Introduction to BigData

Relevel
by Unacademy

Data Science with Pyspark

Class 1 : Introduction to Big Data and Apache Spark

What is Big Data ?

The term Big Data is used to define massive , complex data that are received from various sources high pace which includes structured, unstructured and Semi Structured Data

Big Data is a part of a lot of tech revolutions happened over last few decades such as Cloud computing , IoT , Automation , Artificial Intelligence

The Term Big does not just mean the Size or Volume only , It refers to its complex characteristics which is represented by 5 Vs

Big Data unlocks the value of hidden information which can be used across various fields in terms of estimating the future and making data driven decisions. It means where these is a information , there is more advantage

According to Gartner Definition :

"Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation"

Three Different Types of Data

- **Structured = the data is stored in a structured format (eg. RDBMS).**

  Structured Data is the most commonly available data format , This follows a predetermined structure or Schema , It can be stored as text file , Csv file or in RDBMS SQL or NOSQL Databases

  Ex: Transactional Data , Machine log data , sensor data , Sales Data

- **Unstructured = unorganised data (eg. videos).**

  Unstructured data don't have any predefined structure in place , It was difficult to perform the querying and processing, but . Advancements in AI , ML made this possible these days

  Ex: Images , Text , Audio , Music data

- **Semi-structured = the data is organised in a not fixed format (eg. JSON).**

  This is essentially structured and unstructured data combined. What makes semi-structured data interesting is that it has enough properties to make its analysis fairly manageable. As mentioned by the company HubSpot, "semi-structured data is information that does not reside in a relational database or any other data table."

  Ex: Emails , Web pages , Digital marketing ads.

Sources of Big Data :

Majority of Big Data is generated via 3 Major Resources

1. **Machine Data**
   a. Generated by Industry Equipments , Sensors , Logs of Computers which has the ability to unveil the user / entity behavior
   b. Amount of Machine based Data increases exponentially as application of IoT surges
   c. Data collected from Consumer electronics , Geo Sensing Devices are helpful in solving problems across various sectors such as healthcare , automotive , city planning etc...

1. **Social Data**
   a. Social Data is nothing but Likes , comments , posts on Linkedin , Facebook , instagram , twitter
   b. This kind of data brings valuable attributes about User preference , behaviour , opinion
   c. This helps companies to track target audience , harness the power of ML to improve marketing strategies

1. **Transactional Data**
   a. Transactions that happens in Online , Offline cause the lot of Data Generation.
   b. Payments , Orders , Booking , Invoices , Plans all these are examples for Transactional Data
   c. Most of Industry vertical problems are solved via Transactional Data

Characteristics of Big Data :

Characteristics of Big Data Defined by 5 Vs

1. **Volume**

   It refers the size of the Data , Based on the Volume of the data , the term Big Data is defined.

   In the year 2016, the estimated global mobile traffic was 6.2 Exabytes(6.2 billion GB) per month. Also, by the year 2020 we will have almost 40000 ExaBytes of data.

   The rapidly increasing volume data is due to cloud-computing traffic, IoT, mobile traffic etc

2. **Velocity**

   It refers to high speed of Generation of Data

   Sources like Mobiles , computers , Electronic devices , Social Media Platforms see Massive Data flow and that increases exponentially Year over Year.

   Example: There are more than 3.5 billion searches per day are made on Google. Also, Facebook users are increasing by 22%(Approx.) year by year.

Characteristics of Big Data :

3. **Variety**

   It refers to different types of data that is generated , such as Structured , Semi Structured , Un Structured.

   When a Source generated all of them together , it is challenging to store and process them together. And Storage layer differs for each and every type of data

4. **Veracity**

   Since Data is being generated in Multiple sources and is multi dimensional , It leads to uncertainty , accuracy issues and quality management pitfalls

   It is the assurance of Quality, Integrity , Credibility, Accuracy

   Few Examples of Veracity (Degree of Accuracy) : 1. Error in Data due to Human mistake , 2. Noise in IoT Sensor Measurement Data , 3. Ambiguity in Social Media Profiles

5. **Value**

   Having Bulk data in place does not mean that helps Business , It needs to converted into useful information which add value to the business in access near future and helping in Decision taking