Data Analytics Project

9. Unstructured Data & Sources

Aagam Deolasi

1

WHAT is Unstructured Data?

Definition: Unstructured data *lacks the organizational structure* found in structured data. It exists in a *raw form*, not organized in predefined ways, and *lacks a fixed schema*.

Characteristics:

1. No predefined organization or schema.
2. Lacks a set way of entering, grouping, and analyzing the data.

Aagam Deolasi

2

Examples of Unstructured Data

1. **Multimedia**: Photos, Videos, Audio
2. **Written Content**: Web Pages, Books, Blogs
3. **Text Documents**: Journals, White Papers, PPTs
4. **Informational Content**: Articles, Emails, Wikis
5. **General Texts**: Word Documents, General Text

Additionally, **PDFs** are also considered as Unstructured Data where text is searchable but lacks predefined formats.

Aagam Deolasi

3

Sources of Unstructured Data

Unstructured data is abundant on the internet and takes various forms, including text, images, videos, and audio. Some of the sources of unstructured data are:

1. **Public Web Forums and Blogs**: Generate unstructured data via user interactions and content creation.
2. **Social Media Platforms**: YouTube, Facebook, instant messaging, and Twitter contribute diverse unstructured data.
3. **Web Scraping**: Automates data extraction from web pages, transforming human-readable content into machine-readable formats.

Aagam Deolasi

4

Extraction of Unstructured Data

1. **NoSQL Databases and Data Lakes**: Centralized repositories for raw data, ideal for real-time storage from IoT, websites, mobile apps, and social media.

2. **Web Scraping for Data Transformation**: Automated extraction using bots to transform data from HTML pages for analysis.

3. **Application Program Interfaces (APIs)**: Standardized interfaces (e.g., RESTful APIs) from major providers like Facebook, enabling automatic data collection for analysis.

Aagam Deolasi

THANK YOU!!! FOR YOUR SUPPPORT! For Now...

Keep Learning, Keep Sharing & Keep Following ***Aagam Deolasi***.

Aagam Deolasi