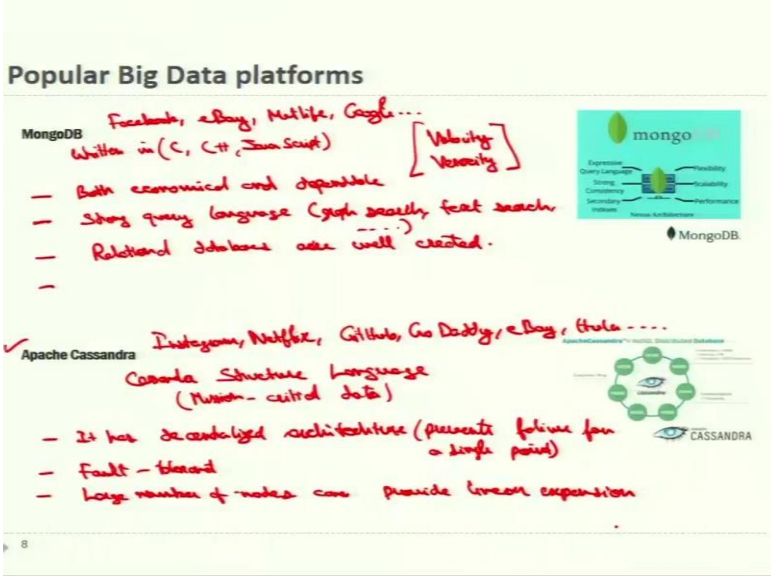
Computer Aided Decision Systems - Industrial practices using Big Analytics

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 24
Big Data Analytics Tools and Software (Part 2 of 4)

Welcome back to the next lecture this week. We were discussing the various software packages or the tools which are widely used in big data analytics.

(Refer Slide Time: 0:26)



Popular Big Data platforms

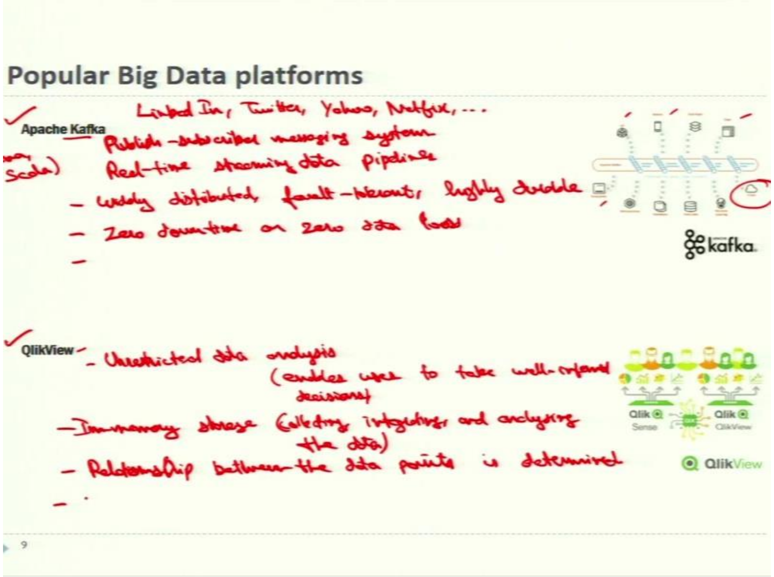
MongoDB Facebook, eBay, Metlife, Google...
Written in (C, C++, Java Script)
[Velocity
Veracity]
- Both economical and dependable
- Strong query language (graph search, text search ....)
- Relational databases are well created.
-

✓ Apache Cassandra Instagram, Netflix, GitHub, Go Daddy, eBay, Hulu....
Cassandra Structure Language
(Mission-critical data)
- It has decentralized architecture (prevents failure for a single point)
- Fault-tolerant
- Large number of nodes can provide linear expansion

8

I have discussed MongoDB and Apache Cassandra in the recent slide.

(Refer Slide Time: 0:31)



Popular Big Data platforms

Apache Kafka

LinkedIn, Twitter, Yahoo, Netflix, ...

- Publish-subscribed messaging system

Scala)

Real-time streaming data pipelines

- widely distributed, fault-tolerant, highly durable
- Zero downtime on zero data loss
-

QlikView

- Unrestricted data analysis (enables users to take well-informed decisions)
- In-memory storage (collecting, integrating and analyzing the data)
- Relationship between the data points is determined
-

9

Next, let me try to continue with another software known as Kafka, Apache Kafka. Apache also developed a messaging system known as Apache Kafka. In this messaging system, Apache, try to have access to the majorly unstructured data. So, we can have messages from the mobile services or from the SAS applications, from edge, from IoT or maybe from data centers.

So, then we have micro-services, databases, data lake so, all these things are developed, when we try to develop a database change, we try to have and processing on the customer interactions, we try to store the data and internet streams or in the inventory streams of the internet, then SAS data, SAS data it is also taken away. So, this can be stored and put in the cloud software.

So, Kafka is an open source distributed streaming platform, so it is a

- Publish subscribed messaging system, which means it can handle a large amount of data that has reliable query or reliable queues. So, the queues of the data from the mobile services, so there is a small graphic here, which is showing the data is coming from the mobile services, it is coming from the database, it is coming from different internet of things. So, it is coming from maybe web services or so, images it has also processed which is all stored in the cloud. So, one can transmit the message from one place to another, and maybe put a thanks note and kind of the timeline that when we could set that, we can maybe schedule the messages as well.

- So, real-time streaming data, I would say or real time streaming data pipelines and applications are constructed using Apache Kafka. It is developed in programs such as Java, or Scala. Kafka is also widely used by many social media tycoons because I am talking about messaging systems. Definitely, LinkedIn is using it. Then along with LinkedIn, we have Twitter that is using it, Yahoo, then Netflix. So, these big business tycoons use Kafka for sending or receiving messages. It is because Kafka can easily handle large amounts of data. So, that is why it is known as big data system, big data platform or big data software.
- Kafka is widely distributed, fault-tolerant and highly scalable, widely distributed, it is again fault tolerant and is highly durable.
- So, it says almost zero downtime or zero data loss are the features that they can offer. So, data is extracted from a searchable repository, correlates it, indexes it, then it produces illuminating graphs, reports, alerts, dashboards. So, features support the processing of data in real time. So, that is majorly Apache has developed a platform for the messages, for the SMS, for the email, for the accessing of the textual data in a different form the Kafka platform is used.

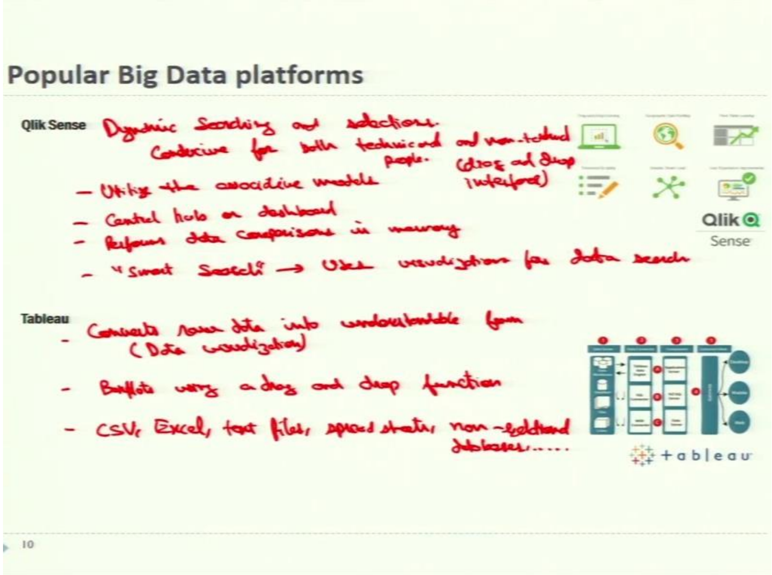
Next comes QlikView. QlikView is a fastest evolving business intelligence softwares, many managers say this and data visualization tool of this is QlikView, it is one of the best business intelligence tools for turning unprocessed data into a knowledgeable one, it is also used by big companies. So, it is simple, clear and uncomplicated user interfaces.

So, QlikView gives existing data storage or easy to trust stores, a whole new level of analysis or values. Users can perform or maybe have a direct performance or indirect searches across the entire population or entire is the application of the data that is available. So, no queries are fired when the user clicks on the data point.

- So, based on the user selection every other field filter itself; it encourages unrestricted data analysis. So, what does this help? This enables users to take well informed decisions, however, the users should have a basic intellect, basic information that how they understand this information that is unselected is very openly available, without much stringent cleaning or so, they can have quick access to it. So, it tries to store the data, the data stores now have a zero queue and new levels of the insights are there.
- So, it offers them in memory storage. So, that speeds up the process by collecting, integrating and analyzing the data.

- ❖ So, associative modeling is how it operates, the relationship between the data is automatically determined by a software between the data points is determined. It offers in a way robust and international data discovery and in assistance with mobile and social data discovery as well.

(Refer Slide Time: 9:02)



Popular Big Data platforms

Qlik Sense Dynamic Searching and selections.
Conducive for both technical and non-technical people. (drag and drop interface)
- Utilize the associative models
- Central hub on dashboard
- Performs data comparisons in memory
- "Smart Search" → Uses visualizations for data search

Tableau
- Converts raw data into understandable form (Data visualization)
- Builds using a drag and drop function
- CSV, Excel, text files, spreadsheets, non-relational databases.....

Next comes Qlik sense, Qlik sense is a tool of data analysis for majorly visualization.

- ❖ So, Qlik sense is like the QlikView, it uses an associative data setup and it links data from various sources and performs dynamic, searching and selections. So, this makes quick sense and associative user interface engine. So, both technical and non-technical users can utilize its data from the data analytics platform.
- ❖ So, it is, I would say, conducive for both technical and non-technical people because data visualization is where the non-technical people can better understand what the graphics are saying, what the graphs which are plotted are saying, so accordingly they can click. So, the user can easily create an analytical report that is in the form of a story and is easy to understand using a drag and drop interface. So, for non-technical people, a drag and drop interface is generally used. So, the client team can share secure data models, export data stories to improve the business and share applications and reports on a central hub.
- ❖ So, Qlik sense features if I tried to jot it down, it utilizes the associative model.
- ❖ All the documents and reports are created with the Qlik sense software and can be shared through a central hub or dashboard. It has a central hub or dashboard, which shares all the information. It performs data comparisons in memory.

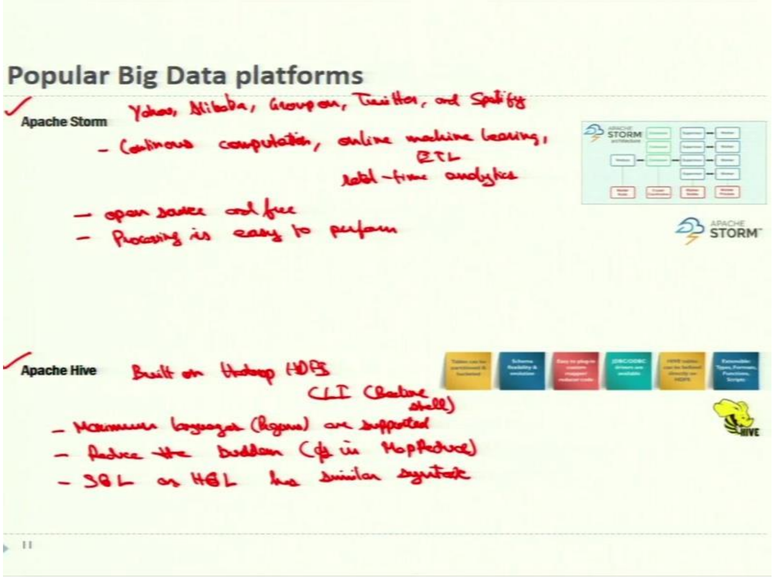
- It has a smart search feature. I will put that feature name here. It has a smart search feature which helps the data analysis through engagement which starts in visualization which uses visualizations for data search.

Now, Tableau in the business intelligence or the analytics sector, is a potent data visualization software solution tool. It is an ideal tool for converting raw data into a format that majorly people who are involved most of them can understand without the need of technical expertise or coding knowledge.

- So, it converts majorly I would say raw data into understandable form because it uses data visualization. So, when I say data visualization, the kind of the streams which are being shown here, it can be created from the data servers, it can use the data connectors to connect to the components and the customers and the clients. So, then it derives the data from different sources and tries to create tables or graphics out of it. So, it is easy to understand by people who are non-technical as well, or people who do not have the coding knowledge. So, tableau enables user to work with real time datasets, and transform raw data into insightful knowledge and improve decision making. It provides a quick data analysis process that yields interactive dashboards and worksheets for visualizations. So, the collaborations with other big data software the tools are also there that Tableau can offer and it helps to function properly even while using or working with the other software systems or so.
- Tableau you create visualization that is already jotted down here. And it can create a bar chart, pie chart, histogram, tree map, Gantt chart, bullet chart, boxplot and many others just using a drag and drop system. So, even I would say the box plots using a drag and drop function.
- So, Tableau provides a wide ranges of data sources, generally very common files such as we have CSV, Microsoft Excel, then we have text files and majorly some spreadsheets, then non-relational databases, it gives you the wide range of these kinds of data sources to develop a usable or readable or reliable and secure source or the visualization from which people can at least understand what are the models they are going to try to build.

So, Tableau is a model preparation software, it is a data preparation software, it is a model building system as well where we can first visualize that and using the visualization one can build them.

(Refer Slide Time: 16:04)



Popular Big Data platforms

✓ Apache Storm Yahoo, Alibaba, Groupon, Twitter, and Spotify

- Continuous computation, online machine learning, ETL real-time analytics
- open source and free
- Processing is easy to perform

✓ Apache Hive Built on Hadoop HDFS CLI (Beeline shell)

- Numerous languages (Regions) are supported
- Reduce the burden (of in MapReduce)
- SQL or HQL has similar syntax

So, along with Tableau we have Apache Storm that is again a distributed real time computation software which was developed using Java and cloture. Apache Storm is used by Yahoo, Alibaba, Group on, Twitter and Spotify. Spotify for instance, if you need to understand what are the new kinds of the songs or the streaming or the podcast a person is trying to more focus upon is a specific area.

So, it can try to have an access to the different data points, places that people are clicking, or the points where the people are trying to stop and try to re listen or re-iterate the songs that you are trying to hear in between one of the lines could be might be multiple times try to be liked by the person. So, those small data points are taken out of the analytics system. So, which uses Apache Storm.

- So, Apache Storm can be used for a continuous competition, because I am using or I am specifying your Spotify or Twitter as well, where people keep on changing their ideas. So, it has to be a continuous competition, which is a feature for the Apache Storm. Also, it is known as online machine learning or online ETL or we can call it real-time analytics.
- Storm features if I tried to jotted down, it is an open source and free software it can scale up very well. It is simple to setup and fault-tolerant.
- The processing of data is guaranteed via Apache Storm. A process is easy to perform, it is capable of handling millions of peoples per node, per second. So, it is widely used by these big business people.

Apache also built a platform known as Apache hive, which was again built upon its Hadoop HDFS. So, without having to create difficult MapReduce jobs it was based upon this. So, where users can communicate with hive using the CLI, that is what they call a beeline shell. It is an open-source database housing tool for big data analysis that is known as hive. It uses the similar SQL hive query language. In Apache hive, all the clients, applications written in various languages are supported.

- So, I would say maximum languages, that is the programming languages, programs are supported in it and it lessens the burden of creating intricate MapReduce jobs. So, what is it in MapReduce? It generally tries to avoid that then SQL or HQL share a similar syntax. It is a hive query language which has similar syntax. So, as a result someone who is familiar with SQL already can write the hive query languages.

So, with this again I would like to have a break and I will continue the software packages in the next lecture, where I will try to have a few examples with the forthcoming software's that I am trying to discuss. Thank you.