Nitya CloudTech
Dream.Achieve.Succeed

TOP 12 IMPORTANT

DATA ENGINEERING

QUESTIONS AND ANSWERS

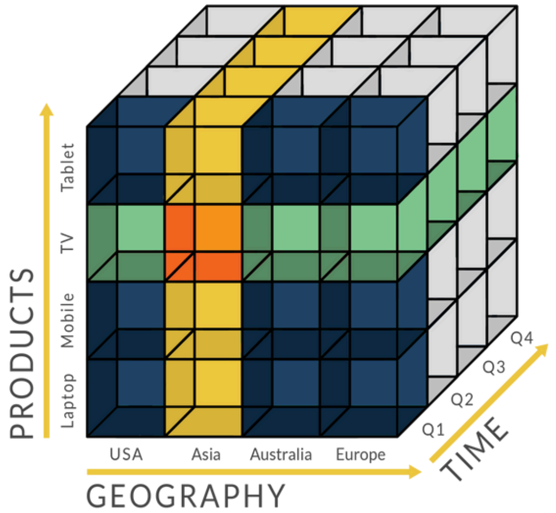
Question - 1

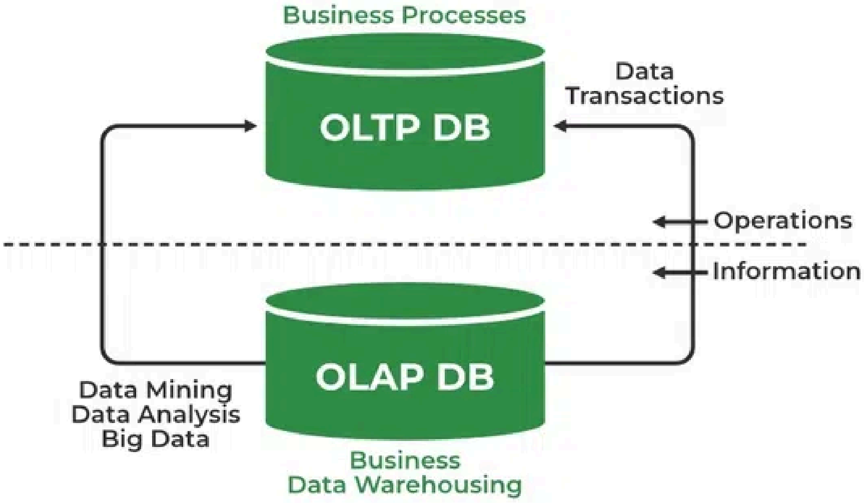
What makes OLTP different from OLAP?

- OLTP (Online Transaction Processing) handles day-to-day transactions, ensuring real-time data entry and retrieval.

- OLAP (Online Analytical Processing) focuses on analyzing large amounts of data, ensuring high integrity in queries and reports for decision-making.

- In short: OLTP is optimized for fast transaction processing, while OLAP is suited for complex data analysis.





Question - 2

How would you approach cleaning a dataset with 10% missing values?

- Assess missing data – Identify which columns have missing values and how many records are affected.

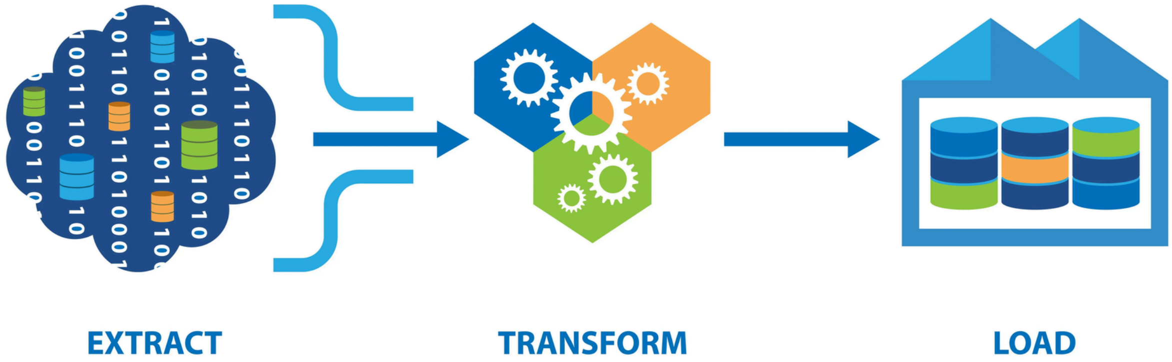
- Choose handling methods:
  - For numerical data: Use imputation (mean, median, or model-based methods) or remove rows/columns if necessary.
  - For categorical data: Use mode imputation or introduce a new category like 'Unknown.'

- Ensure no data bias – Maintain data integrity and avoid losing significant patterns.

Question - 3

How do you design an ETL pipeline for real-time analytics?

- Extract: Utilize message queues like Kafka or APIs to fetch real-time data.
- Transform: Perform on-the-fly operations like filtering, aggregation, and enrichment using stream processing engines like Apache Flink or Spark Streaming.
- Load: Store transformed data in a real-time data warehouse such as AWS Redshift or Google BigQuery.



Question - 4

How do you ensure data quality in a project?

- Clear Data Collection Standards – Define structured guidelines for data gathering.
- Data Validation – Regularly validate data using automated tools.
- Data Cleaning – Remove duplicates and irrelevant data.
- Timely Updates – Keep the data refreshed and up to date.
- Regular Audits – Periodically review data for accuracy and completeness.

Question - 5

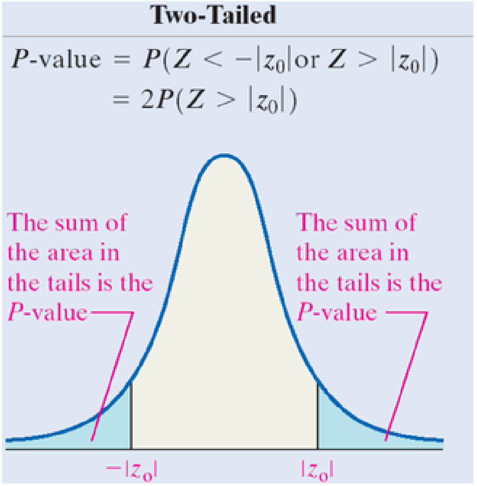
What is the importance of p-values in hypothesis testing?

- A p-value determines the statistical significance of test results.

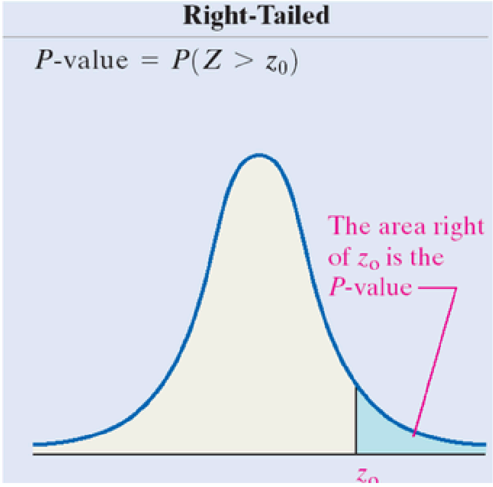
- Low p-value ($p < 0.05$): Rejects the null hypothesis, supporting the alternative hypothesis.

- High p-value: Indicates insufficient evidence to reject the null hypothesis.

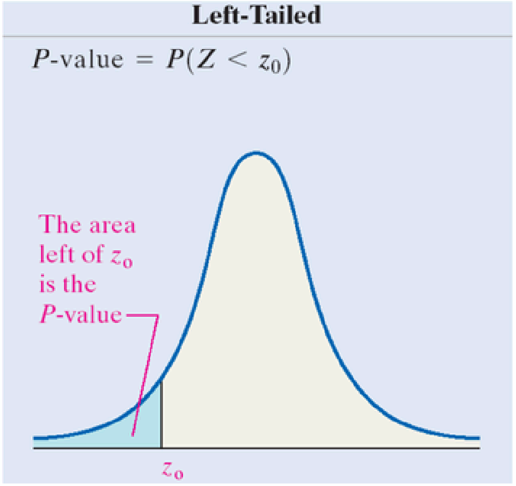
Two-tail


Two-Tailed
$P\text{-value} = P(Z < -|z_0| \text{ or } Z > |z_0|)$
$= 2P(Z > |z_0|)$
The sum of the area in the tails is the P-value
The sum of the area in the tails is the P-value
$-|z_0|$
$|z_0|$

Right Tail


Right-Tailed
$P\text{-value} = P(Z > z_0)$
The area right of z_0 is the P-value
z_0

Left Tail


Left-Tailed
$P\text{-value} = P(Z < z_0)$
The area left of z_0 is the P-value
z_0

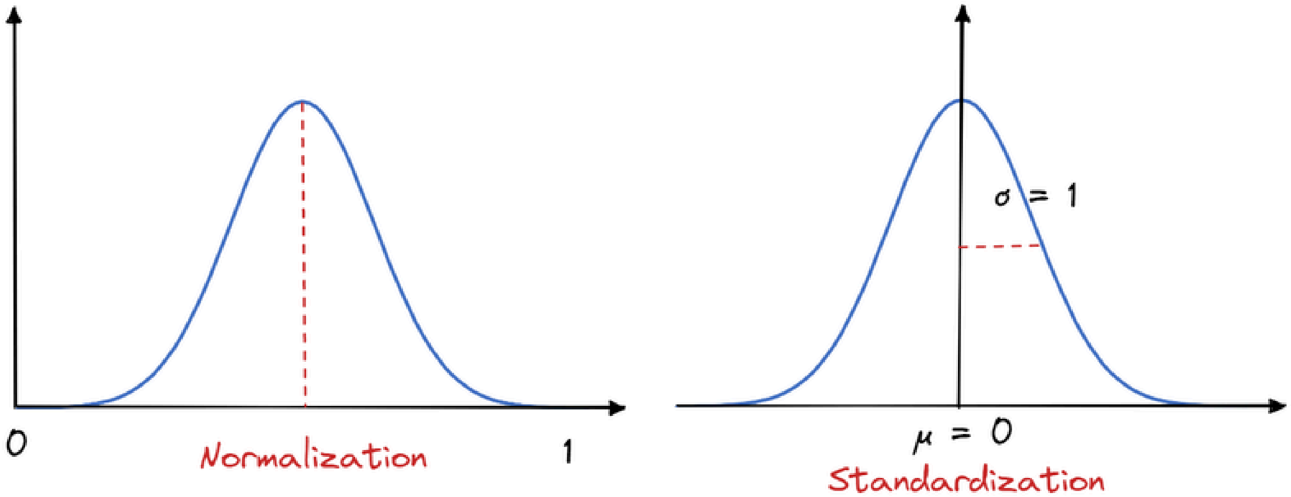
Question - 6

What is the difference between normalization and standardization?

- Normalization: Scales data within a specific range (e.g., 0 to 1).

- Standardization: Adjusts data to have a mean of 0 and a standard deviation of 1.

- When to use:
    - Use normalization when feature values have different units.
    - Use standardization when features have different scales but need uniformity.



Question - 7

How do you optimize a SQL query for large datasets?

- Use Indexes – Index frequently queried columns and JOIN keys.
- Limit Result Set – Use LIMIT or TOP to reduce processing time.
- Avoid SELECT * – Fetch only necessary columns.
- Use Efficient Joins – Prefer INNER JOIN over OUTER JOIN when possible.
- Apply WHERE Filters Early – Minimize the number of rows processed.
- Optimize Subqueries – Replace subqueries with joins where possible.
- Analyze Execution Plan – Use EXPLAIN to identify performance bottlenecks.

Question - 8

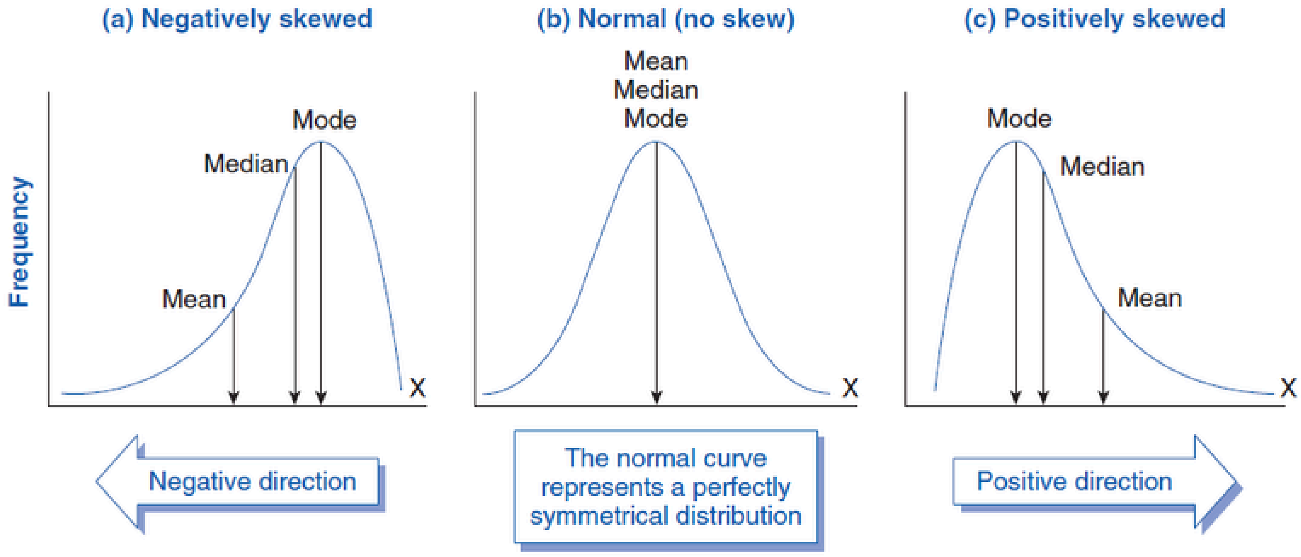
How do you handle skewed data distributions?

- Log Transformation – Apply log or square root transformation to normalize skewed data.

- Winsorization – Cap extreme values to reduce the impact of outliers.

- Resampling – Use oversampling or undersampling for imbalanced data.

- Model Selection – Use robust models like tree-based algorithms that handle skewed data well.

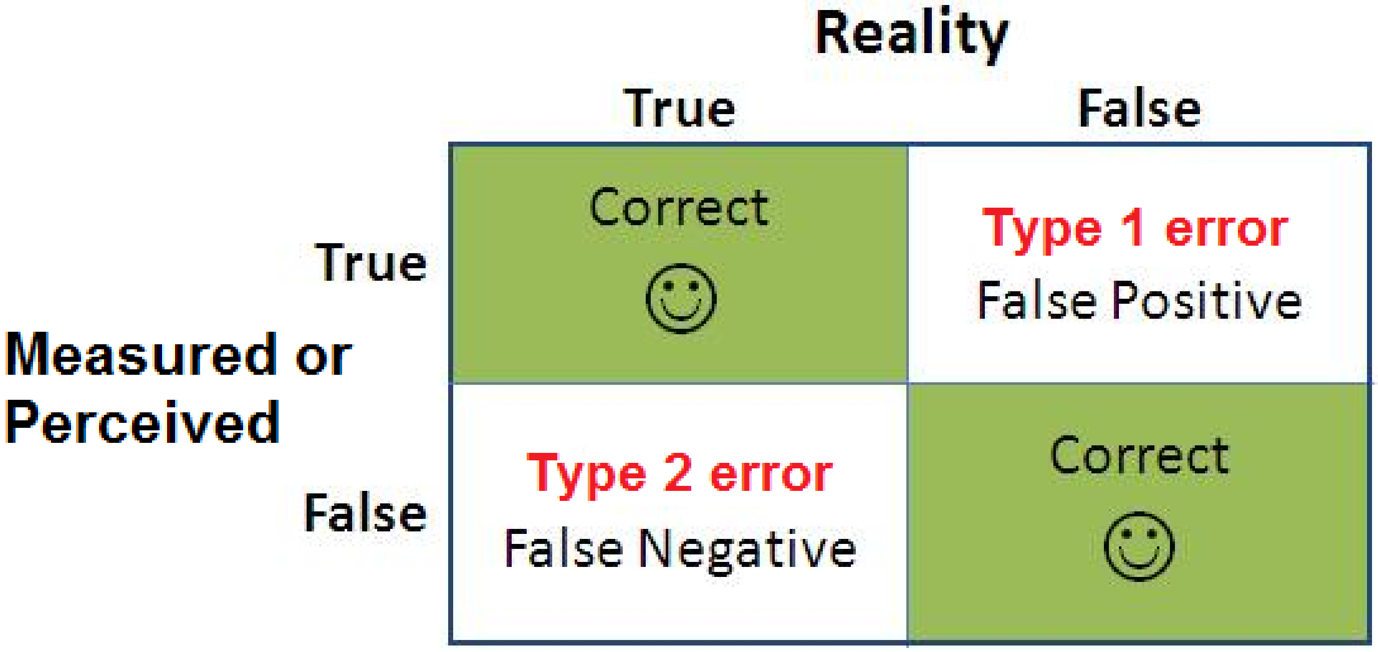


Question - 9

What are Type I and Type II errors?

- **Type I Error (False Positive):** Rejecting a true null hypothesis.
  - Example: A medical test wrongly detects a disease in a healthy person.

- **Type II Error (False Negative):** Failing to reject a false null hypothesis.
  - Example: A medical test fails to detect a disease in an infected person.



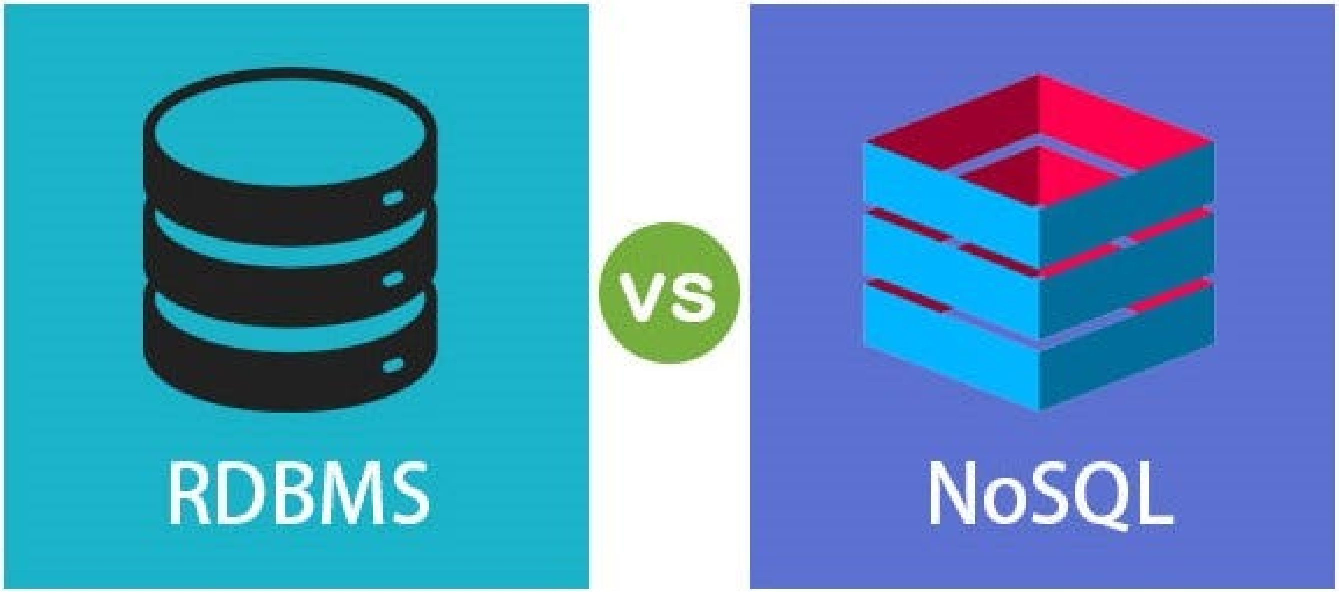
		Reality	
		True	False
Measured or Perceived	True	Correct ☺	Type 1 error False Positive
	False	Type 2 error False Negative	Correct ☺

Question - 10

How do you decide between RDBMS and NoSQL for a project?

- RDBMS (e.g., MySQL, PostgreSQL) – Best for structured data, complex relationships, and transactional consistency.

- NoSQL (e.g., MongoDB, Cassandra) – Ideal for semi-structured or evolving data with scalability needs.



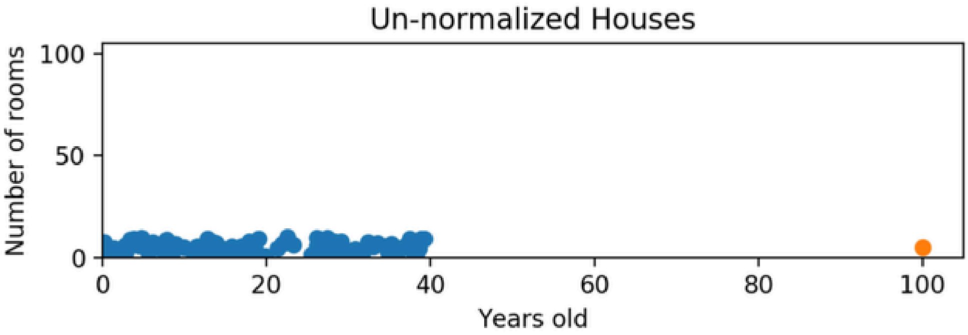
Question - 11

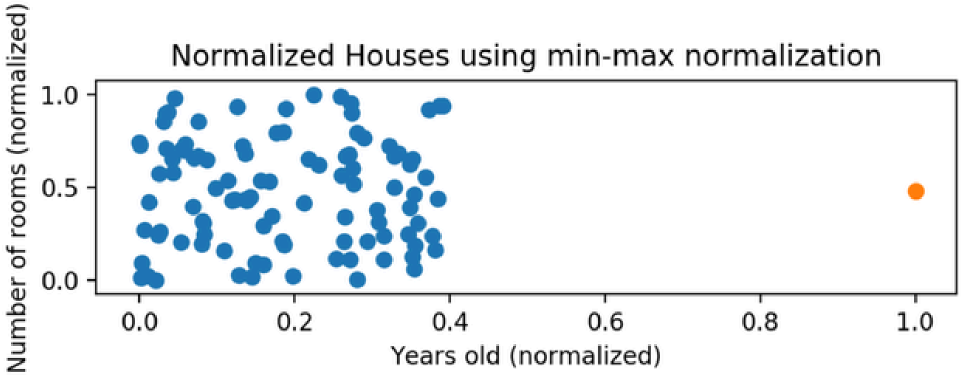
What is data normalization in databases?

- Data normalization reduces redundancy and improves data integrity.

- It involves breaking large tables into smaller ones and establishing relationships using foreign keys.

- Normalization improves database efficiency and ensures consistency.



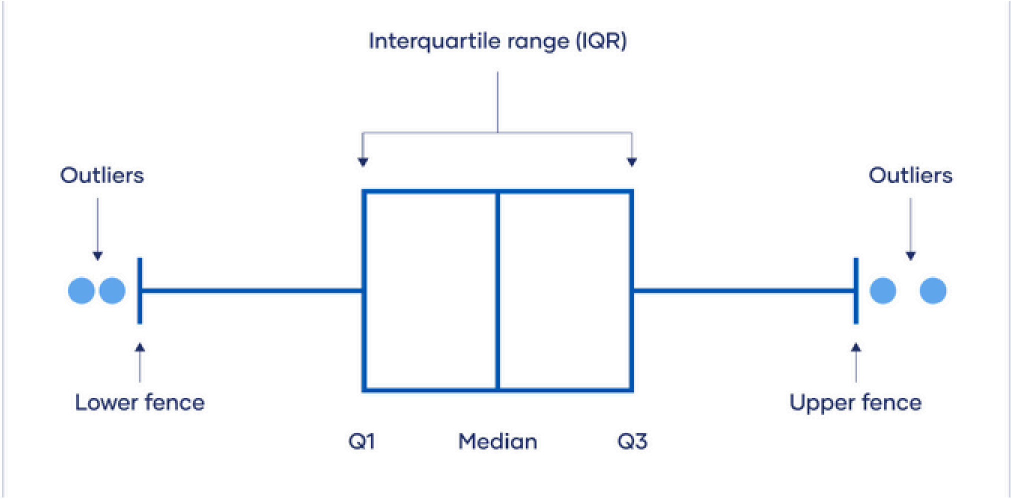


Question - 12

How do you detect and handle outliers in a dataset?

- Detect Outliers:
  - Use visual methods like box plots and scatter plots.
  - Use statistical methods like the IQR rule or Z-score.

- Handle Outliers:
  - Remove – If due to errors or irrelevance.
  - Transform – Apply log transformations.
  - Cap/Impute – Replace outliers with median or reasonable limits.



Nitya CloudTech

Dream.Achieve.Succeed

FOR CAREER GUIDANCE, CHECK OUT OUR PAGE

www.nityacloudtech.com

Follow Us on Linkedin:

Aditya Chandak

Nitya CloudTech

Dream.Achieve.Succeed