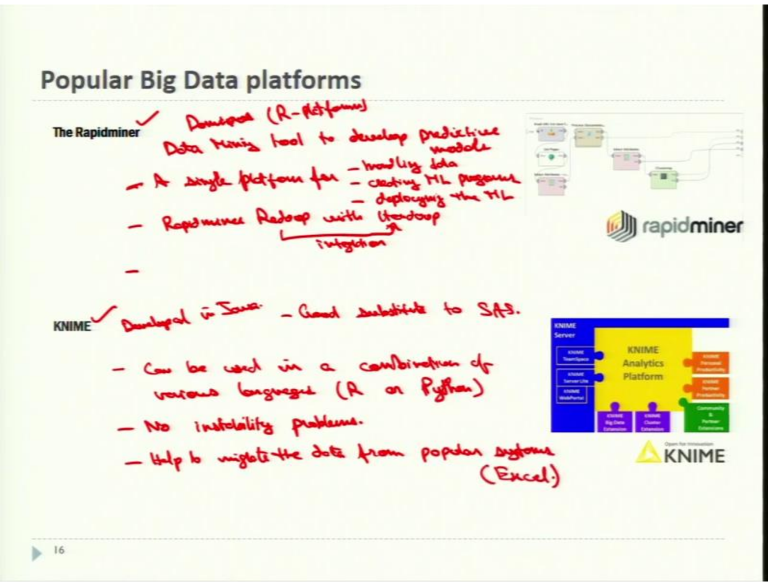
Computer Aided Decision Systems - Industrial Practices using Big Analytics

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

Lecture 26
Big Data Analytics Tools and Software (Part 4 of 4)

(Refer Slide Time: 0:18)



Popular Big Data platforms

The Rapidminer ✓ Developed (R-platform)
Data Mining tool to develop predictive models
- A single platform for - handling data
  - creating ML programs
  - deploying the ML
- Rapidminer Radoop with Hadoop
  integration
-

KNIME ✓ Developed in Java. - Good substitute to SAS.
- Can be used in a combination of various languages (R or Python)
- No instability problems.
- Help to migrate the data from popular systems (Excel)

16

Welcome back to the course on Computer Aided Decision Support systems. Let us continue our discussions on the different software platforms of Big Data. The next software I have is 'Rapidminer'. Rapidminer has become one of the most popular tools for applying data science. So, there is a case study by Domino’s where they improved their prototype, meaning the kind of the food that they were producing and their demand forecast while using the Rapidminer software.

So, it held the top spot in one of the data science platforms in 2017. It also got an award for that. It is an effective tool for data mining that helps to create predictive models.

- ❖ It is a data mining tool to develop predictive models. When you say predictive models that means forecasting is the best that it could handle. Data preparation, machine learning, and deep learning are all features which are included in all-in-one tool that is known as Rapidminer, which means Rapidminer is a tool that helps to have better AI (Artificial Intelligence).

There are multiple case studies a Rapidminer work with like, Future Bright Analytics or International Education Company use Rapidminer systems or Rapidminer analytics to have the growth in the student success at scale, then for cyber security itself there is application which in which they identified the mitigating threats by using Rapidminer. So, Rapidminer is used as one of the supply chains as I said in Domino's case as well with the forecasting of the demand across the supply chain was crucial and how the data science team at Domino's tackle the challenge and work through complex time series forecasting from prototype to delivery.

This was developed in a 'R' software only. So, Dominos use the R platform as well. So, there were two major things: prototype and delivery for both of them. So, the reduction in the errors and quickening the runtime. This was taken through the Rapidminer.

- ❖ Rapidminer provides a single platform for handling data, creating machinery programs, then creating machine learning programs and deploying the programs.
- ❖ We have Rapidminer Radoop with Hadoop. It uses automated modeling to produce predictive models. It is already mentioned.

So, Rapidminer case study is one that I would put a video link in the reference slide, so you can just have a quick look that how Domino's manager for the data science Ryan Frederick praised this, that is how the software is. It is not about Rapidminer, it is not about the specific case study or so. It is how database analytics, the big database analytics and how artificial intelligence is being enhanced using these small platforms which were developed.

Now 'KNIME' is another platform. KNIME is an open-source data analytics platform for data science or data analysis and business intelligence. It was developed in Java. It enables users to interactively create views and models, visually create data flows, then execute the analysis steps which are needed.

- ❖ It is a good substitute to SAS being open sourced.
- ❖ It provides the basic ETL operations and it can be used in a combination with various languages.

So, it provides a wide range of integrated tools and cutting-edge algorithms. So, setting up KNIME is also simple.

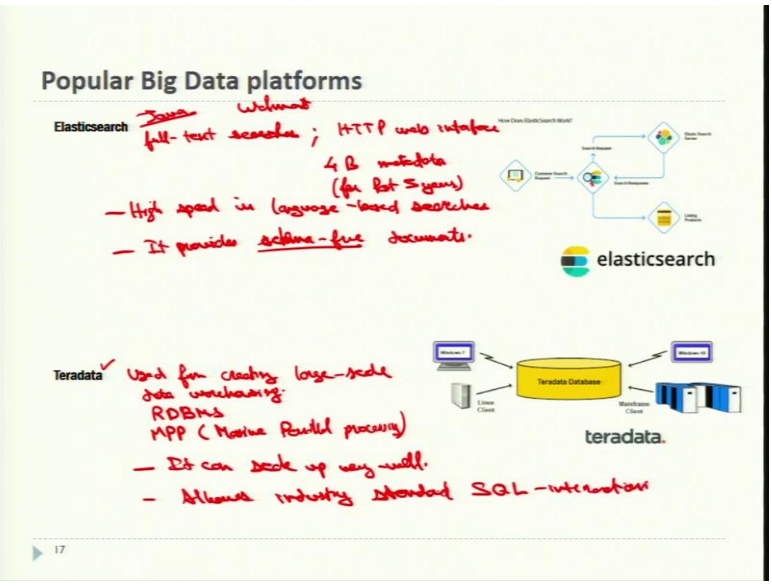
- ❖ It has no instability program, no instability problems are there. So, KNIME is an interface which can have a graphics analytics platform.

It can be used or combined with various languages as I said combination of various languages, for example R or python. The workflow interface is graphic based which means it can have a better user interface, that what we are trying to present to the people or a team that we have already created.

It can also be used in migrating data from very simple programs like from the excel sheet itself. We can migrate the program to KNIME and try to create the visual views of it. So, from the excel itself, we can get the program and try to run a small analysis that we require.

❖ It can help to migrate the data from popular systems, for example MS Excel.

(Refer Slide Time: 7:42)



Popular Big Data platforms

Elasticsearch ~~Java~~ Walmart
full-text searches ; HTTP web interface
4 B metadata
(for last 5 years)
– High speed in language-based searches
– It provides schema-free documents.

Teradata ✓ Used for creating large-scale data warehousing.
RDBMS
MPP (Massive Parallel processing)
– It can scale up very well.
– Allows industry standard SQL-interaction

17

'Elastic search' is another platform which is developed in Java and it was used by 'Walmart'. So, what is this? It is an open-source database server, which is used to conduct full text searches. Not only full text searches, it also does analysis on the searches that data is conducted, which is conducted on an HTTP web interface. It collects the unstructured data and organizes them into an algorithm or a program which are complex and suited for the language-based searches. Language based searches that are the full text of the language could be searched or that could be then programmed to have a useful output of it. How does it work?

In the Walmart case study, Walmart ingested 4 billion metadata records. The metadata records; who is the person? What was the item number? the item number, where was it sent? What was the expiry date of the data? What date before expiry versus sold? How many items were sold in what store? So, this could be all metadata points. So, these records were put for the past 5 years.

So, this was put in the Elasticsearch and Walmart Global risk analysis team can identify instances of fraud in real time that is especially when gift card scams were targeted at senior citizens. So, how they could hear using elastic security. So, alerting the people, the graph features to save consumers and millions of fraudulent gift card requests when the people try to use the Walmart's name and try to do swamp frauds or so.

- Elasticsearch is a trustworthy and simple to scale program so even when searching through a very large data set, it provides high speed in language-based searches.

Elasticsearch makes searching, indexing, querying of the data simple by providing straightforward result full API.

- When I say schema free documents, it means that the raw data is not tested. A schema free database makes almost no changes to your data, that is each entity in the data is saved in its own document with the partial schema, leaving the raw data information untouched. So, the beauty of Elasticsearch is that your raw data, that your raw language files or language program that attaches full text searches that were made are all safe.

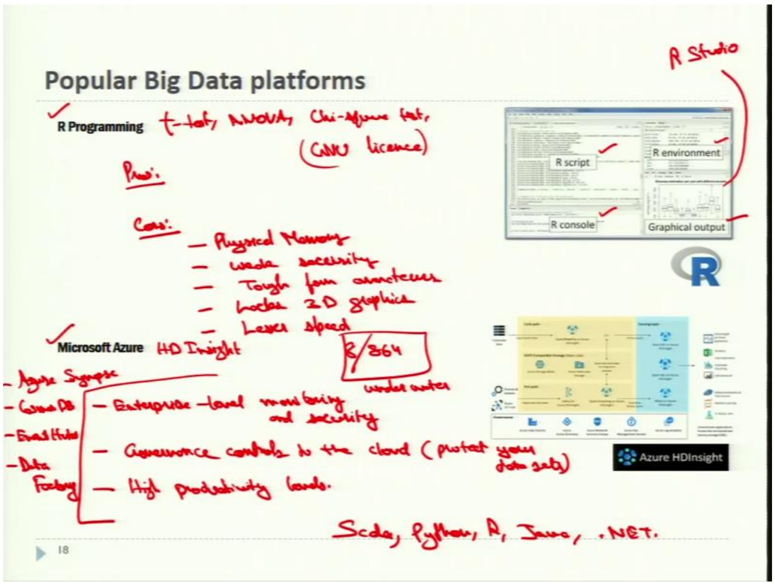
Next comes 'Teradata', Teradata is a tool for creating large scale data warehousing. So, in the large-scale data warehousing, where applications are developed using Teradata. It is again a Relational Database Management System (RDBMS) and it typically provides a comprehensive database warehousing solution.

- Massive parallel processes could have run.

Network attached systems or mainframes can be connected using Teradata in nodes, a passing engine, a message passing layer, an access module processor, make up its important parts.

- It can scale up very well.
- It allows for industry standard SQL interaction.

(Refer Slide Time: 12:58)



Popular Big Data platforms

✓ R Programming t-test, ANOVA, Chi-square test, (GNU licence)

Pros:

Cons:
- Physical Memory
- Weak security
- Tough for amateurs
- Lacks 3D graphics
- Lesser speed

✓ Microsoft Azure HD Insight 8/364 under water

- Azure Synapse
- Cosmos DB
- Event Hubs
- Data Factory

- Enterprise-level monitoring and security
- Governance controls to the cloud (protect your data sets)
- High productivity tools.

Scala, Python, R, Java, .NET.

18

Next comes the two most important softwares which are available. I will call the big data platforms which are most widely used. 'Microsoft', we cannot ignore Microsoft has its Azure data analytics platform. R programming is one, which is the highly or the most highly sought of programming languages because it is free and it has a good R Studio.

It has an R script where we can write a script of the program and it also has a console where you can just download these 10 different points and the R environment it tells you what are the programs which are running and also, we have R studio, the graphical output, here we call it as R Studio. So, R can help us to develop many kinds of different visualizations. For instance, we can have graphical tools for just plotting a simple histogram, a pie chart, a line diagram. Here you can see a small box and whisker plots are there. So, the R tool is very widely used. Statistical computing graphics and big data analysis programming languages are used here.

It offers a huge selection of statistical tests. For instance, if we need to do a T Test, if we need to conduct an ANOVA, if we need to conduct a classification test that is the chi square test. So, these all could be done using R software. The tools for R programming offer efficient data handling and storage. It offers a logical and comprehensive set of big data tools for data analysis. The graphics are unmatched for any other freeware platforms which are available.

Software engineers commonly use R programs for data analysis, for statistical tests. R programming language is extensible and offers an open-source environment allowing programs to access it free.

However, this language must be used with the terms provided in the free software foundation GNU, that is there is GNU license. So, GNU license is the general public license, that is the free software foundations, GNU is there. So, R programming language could be used with different kinds of the operating systems like it could be used in Linux, Windows, Mac operating system. The language has a major benefit that it is an effective statistical programming language while considering its drawbacks even.

So, when we start using it the main benefits of the R program would definitely be overlooked or would be overrun, whatever small drawbacks are there. So, it has a broad range of libraries, excellent statistical calculations and data analysis could be conducted.

It supports across different platforms and it supports across different types of the data, for example wide range of the vectors, the matrices, arrays then data objects of (different) diverse sizes could be ingested in it, data cleansing or data transforming is one of the main features that is R programming allows one to do the web scraping to collect the data, to collect the most useful data for the research that the one is trying to work upon.

It also allows one to have data cleansing while allowing to identify and remove the correct or arrays or corrupt data. It has powerful graphics, definitely, that is talked upon, there are small drawbacks for the R program. So, there are certain pros and cons. The pros I have just mentioned, the cons would be in data handling, it uses physical memory. So, in contrast the language is like python or so, the R uses memory, that is little more than python.

Basic security is lacking in R because it is weak software. It is open access software. It is an essential part of the most prominent languages such as python, that the basic security is there because many restrictions are there. So, R cannot be embedded in a web application. To some people, it is a complicated language in a way that it has a deep learning curve. The people who do not have prior knowledge of programming or maybe the basic syntax of C or C ++. Sometimes I may find it difficult to learn R in the beginning itself.

So, like the excel sheet or maybe putting the data in just a graphical system may be just like KNIME. So, it is easy to work upon so R program one has to have basic understanding of the programming languages.

- It uses physical memory
- It has weak security
- It is considered little tough language or little hard language for amateurs.

Then the difference between R and python or other big languages is that R also has a weak origin, that is it does not have a support for dynamic or 3D graphics. Though the graphics, whatever statisticians would like to see do come there, but the graphics in the 3-dimensional form are still lacking here,

- Lacks 3D graphics.

Also its speed is little slower than the other set of the languages such as Matlab, python they are faster than what R does, but still because of the advantages of the R, which I just mentioned R is still most highly used programming languages and the people who would like to just start with the data analytics always majorly focus on the R program, MS Excel or Microsoft Excel sheets or these do help but R is a free license, so we can develop even the big case studies which I discussed about the Netflix also used its program in the beginning in R itself. Walmart also used R to store the programs and to have the program run upon them.

So, next is 'Microsoft Azure' because the big data analytics business was all coming up with a big boom Microsoft also developed its program known as HD insight, which is now known as Azure software. So, Spark and Hadoop service in the cloud is called Azure HD insight. So, both Spark and Hadoop, when it is run in the cloud and the data centers where those are located. Nowadays, I will just let you know that the data centers that Microsoft have located are underwater since 2015 it ran a program where they store the data sets or the data centers the other servers underwater, with the underwater they also observed that the failure that happens in the next four years one minimal or one about one eighth of that those run in the data centers on the surface.

So out of 864 terminals, only 8 were crashed underwater, which means the human interaction is one of the factors which was completely absent underwater so that also sometimes increase your crashing of the data points also or data centers service points, so, HDInsight is a software that helps you to have a scale cluster for the organization to use or run its big data workloads.

- It provides enterprise level monitoring and security, then it stands down the premises security.
- Governance control to the cloud that means it protects your data sets.
- It offers developers and scientists a platform with a higher level of productivity, because it is an enterprise-based software, enterprise level if I can say that means all parts of the system or the enterprise. When we say the enterprise research management, whether it

is purchasing of the material, whether it is the transfer of the material to the machine shop or to the processing system, in a manufacturing concern or it is the salaries of the people who are working upon and they are the timely or workload of the people who have to be designed, all the datasets could be there. Then the forecasting, the finances from where and how are the funds coming? How do the funds flow? So, what are the different rates of return payback period, what are the net present value systems, these are all in finances at the enterprise level. It tries to support you at different kinds of the levels and different kinds of the functions of organization, whether it is production, finance, human resource or marketing everywhere. Azure Microsoft does help you to ingest the data and try to transform into the useful form.

So, Microsoft keeps on updating the Azure system or Azure platform with the latest open-source ecosystems. For example, the newest releases of the open-source frameworks such as Kafka then Hive or so, HD insight on Azure supports a list of latest open-source projects from the Apache Hadoop or Spark ecosystems.

It also helps you to integrate natively with the Azure services, that is you can build your own data lake, it is called Data Lake here or you call it analytics sandbox. Through seamless integration with the Azure data storage solutions and services including. There are different modules of Azure, they call it Azure synapse or you call it Azure Cosmos database or you call it Azure data Lake storage, Azure blob storage, Azure event hubs or Azure data factory.

Integration of the different functions and or the horizons of the enterprise could be done here and it provides flexibility to use multiple language tools such as Scala, python, then R, JavaScript or .NET. So, this is how Azure or Microsoft has also taken its lead while including the Apache Hadoop, Spark, Hive, Kafka and more and using Azure HD insight as a customizable enterprise great service for open-source analytics.

So, a massive amount of data can be taken and benefits of the broad open-source project ecosystem with a global scale of Azure, it easily migrates your big data workloads into the cloud and you can work upon them. With this I would like to take a rest here. With this the major softwares or major platforms which are available for the big data are discussed. I would like to discuss a case study on the big data analytics life cycle, the life cycle which I discussed in the previous week, the case study would be taken on a network service provider application. Thank you.