Topic for the class:– Benefits and uses of data science and big data - facets of data

Unit _1 : Title-Data Evolution
Date & Time : 22.7.22 11.00 AM – 11.50 AM

Dr. Bhramaramba Ravi
Professor
Department of Computer Science and Engineering
GITAM School of Technology (GST)
Visakhapatnam – 530045
Email: **bravi@gitam.edu**

Unit1-syllabus

- **UNIT 1 Data Evolution 9 hours, P – 2 hours Data Evolution**: Data to Data Science – Understanding data: Introduction – Type of Data, Data Evolution – Data Sources. Preparing and gathering data and knowledge - Philosophies of data science - data all around us: the virtual wilderness - Data wrangling
-
- : from capture to domestication - Data science in a big data world - Benefits and uses of data science and big data - facets of data.
- https://www.coursera.org/learn/intro-analyticthinking-datascience- datamining?specialization=data-science-fundamentals

Big Data and Data Science: A Comparative Analysis

- Big Data Definition
  - Big data refers to large, complex data sets that are difficult to process using traditional data management techniques.
  - Data science involves methods to analyze and extract knowledge from massive amounts of data.

Big Data Characteristics

- Big data is referred to as the five Vs: volume, variety, velocity, and veracity and value
  - These properties distinguish big data from traditional data management tools, causing challenges in data capture, curation, storage, search, sharing, transfer, and visualization.

Data science and big data

- Data science is an evolutionary extension of statistics, incorporating methods from computer science.
  - Data scientists are distinguished from statisticians by their ability to work with big data and experience in machine learning, computing, and algorithm building.
  - Their tools include Hadoop, Pig, Spark, R, Python, and Java.

Python in data science

- Python is a popular language for data science due to its numerous libraries and support by specialized software.
  • Python's ability to prototype quickly with Python while maintaining acceptable performance is growing in the data science world.
-

BENEFITS AND USES OF DATA SCIENCE AND BIG DATA

- Commercial companies use data science and big data to gain insights into customers, processes, staff, and products.
  • Companies use data science to offer better user experience, cross-sell, up-sell, and personalize their offerings.
  • Human resource professionals use people analytics and text mining to screen candidates, monitor employee mood, and study informal networks.
  • Financial institutions use data science to predict stock markets, determine lending risk, and attract new clients.

BENEFITS AND USES OF DATA SCIENCE AND BIG DATA

- • Governmental organizations rely on internal data scientists to discover valuable information and share their data with the public.
  • Data scientists work on diverse projects such as detecting fraud and optimizing project funding.
  • Nongovernmental organizations (NGOs) use data to raise money and defend their causes.
  • Universities use data science in their research and to enhance the study experience of their students.
  • Massive open online courses (MOOCs) produce a lot of data, allowing universities to study how this type of learning can complement traditional classes.
-

Data Types and Their Implications in Data Science

- Facets of Data
  - Structured data: Depends on a data model and resides in a fixed field within a record. It can be stored in tables within databases or Excel files. SQL is the preferred way to manage and query data that resides in databases.
  - Unstructured data: Content is context-specific or varying, making it difficult to fit into a data model. Examples include regular emails, which contain structured elements but are challenging to find due to the numerous ways to refer to a person and the thousands of different languages and dialects.

Natural Language Processing Challenges

- Natural language is a unique type of unstructured data that requires specific data science techniques and linguistics.
  - Successful methods include entity recognition, topic recognition, summarization, text completion, and sentiment analysis.
  - Techniques trained in one domain may not generalize to other domains.
  - Even advanced techniques struggle to decipher the meaning of every text.
  - Humans also struggle with natural language due to its ambiguity and questionable concept of meaning.

Machine-generated data overview

- Machine-generated data is information created automatically by machines without human intervention.
  • It is a significant data resource, with the market value of the industrial Internet projected to be around $540 billion in 2020.
  • The internet of things, a network of 26 times more connected things than people, is expected to grow.
  • Machine data analysis requires scalable tools due to its high volume and speed.
  • Examples of machine data include web server logs, call detail records, network event logs, and telemetry.
  • Classic table-structured databases may not fit machine data, which requires interconnected relationships.

Graph-based or network data overview

- • Graph data refers to mathematical structures that model pair-wise relationships between objects.
  • Graph structures use nodes, edges, and properties to represent and store graphical data.
  • Graph-based data is a natural representation of social networks, allowing for calculation of specific metrics like influence and shortest path.
  • Examples of graph-based data include LinkedIn's company list and Twitter's follower list.
  • Power and sophistication come from multiple, overlapping graphs of the same nodes.
  • Graph databases store graph-based data and are queried with specialized query languages like SPARQL.
  • Graph data presents challenges, especially for computer interpretation of additive and image data.

Data Scientist Challenges in Audio, Image, and Video

- Audio, image, and video data types pose specific challenges for data scientists.
  - Human tasks like object recognition in pictures are challenging for computers.
  - MLBAM plans to increase video capture for live, in-game analytics.
  - High-speed cameras capture ball and athlete movements for real-time calculations.
  - DeepMind developed an algorithm for learning video game play, prompting Google to purchase the company for AI development.
  - The learning algorithm takes in data as it's produced by the computer game.

Streaming data overview

- Streaming data flows into the system during an event, not in batches.
  - It requires adaptation in processes to handle this type of information.
  - Examples include "What's Trending" on Twitter, live sporting events, and stock market.

THANK YOU