U-3 Big Data Analytics Life Cycle

Page No.
Date

* Big Data & sources -

Large volumes of data available at various sources in varying degree of complexity, generated at different speed ie. velocities & technologies, processing modules.

- term → describe collection of data ie huge in size & yet growing exponentially with time.
- data → large & complex
  → no traditional method/ tools work or to process/store it.
- consists of extensive datasets that require a scalable architecture for efficient storage, manipulation & analysis.

Sources of Big Data -

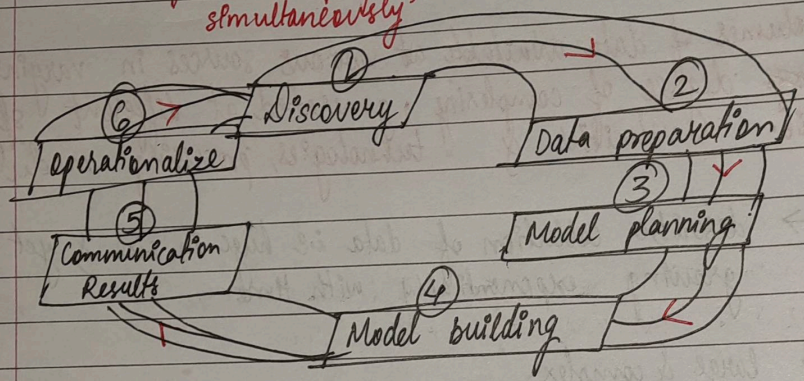
1) Social Media - Fb → 500+ TB of data everyday
   → status msg, photos & videos.
2) stock exchange → TB/sec → trade data of users & companies
3) aviation industry → Jet engine → 10 TB of data → 30 min flight
4) Survey Data → online or offline survey → 100 & 1000's response
   → processed for analysis & visulation
5) Compilance Data → Healthcare, hospitals, life sciences & finance, have file compilance reports, etc.

appln of Big data →
- govt
- medical
- energy
- finance
- manufacturing & transportation
- media & entertainment
- weather
- ecommerce.

Page No.
Date

* Data Analytics Lifecycle -

→ data problems & data science projects.
→ six phases the project work can occur in several phases simultaneously.



1) Discovery →
- defining data's process.
- critical objective of business → mapping out data.
- term learns → business domain & previous similar project
- team evaluates - technology, people, data & time.

2) Data prep →
- business req → data req
- collecting, processing & clean data
- analytics → workspace.
- sandbox → collects all king of data.
- ETLT → extract, transform, load & transform.
- data → sandbox.
- data acquisition → existing data from outside.
- data entry → creating new data values from data.
- data conditioning → cleaning data performing transformations
- common tools for data prep.

Tools →
- Hadoop
- Alpine Miner
- Open Refine

3) Model Planning -
- asses structure of data.
- model selection
- common tools → SQL/SAS/ACCESS.

4) Model Building -
- design model
- execute model
- SPSS modular (IBM) → enterprise level computing.
- STATISTICA & MATHEMATICA → data mining & analytics tools.

5) Communicate Results -
project ~~repor~~ results → success or faliure
vital findings → analysis
communicate / document → key findings & major insights.

6) Operationalize -
→ sandbox → live environment.
→ team → communication → benefit of project
→ risk management → efficiently.
→ pilot project → execute algorithms more efficiently.
→ to test model → produc° envi → monitor accuracy is retrain model if nescessary.

5V's of Big Data -

- **Volume →**
  - Big Data characterised → enourmous volume → petabytes or exabytes
  - large data needs specialized tools & technologies to store process & analyze.
  - observes & tracks data from various sources

- **Velocity →**
  - data streams → high speed & dealt timing
  - processing of data → streamed data → real time results → fast
  - insights generated → relevant & actionable.
  - speed of generation of data.

- **variety →**
  - heterogenous sources
  - nature of data → structured & unstructured.
  - social media data → sensors, audio, video, etc
  - data → insights → manage & analyze.

- **value →**
  - business value → Big Data.
  - generate some sort of value → company doing analysis
  - insights → formal decisions
    - optimize operations
    - gain competative edge in marketplace

- **veracity →** led to
  - inconsistency / uncertainity in data
  - challenges in data quality & Reliability.
  - insights → accurate & trustworthy.

Page No.
Date

Q3 Exploratory Data Analysis (EDA)-

- crucial step in Data Analysis → understand data before modelling
- visually & statistically exploring data → gain insights
  identify patterns
  detect anomalies
  formulate hypothesis
- analysts to develop a intuition about dataset & informed decisions

EDA's → 5 M framework

Mindset
- curious, open about dataset
- allow data to reveal characteristics & patterns

Methods (to explore & summarize data)
- summary statistics
- Data Visualization
- Dimensional Redu^n
- Data Profiling
- Correlation Analysis

Missing Data
- missingness
- pattern (understand)
- strategies to handle

Multivariant Analysis
- Rel^n b/w multiple variables
- diff variables interaction
- identify clusters

Models
- potential predictors
- variable importance
- formulate hypothesis
- ML models

→ uncover pattern & anomalies
→ generate hypothesis
→ guide analysis
→ modelling decisions.