Big Data

We have big data today. But we can't simply store analyze it.
Reasons: Large amount of storage and access speed.

Storing data in multiple drives equally distributed and using them in parallel can make it easy, fast and possible.

Types of Data:

~~structured Data: Data can be stored in~~

structured Data	Semi-structured	Unstructerd
• They have relational key.	• They can be made into relational after process but not always.	• They are not relational.
• Matured transaction control	• Transaction is adapted from DBMS but not mature	• No transactions.
• eg: Relational database	• XML files	• Word, PDF, Media, etc.
• Arranged in rows and cols	• There is no convent ional formal	• Contains text or binary. No formal.
• Flexible: Schema dependent but less flexible	• Schema ~~dependent~~ less but more flexible than structured	• Schema less and most flexible.

• Scalibility: Difficult to scale DB because of schema.	• More scaleable than structured data	• More scaleable
• Quries: Allows complex joins	• Quries using nodes:	• Only text queries.
• Very robust	• not very spread	

Vs of Big Data:

Velocity: Data is generating with high speed and need to be handled on time with speed. Late decisions Leads to missing opportunities. e.g: 3.5B+ searches on goole daily.

Variety: Data is of different Types. It can be structured, unstructures or semi-structured.

Volume: Volume refers to the amount of data. We have to deal with huge amount of data. Eg: Mobile traffic in 2016 was 6.2 exabytes per month. It will be 40,000 exabyte exabyte of data in 2020.

Varicity: Data can be inconsistent, incomplete. Eg: Larger data set may be inconsistent can create ambigutie while less amount of data will be incomplete information or half info.

Value: Data itself has no value. We have to make it valueable.

Variability: How fast the data changes or its shape changes. Eg: you are drinking the same drink daily but it taste changes.

Venue: Location of data.

Vocabulary: Semantics.

Harnessing Big Data.

OLTP	OLAP
○ Online Transaction processing.	○ Online Analytical processing.
○ Provides transactions based applications.	○ Use different databases from different OBS to get insights
○ ATM transaction	○ Netflix recomendation system
○ It provide security restrictions. ⟷	○ Better security features.
○ It cant be used For decision making.	○ It cant be used for decision making.
○ Read, write, delet operations.	○ Keep data consistent.
○ Data consistency, integrity.	○ Handles large data.

Real Time Analytical Processing → RTAP

It process Large amount of data with less reaction time.

Predictive Analytics and Data Mining

- More of a real time analysis
- Large datasets
- Any type of data
- Complex statistical analysis

Business Value

- Deals in Data mining techniques
- Small to mid sized
- Structured Data
- Ad-hoc queries and reporting

Application of BD:

- Recomendation system
- SNA

Recomendation system:

- Track each visit of the user.
- Track content and metadata about the visited pages.
- Track meta data of the user.

eg: News Recomendation

- Can be based on current article
- User history
- Clickstream:

Click stream Analyze: Analyze the collected data to get insights

Click stream Analytics: Collect data from visited pages and their content

- In general, combination of text stream (Article) with click stream.
- To make a good model that describes the user.

Why is it Hard?

- Cold start Problem
- Providing Accurate recommendations.

Cold start:

No existing data of the user to predict recommendations.

System Network Analysis : SNA

observe social and communication phenomina at a planetary scale.

- Small World problem.
  - Average distance between two random users is 6.6 degrees.
  - Facebook reported in 2016, it is 3.5 degrees of separation.