Topic for the class:– Data all around us: the virtual wilderness

Unit _1 : Title-Data Evolution

Date & Time : 15.7.22 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

Data Science: A Unique Approach

- Data science is the extraction of knowledge from data, a concept not found in other fields like operations research, decision sciences, analytics, data mining, mathematical modeling, or applied statistics.
  • The term is often used to describe the unique tasks data scientists perform, which previous applied statisticians and data-oriented software engineers did not.
  • The data science process involves setting goals, preparing, building, finishing, exploring, wrapping up, wrangle, revising, assessing, delivering, planning, analyzing, engineering, optimizing, and executing.

The Information Age: From Computing to Data Generation

- The Information Age, from the second half of the 20th century to the beginning of the 21st century, is characterized by the rise of computers and the internet.
  - Early computers were used for computationally intensive tasks like cracking military codes, navigating ships, and performing simulations in applied physics.
  - The internet developed in size and capacity, allowing data and results to be sent easily across a large distance, enabling data analysts to amass larger and more varied data sets for study.

The Information Age: From Computing to Data Generation contd.

- Internet access for the average person in a developed country increased dramatically in the 1990s, giving hundreds of millions of people access to published information and data.
- Websites and applications began collecting user data in the form of clicks, typed text, site visits, and other actions a user might take, leading to more data production than consumption.
- The advent of mobile devices and smartphones connected to the internet allowed for an enormous advance in the amount and specificity of user data being collected.
- The Internet of Things (IoT) includes data collection and internet connectivity in almost every electronic device, making the online world not just a place for consuming information but a data-collection tool in itself.

Data Collection and its Purpose in the Information Age

Collecting User Data

- As businesses realized the potential for data sale, they began collecting vast amounts of user data.
  - Online retailers, video games, and social networks stored every item, link, and activity.

  Data Collection and Its Use
  - Major websites and applications use their own data to optimize user experience and effectiveness.
  - Publishers often struggle to balance the value of the data sold and its internal use.
  - Many keep their data to themselves, hoarding it for future use.

Facebook and Amazon's Data Collection

- • Facebook and Amazon collect vast amounts of data daily, but their data is largely unexploited.
  • Facebook focuses on marketing and advertising revenue, while Amazon has data that could potentially revolutionize economic principles or change industry processes.
  • Despite their vast data sets, these companies focus on their own use and do not want others to take their profits.

Access to data

- Some companies, like Twitter, provide access to their data for a fee.
- An industry has developed around brokering the sale of data for profit.
- Academic and nonprofit organizations often make data sets available publicly and for free, but there may be limitations on their use.
- There has been a trend towards consolidation of data sets within a single scientific field.

Data’s role and value

- Data is now ubiquitous and has become a purpose of its own.
  - Companies collect data as an end, not a means, and many claim to be planning to use it in the future.

Data Scientist as an Explorer in the 21st Century

Collecting and Exploration of Data

- Data sets are increasingly being collected at unprecedented rates, often not for specific purposes.
  - Data analysts are now collecting data first and then deciding what to do with it.
  - The internet, ubiquitous electronic devices, and a fear of missing out on hidden value in data have led to the collection of as much data as possible.

Big Data Innovations

• Big data refers to the recent movement to capture, organize, and use any and all data possible.

Innovation type	The stages of computing innovation				
	Problem	Innovation	Proof/ recognition	Adoption	Refinement
Computing	Cracking codes High-powered physics Ship navigation	Pre-1950s • Purpose-built computing machines	~1950s • Enigma • ENIAC	~1970s • First PCs • Computers in schools and libraries	~1980s • Supercomputers • Consumer PCs
Networking	Communicating and sending text and files	~1970s • pre-internet • ARPANET	~1980s • Academic networks • IRC	1990s • Prodigy • Compuserve • AOL	2000s • Mobile devices • Social networks • Cloud services
Big data collection and use	Too much useful data being thrown away	~2000 • Web crawling • Click tracking • Early, big social networks	2000s • Google search • Big retailers tracking users	2010s • Twitter firehose • Hadoop	2015+ • Massive API development • Format standardization
Big data statistical analysis	Even basic statistics are hard to calculate on large data sets	2000s • Google search • Amazon streamlining processes	2010s • Netflix challenge • Kaggle.com	2015+ • Google Analytics • Budding analytics start-ups	2025+? • Ubiquitous intelligent, integrated systems

Big data innovation

- Each innovation begins with a problem that needs to be addressed and goes through four phases of development: problem, invention, proof/recognition, adoption, and refinement.
  - The current phase of big data collection and widespread adoption of statistical analysis has created an entire data ecosystem where the knowledge extracted is only a small portion of the total knowledge contained.

Exceptions to the data ecosystem

- Companies like Google, Amazon, Facebook, and Twitter are ahead of the curve in allowing access to their entire data set and analyzing their data rigorously.
  • Google's work on search-by-image, Google Analytics, and its basic text search are examples of solid statistics on a large scale.

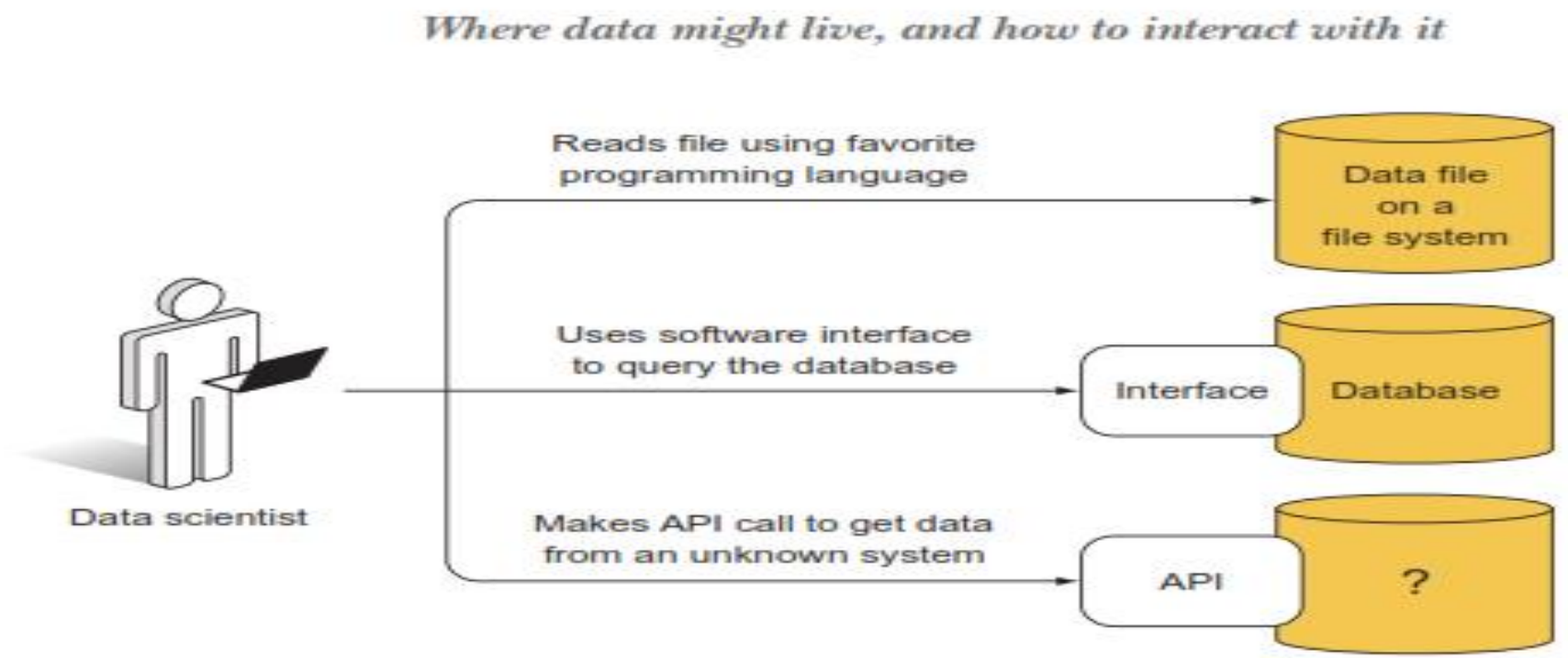
The role of data scientists

- • Data scientists are like early European explorers, accessing interesting areas, recognizing new and interesting things, handling new, unfamiliar, or sensitive things, evaluating new and unfamiliar things, drawing connections between familiar and unfamiliar things, and avoiding pitfalls.
  • Data scientists must survey the landscape, take careful note of surroundings, and dive into unfamiliar territory to see what happens.
  • The existence of data everywhere enables us to apply the scientific method to discovery and analysis of a preexisting world of data.
  • To get real truth and useful answers from data, we must use the scientific method, or the data scientific method: ask a question, state a hypothesis, make a testable prediction, test the prediction via an experiment involving data, and draw the appropriate conclusions through analyses of experimental results.
-

Data storage and interaction in data science

- Discusses various data formats such as flat files, XML, and JSON.
  - Each format has unique properties and idiosyncrasies, making it easier to access and extract data.
  - The discussion includes a discussion of databases and APIs, as they are essential for data science projects.
  - Data can be accessed as a file on a file system, in a database, or behind an API.
  - Data storage and delivery are intertwined in some systems, making them a single concept: getting data into analysis tools.
  - The goal is to provide descriptions that make readers comfortable discussing and approaching each data format or system.
  - The section aims to make data science accessible to beginners, allowing them to move on to the most important part: what the data can tell.

Role of data scientist



Understanding flat files

- Flat Files Overview
  - Flat files are plain-vanilla data sets, the default data format.
  - They are self-contained and can be viewed in a text editor.
  - They contain ASCII (or UTF-8) text, each character using 8 bits of memory/storage.
  - A file containing only the word DATA will be of size 32 bits.
  - If there is an end-of-line character after the word DATA, the file will be 40 bits.

Types of flat files

- • Plain text: Words, numbers, and some special characters.
  • Delimited: Plain text with a delimiter appearing every so often in the file.
  • Table: Data in the file can be interpreted as a set of rows and columns.
  • Most programs require the same number of delimiters on each line to ensure consistency in the number of columns.

Reading flat files

- Any common program for manipulating text or tables can read flat files.
  • Popular programming languages all include functions and methods that can read such files.
  • Python (csv package) and R (read.table function and its variants) contain methods that can load a CSV or TSV file into the most relevant data types.

Limitations and considerations

- • Flat files are the smallest and simplest common file formats for text or tables.
  • They provide no additional functionality other than showing the data, making them inefficient for larger data sets.
  • In cases where reading flat files is too slow, alternative data storage systems are designed to parse through large amounts of data quickly.

Understanding HTML and web scraping

- HTML is a markup language used to interpret plain text.
  • It is widely used on the internet and is used to create web pages.
  • The body of the document is considered by an HTML interpreter, with everything between the tags considered as HTML.
  • HTML tags are usually of the format to begin and end the annotation, for an arbitrary TAGNAME.
  • The class "column" is applied to the div, allowing the interpreter to treat a column instance in a special way.
  • Web scraping involves writing code that can fetch and read web pages, interpret the HTML, and scrape out specific pieces of the HTML page.
  • Web scraping can be useful if the data needed isn't contained in other formats.
  • It's important to check the website's copyright and terms of service before scraping.

XML overview

- • XML is a more flexible format than HTML, suitable for storing and transmitting documents and data.
  • XML documents begin with a tag declaring a specific XML version.
  • XML works similarly to HTML but without most of the overhead associated with web pages.
  • XML is used as a standard format for offline documents like OpenOffice and Microsoft Office.
  • XML specification is designed to be machine-readable, allowing for data transmission through APIs.
  • XML is popular in applications and documents using non-tabular data and other formats requiring flexibility.
-

JSON Overview

- JavaScript Object Notation (JSON) is a functionally similar language for data storage or transmission.
  • JSON describes data structures like lists, maps, or dictionary in programming languages.
  • Unlike XML, JSON is leaner in terms of character count.
  • JSON is popular for transmitting data due to its ease of use.
  • It can be read directly as JavaScript code, and many programming languages like Python and Java have natural representations of JSON.
  • JSON is highly efficient for interoperability between programming languages.

Relational databases: An overview

- • Relational databases are data storage systems optimized for efficient data storage and retrieval.
  • They are designed to search for specific values or ranges of values within the table entries.
  • A database query can be expressed in plain English, with the most common basis query language being Structured Query Language (SQL).
  • A well-designed database can retrieve a set of table rows matching certain criteria much faster than a scan of a flat file.
  • The main reason for databases' quick retrieval is the database index, which is a data structure that helps the database software find relevant data quickly.
  • The administrator of the database needs to choose which columns of the tables are to be indexed, if default settings aren't appropriate.
  • Databases are also good at joining tables, which involves taking two tables of data and combining them to create another table that contains some of the information of both the original tables.

Relational databases: An overview contd.

- Joining can be a large operation if the original tables are big, so it should be minimized the size of those tables.
  - It's a good general rule to query the data first before joining, as there might be far less matching to do and the execution of the operation will be much faster overall.
  - For more information and guidance on optimizing database operations, practical database books are available.
  - If you have a relatively large data set and your code or software tool is spending a lot of time searching for the data it needs at any given moment, setting up a database is worth considering.

Nonrelational databases: efficiency and flexibility

- NoSQL (Not only SQL) allows for database schemas outside traditional SQL-style relational databases.
  • Graph and document databases are typically classified as NoSQL databases.
  • Many NoSQL databases return query results in familiar formats, like Elasticsearch and MongoDB.
  • Elasticsearch is a document-oriented database that excels at indexing text contents, ideal for operations like counting word occurrences in blog posts or books.
  • NoSQL databases offer flexibility in schema, allowing for the incorporation of various data types.
  • MongoDB is easy to set up but may lose performance if not optimized for rigorous indexing and schema.

Understanding APIs and data collection

- APIs as Communication Gateways
  - APIs are rules for communicating with software.
  - They define the language used in queries to receive data.
  - Many websites, like Tumblr, have APIs that allow users to request and receive information about Tumblr content.

Understanding APIs and data collection

- Tumblr's API
  - Tumblr's public API allows users to request and receive information about Tumblr content in JSON format.
  - The API is a REST API accessible via HTTP.

  API Key and Response
  - An API key is a unique string that indicates the user's use of the API.
  - The API key can be obtained as a developer and used to request information about a specific blog.

Understanding APIs and data collection

- Programmatic API Use
  - To capture the Tumblr API response programmatically, an HTTP or URL package in a programming language is needed.
  - The request URL is assembled as a string object/variable and passed to the appropriate URL retrieval method.
  - The response should contain a JSON string similar to the response shown.

- Aptitude in APIs
  - Accurate API usage can be a powerful tool in data collection due to the vast amount of data available through these gateways.

Common Bad Formats

- Avoids typical office software suites like word processing programs, spreadsheets, and mail clients.
  - Avoids these formats when data science is involved.
  - Uses specialized programs for data analysis, as these programs are usually incapable of the analysis needed.
  - OpenOffice Calc and Microsoft Excel allow for exporting individual sheets into CSV formats.
  - Exports text from Microsoft Word documents into plain text, HTML, or XML.
  - Exports text from PDFs into plain text files for analysis.

Unusual formats

- This category includes data formats and storage systems unfamiliar to the user.
  - Some formats are archaic or have been superseded by another format.
  - Some formats are highly specialized.
  - When encountering unfamiliar data storage systems, the user searches online for examples of similar systems and decides if it's worth the trouble.
  - If the data is worth it, the user generalizes from similar examples and gradually expands from them.
  - Dealing with unfamiliar data formats or storage systems requires exploration and seeking help.

Data formats and scouting for data

- Choosing Data Formats
  - Data formats can be inefficient, unwieldy, or unpopular.
  - Secondary data stores can be set up for easier access, but at a cost.
  - For applications requiring critical access efficiency, the cost may be worth it.
  - For smaller projects, it may not be necessary.

  General Rules for Data Formats
  - Export for spreadsheets and office documents.
  - Common formats are better for the data type and application.
  - Don't overspend on converting; weigh the costs and benefits first.

Some common types of data and formats

- Tabular data: small amount delimited flat file.
  - Relational database: large amount with lots of searching/querying.
  - Plain text: small amount flat file.
  - Non-relational database with text search capabilities.
  - Data transmission between components: JSON.
  - Document transmission: XML.

Scouting for data

- Data can be found in various forms, from file formats to databases to APIs.
  - It's important to find data that can help solve problems.
  - If internal system data doesn't answer major questions, consider finding a data set that complements it.
  - There's a vast amount of data available online, and a quick search is worth it.
  - Highlights the difficulty in finding data that can help solve problems.
  - Emphasizes the importance of not taking internal system data for granted.
  - Suggests that external data sets can complement existing data and improve results.
  - Emphasizes the value of quick searches for potential data aids.
  - Encourages focusing on content rather than format in data search.

Example Google search and data usage

- Google searches are not perfect and require understanding of what to search for and what to look for.
- Searches for "Tumblr data" and "Tumblr API" yield different results.
- The former returns results involving data as used on Tumblr posts and third parties selling historical Tumblr data.
- The latter deals almost exclusively with the official Tumblr API, providing up-to-the-minute information about Tumblr posts.
- Terms like data and API significantly impact web searches.
- When searching for data related to your project, include modifying terms like historical, API, real time, etc.

Copyright licensing

- Data may have licensing, copyright, or other restrictions that can make it illegal to use it for certain purposes.
  - Academic data often has restrictions that the data can't be used for profit.
  - Proprietary data, like Tumblr or Twitter, often comes with restrictions that can't be used to replicate functionality.
  - It's best to read any legal documentation offered by the data provider and search for examples of similar use.
  - Without confirming that your use case is legal, you risk losing access to the data or a lawsuit.

Data Science: Choosing the Right Data for Your Project

- The decision to use existing data or seek more is complex due to the variability of data sets.
  • An example of this is Uber's data sharing with the Taxi and Limousine Commission (TLC).
  • The TLC required ZIP codes for pick-up and drop-off locations, which are not specific enough to cover large areas.
  • Addresses or city blocks would be better for data analysis, but this poses legal issues regarding user privacy.
  • After initial disappointment, it's important to check if the data will suffice or if additional data is needed.
  • A simple way to assess this is to run through specific examples of your intended analyses.
  • The decision should be based on the project's goals and the specific questions you're aiming to answer.

Combining data sources

- If your current data set is insufficient, it might be possible to combine data sets to find answers.
  - This can be likened to fitting puzzle pieces together, where each piece needs to cover precisely what the other pieces don't.

THANK YOU