UNIT - II

Big data

Big data is a term for data sets that are so large or complex that traditional data processing application softwares are inadequate to deal with them.

Challenges include capture, storage, analysis, data curation, search, sharing, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

"There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem

What is big data?

Big data describes a holistic information management strategy that includes and integrates many new types of data and data management alongside traditional data.

Big data has also been defined by the four Vs:

Volume:

The amount of data. While volume indicates *more* data, it is the granular nature of the data that is unique. Big data requires processing high volumes of low-density, unstructured Hadoop data—that is, data of unknown value, such as Twitter data feeds, click streams on a web page and a mobile app, network traffic, sensor-enabled equipment capturing data at the speed of light, and many more.

It is the task of big data to convert such Hadoop data into valuable information. For some organizations, this might be tens of terabytes, for others it may be hundreds of petabytes.

Velocity :

The fast rate at which data is received and perhaps acted upon. The highest velocity data normally streams directly into memory versus being written to disk.

Some Internet of Things (IoT) applications have health and safety ramifications that require real-time evaluation and action. Other internet-enabled smart products operate in real time or near real time. For example, consumer eCommerce applications seek to combine mobile device location and personal preferences to make time-sensitive marketing offers.

Operationally, mobile application experiences have large user populations, increased network traffic, and the expectation for immediate response.

Variety :

New unstructured data types. Unstructured and semi-structured data types, such as text, audio, and video require additional processing to both derive meaning and the supporting metadata.

Once understood, unstructured data has many of the same requirements as structured data, such as summarization, lineage, auditability, and privacy.

Further complexity arises when data from a known source changes without notice. Frequent or real-time schema changes are an enormous burden for both transaction and analytical environments.

Value:

Data has intrinsic value—but it must be discovered. There are a range of quantitative and investigative techniques to derive value from data—from discovering a consumer preference or sentiment, to making a relevant offer by location, or for identifying a piece of equipment that is about to fail.

The technological breakthrough is that the cost of data storage and compute has exponentially decreased, thus providing an abundance of data from which statistical analysis on the entire data set versus previously only sample. The technological breakthrough makes much more accurate and precise decisions possible.

However, finding value also requires new discovery processes involving clever and insightful analysts, business users, and executives. The real big data challenge is a human one, which is learning to ask the right questions, recognizing patterns, making informed assumptions, and predicting behavior.

Why big data?

Many of today's data processing platforms let data scientists analyze, collect and sift through various types of data. While it does take some technical know-how to define how the data is collected and stored, many of today's big data and business intelligence tools let users sit in the driver's seat and work with data without going through too many complicated technical steps.

This added layer of abstraction has enabled numerous use cases where data in a wide variety of formats has been successfully mined for specific purposes. One example is real-time video processing. The 2012 Summer Olympic Games in London made heavy use of closed-circuit video, with 1,800 cameras monitoring Olympic Park and the athletes' village.

Teams of analysts used applications to process data pertaining to those who were filmed and flag any individuals behaving suspiciously.

How-to: 5 Tips to Find and Hire Data Scientists

Another example is medical transcription. As electronic health record (EHR) use grows, healthcare organizations are increasingly using natural language processing systems to transcribe, extract and process data within a clinical context.

You'll Benefit From Speed, Capacity and Scalability of Cloud Storage Organizations that want to utilize substantially large data sets should consider third-party cloud service providers, which can provide both the storage and the computing power necessary crunch data for a specific period.

Cloud storage presents two clear advantages. One, it lets companies analyze massive data sets without making a significant capital investment in hardware to host the data internally.

Two, as internal IT departments recognize that big data hosting platforms require new skills and training, they find that a hosted model tends to abstract that complexity, enabling more immediate deployment of big data technology. This also lets developers build a sandbox environment that's preconfigured and ready to go without having to set up the necessary configurations from scratch.

Your End Users Can Visualize Data While the business intelligence software market is relatively mature, a big data initiative is going to require next-level data visualization tools, which present BI data in easy-to-read charts, graphs and slideshows.

Due to the vast quantities of data being examined, these applications must be able to offer processing engines that let end users query and manipulate information quickly—even in real time in some cases. Applications will also need adaptors that can connect to external sources for additional data sets.

Analysis:

Four Barriers Stand Between You and Big Data Insight Usability is another consideration. CFOs, CMOs and other non-IT executives are looking to leverage data, so they need access to charts, infographics and dashboards.

Fortunately, leading BI vendors are shifting from an IT-driven to self-service analytics model that puts business users in the driver's seat.

This accelerates adoption as well as return on investment and expands analytics' reach beyond report writers and more technical end users.

Your Company Can Find New Business Opportunities As big data analytics tools continue to mature, more users are realizing the competitive advantage to being a data-driven enterprise. The 2012 presidential election demonstrated this.

Campaign managers in both the Democratic and Republican parties saw a critical need for information on voters and their specific interests; taking this info and addressing an issue through a customized email or flyer meant the potential to gain or sway a vote.

Analysis:

2012 Presidential Election a Victory for Quants Information regarding our preferences, likes and dislikes is critical to more than just political candidates. Social media sites have identified opportunities to generate revenue from the data they collect by selling ads based on an individual user's interests.

This lets companies target specific sets of individuals that fit an ideal client or prospect profile.Finally, big data use cases in about in retail, where the focus is on gaining insights by studying consumer behavior in online stores or physical shopping centers.

Your Data Analysis Methods, Capabilities Will Evolve Data is no longer simply numbers in a database. Text, audio and video files can also provide valuable insight; the right tools can even recognize specific patterns based on predefined criteria. Much of this happens using natural language processing tools, which can prove vital to text mining, sentiment analysis, clinical language and name entity recognition efforts.

One example that highlights the use of audio analysis and big data comes from MatterSight. This call center tool can match incoming caller to the appropriate customer agent by using predictive behavioral routing and other analytics technology.

MatterSight performs audio analysis to identify and score the calls based on specific criteria and then match customers with the best department to ensure the best experience. These advanced capabilities highlight some of the advancements we continue to see in unstructured data analysis and Big Data capabilities.

Other characteristics of data but not definitional for big data

VERACITY

Although there's widespread agreement about the potential value of Big Data, the data is virtually worthless if it's not accurate. This is particularly true in programmes that involve automated decision-making, or feeding the data into an unsupervised machine learning algorithm.

The results of such programmes are only as good as the data they're working with. Sean Owen, Senior Director of Data Science at CloudEra, expanded upon this: 'Let's say that, in theory, you have customer behaviour data and want to predict purchase intent. In practice what you have are log files in four formats from six systems, some incomplete, with noise and errors.

These have to be copied, translated and unified.' Owens' US counterpart, Josh Wills, said their job revolves so much around the cleaning up of messy data that he was more a 'data janitor' than a data scientist. What's crucial to understanding Big Data is the messy, noisy nature of it, and the amount of work that goes in to producing an accurate dataset before analysis can even begin.

VISUALISATION

Once it's been processed, you need a way of presenting the data in a manner that's readable and accessible- this is where visualisation comes in.

Visualisations can contain dozens of variables and parameters- a far cry from the x and y variables of your standard bar chart- and finding a way to present this information that makes the findings clear is one of the challenges of Big Data.It's a problem that's spurned a burgeoning market- new visualisation packages are appearing all of the time, with AT&T announcing their offering, Nanocubes, just this week.

VALUE

The potential value of Big Data is huge. Speaking about new Big Data initiatives in the US healthcare system last year, McKinsey estimated if these initiatives were rolled out system-wide, they “could account for $300 billion to $450 billion in reduced health-care spending, or 12 to 17 percent of the $2.6 trillion baseline in US health-care costs”. However, the cost of poor data is also huge- it’s estimated to cost US businesses $3.1 trillion a year. In essence, data on its own is virtually worthless. The value lies in rigorous analysis of accurate data, and the information and insights this provides.

Challenges with big data

In addition to buying the right software, recruiting the right talent ranks among the most important investments an organization can make in its big data initiative. Having the right people in place will ensure that the right questions are asked—and that the right insights are extracted from the data that's available.

Keep in mind that data scientists, as many refer to those working with big data, are in short supply and are being quickly snapped up by top firms.

Every CIO wants to keep his finger on the pulse of innovations that can transform his company, enhance existing business models and identify potential revenue sources. Enabling this business transformation means adopting the right tools, hiring the right people and—most of all—convincing executive leadership to embrace new models for using existing and brand-new data.

A successful big data initiative, then, can require a significant cultural transformation that's driven by the IT department. Highlight these five advantages of pursuing a big data initiative, though, and your executives are more likely to give you the resources, and the talent, you need to rise to the challenge.

Big data stack

Here's a closer look at what's in the image and the relationship between the components:

Interfaces and feeds: On either side of the diagram are indications of interfaces and feeds into and out of both internally managed data and data feeds from external sources. To understand how big data works in the real world, start by understanding this necessity.

In addition, keep in mind that interfaces exist at every level and between every layer of the stack. Without integration services, big data can't happen.

Redundant physical infrastructure: The supporting physical infrastructure is fundamental to the operation and scalability of a big data architecture. Without the availability of robust physical infrastructures, big data would probably not have emerged as such an important trend.

To support an unanticipated or unpredictable volume of data, a physical infrastructure for big data has to be different than that for traditional data. The physical infrastructure is based on a distributed computing model.

This means that data may be physically stored in many different locations and can be linked together through networks, the use of a distributed file system, and various big data analytic tools and applications.

Security infrastructure: The more important big data analysis becomes to companies, the more important it will be to secure that data. For example, if you are a healthcare company, you will probably want to use big data applications to determine changes in demographics or shifts in patient needs.

Exercises – Puzzle & fill in the blanks

a)The dot over the letter “i” is called ?

The strongest muscle in the body is ______.
2. ______ is the only animal that can not jump.
3. The dot over the letter “i” is called ______.
4. ______ invented the word “bump”.
5. _____ taste with their feet.
6. An _____eyes is bigger than his head.
7. _____ never sleeps.
1 word in 7 blanks? Guess the word.

b) Fill in the blank to make two words

Add an English word to the following set of words in such a way that the first word is completed and the second word starts.

Example – foot——— pen
Answer: ball
As it makes football and Ballpen.

c) Fill in the blanks with words starting with letter ‘W’

Example: dry fruit ….

Answer: Walnut

1 direction__

2. Vehicals have __

3 selling goods in large quantities__

4 color __

5 seven days __

6 buildings have __

7 term used in cricket __

8 purse __

9 portable music system __

10 domestic animal _

11 place for storing clothes __

12 noise made by pressure cooker __

13 fight _

14 determination _

15 march 8 _

16 aquatic mammal _

17 biscuits _

18 salary _

19 exercise _

20 health is __

21 two third of earth is___