03 Big Data analytics
DSBDA Life cycle.

Big Data :-
Collection of datasets, which are large & complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Data Science :-
is a field that uses Scientific methods, algorithms, & tools to extract knowledge & insights from structure/unstructured data involves techniques from statistics, ML, CS, to analize data

Business Intelligence.
turning data into useful info to help business make better decisions.

Data Deluge : overwhelming flow of large amount of data being generated constantly from various sources.
Reasons for Data Deluge
1. widespread use of internet & smart devices.
2. Social media & streaming platform
3. Sensors & IOT Devices
4. Digital transactions & Businesses.
5. cloud computing & Big data technologies.
6. Healthcare products & Security monitoring

Eg :- fitness tracker watch, every day x millions of users = terabytes of data every day.

Data Analytics lifecycle
1. Discovery
2. Data preparation
3. Model planning
4. Model Building
5. Communicate results
6. operationalize

— Iterative (to cover new informations or data)

(1) Discovery :-
- first phase, goal is to
  - understand business problem, identify project goals & gather all necessary background info.
  (1) Learning the Business Domain :-
    understand company goals & business problems to solve
  (2) Resources :-
    identify tools, tech, team, data infrastructure for proje
  (3) Framing the problem :-
    convert the business challanges to clearly defined analytical problem
  (4) Identifying key stakeholders :
    who's involved & who is benifited.
  (5) Interviewing the Analytics Sponsor.
    discuss with clients their objectives, expectations.
  (6) Developing initial hypotheses.
    assumptions about what might be happening in data.
  (7) Identify potential data sources.
    external / internal sources from which data can be collected.

Activities involved. identify potential resources.
(1) understanding data Needs.
(2) Identify internal data sources.
(3) Identify external data sources
(4) check data availability & access.
(5) Assess data Quality. (accurate, complete & up to date)
(6) understand data formats & storage
(7) Document data source. (record)

- right data found, verified & prepared.

(2) Data Preparation
- includes ~~prior~~ steps to explore, preprocess & condition data
- most iterative phase.

Steps in data preparation
(1). Data collection. - gather data internal/external source
(2). Data integration :- combine data from multiple sources
(3). Data cleaning :- fixing issues like, missing val, duplicates, outliers.
(4). Data transformation :- convert data to suitable form ie normalize, encoding, aggregation.
(5). Data Reduction :- Reduce data size by removing unnecessary feature
(6). Data formatting :- data in standard format.
(7) Data Sampling :- selecting a small subset for working on data
(8). Feature Engineering :- new feature from existing to increase performance

Preparing analytic Sandbox.
(1). It is a separate environment or workspace for data analysis & experimentation.
purpose of sandbox.
(1). Safe Exploration :- Team can perform testing, analysis & model building. without affecting live production systems.
(2). Expansive Data collection.
Sandbox gathers diverse & large dataset from multiple source for deeper analysis
(3). Supports Advanced Analytics.
complex tasks. like AI, ML not just BI
(4). Encourages Innovation :- provides risk free zone for testing new ideas.

⑤. Separation from production Data
⑥. Gaining acceptance.

Performing ETLT (Extract, Transform, load, Transform
- Data Extracted from source.
- transformed (cleaned / structured) & then
- loaded to targeted systems.

ETLT in Analytics Sandbox
in sandbox env. Data is ~~extr~~ E & L first & then transformed.
- early load to prevent the original form of Data.

use case - Fraud Detection in credit card transaction.
- outlaws can signal fraud.

- Hadoop is used. since it can process large amount of data

Data Conditioning
- includes cleaning data, normalizing dataset & performing transformations.
- viewed as data preprocessing step.
- Conditional Question like what source etc is asked.

Survey & Visualise:-
analysts use graphs, charts, visual tools to get a clear overview of data., better understanding data. patterns, quality & readyness of data.
need. ① Quick understanding of data.
② Data Quality check
③. Distribution consistancy
④. Assess granularity
⑤. population representation.

Tools for Data preparation.

1. Hadoop. - perform parallel insight & analysis
2. Alpine Miner. - provides GUI for creating analytic workflow.
3. OpenRefine - opensource tool to work with data.
4. Data Wrangler - tool for data cleaning & transformation.

(3). Model planning.

- Deciding how to approach the data analysis what techniques to use & planning the modelling process. to solve business problems.

Key activities in model planning.

1. Assess Data Structure. (structure / semi-structured / unstructured)
   - this helps choose right tools & technique for modeling
2. Choose suitable Analytical Techniques.
   select ~~model~~ methods that helps meet business goals. & test hypotheses effectively
3. plan modeling Aproach.
   Decide if single model is enough, or series of them is needed.
4. Learn from past work.
   Research similar problems from past & use those insights.

need.

1. ensures right tool selection.
2. align model work with business goals.

Data Exploration & Variable Selection.

- exploring data to understand relationship b/w variables & selecting most relevant ones for modeling.

Activities involved in Exploration & variable selection

1. Exploring variable Relationship
2. use Data Visualization tools. to find patterns, trends.
3. consider stack holders input to guide variable selection.
4. identify key predictors :- select most imp. variables.
5. Veriable selection needs trial & error hence iterative.
6. plans for specific models.
7. Remove irrelevant or Redundant Data.

Model Selection.

step where the team chooses most suitable analytical techniques to solve business problems. based on type of data & project goals.

key points.

1. Goal oriented selection. - choose best analytical technique
2. Real would representation - models are simplified versions of real-would behavior., created using ~~###~~ rules & logic to mimic real events.
3. Type of Data. Matters :- structured / un / hybrid. & select model accordingly.
4. use of statistical Tools :- initial model build using R, SAS, Matlab for quick testing & validation.
5. Tool ~~#~~ limitations :- tools dont work well with lauge datasets.
6. Candidate Models :- select & see which performs best
7. prepare for next phase

Tools :

1] R :- complete set of modelling capabilities., 5000 packets for. D.A & graphical representation

2] SQL Analysis services :- can perform in database analysis of common data mining functions, involved aggregations,

& basic predictive models.

5] SAS / ACCESS provides integration b/w SAS & the analytics sandbox via multiple data connections.

(4). Model Building.

- Team creates, test & validate the model using training & testing data, based on techniques selected in previous phase.

key activities

(1.) Execute the chosen models.
- Run models planned in phase 3 - model selection.

(2.) Prepare Datasets.
Divide data into training, testing & production dataset.

(3.) Train the model. using training dataset.

(4.) Test the models. using testing dataset. to eval. model.

Questions to consider.

(1). model validity & Accuracy
(2). Domain Expert validation
(3). Parameter Meaningfulness
(4). Readiness for use. (accurate, less mistakes, desired env).

Tools.

(1) SAS Enterprise Miner - enterprise level computing & analytics.
(2). SPSS Modeler (IBM) - enterprise level
(3) Matlab :- high level data analytics, algo, data exploration.
(4). Alpine Miner - provides GUI & backend analytics tools statistical & mathematical. - for data mining & analytics tools.
(5). R & PL/R - procedural lang for PostgreSQL with R
(6) Octave - for computational modeling.
(7) Weka - data mining software package with analytic workbench
(8). Python - toolkit for ML & analysis

SQL

(5) Communicate Results.
- determine if the team succeeded or failed in objectives
- Assess if results are statistically significant & valid
- Identify aspects of results that present salient features
- communicate & Document the key findings & major insights of analysis
- Satisfaction of customer

(6) Operationalize
model is deployed in real world setting & its performance is monitored in production
key points.
(1). communicate Project Benifits. (Business stakeholders)
(2). Pilot Deployment :- launch moden in small controled environment to minimize risk
(3). Risk management :- pilot run helps detect risk early & reduce risk
(4). Database Execution for Efficiency.
for large Dataset. the algo needs to be executed inside database. instead of memory
(5). Live Testing :- Run model in Real time conditions. & test actual performance
(6). performance monitoring (model accuracy, speed & o/p)
(7) model Retraining. with new Data.
(8). Ready for full deployment

Four Main Deliverables.
(1). Presentation for project Sponsors.
- executives & senior management.
- focused on key takeaways
- high level overview.

(2). presentation for Analysts.
- aimed at business analysts & operational user.
- Explains business processes & reports.

(3). Code for Technical Team.
- complete & cleaned codebase used to build & deploy the model.
- Replication & scaling of model.

(4). Technical specifications:
- Detailed dock of how code works & how to implement it
- helps in future maintenance & upgrades.

BI	Data science.
(1). Analyzes past & present data.	(1) predicts future trends & outcomes.
(2). Reporting & decision making based on historical data.	(2) Building productive models & uncovering patterns.
(3) Mainly structured data used	(3). structured/ unstructured data used
(4) Tools:- Power BI, Tableau	(4). Tools:- Python, R
(5) used for Business analysts, managers.	(5). Data Scientists, researchers.
(6). less complex., UI based.	(6). more complex., stats & ML
(7) Eg. Monthly salary Report	(7) predict next month customers churn

Descriptive Analytics.
- looks at past data to understand what has already occured.
- helps in identifying trends, patterns & summaries.
- answers the Question "what has happend"?

Purpose:- Summarize historical data

tools & techniques.
1. data aggregation.
2. data visualization.
3. statistical summaries.

Eg. Monthly salary Report.

Diagnostic Analytics. - why did it happens
- invistigates causes behind past outcomes.
- helps in identifying reasons or root cause of trends or problem.

Purpose.
To invistigate causes behind trends or outcomes found in descriptive analytics.

Tools.
① drill down analysis
② Data mining
③ correlation & root cause analysis

Eg. Analyzing why sales dropped in a specific region.

Predictive Analytics. - What is likely to happen?
- uses past data, statistics, & ML to make predictions
- helps in forcasting future events.

Tools. ① Regression
② classification.
③ forcasting models.

Eg. Predicting which customers are likely to cancel a subscription.

Key Roles for successful Analytics project.

1. Business Users.
   - understand business domain
   - provides practical insights & context for problem
2. Project Sponsor.
   - initiates & funds the project.
   - provides high level requirements & business goal.
3. Project manager.
   - manages project timelines & resources.
4. Business Intelligence Analyst.
   - Bridges gap b/w business & data.
   - understand data trends, reporting & KPI's
5. Database Administrator (DBA)
   - Design & manage database infrastructure
6. Data Engineer.
   - Handles data ETL (Extract, Transform, Load)
7. Data Scientists.
   - Apply statistical & ML models.
   - generates insights & predictions to solve problems.