Week 1: Tutorial
6CS030-Big Data

Big Data Overview

Module Assessments

ASSESSMENT	DESCRIPTION	% COVERAGE
Coursework	Implementation based research work involving handling and analysing of big data in your interest specific subject domain.	Report (80%) Code (20%) out of total 70%
Timed Constrained Assessment.	MCQ Exam based on Big Data subjects covered over the weeks.	30%

Why the term Big Data?

- Because its data that is **higher in volume as its** exponentially growing **& complex in nature** .
- Given the nature, it gets difficult to be handled using

  Traditional Data Management Tools.

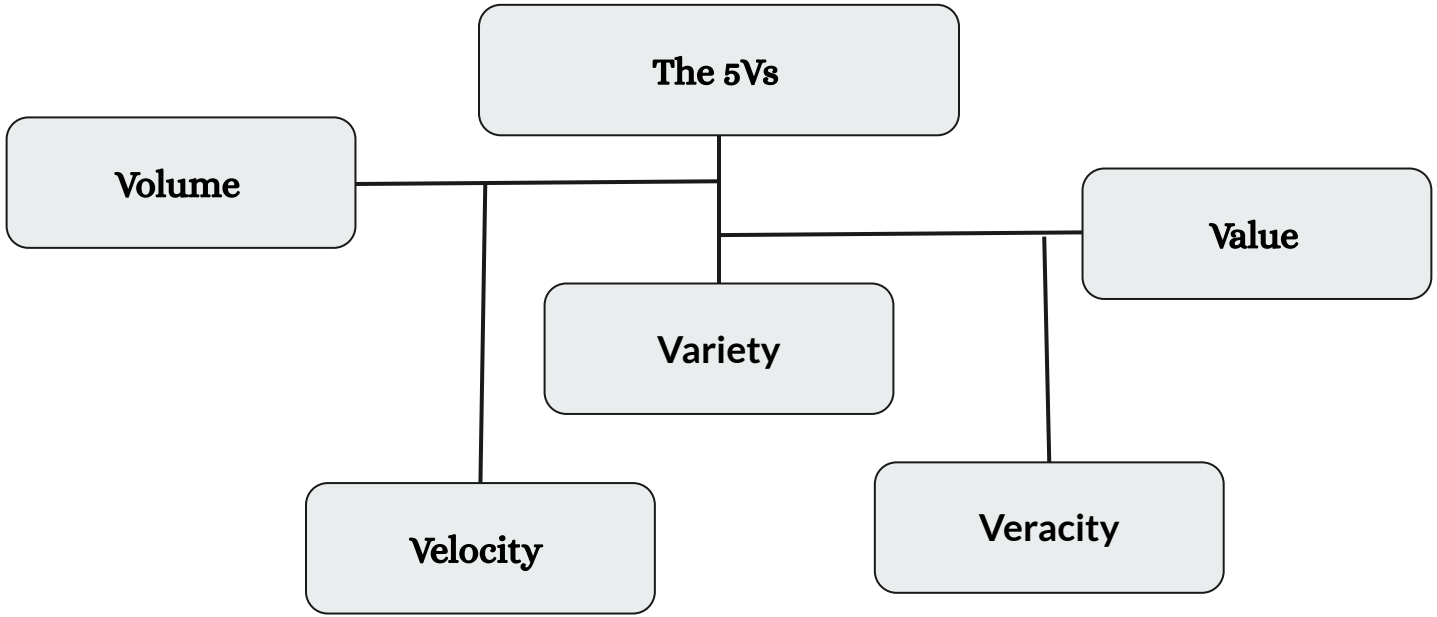
  Example : Excel, RDMS management systems.

  Some Examples :

  Data Ingested in Popular Social Media Platforms.

  But what makes it Big Data and not just Data ?

CHARACTERISTICS OF BIG DATA

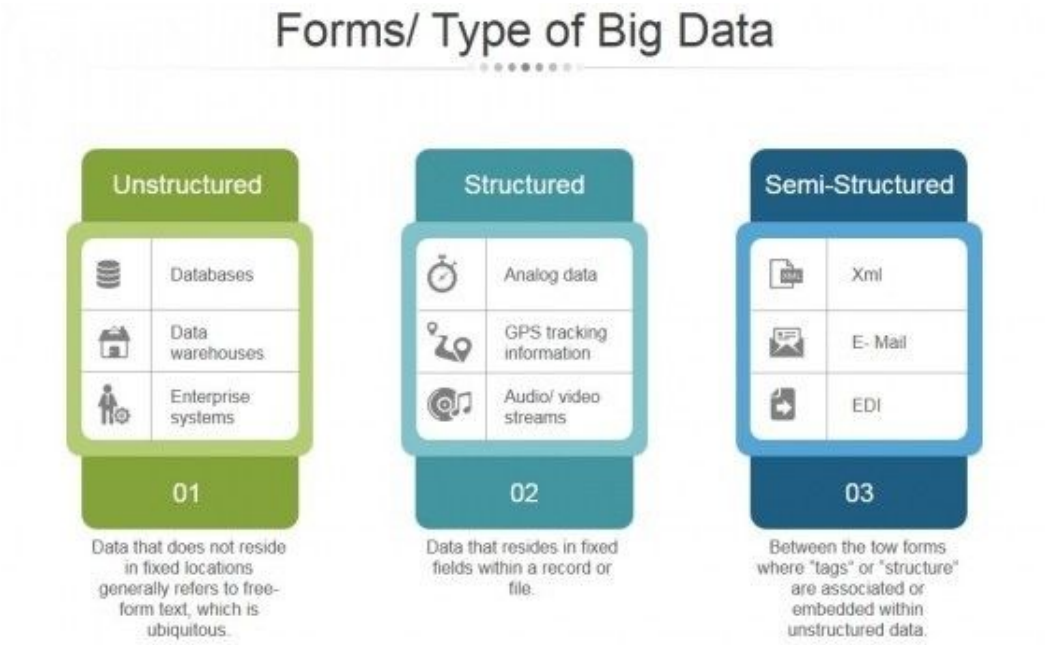


VOLUME

- The amount of data generated every second.
- Its volume ranges from Petabytes to Zettabytes or Exabytes.
- High volume leads to high system configuration to process this volume for effectiveness.

VARIETY

- Big Data unlike traditional structured format can come in many forms.
- Given the variety of dataset it can be **heterogeneous, complex, incompatible.**
- **It is also variable in nature.**



VALUE

- It refers to the benefits the retrieved data can provide.
- The value of big data is related to the **amount of insights** that can be gained from it.
- It is also measured with an estimation made based on the positive impacts it helps to bring in a subject.

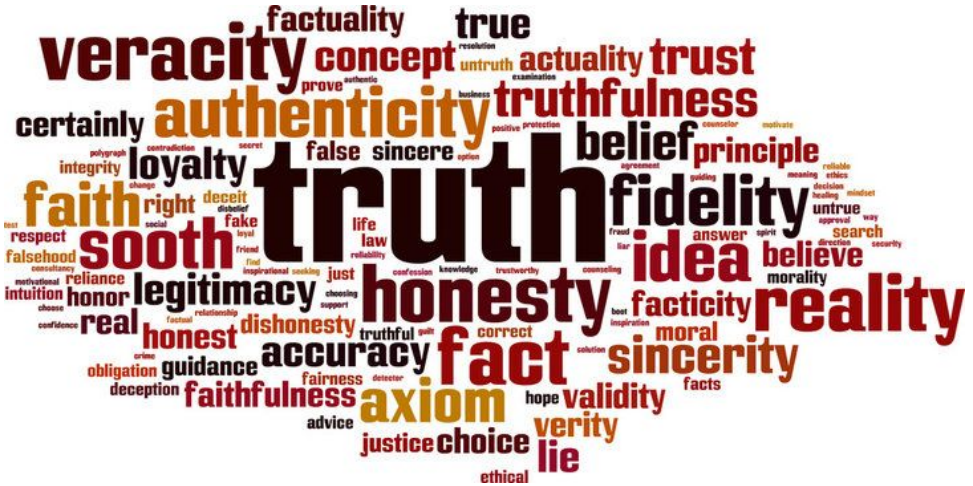
VELOCITY

- This refers to the rate at which the data is being generated or renewed.
- And at what rate it moves from various source to data repositories.
- Larger Velocity refers to high volume of data and an effective measure for managing massive volumes of data.
- High velocity of the data demands adequate processing strategies for timely insight generation.

Case	Data Processed	Need for Big Data Velocity
Twitter	500 million tweets per day	Hashtag trending, ad targeting, user engagement tracking
Uber	15 million rides per day, operating in 40 countries	Estimating arrival times, dynamic pricing, balancing supply-demand

VERACITY

- It refers to the **certainty or reliability** factor of a data.
- The **accuracy** with which it represents the real-world scenarios play an important role.
- With data changing every millisecond, it is important that it is obeys the **timeliness** too. (Upto date data) .
- Noise in data affects its veracity.



Big Data Statistics

45% of IT professionals anticipate the usage of several technologies-including 5G, edge computing and machine learning to manage data. (Dell Technologies)

Netflix saves **$1 Billion** per year with the usage of Big Data for extraordinary customer retention.

70% of the world's data is user-generated.

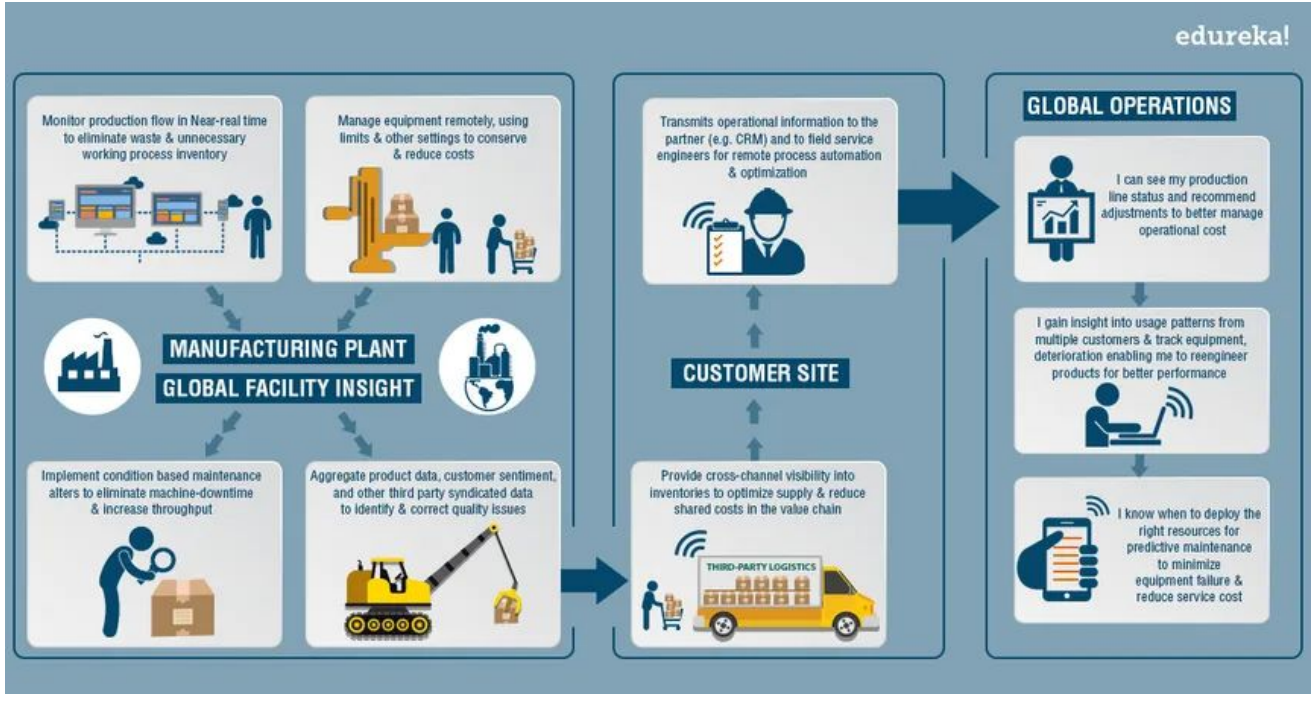
~570 new websites spring into existence every minute every day.

Google alone processes **40k** search queries per second.

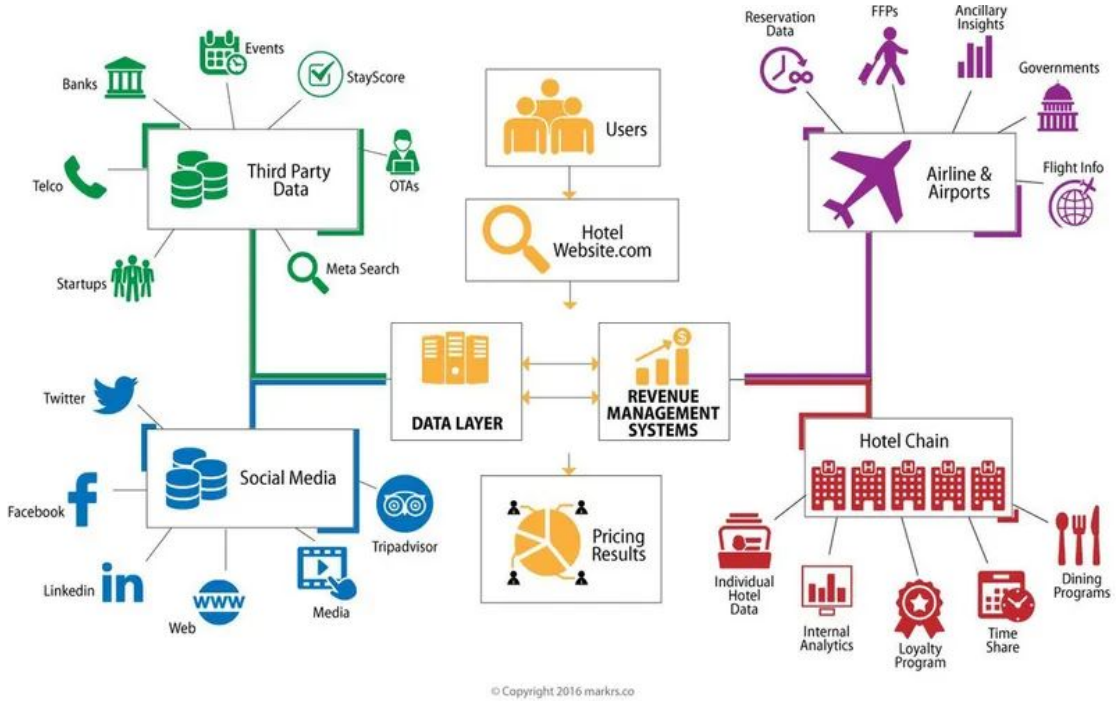
~100 hours of videos are uploaded to Youtube every minute.

BIG DATA APPLICATION IN HEALTHCARE

BIG DATA APPLICATIONS IN MANUFACTURING



BIG DATA APPLICATIONS IN AIRLINES



CHALLENGES OF BIG DATA

Storage Contrataints ⟶ Scalable Infrastructure

Real Time Analysis ⟶ Investing in Real-Time Analytics with tools like Apache Spark

Data Value ⟶ Data Preprocessing through partitioning, cleaning and memory management

Unproportionate Demand & Talent Ratio ⟶ Awareness & Education in Big Data Analytics

TOOLS & TECHNOLOGIES USED IN BIG DATA

CAREERS WITH BIG DATA ANALYTICS

So, what's up for workshop this week?

Remember Python ?

We will explore Python for data handling and explore its use cases because Python provides following benefits in Big Data

Ease of Use : First things first, its super easy and comprehensible

Rich Ecosystem : Diverse libraries for handling data kinds and situations. Example : PyMongo, Pyspark.

Flexible Integration : Its rich ecosystem allows its integration with variable data pipelines and databases envs.