Fundamentals of Big Data Analytics

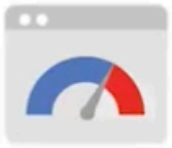
Lecture 2- Data & Exploratory Data Analysis

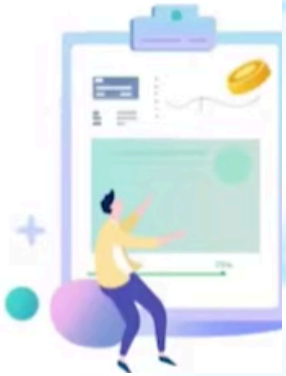
Dr. Iqra Safder
Assistant Professor
FAST NUCES, Lahore

WHAT IS BIG DATA

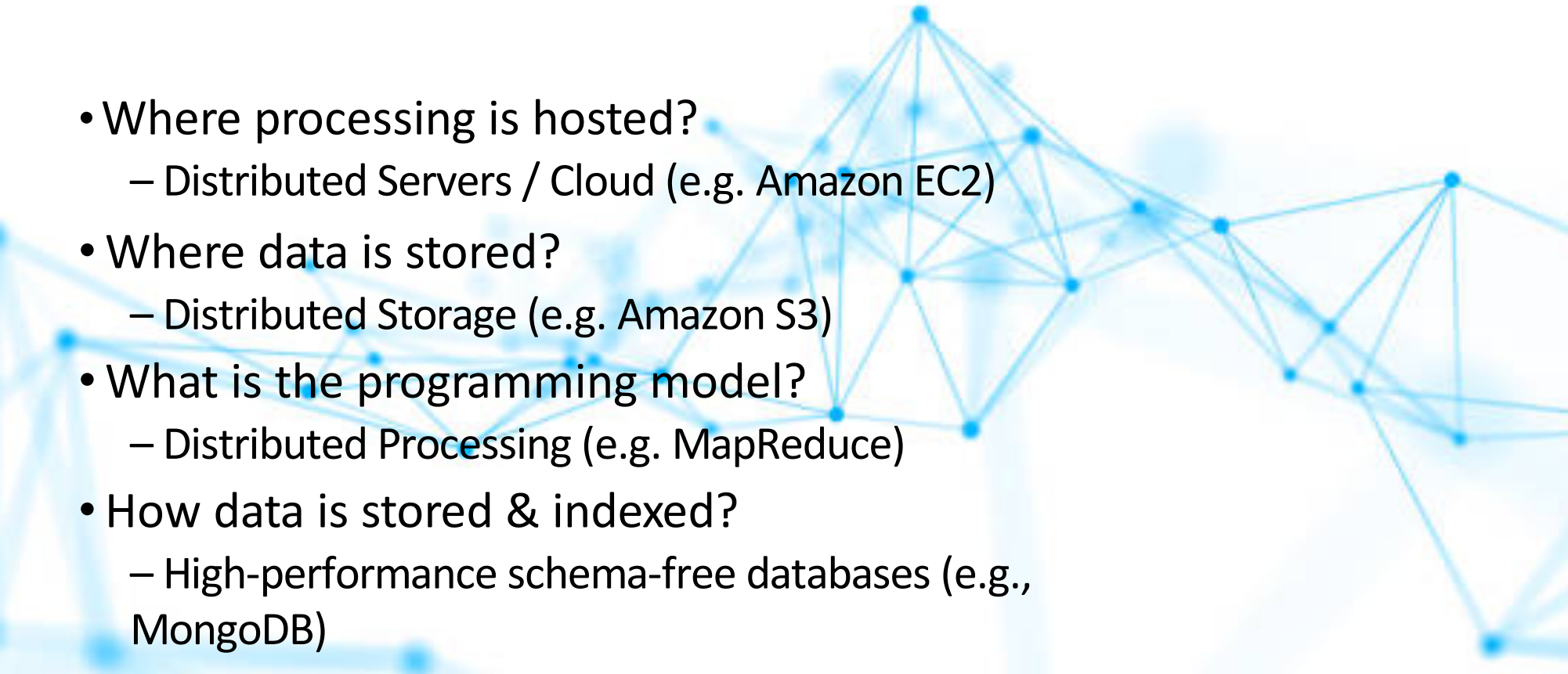
UNIQUE

- **Huge Amount of Data:** We're aware of the expenses in storing and handling huge amounts of data
- **Heterogeneous Data:** They are unstructured, semi-structured, and structured data. Lots of variety from lots of sources
- **Accessing and processing speed:** If you have a 100 Mbps I/O channel and you need to process 2TBs of data – it will take you nearly 6 hours to process the data





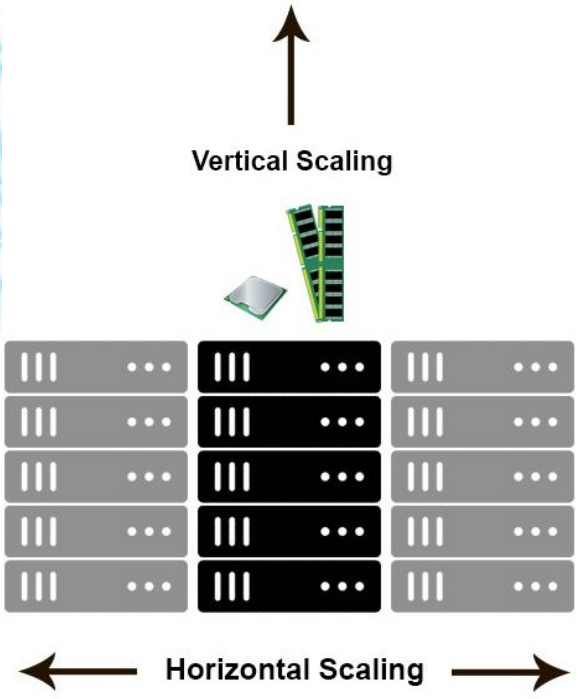
Big Data Analytics



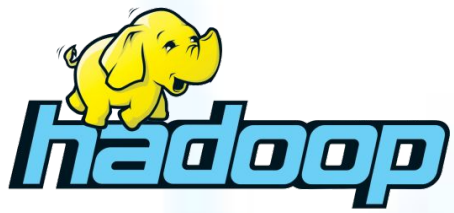
- Where processing is hosted?
  - Distributed Servers / Cloud (e.g. Amazon EC2)
- Where data is stored?
  - Distributed Storage (e.g. Amazon S3)
- What is the programming model?
  - Distributed Processing (e.g. MapReduce)
- How data is stored & indexed?
  - High-performance schema-free databases (e.g., MongoDB)
- What operations are performed on data?
  - Analytic / Semantic Processing

Vertical Scaling Vs. Horizontal Scaling?

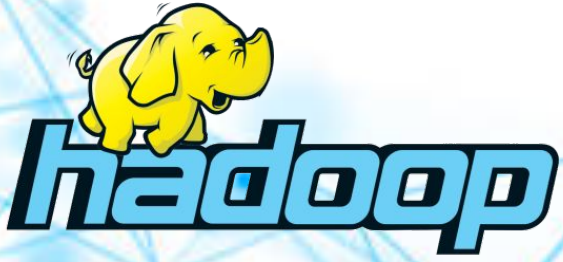
- **Horizontal scaling** means that you scale by adding more machines into your pool of resources
- **Vertical scaling** means that you scale by adding more power (CPU, RAM) to an existing machine.



Big Data Platforms



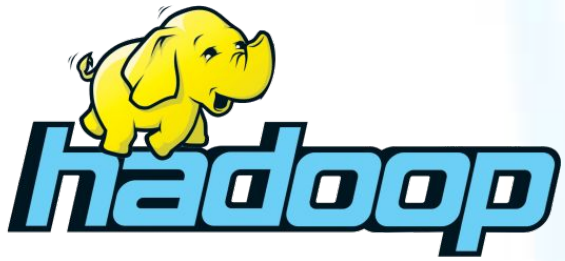
Big Data Platforms



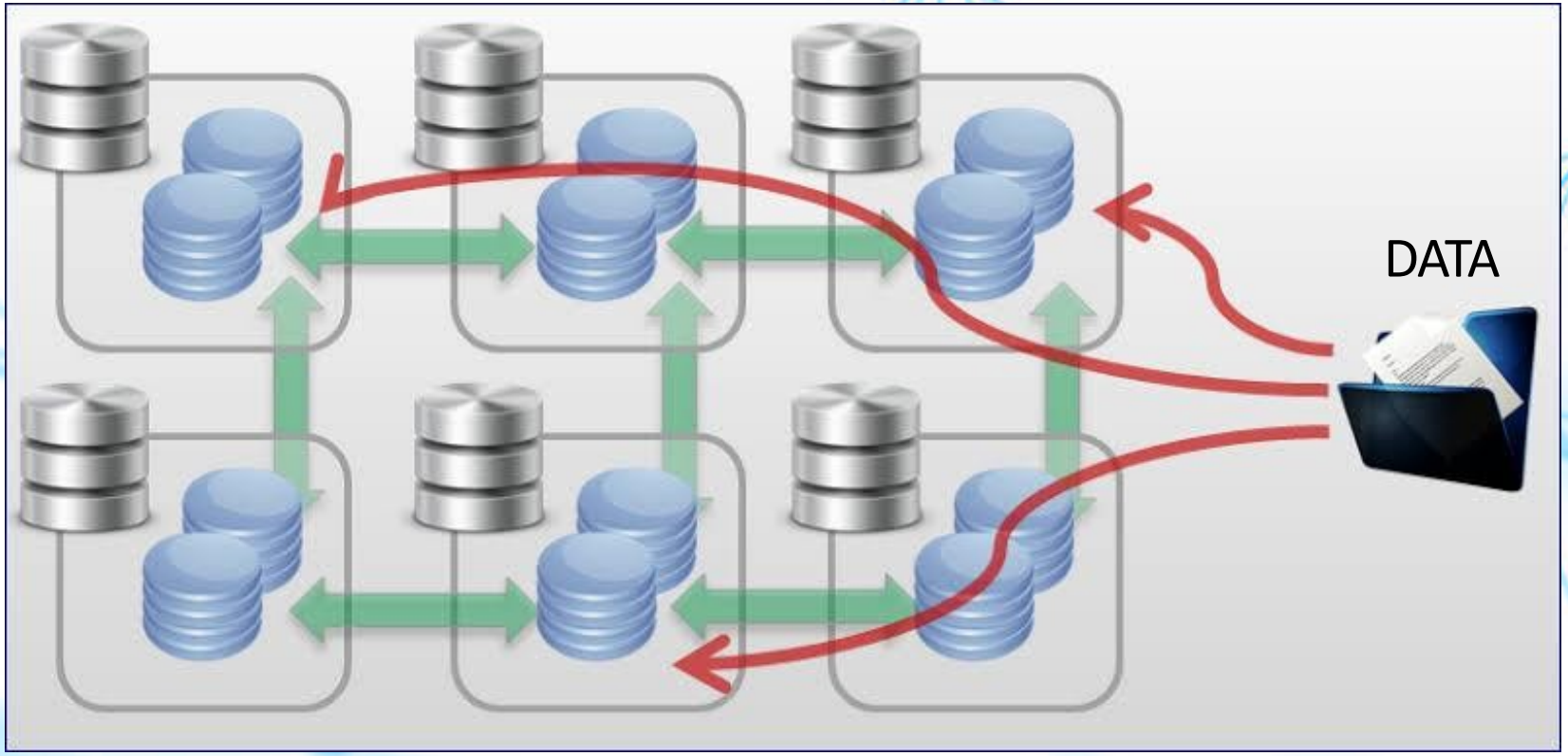
- Hadoop
  - HDFS (Hadoop Distributed File System)
  - Map-reduce
- Spark
  - RDD

What is Hadoop?

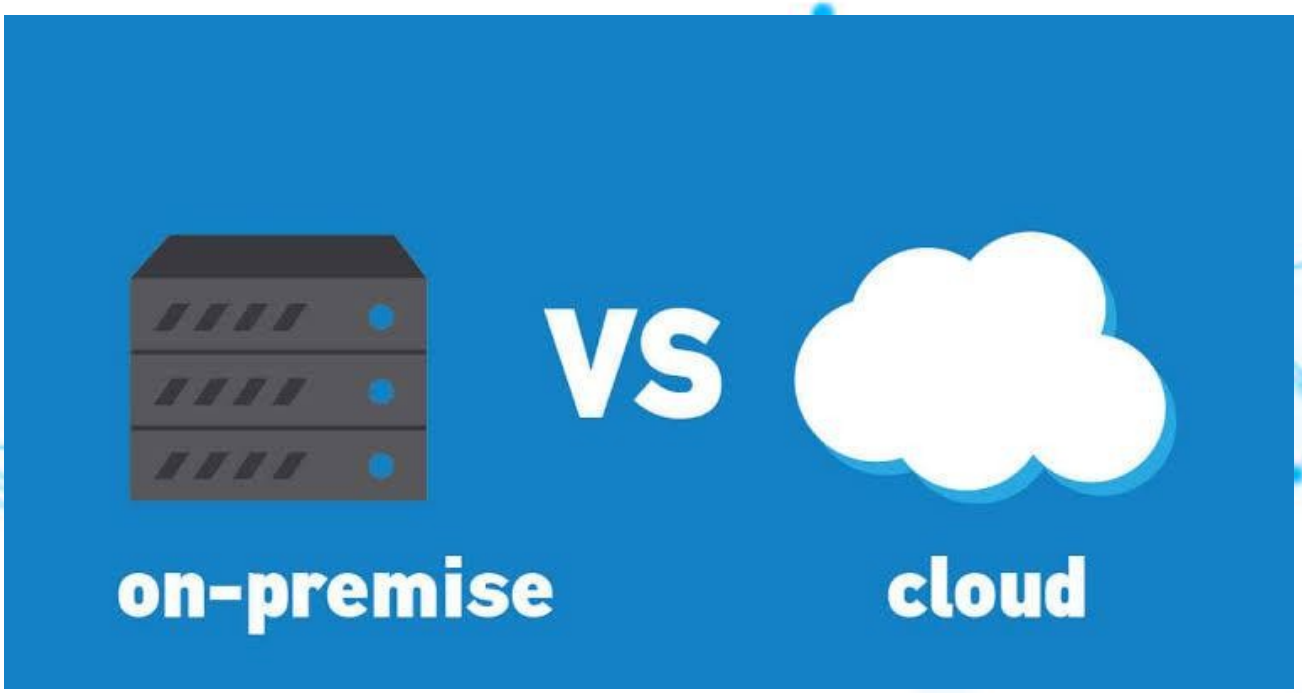
Apache Hadoop is an open source software framework for storage and large scale processing of data-sets.



Moving Computation to Data



Source: UC San Diego, Big Data


VS
on-premise
cloud

On Premises

- Software and technology
- located within the physical confines of an enterprise
- Often in the company's data centers – as opposed to running remotely on hosted servers or in the cloud.

Cloud Computing

- Differs from on-premises software in one critical way.
- Third-party provider hosts all that for you.
- Allows companies to pay on an as-needed basis.
- Allows companies to effectively scale up or down depending on overall usage, user requirements, and the growth of a company.

Data Engineering

What is Data Engineering?

"Data engineering comprises all engineering and operational tasks required to make data available for the end-user, whether for the purpose of analytics, model building or app development etc."

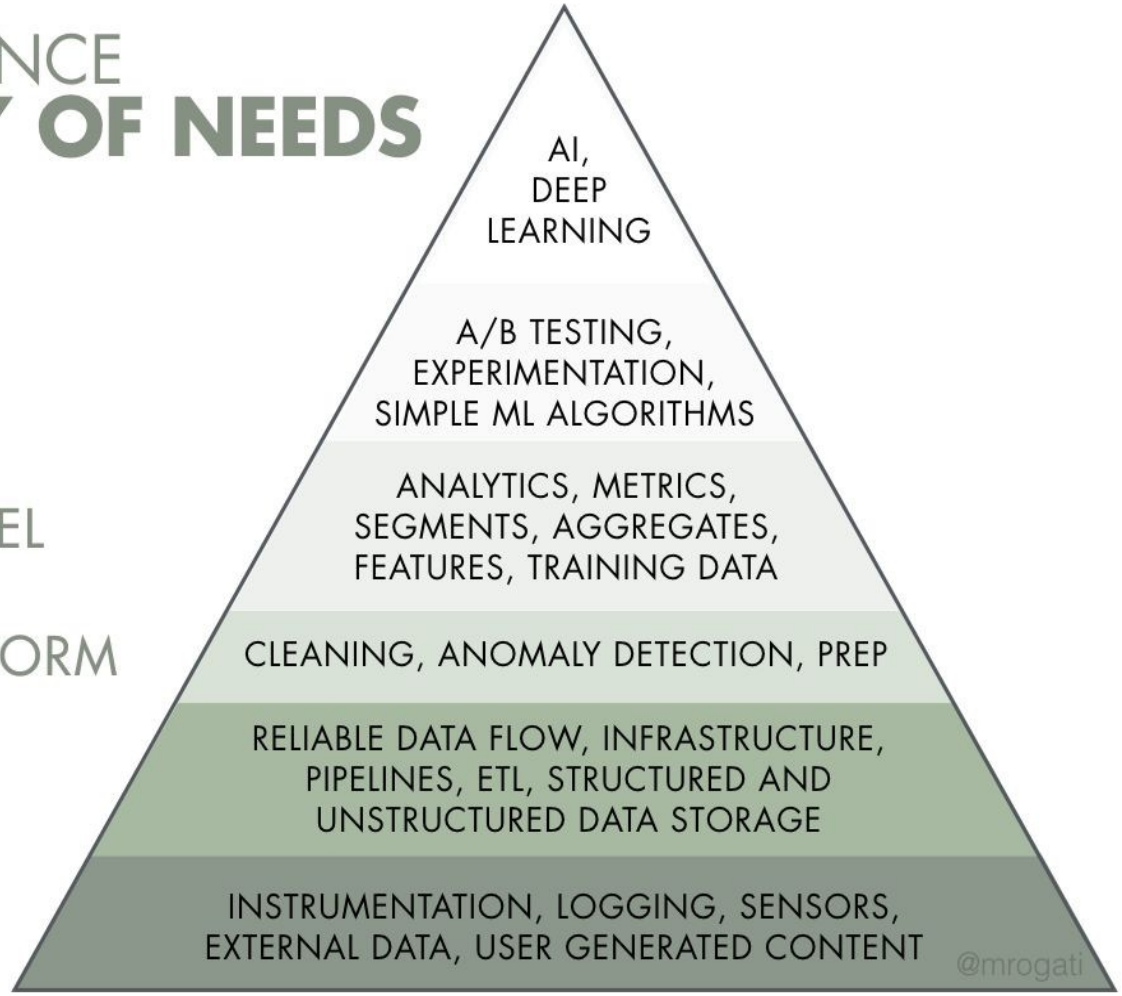
3 step process in layman's terms.

1. Taking raw data.
2. Doing a bunch of work to it.
3. Delivering a clean dataset of database.

What do Data Engineers do?

- Model data
- Build production ready data warehouses and data lakes.
- Tools that process massive amount of data.
- Automate data pipelines.

THE DATA SCIENCE HIERARCHY OF NEEDS



LEARN/OPTIMIZE

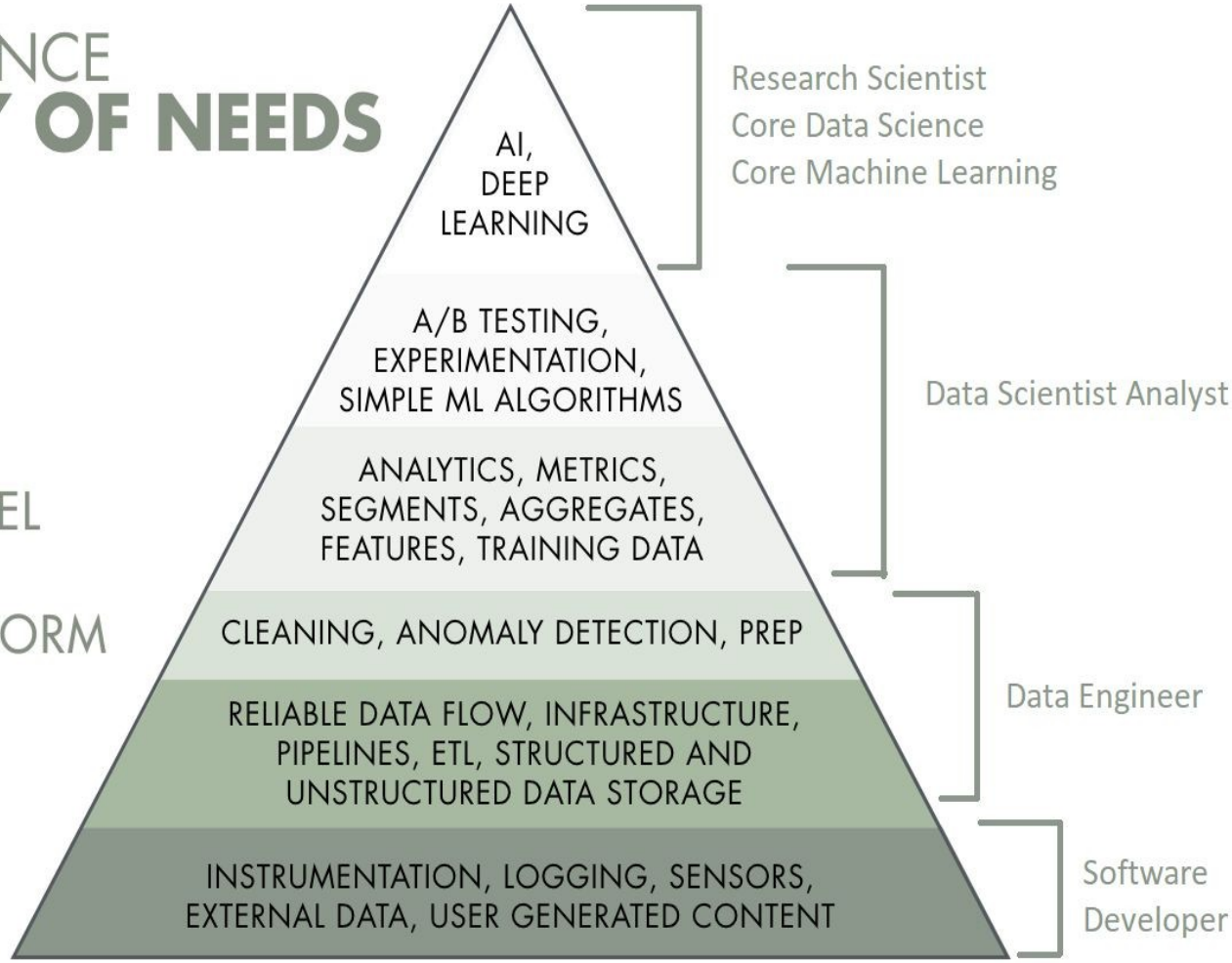
AGGREGATE/LABEL

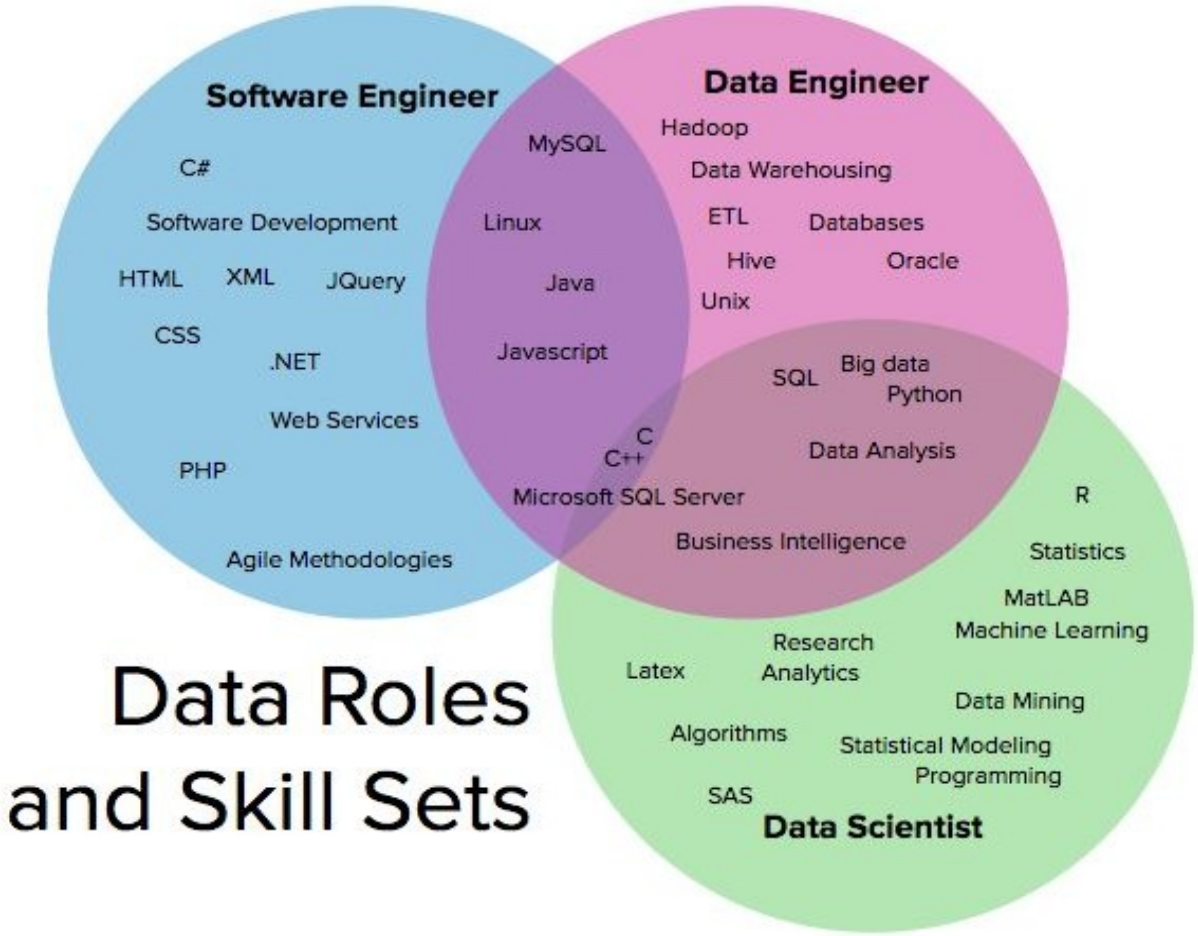
EXPLORE/TRANSFORM

MOVE/STORE

COLLECT

THE DATA SCIENCE HIERARCHY OF NEEDS




Software Engineer
C#
Software Development
HTML
XML
JQuery
CSS
.NET
Web Services
PHP
Agile Methodologies
Data Engineer
MySQL
Hadoop
Data Warehousing
Linux
ETL
Databases
Hive
Oracle
Java
Unix
Javascript
SQL
Big data
Python
C
C++
Data Analysis
Microsoft SQL Server
R
Business Intelligence
Statistics
MatLAB
Machine Learning
Research
Latex
Analytics
Data Mining
Algorithms
Statistical Modeling
Programming
SAS
Data Scientist
Data Roles
and Skill Sets

Common data engineering activities.

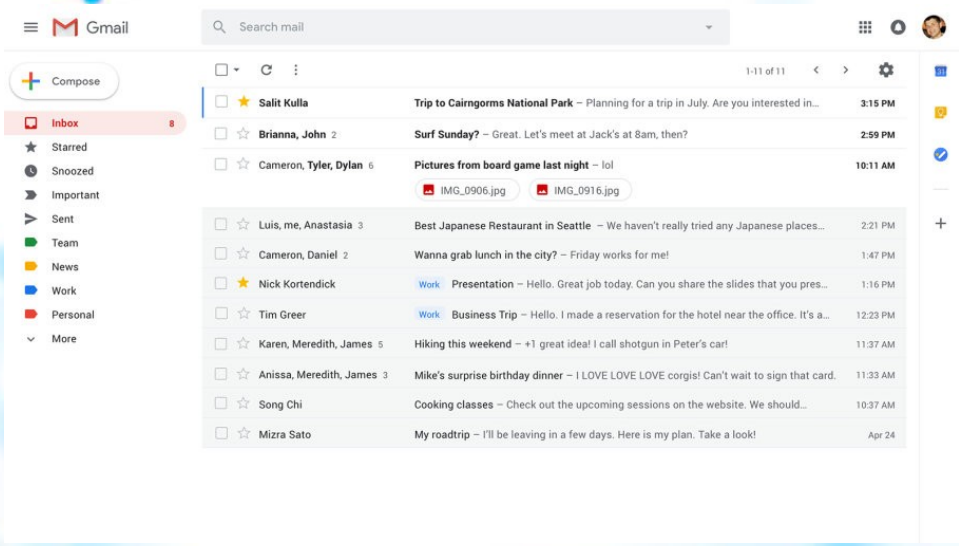
1. Ingest data from a data source.
2. Build and maintain data warehouse.
3. Create a data pipeline.
4. Create an analytics table for a specific use case.
5. Migrate data to the cloud.
6. Schedule and automate pipelines.
7. Debug data quality issues.
8. Optimize Queries.
9. Design a database.

Types of Data

- Relational Data
- Text Data
- Multimedia Data
- Time Series Data
- Sequential Data
- Streams
- Graphs and Homogeneous Networks
- Graphs and Heterogeneous Networks

Types of Data: Text

- blogs, webpages, tweets, documents, emails
- High dimensionality, vocabulary, information retrieval, natural language processing
- Latest search engine for Walmart.com uses text analysis, machine learning and even synonym mining to produce relevant search results. Wal-Mart says adding semantic search has improved online shoppers completing a purchase by 10% to 15%. "In Wal-Mart terms, that is billions of dollars,"



Types of Data: Multimedia

- image, audio, video
- 'Fast food and video' company is training cameras on drive-through lanes to determine what to display on its digital menu board. When the lines are longer, the menu features products that can be served up quickly; when the lines are shorter, the menu features
  higher-margin items that take longer to prepare

Here's why some McDonald's restaurants are putting cameras in their dumpsters

By Rachel Metz, CNN Business

Updated 1736 GMT (0136 HKT) December 18, 2020

CNN BUSINESS

Types of Data: Time Series

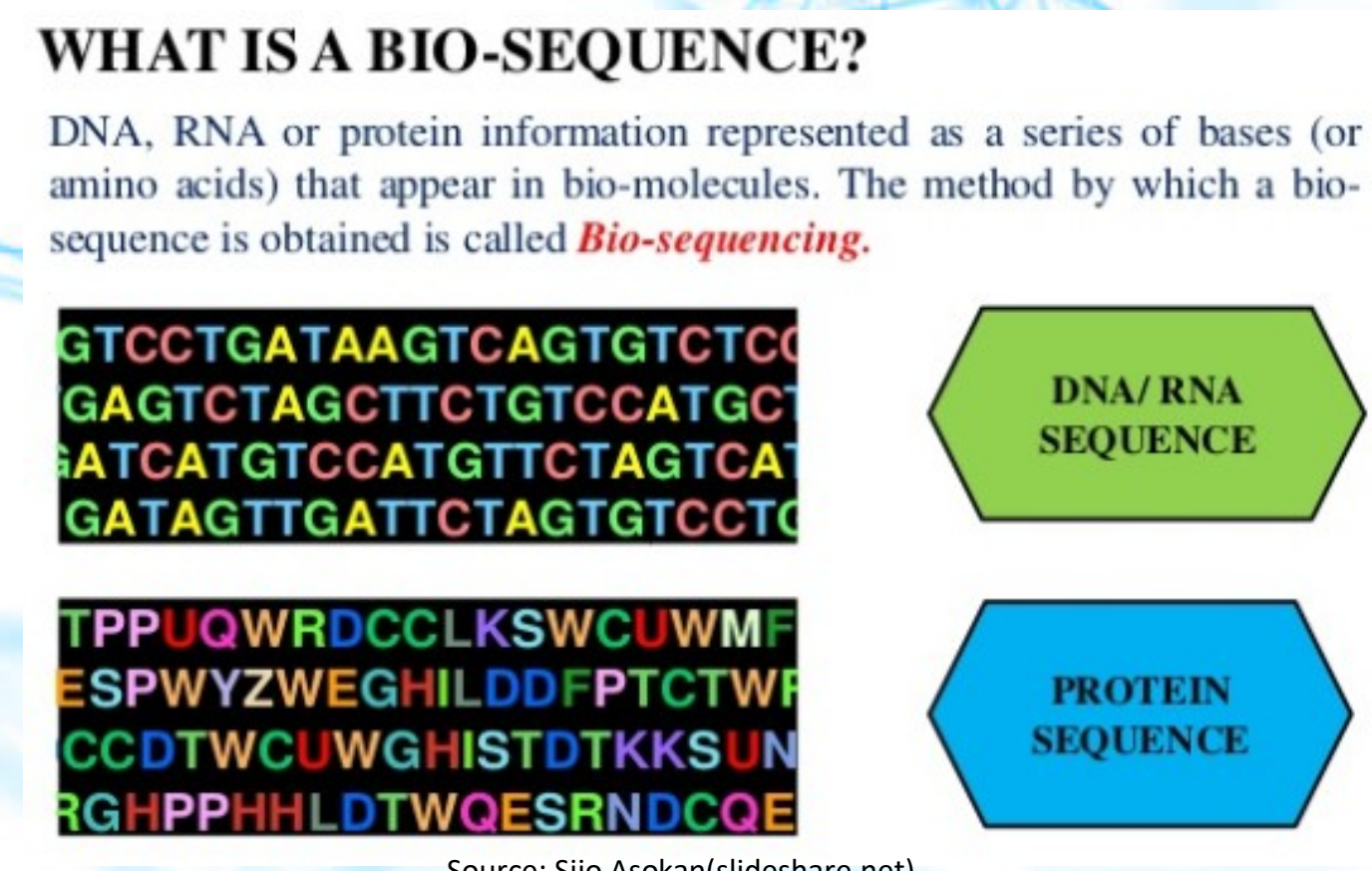
- Sequence of data points at equally spaced time intervals
- Sensor data, Stock market data, Forex rates, Temporal tracking (GPS), Smart Meters Data (AMI)
- Understanding the underlying forces and structure of observed data and fit a model to forecast, monitor or control
- Economic Forecasting, Sales Forecasting, Stock Market Analysis, Yield Projections, Process and Quality Control, Inventory Studies, Workload Projections, Census Analysis

market momentum

Application of Time Series Analysis in Financial Economics by @Statswork https://link.medium.com/n3FJPzhIadb

Types of Data: Sequential Data

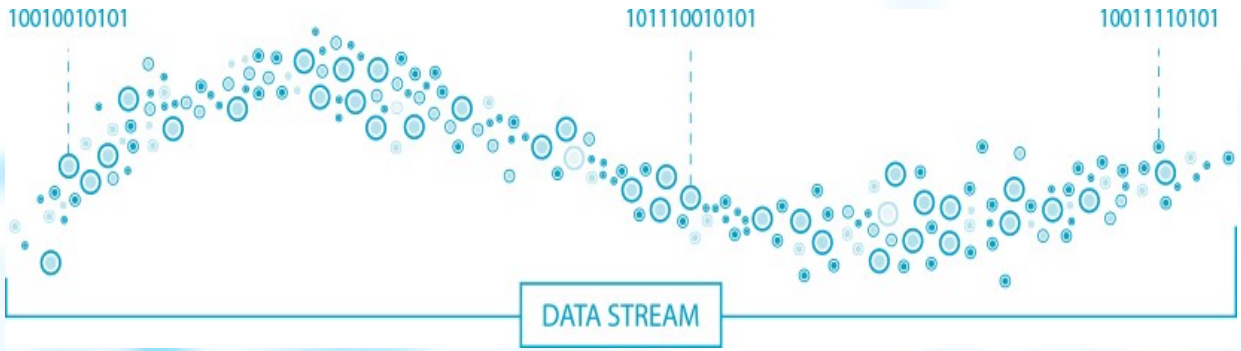
- Bio-sequences
- Discretized music and audio data
- Text



Source: Sijo Asokan(slideshare.net)

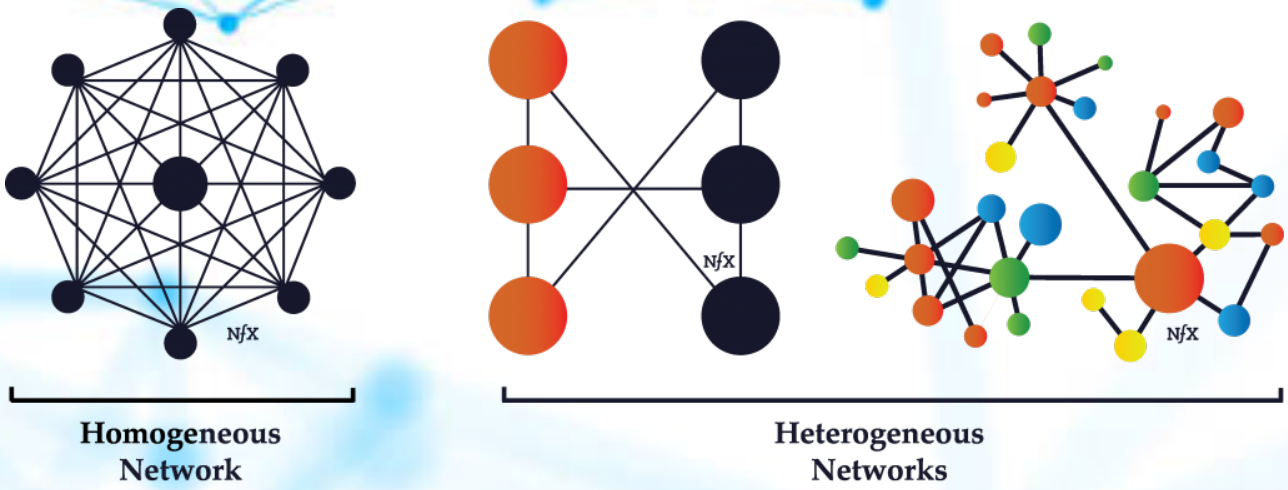
Types of Data: Streams

- Real time data
- Single pass algorithms/online
- Algorithms Irreversible decisions
- Small memory algorithms



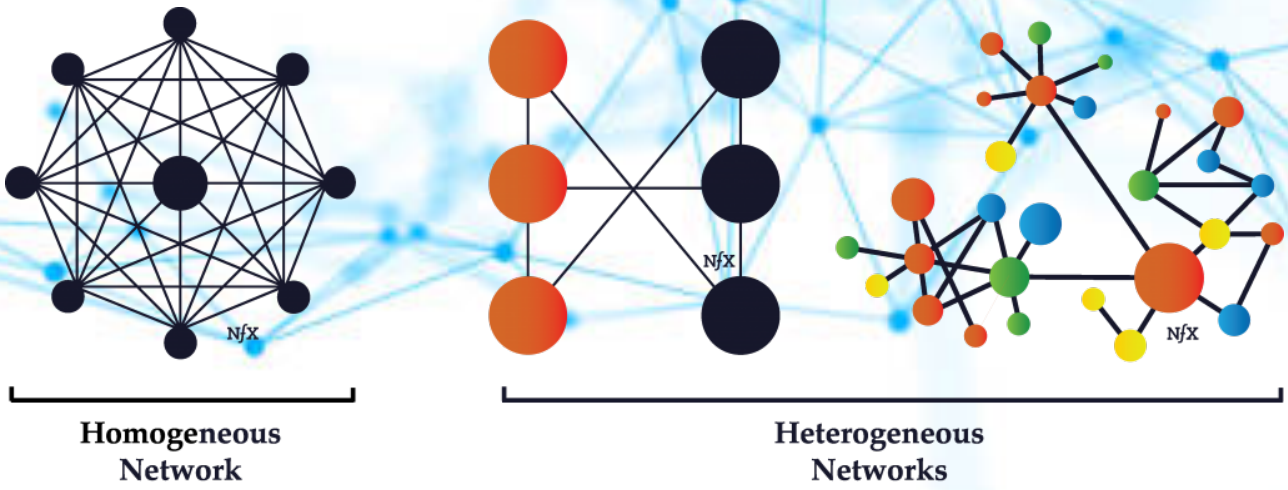
Types of Data: Graphs/Homogeneous Networks

- $G = (V,E)$, data items represented as graphs
- Could have similarity on edges
- Could have weights on vertices, edges or both
- Facebook, webgraph, twitter, co-authorship graphs (bibliometric), citation networks



Types of Data: Heterogeneous Networks

- Nodes represent different entities
- Authors and conferences



Find me @

Dr. Iqra Safder
Assistant Professor
iqra.safder@nu.edu.pk
Office: Ground floor, Civil block,
NUCES, Lahore

National University
Of Computer and Emerging Sciences