BigData – Intro

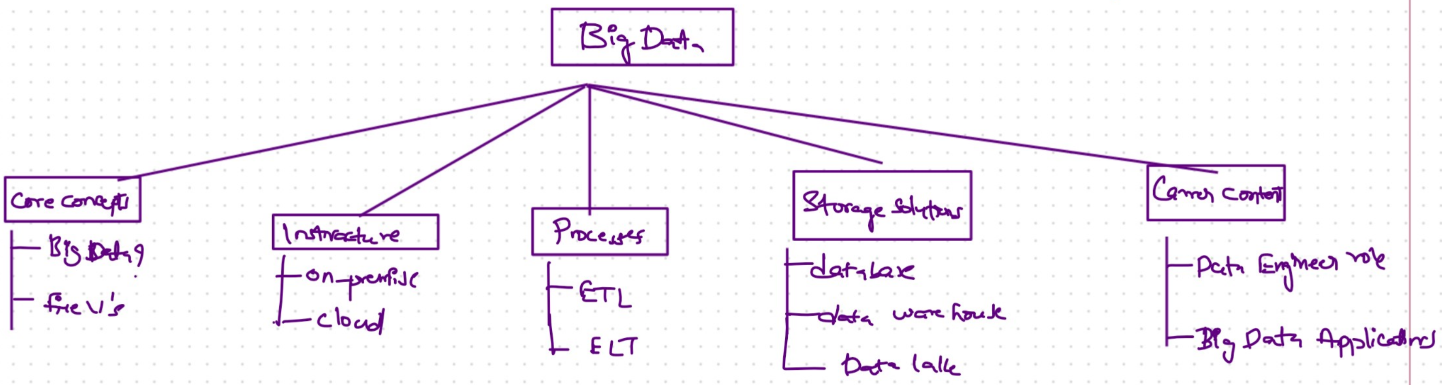
Pre-requisites

→ python
→ Sql

Carrer

→ role of a D.E ?

→ where BigData fits in the data engineering ecosystem



What is Big Data ?

Bigdata refers to datasets that are too large, fast, or complex for traditional databases to process efficiently

A problem — handling massive data

A solution — technologies developed to address these challenges

Traditional RDMS limitations (SQL)

Case 1 small online store : customers order, inventory, basic analytics } RDMS ✓

(1) well struct data & limited volume

(2) queries efficiently retrieve data → order history, product details ...

Case 2 Company like walmart

→ billion of transactions across thousands of stores

→ need to analyze customer behaviour, inventory trends & product performance

* Hadoop / BigQuery instead of RDMS to process

Why RDBMS fails for BigData:

5TB ← 100GB

64

2TB

(100TB) → 200TB

(1) Scaling Limitations:

→ Vertical scale has physical/pratical limits

→ cost is expensive if we infinitely expand storage

(2) Performance Issues:

* Large datasets requie complex partioning

* joins → time-consuming
→ prone to error

* operations becomes extremly slow with growth of data

Structured

(3) Media Storage Challenges

poor to handle/process → images, videos etc, (unstructured

(4) Cost Inefficiency

hardware costs increases for enterprise-grade solutions

Why BigData Emerged:

Internet Growth → Social Media → IOT Devices → Big Data Need
(Fb, wh, Instagram, Youtube, Linkedin)

Data Penetration:

Storage

Facebook: 4 + petabytes of data generated daily

Twitter: 500 + million tweets daily

Youtube: 500 + hours of video uploaded every minutes

requires specialized storage & processing solutions ←

IOT Devices: estimated to reach 75 billion by 2025

Processing

Netflix: 450 + billion events daily for content recommendations

Uber: handles 100 + petabytes of data for real-time data

Weather forecasting: Analyzes 15 + terabytes of data daily from satellites & sensors

Genomics: human genome → generates 200 + gigabytes per person

Big Data Scale: → nearly half of all the worlds popupulation in just last 3-5 years

→ grow to 175+ zettabytes by 2025

Five V's of Bigdata

* framework to understand the challenges & characteristics of BigData
* helps us to determine whether a problem requires Big Data technologies

5 V's are

Volume
Velocity
Variety
Veracity
Value

1, Volume: refers to the size or amount of data being generated

* volume determines if Big Data technologies are required
(Hadoop) → Store & process volume of data
Company → 30GB (Startup)

example

Grocery store → RDMS
Amazon → Big Data

* volume alone does not define BigData problem

(2) Velocity: The speed at which data is generated & processed

Types

(1) Real Time: data processed instantly

ex: (1) Cred-card fraud detection

(2) Youtube feed updates

Spark Streaming, Spark

(2) Batch Processing: Data processed at scheduled intervals

ex:

monthly credit bill

daily reports (24 hrs) (1 hr)

Hadoop tools

(3) Near Real Time: data is processed with a slight delay (2-5 minutes)

ex: Bank transaction updates

Payment status

→ delivered

3. Variety: refers to the different forms of data

Types

Structured

rows & columns (SQL DB)

tabular format

Semi-Structured

JSON, XML

CSV files

some structure but not fully tabular

Unstructured

data without a predefined format

* videos, images, text files

(4) Veracity: trustworthiness or quality of data

* data can have issues like inaccuracies / inconsistencies

eg: -ve age value

* poor data quality leads to bad decisions

Solution: Data cleaning & validation process

Ex: Student / employee records

(5) Value: The insightful or business value derived from data.

* It should provide insights & help in decision-making

ex: Sales data → analysis to identify trends [iphones ——

* Big Data technologies turn raw data into meaningful insights

* Data without value is meaningless, even if it has high volume / velocity