What is Big Data?

Saptarshi Pyne
Assistant Professor
Department of Computer Science and Engineering
Indian Institute of Technology Jodhpur, Rajasthan, India 342030

CSL4030 Data Engineering Lectures 2 and 3
August 4th and 7th, 2023

What we discussed in the last class

- Who are Data Engineers?

- What is the difference between a Data Scientist and a Data Engineer?

Development of the World Wide Web

Web (1994)

- Static HTML pages

Web 2.0 (2004)

- No need for HTML programming
- Users can upload content on social media

Web 3.0

- You guys will build
- Decentralized wealth: Public blockchains, Digital currencies
- Synthetic General Intelligence that “understands” human contents

Development of the World Wide Web

Web (1994)

- Static HTML pages

Web 2.0 (2004): Advent of Big Data

- No need for HTML programming
- Users can upload content on social media

Web 3.0

- You guys will build
- Decentralized wealth: Public blockchains, Digital currencies
- Synthetic General Intelligence that “understands” human contents

Web 2.0 and the Big Data Revolution: The Problem

- Feb, 2004: Mark Zuckerberg and colleagues founded Facebook.
- Feb, 2005: Jawed Karim and colleagues founded YouTube.
- Users started uploading large volumes of content on social media and other online services.

- Google and Yahoo realized that they can not **economically** manage the flow of such massive amounts of data with the traditional data management technologies.

Web 2.0 and the Big Data Revolution: A solution

- In 2004, Google started experimenting with a novel Distributed Computing paradigm which they called **MapReduce**.

- In 2008, Jeffrey Dean and Sanjay Ghemawat of Google published **the MapReduce paper** in Communications of the ACM.
  - MapReduce automatically identifies parallelizable tasks/jobs
  - Distributes them to a large cluster of computing nodes for parallel processing
  - Manages inter-node communication to make efficient use of their processing power, network bandwidth, and secondary storage
  - Also, handles node failures. For example, if a node gets disconnected, MapReduce detects it and assigns its job to an available node.
  - Paper Link: https://dl.acm.org/doi/pdf/10.1145/1327452.1327492

Web 2.0 and the Big Data Revolution: A solution (contd.)

- In 2008, Doug Cutting and colleagues at Yahoo! developed a general-purpose implementation of the MapReduce paradigm which was named after a toy elephant named **Hadoop**. They shared Hadoop with the Apache Software Foundation, a community of open-source developers.

- In 2011, the Apache Software Foundation publicly released **Apache Hadoop 1.0**, an open source implementation of Hadoop.

Web 2.0 and the Big Data Revolution: A solution (contd.)

- In 2008, Doug Cutting and colleagues at Yahoo! developed a general-purpose implementation of the MapReduce paradigm which was named after a toy elephant named **Hadoop**. They shared Hadoop with the Apache Software Foundation, a community of open-source developers.

- In 2011, the Apache Software Foundation publicly released **Apache Hadoop 1.0**, an open source implementation of Hadoop.

- In 2014, they released **Apache Spark**, specialized for streaming apps. It processes streaming data in main memory and avoids access to slower secondary storage.

Web 2.0 and the Big Data Revolution: A solution (contd.)

- In 2008, Doug Cutting and colleagues at Yahoo! developed a general-purpose implementation of the MapReduce paradigm which was named after a toy elephant named **Hadoop**. They shared Hadoop with the Apache Software Foundation, a community of open-source developers.

- In 2011, the Apache Software Foundation publicly released **Apache Hadoop 1.0**, an open source implementation of Hadoop.

- In 2014, they released **Apache Spark**, specialized for streaming apps. It processes streaming data in main memory and avoids access to slower secondary storage.

What is Big Data made up of?

- Structured Data
- Unstructured Data
- Semi-structured Data: Not a widely recognized term

Structured Data

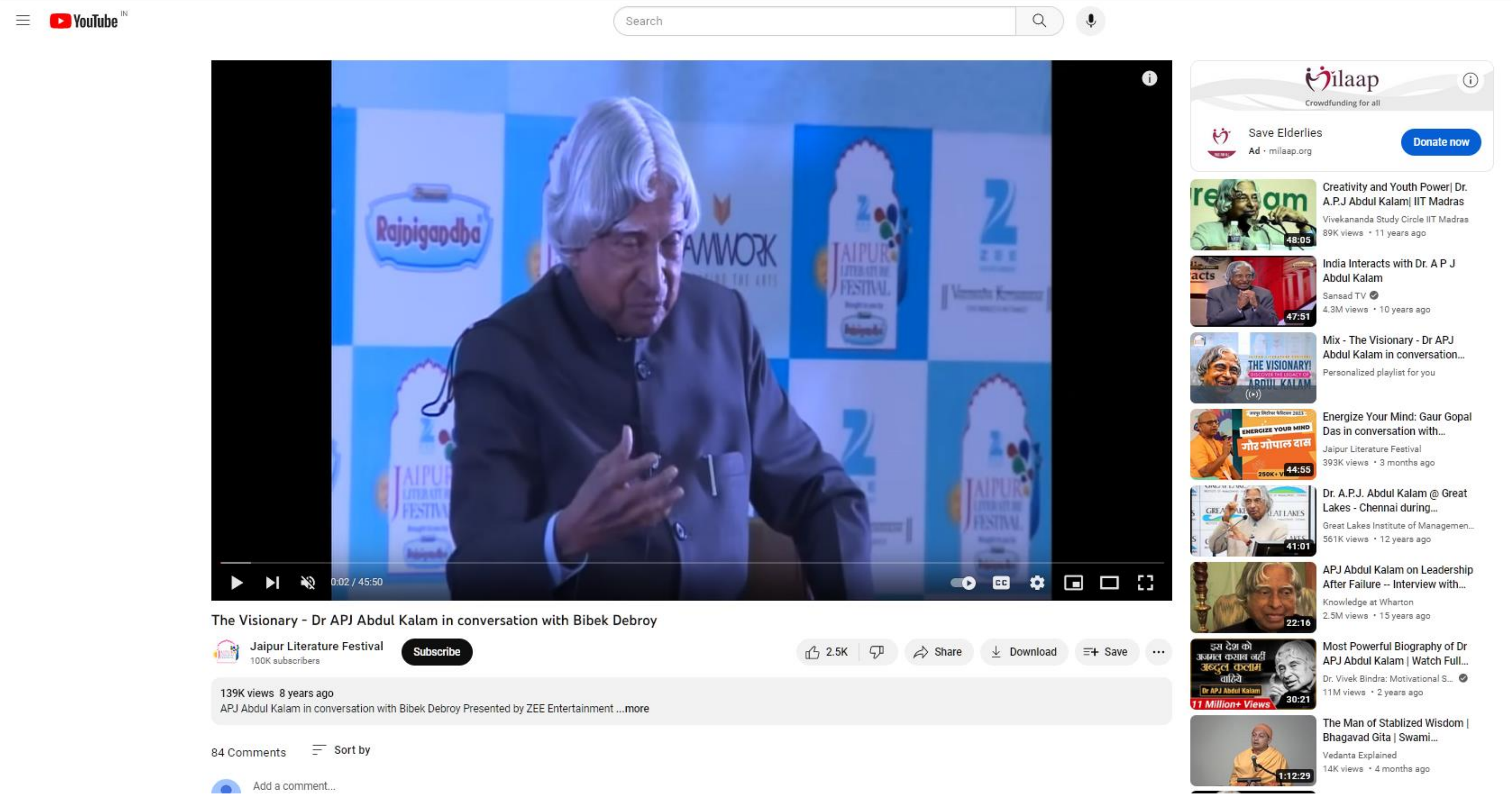
Data that follows a predefined **schema**.

Example: The student database of this course.

#	Roll No	Student Name	eMail	Registered_as	Course Type
1	B21AI001	KAMUJU AASHISH	kamuju.1@iitj.ac.in	Credit	PC
2	B21AI002	ABHISHEK ARYA	arya.7@iitj.ac.in	Credit	PC
3	B21AI003	ADARSH RAJ SHRIVASTAVA	shrivastava.10@iitj.ac.in	Credit	PC
4	B21AI004	ADEEM HARIS	haris.1@iitj.ac.in	Credit	PC
5	B21AI005	AKRITI GUPTA	gupta.97@iitj.ac.in	Credit	PC
6	B21AI006	ARVIND KUMAR SHARMA	sharma.126@iitj.ac.in	Credit	PC
	…	…			

Unstructured Data

Data that does **not** follow a predefined **schema**, e.g.,



The Visionary - Dr APJ Abdul Kalam in conversation with Bibek Debroy

Jaipur Literature Festival
100K subscribers

139K views 8 years ago
APJ Abdul Kalam in conversation with Bibek Debroy Presented by ZEE Entertainment ...more

84 Comments Sort by

Add a comment...

milaap
Crowdfunding for all
Save Elderlies
Ad · milaap.org

Creativity and Youth Power| Dr. A.P.J Abdul Kalam| IIT Madras
Vivekananda Study Circle IIT Madras
89K views • 11 years ago
48:05

India Interacts with Dr. A P J Abdul Kalam
Sansad TV
4.3M views • 10 years ago
47:51

Mix - The Visionary - Dr APJ Abdul Kalam in conversation...
Personalized playlist for you

Energize Your Mind: Gaur Gopal Das in conversation with...
Jaipur Literature Festival
393K views • 3 months ago
44:55

Dr. A.P.J. Abdul Kalam @ Great Lakes - Chennai during...
Great Lakes Institute of Managemen...
561K views • 12 years ago
41:01

APJ Abdul Kalam on Leadership After Failure -- Interview with...
Knowledge at Wharton
2.5M views • 15 years ago
22:16

Most Powerful Biography of Dr APJ Abdul Kalam | Watch Full...
Dr. Vivek Bindra: Motivational S...
11M views • 2 years ago
30:21

The Man of Stablized Wisdom | Bhagavad Gita | Swami...
Vedanta Explained
14K views • 4 months ago
1:12:29

Can we generate Structured Data from Unstructured Data?

Yes.

Video

-> Speech-to-Text

-> Keyword Mining

-> Map the keywords to hashtags

-> Add the hashtags to the video.

Thus we can generate Structured Data from Unstructured Data, and then combine them.

A case study on Amazon.in

Search for **“Echo Dot (3rd Gen) - Smart speaker with Alexa (Black)” on Amazon.in**
and
analyze the types of contents on the page.

How do organizations store their Big Data?

They store Big Data in their **Data Lakes**.

All structured and unstructured data are stored in the raw format in an organization's Data Lakes. When they want to analyze a particular subset of data (say, their HR data), they use advanced Data Lake "query processing" software like **Apache Pig**.

Earlier, organizations used to store their data into **topic-specific** storages for topic-specific querying. Such storages are called **Data Warehouses**.

Data Warehouses and Data Lakes both are critical for generating **Business Intelligence**.

Where does an organization host its Data Lakes?

- Big companies create their own **Data Centers** to host their Data Lakes, e.g., Amazon, Google.

- Smaller companies subscribe to big companies' data centers to host their Data Lakes. Such subscription-based storing mechanism is known as
  storing in the Cloud.
  A big portion of big companies' revenues comes from providing such **cloud services**.

Finally, let us formally define Big Data

What is **Big Data**?

- "Big data is **high-volume**, **high-velocity** and **high-variety** information assets that demand <u>cost-effective</u>, innovative forms of information processing for enhanced insight and decision making." ~ Gartner, Inc.

Three Vs of Big Data

- Volume: Quantity
  - A typical PC might have had 10 gigabytes of storage in 2000.
  - Today, Facebook ingests 500 terabytes of new data every day.
  - Boeing 737 will generate 240 terabytes of flight data during a single cross-country flight
  - Smart phones and IoT => Continual generation of data
- Variety: Type of data
  - Big Data beyond numbers, dates, and strings; may be structured, semi-structured or unstructured
  - Big Data is multimodal: geospatial, temporal, 3D data, audio, video, unstructured text, including log files and mixed media.
  - Traditional database systems were designed to address smaller volumes of structured data, had fewer updates, and operated on, consistent data structures.
- Velocity: Operational speed & Data speed
  - Clickstreams and ad impressions capture user behavior at millions of events per second
  - High-frequency stock trading algorithms reflect market changes within microseconds
  - Machine to machine processes exchange data between billions of devices
  - Infrastructure and sensors generate massive log data in real-time
  - On-line gaming systems support millions of concurrent users, each producing multiple inputs per second.
- Also, please go through the “Case studies” section on https://en.wikipedia.org/wiki/Big_data

Parameters of Big Data

- **Veracity:** Low signal-to-noise ratio. The correctness of captured data can vary greatly, affecting the correctness of the analysis.
- Exhaustive: Whether data pertaining to all possible use-cases of the system or the problem concerned are recorded or not
- Fine-grained and uniquely lexical: The proportion of specific data of each element, per element collected, and if the element and its characteristics are properly indexed or identified, respectively
- Relational: If the data collected contains commons fields that would enable a conjoining, or meta-analysis, of different data sets
- Extensional: If new fields can be incorporated or changed easily
- Scalability: Rate of expansion of data
- **Value:** The utility that can be extracted from the data
- Variability: It refers to data whose properties are context-sensitive.

References

- https://www.oracle.com/in/a/ocom/docs/big-data/big-data-evolution.pdf

- https://www.oracle.com/big-data/structured-vs-unstructured-data/

What we discussed today

- How did the Big Data Revolution happen?
- What are structured and unstructured data?
- What is the difference between a Data Lake and a Data Warehouse?
- Finally, what is the formal definition of Big Data?

What we will discuss in the next class

- What are Data Models?

Thank you