Data Analytics

Unit - 1 [One - Shot]

Most important topics:

1. Data analytics (definition, need and application)
2. Big data platforms, sources of data.
3. Process and tools of analytics.
4. Structured vs Semi-structured vs Unstructured data.
5. Data analytics life cycle.

(#) Data analytics:

Definition: DA is the science of examining raw data with the purpose of drawing conclusions about that information. It involves various techniques and process for inspecting and analysing the data.

* Need of Data Analytics:

- Informed decision-making:

Organizations rely on data analytics to make data-driven decisions.

- Competative advantage:

Companies use data analytics to gain insights into market trends, customer preferences, which can give them an edge over competitors.

- Cost reduction:

By analyzing data, organizations can identify inefficiencies, optimize resource use, and cut costs.

- Risk management:

DA helps in predicting and mitigation of risks by identifying patterns that could lead to financial, operational or strategic pitfalls.

* Applications of DA:

1. Customer analysis
2. Risk assesement
3. Fraud detection
4. Healthcare
5. Personalized advertisement

Big data platforms:

- These are comprehensive frameworks that integrate various tools and technologies to manage, process and analyze large volumes of data that traditional data processing software cannot handle efficiently.
- These platforms provides the infrastructure necesary for storing vast amount of data, managing it across distributed systems and performing complex analysis at high spud.

* Key features of Big Data platforms:

- Scalability: Can add more nodes to handle increased data volumes
- Flexibility: Flexible to multiple types of data.
- Real-time Processing: Capability to handle process, and analyze data in real-time for timely insights.
- Distributed Computing: Utilizes distributed computing architectures to parallelize the data processing tasks across multiple nodes.
- Data integration: Combines data from various sources into an analytical format.

* Big Data platforms:

1. Hadoop: An open-source framework that allows for the distributed processing of large

data sets across clusters of computers using simple programming models.

2. Spark:

An opensource unified analytics engine for large-scale data processing, known for its speed and ease of use.

3. Google BigQuery:

A serverless, highly scalable, and cost-effective multi-cloud data warehouse.

4. Amazon Redshift: A fully managed data warehouse service that makes it simple and cost-effective to analyze all your data.

* Sources of Big Data:

- Social media
- IoT devices
- Log files
- Bank transactions
- Government data
- Scientific research
- Streaming data
- Wearable health devices.

@brevilearning YT

Process and tools of DA:

The data analytics process involves several stages, each critical for transforming raw data into actionable insights.

1. Data Collection:

- Gathering raw data from various sources, such as databases, sensors, social media, and transactional records.

Tools: Apache Kafka, Flume, Google analytics, data APIs.

2. Data Cleaning (data preparation):

- Cleaning the data to remove errors, handle missing values, and ensure consistency.
- This step also involves transforming data into a suitable format for analysis.

Tools: OpenRefine, Trifacta, Python (pandas library)

3. Data Exploration :

- Exploring the data to understands its structure, patterns, and relationships.

Tools: Python, R, Tableau, Power BI, etc.

4. Data modelling :

- Applying statistical and machine learnings models to the data to extract insights and make predictions.

Tools: Apache spark MLlib, SAS, R, python, etc.

5. Data analysis :

- Analyzing the data using various techniques such as ducriptive diagnostic, predictive and pruscriptive analytics to derive meaningful insights.

Tools: SQL, Python (Numpy, Scipy), R, Excel.

ARFAT Date Page

6. Data Visualization :

- Creating visual representations of the analyzed data to communicate findings effectively.

Tools: Tableau, Power BI, R (Shiny), etc.

7. Deployment and Monitoring :

- Deploying the models and insights into production systems where they can be used for decision making.
- Monitoring involves tracking the performance of deployed models and ensuring they continue to provide accurate predictions.

Tools: Docker, Kubernetes, Apache airflow, etc.

8. Decision making and reporting :

- Using the insights derived from the data to make informed decision This stage also involves creating detailed reports to share insights with relevant stakeholders.

Tools: Power BI, Tableau, Looker, D3.js, etc.

ARFAT Date Page

Structured vs Semi-structured vs Unstructured data:

Structured data	Semi-structured data	Unstructured data
• Organized in rows and columns.	• Partially organized with tags or markers.	• Don't have a predefined format or structure.
• Have pre-defined schema.	• Flexible schema, uses tags.	• No pre-defined schema.
• SQL based.	• XQuery for XML, JSON Path for JSON.	• NoSQL based.
• Easy to search and retreive data.	• Comparitively more complicated search and retrieval.	• Difficult to search and retreive.
• Limited scalability.	• Moderately scalable.	• Highly scalable.
• Least flexibility.	• Moderately flexible.	• Highly flexible.
• High data integrity and consistency.	• Moderate data integrity.	• Low data integrity.
• Used to store financial records, customer information, etc.	• Used for metadata management, configuration files, etc.	• Used to store e-mail content, social media data, multimedia content, etc.
• e.g: SQL Databases, Excel Sheets, etc.	e.g: XML files, JSON files, etc.	e.g: text, images, videos, etc.

ARFAT Date Page

Data Analytics life Cycle (DALC):

DALC consists of several phases that guides the process of transforming raw data into actionable insights.

i) Discovery: This phase involves understanding the bussiness problem, defining objectives, and identifying the data sources required to solve the problem.

ii) Data preparation : In this phase, data is collected, cleaned, and transformed into a suitable format for analysis.

iii) Model Planning : This phase involves designing the analytical approach and selecting appropriate techniques and algorithms for modelling.

iv) Model Building : In the model building phase, the planned models are developed and tested using the prepared data.

v) Communicating Results : This phase involves interpreting and visualizing the results of the analysis to communicate insights to stakeholders.

vi) Operationalizations : In this phase, the final model is deployed into the production environment where it can be used for making decisions.

Thankyou for watching !!