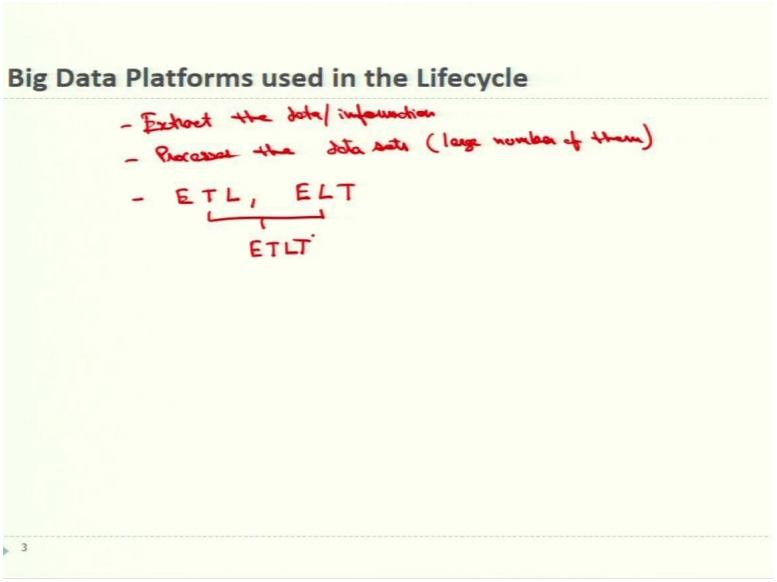
I will discuss Big Data Analytics tools and software in this lecture and I will also like to talk about how these are connected to the different lifecycle phases that we have discussed. There are multiple technologies, maybe tons of technologies, which are available nowadays, but few of them are very prominent, their names I have already suggested in the previous lectures R Studio, Tableau or so.

(Refer Slide Time: 0:51)



I will try to discuss a few of them and I will also try to discuss some of the popular ones.

(Refer Slide Time: 0:57)



So, big data platforms or software which is used in a lifecycle, there are multiple softwares which are used. So, that use of the software is to majorly

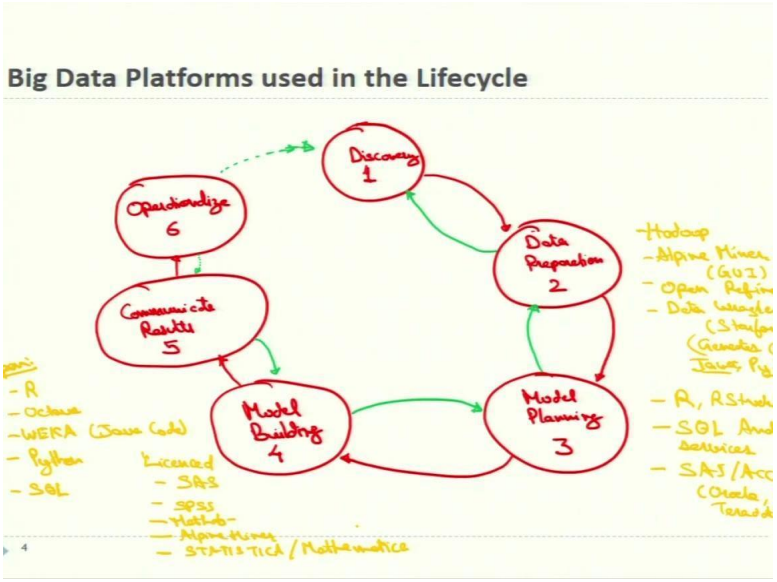
- Extract the information/ data

Sometimes it is data, if the data is in the structured form and this some meaningful understanding could be taken from it in the initial phase itself that is why it is called information. So, we extract the information and also extract the data. That is the major thing.

Then it

- Processes the datasets, when they process datasets, it could be a (large number of them.) Major things that we have discussed it performs 'Extract Transform Load' or 'Extract Load Transform', combined it does 'Extract Transform Load Transform'.

(Refer Slide Time: 2:16)



Let us have a quick look at the Lifecycle phases that we discussed in the previous lectures. So, here in phase 1, when the discovery of the data has to be there. Discovery of the data is Discovery of the program of the Life Cycle. When we say Discovery, we are trying to only say what is our statement, what are the hypotheses that we could define? Those could also be put in a formal manner. We can use a general Google document or we can use maybe Microsoft Word or so to draft them. But to prepare the data, the software which is required, where we try to set the dataset infantry, we try to develop tables on them.

So, majorly I would say

1) Hadoop- we will discuss Hadoop. It can perform massively parallel ingestion of the Data and also parallel processing including if the web traffic is very heavy. So, GPS location analysis, genomic analysis, combining of massive unstructured data feeds or so, it can do multiple things.

Also in data preparation, we use

2) Alpine Miner- So, a software Alpine miner, what does it provide? It provides a major Graphic User Interface (GUI) so that the person who is trying to ingest or use the data is able to use that in a very easy manner. So, it includes analytical workflows, it includes the various data manipulators, and includes various series of analytical events such as stage data men techniques like for investing, maybe first ten customers, those who have come. Then to have descriptive statistics or clustering of the system. So, in SQL or other big data sources, this could also be developed.
3) Open Refine- the word is open; it is also known as Google Refine. Google is providing its services in that data mining practice itself. So, in Google Refine, we have a free open-source powerful tool, which also helps us to work with messy data because it is Google. So, Google has an excellent GUI as I said Graphical User Interface. So, why User Interface, I say what is an example could be, what are the number of minimum clicks a person has to do to get to the 'login' page (login should be highlighted) or to search something what are the number of steps the person has to go through when searching the keyboard or so. So, it is Open Refine. Google provides an excellent GUI and it is again a performing data transformation tool.
4) Data Wrangler- it is an interactive tool for cleansing the data, transformation of the data. It was developed by Stanford University and it can perform many transformations on the given dataset. For example, the data transformations tool output can be put into Java or Python, the other languages that generally data scientists learn in their initial or in the early stages of their career or so. Advantage of this feature or the Data Wrangler software is that a subset of data can be manipulated in Wrangler via its GUI and the same operations can be done out as a Java or Python code. So, the code is generated by itself.
5) R programming- as an open-source software or 'R studio', this was majorly used by many of the applicants. For instance, the big companies like GINA Data will take it as a case study and also use this. So, R contains nearly 5000 or more packages of detailed data analysis and graphical representation. R studio is a tool that helps in data visualization. So, such as we train, we instruct, we try to employ the best practices. So, as well as packaging it in such a way that, it is easy to use and more robust in a way. The phenomena is what happened in Linux in the 1980s or 90s. So, R is also trying to use a similar kind of platform. So, Linux was easy to employ in those times. Now people are using R, everywhere R is in demand.

6) SQL analysis and services- it can perform very basic database analytics or maybe common data mining of functions involving aggregations, basic predictive models.
7) SAS/ACCESS- SAS provides integration between the SAS and analytic sandbox via multiple data connectors. It itself generally is used for file extracts, but SAS/ACCESS users can connect to the relational databases such as it can connect to 'Oracle', then it can connect to 'TeraData' and many data warehouse appliances such as maybe Green Plum is one of them, the enterprise applications such as SAP on Salesforce or so. So, these are used majorly for model planning, when we have not built the model, but we were only planning for it. So, these kinds of software are majorly used.

❖ Now, when we are building the model, I would just list a few of them which are more used here, and are definitely for building the model itself helps us a lot. So, these are open-source software, I would say open source, one is

1) R or R studio
2) Octave- Octave is also a free programming software language for computational modeling. It has some functionality of MATLAB as well, but MATLAB is a lesson software, Octave is a free version. So, that is freely available. So, major universities teaching use this to have learning in machine learning.
3) WEKA- it is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed in Java code. The code that could be generated was Java or Python.
4) Python can directly be used as well in the model building. It is a programming language that has multiple toolkits. It helps or is used in machine learning and analysis, such as various kits which are there which are helping data visualization. So, mat plot library is also used here.
5) SQL can also be used.

Now we have some license software. License means when we have to purchase them. In those we have:

1) SAS Enterprise miner- it permits the users to run the predictive or descriptive analytics, which are based upon the big volumes of the data across the enterprise. It

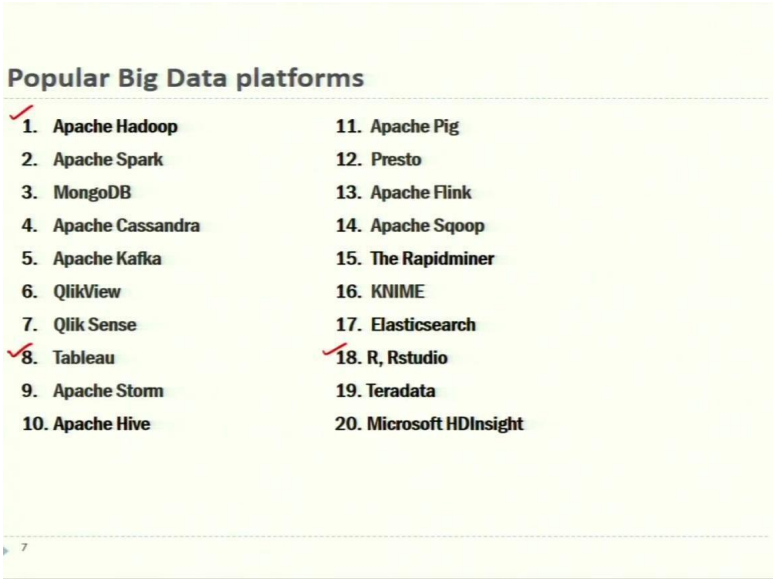
tried to interpolate and extrapolate the other data stores and has many partnerships and is bidding for the enterprise level computing analytics.

2) SPSS- SPSS is provided by IBM (IBM SPSS Modeler) so it offers to explain analysis through the GUI. SPSS is used for the design of experiments. It is used for conducting various analytics tests such as very simple T tests or analysis of variance or maybe multiple analysis of variance, analysis of covariance, we can have cluster analysis there, we can build factors. Factor analysis could also be conducted. SPSS is widely used by the management students in Data science.
3) MATLAB- MATLAB is a one of the most widely used software, which helps in data analytics, algorithms, data exploration. MATLAB has multiple toolboxes within itself. It can be worked up in the financial analysis, it can be worked upon the predictive analytics, maybe in forecasting, then it has a mathematical toolbox in it. So, multiple work boxes or toolboxes are available. So, the Statistics and Machine Learning Toolbox is one, then we have a Control System Toolbox, then comforting is also there, mapping or signal processing, then deep learning that DataFeed Toolbox, Image Processing, Financial Predictive Maintenance Toolbox. So, multiple toolboxes are there in the MATLAB. One cannot learn all the toolboxes definitely, if one of the toolboxes is maybe the person who is in finance, is trying to have the basic information of the financial flows or so, can use the financial tool box to ingest the data, to get the data output out, then to visualize simple moving averages out of the data, then create maybe a candle plot or customized data access, then maybe plot an indicator system also. So, this could also be used. In image processing, it is very important. Image processing detects the measure of the circular objects in an object, or maybe a similar kind of object in an image.

So, image processing is widely used in metallurgy, in face recognition, in recognition of the faults or maybe the truth or lying software. For example, in US immigration services, they use a model or a software that they have called ‘AVATAR’ in which they try to have a digital model of the person and based upon the person the base person is speaking the kind of pauses person is having the kind of expressions of the person’s face. It tries to identify whether the person is lying or is it all flow. Then this also helps them to determine whether the person is suspected maybe to do some illegal activity in the airport or so. US immigration services use these kinds of setups in image processing as well. In MATLAB, multiple tools could be used.

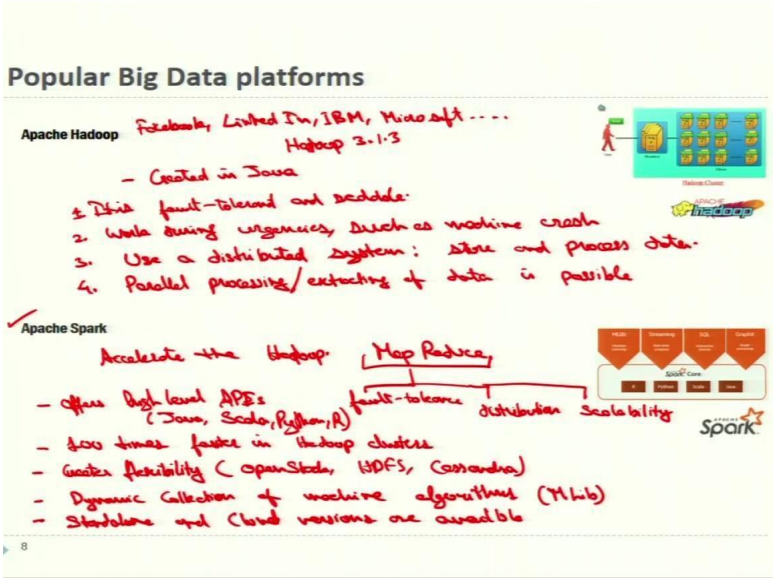
4) Alpine miners- it provides an excellent GUI front end for users to develop analytic workshops and interact with big data tools and platforms at the back end.
5) Statistica/Mathematica- These are also well used and popular procedures for data mining and using analytics.

(Refer Slide Time: 14:39)



Next, I have a list of the software which are common or which are widely used. Some of them are very mature technologies. For instance, Hadoop is very mature and is widely used, R studio is very widely used. I have not listed SPSS or SAS here, but they are majorly used for big data analytics and known for their work. So, those are being listed here. Tableau is very widely used. A few of them are raising technologies only, but they are expected to be great technologies in the future. So, let me try to start discussing all these technologies, the features of them and where they are used.

(Refer Slide Time: 15:21)



Apache Hadoop is the first that I will try to discuss. So, as I said, Hadoop is one of the most mature softwares. It is called as one of the best big data tools available. When people say Hadoop, the word Hadoop itself is now synonymous with big data analytics. So, it is used by the big business tycoons such as Facebook. Then we have LinkedIn, IBM who were the initial users of this system, Microsoft and many more business technologies, Hadoop is a low-cost system. It is a tolerant and highly available framework that is capable of handling data of all shapes and sizes. Hadoop is the most recent stable version that is Hadoop 3.1.3. This version is the most recent one. So, it was created in Java, I am just putting down the features of this.

- ❖ It was created by Apache Software Foundations only, that is why it is called Apache Hadoop. So, this is an open-source software framework for storing and handling big data. So, major attributes of Apache Hadoop system.

a) It is fault-tolerant and scalable. Further in fault tolerance that means if the fault is there, the mistakes which are not tolerable and which should be tolerable, the source software should be flexible and should be sensitive accordingly. So, it can be designed in such a way. And, scalable means analytics sandbox that we have developed. It can be doubled in size. It can be made multiple times of its size. It can also be reduced later. So, it is flexible. So, that is why it is widely used.

b) This framework is made to function even in unfavorable circumstances such as in a machine crash. Now, the framework makes Hadoop economical by storing the data across the commodity hardware, Hadoop uses a distributed system to store and

process data. So, data processing is fast and the results are fast because parallel processing of the data happens.

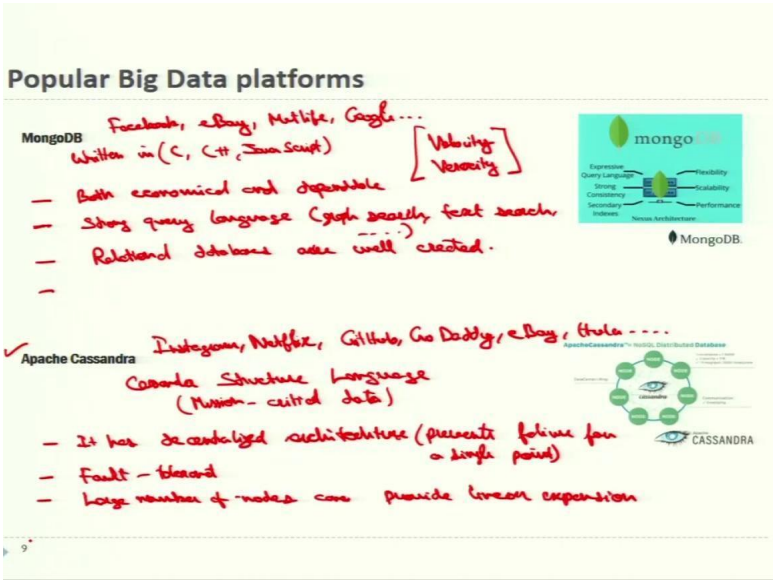
1) Apache did not stop with Hadoop as and when things kept on coming up. It also developed the software known as Apache Spark.

- It was created with the intention of accelerating the Hadoop big data processing only.

The main goal of the Apache Spark project was to improve upon the distributed, scalable, fault-tolerant processing framework of Map Reduce while maintaining expected benefits. So, it was built upon Map Reduce and putting in the same functions like fault-tolerance then Distribution. Distribution means the distributed system and scalability. While retaining the basic function that Map Reduce was already doing, it offers the ability for in memory computing to deliver speed. The spark offers high level APIs features. API is an application programming interface. So, it offers APIs such as Java, scala, Python, R and so on.

- So, in Hadoop structures or in Hadoop cluster itself, spark can run applications 100 times faster.
- Due to its ability to work with various data sources including OpenStack, HDFS or Cassandra, Apache Spark offers greater flexibility than Hadoop because it can work with various data stores including OpenStack, HDFS, Cassandra or so.
- So, it has a dynamic collection of the machine algorithms which includes clustering, collaborative, filtering, classification, regression. So, these are offered by the M library found in Spark. It can be used as a standalone on Hadoop or Apache mezzos or maybe some other softwares or it can be used with cloud.
- So, both standalone and cloud versions are available.

(Refer Slide Time: 23:25)



Now comes another software that is MongoDB (Mongo Database). It was created in 2009 and is widely nowadays also used by Facebook, eBay, MetLife, Google and many other systems like these because it is a NoSQL Database with an easy setup environment

- It was written in C ++, C or Java (JavaScript).

So, this open-source data analysis program MongoDB is one of the lightest databases for big data. That is why it is so because it makes it easier to manage frequently changing data as well as unstructured or semi structured data. So, the velocity (that is the Data that is coming quickly) and veracity, these features of big data are quite well handled by Mongo Database.

- So, it is a feature that is both economical and dependable.
- It has a strong query language because it has a NoSQL based database that supports graph search, text search, geometry-based search, aggregation or so. So, ad hoc queries such as indexing, sharking, replication are also supported.
- It possesses all the capability of a relational database and is well created in this. It is, as I said, used already by the big business tycoons here.

Another important software that I would like to discuss is Apache Cassandra. So, it is distributed open source, NoSQL or not just SQL so, both of them are a work being used in Apache Cassandra. It offers high availability and scalability without sacrificing the performance effectiveness. So, it is one of the most powerful big data tools. It can handle again, both structured and unstructured data in order to communicate with the database.

- It uses Cassandra structure language.

So, Cassandra's linear scalability and fault tolerance and low-cost hardware or cloud infrastructure makes it an ideal platform for Mission- critical data or the data that is required for this specific target that is very well handled in Cassandra.

- Cassandra is also being used by big companies like Instagram, Netflix, I also discussed this as an example in the previous lectures, then GitHub, GoDaddy, eBay, Hulu etcetera.
- It has decentralized architecture that prevents failure from a single point, but decentralized if I say it is one of the data points in the cluster is even outlier or if something is that is unwanted or so, that will try to affect the whole cluster itself, that is removed whole clusters sometimes in the centralized system is not kept meaningful. So, for instance, if the class heights of the students in the class is one of the datasets that we are creating, and there is one person who is maybe 6 feet 10 inches tall, so this tries to hamper my overall system, overall cluster that I am creating, it will be trying to skew it towards the left.

Now decentralized architecture means the single database can be taken off and it can be rebuilt. So, it is a decentralized architecture that prevents a (failure) from a single point. So, that is why Cassandra is widely used because it is again a flexibility feature that is widely or very prominent in this. So, it is again extremely durable and fault-tolerant. Definitely, we have to define the boundaries that are the faults, what are the systems that it could accept, it can help in trying to identify that.

- Large number of nodes can provide linear expansion.

In actual applications it performs better than a well, light, NoSQL substitute or so. So, Cassandra is widely used by big programmers or big companies to prepare the data majorly. To prepare the data means, how to store the data, the data that could be extracted from the sandbox when analytics sandbox is built. So, what kind of tools are to be used there that are put here, if Cassandra is one of the two selected, tend to extract the data whether it is in a structured or unstructured form that can be taken off. If we even take a few data points or outliers out of the study, the Cassandra tool still allows you to perform while processing itself.

So, with this, I will let us look to have a break and I will continue the different technologies, the platforms or the software, which are there in Big Data Analytics in the next lecture. Thank you.