Introduction to Big Data

By Dinesh Amatya

Big Data

Size of data ?

Technology ?

A method ?

Big data

“Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store,manage and analyze.”

- till 2003 was 5 billion gigabytes (exabytes)

- same amount was created in every two days in 2011 and

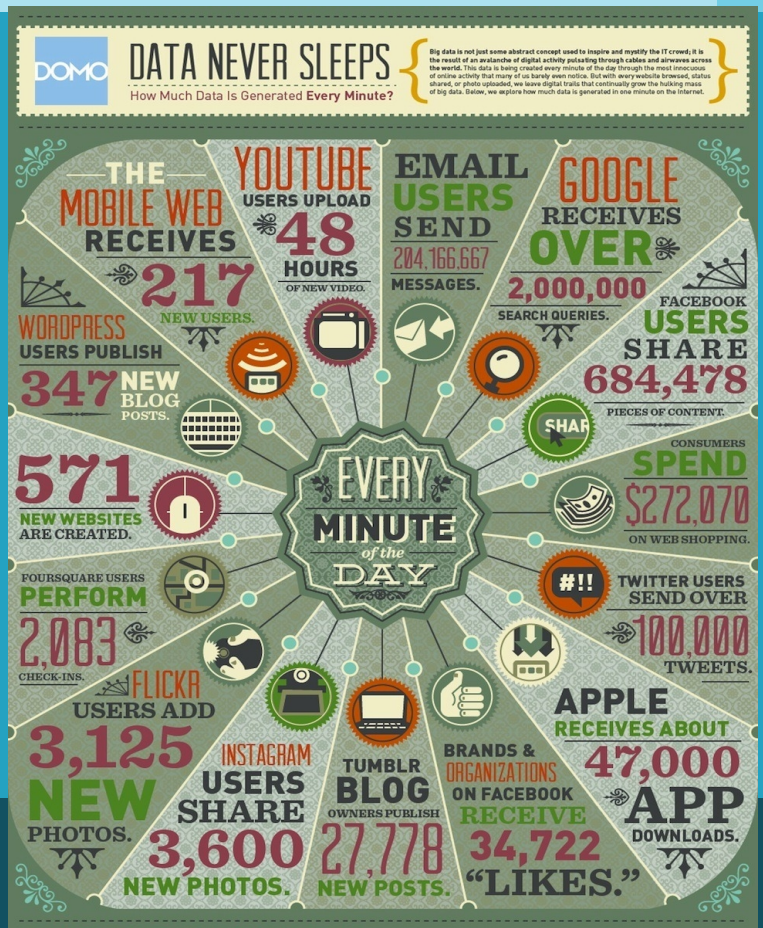
  every 10 minutes in 2013

Big data

Upload videos , Take pictures ,
Update status, Leave comments,
Web activities, Email

Machine generated data

Google, Yahoo,
Amazon, Microsoft



Big data characteristics

Volume

Velocity

Variety

Veracity

Technical Challenges

Fault tolerance

Scalability

Heterogeneous data

Data Analytics

- Big data analytics is the process of examining large amounts of data of a variety of types.
- The primary goal of big data analytics is to help companies make better business decisions and gain a competitive advantage. *e.g Amazon*
- analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence (BI) programs.

Data Analytics :Using Big Data to Get Results

- What problem are you trying to solve?
- Are you interested in predicting customer behavior to prevent churn?
- Do you want to analyze the driving patterns of your customers for insurance premium purposes?
- Are you interested in looking at your system log data to ultimately predict when problems might occur?

Data Analytics :Using Big Data to Get Results

Basic analytics

Advanced analytics

Operationalized analytics

Monetized analytics

Using Big Data to Get Results: Basic analytics

can be used to explore your data, if you're not sure what you have, but you think something is of value

- Slicing and Dicing
- Basic Monitoring
- Anomaly Identification

Using Big Data to Get Results: Advanced analytics

provides algorithms for complex analysis of either structured or unstructured data.

includes sophisticated statistical models,machine learning, neural networks, text analytics and other advanced data-mining techniques

can be deployed to find patterns in data, prediction, forecasting, and complex event processing

- Predictive modeling
- Text Analysis

Using Big Data to Get Results: Operationalized analytics

When a company operationalize analytics , it make them part of a business process

Example : Model for fraudulent claims in Insurance Company

Using Big Data to Get Results: Monetizing analytics

can be used to derive revenue above and beyond the insights it provides just for your own department or company

You might be able to assemble a unique data set that is valuable to other companies, as well

Example : Location based data , web-browsing data

Modifying Business Intelligence Products to Handle Big Data

Traditional business intelligence products were designed to work with highly structured, well-understood data, often stored in a relational data repository and displayed on your desktop or laptop computer

- Data
- Analytical algorithms
- Infrastructure support

Modifying Business Intelligence Products to Handle Big Data : Data

- It can come from untrusted sources
- It can be dirty
- The signal-to-noise ratio can be low
- It can be real time

Modifying Business Intelligence Products to Handle Big Data : Analytical algorithms

the algorithms need to be refactored, changing the internal code without affecting its external functioning

the algorithm needs to be data aware

analytics designed to be placed close to the big data sources to analyze data in place rather than first having to store it and then analyze it

Modifying Business Intelligence Products to Handle Big Data : Infrastructure support

- Integrate technologies
- Store large amount of disparate data
- Process data in motion
- Warehouse data

Data Scientist

- High ranking professional with training and curiosities to make discovery in the world of big data
- The people who understand how to fish out answers to important business questions from today's tsunami of unstructured information
- Newly coined term , in 2008 by D.J Patil and Jeff Hammerbacher
- A hybrid of data hacker, analyst, communicator,and trusted adviser. The combination is extremely powerful—and rare

Data Scientist

Sudden appearance of Data Scientist on the business scene reflects the fact that companies are now wrestling with information that comes in varieties and volumes never encountered before

If the organization stores multiple petabytes of data, if the information most critical to the business resides in forms other than rows and columns of numbers, or if answering the biggest question would involve a “mashup” of several analytical efforts, it has got a big data opportunity.

Role/Skill of Data Scientist

Data Scientist should have skill set to

- use technologies that make taming big data possible, including Hadoop (the most widely used framework for distributed file system processing) and related open-source tools, cloud computing, and data visualization.
- make discoveries while swimming in pool of data
- bring structure to large quantities of formless data and make analysis possible
- identify rich data sources, join them with other, potentially incomplete data sources, and clean the resulting set

Role/Skill of Data Scientist

- Data Scientist should have skill set to
  - communicate what they’ve learned and suggest its implications for new business directions
  - be creative in displaying information visually and making the patterns they find clear and compelling
  - fashion their own tools and even conduct academic-style research
  - write code
  - desire to go beneath the surface of a problem, find the questions at its heart, and distill them into a very clear set of hypotheses that can be tested

Current trend in Big data Analytics

Homework

References

http://www.dummies.com/how-to/content/distributed-computing-basics-for-big-data.html

http://www.business2community.com/digital-marketing/4-vs-big-data-digital-marketing-0914845

http://www.dataversity.net/distinguishing-analytics-business-intelligence-data-science/

http://hmchen.shidler.hawaii.edu/Chen_big_data_MISQ_2012.pdf

http://www.sciencedirect.com/science/article/pii/S0743731514000057

http://ac.els-cdn.com/S0743731514000057/1-s2.0-S0743731514000057-main.pdf?_tid=b37d8a0e-1742-11e5-92ad-00000aacb361&acdnat=1434801275_d30803ca170f0bb57decc236f04f6a14