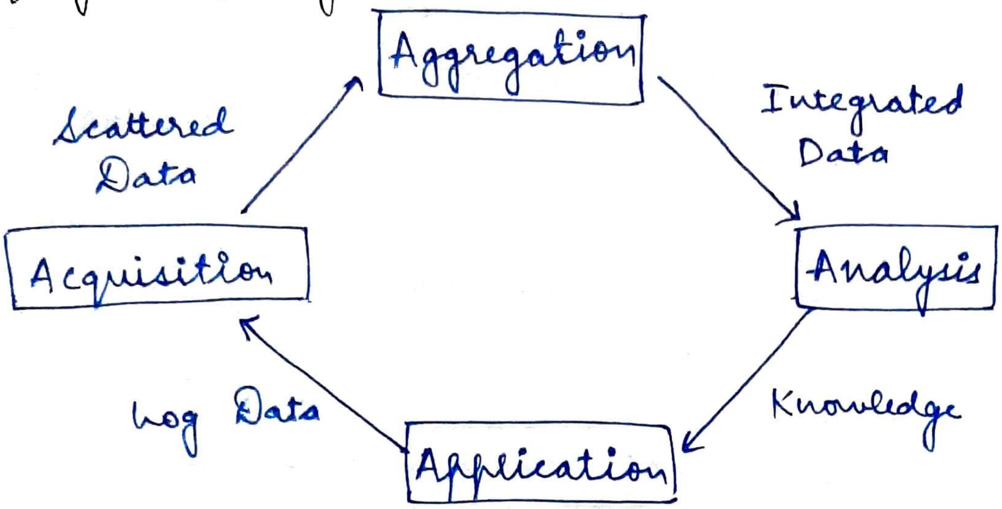
UNIT 2 : Introduction to Bigdata

→ Big Data

Big Data are high-volume, high-velocity and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

Lifecycle of Big Data



Type Of Data

- Relational Data (Tables / Transaction / Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF),
- Streaming Data

Uses Of Data

- Aggregation and Statistics
- Data Warehouse and OLAP
- Indexing, Searching and Querying
- Keyword based search
- Pattern matching (XML/RDF)
- Knowledge discovery
- Data Mining
- Statiscal Modelling

Data Mining

- Data Mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems.
- Discovery of useful, possibly unexpected, patterns in data
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data.
- Exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns.

Data Mining Tasks

- Predictive - helps you to identity what kind of data you are looking for.
  - Classification
  - Regression
  - Deviation Detection
  - collaborative Filter.
- Descriptive - helps with the detailing of the information we are looking for.
  - Clustering
  - Association Rule Discovery
  - Sequential Pattern Discovery
  -

5 V's of Big Data

→ Volume - Data
(Terabytes, Records / Arch, Tables, Files, Distributed)

→ Velocity
(Batch, Real / near-time, Processes, Streaming)

→ Value
(statistical, Events, Correlations, Hypothetical)

→ Variability
(changing data, changing Model, Linkage)

→ Veracity
(Trustworthiness, Authenticity, Origin, Reputation, Availability, Accountability)

→ Variety
(Structured, Unstructured, Multi-factor, Probabilistic, Linked, Dynamic)

Advantages Of Big Data

- Big data analysis derives innovative solutions.
- Big data analysis helps in understanding and targeting customers.
- It helps in optimizing business processes.
- It helps in improving science and research
- It improves health care and public health with availability of record of patients.
- It helps in financial trading, sports, polling, security / law enforcement etc.
- Any one can access vast information via surveys and deliver answer of any query
- Every second additions are made.
- One platform carry unlimited information.

Disadvantages Of Big Data.

- Traditional storage can cost lost of money to store big data.
- Lots of Big data is unstructured.
- Big data analysis violates principles of privacy.
- It can be used for manipulation of customer recor
- It may increase social stratification.
- Big data analysis is not useful in short run.
- It needs to be analysed for longer duration to leverage its benefits.

- Big data analysis results are misleading sometimes
- Speedy updates in big data can mismatch real figures.

Classification Of Types Of Big Data

→ Social Networks (human - sourced information)

- Social Networks : Facebook, Twitter, Tumbler, etc
- Blogs and comments
- Personal documents
- Pictures : Instagram, Flicker, Picasa, etc
- Videos : Youtube, etc
- Internet searches
- Mobile data ~~searches~~ content : text messages
- User generated maps.
- E-mail

→ Traditional Business Systems (process - mediated data)

- Data produced by Public Agencies
    - Medical Records
- Data produced by business
    - Commercial transactions
    - Banking / stock records
    - E- commerce
    - Credit Cards

→ Internet Of Things (machine - generated data):

- Data from sensors
    - Fixed sensors
    - Home Automation
    - Weather / pollution sensors
    - Traffic sensors / Webcam
    - Scientific sensors
    - Security / surveillance videos / images
- Mobile sensors (tracking)
    - Mobile phone location
    - Cars
    - Satellite images

- Data from computer systems
  - Logs
  - Web Logs.