re that stores
ta. A faster
ity and efficiency

es and triggers

n a query
script that

natically when

E L T.
transform before

ansform inside

cessing and

e of data at

tinuously in

a list types

tables
ght join, full

Bright Light Data Analytics
Research Assignment
Due date: 20 October 2025

Section A: Database fundermentals

1. what are the main type of databases?
Relational database, NoSQL databases, object-oriented databases, time-series database

2. What is Relational Database Management System (RDBMS):
is a program used to create, update, and manage relational database

3. what is a primary key and a foreign key in a database
A primary key uniquely identifies a record in a table. a foreign key is a field that is used to linkee one table to the primary key of another table

4. what is database normalization and why is it important.
is the process of organizing data to reduce redundancy, improve data integrity and make it more efficient

5. How is database schema?
A schema is the blueprint or structure of a database that defines it's tables

fields and relationships

6. Differentiate between structured, semi structured and unstructured.

- Structured: organized into rows and columns (eg SQL tables)
- Semi-structured: has some structure but not fixed
- unstructured: No predefined format (eg images)

7. Difference between a fact table and a Dimension table

fact table: Stores numerical measures
Dimension table: store descriptive information (eg product, date)

8 what is a data model and why is it important?
A data model defines how data is structured and related. it guides database design and ensures data is consistent.

9. Difference between a database, a data warehouse and a data lake.
- Database: Store current operation data.
- Data warehous- Store historical, structured data for analysis
- Data lake: Stores all types of raw data- structured and unstructured.

10 what is dat
it differ from

A data mart
of a data war
specific team

section B: SQ

11 what is a qu
SQL the mos
A query lan
from a datab
because it i
supported by

12 what are ind
they impr
indexes a
searching
much data

13 what are
what are
A transac
run as
A - Atomic
C - Consis
I - Isola
other
D - Durab

14: what is

nuctured, semi
ured.

s rows and
SQL tables)
ne structure

d format eg image

table and

measures
iptive
t, date)

hy is it

a is structure
tabase
nsistent.

se, a data

n data.
, structured

data-

10 what is data mart, and how does it differ from a data warehouse?

A data mart is a smaller, focused version of a data warehouse, created for a specific team or department.

section B: SQL and data processing.

11. what is a query language, and why is SQL the most commonly used?
A query language is used to ask questions from a database. SQL is the most common because it is standardized, powerful and supported by all major database.

12 what are indexes in databases and how do they improve performance?
Indexes are structures that speed up searching and filtering by reducing how much data the database scans

13 what are transactions in database and what are the ACID properties?
A transaction is a group of operations that run as one:
A - Atomicity All or nothing
C - Consistency: Must follow rules
I - Isolation: Transaction don't affect each other
D - Durability: changers are permanent

14: what is a database engine, and how does

it impact performances?
The engine is the core software that stores reads, writes and manages data. A faster engine improves speed, reliability and efficiency

b) What are views, Store procedures and triggers in SQL!
view: A virtual table based on a query
Store procedure: A saved sql script that can be run anytime
trigger: Code that runs automatically when an event happens

16. Difference between ETL and ELT.
ETL: Exact. transform - Load (transform before loading)
ELT: Exact - load - transform (transform inside Data warehouse

17. Difference between batch processing and stream processing.
Batch: processing large chunks of data at Scheduled times
. Stream: processes data continuously in real time

18. Explain what a join is in SQL a list types of joins

A join combines rows from 2 tables
types: INNER Join, Left join, Right join, Full

outer join, cross join

19. what is referential integrity and why is is important?
it ensures relationship between tables remain valid preventing orphan records and maintaing accuracy

20 How does data redundancy affect performance and storage?
Redundant data uses more storage, slow queries, and increases inconsistency

Section C: Data Management and Analytics Concepts

21. How does cloud database management differ from on-premise database.
Cloud databases are hosted online and managed by providers, while on-premise databases run on local servers and require interal maintence

22. what is data governance, and why is it important?
Data integrity means data is accurate and reliable. it is maintained through validation, constraints, backups and access control

23. what is data integrity and how can it be mainted.

Data integrity means data is accurate and reliable it is maintained through validation, constraints, backups and access control.

24. What is data quality and why it is critical?
Data quality means data is accurate and reliable it is maintained through validation, contraints, backups and access control.

25. Explain the role of a Data analyst.
A Data Analyst Collect, Cleans, organized and analyzes to give insights to support business decisions

26. What are the key responsibilities of a database administrator (DBA)?
installing database, managing users, performing backups, monitoring performance, securing data and optimizing queries

27. what are the main steps involved in designing a data pipline.
Requirement - Data collection - processing - transformation - storage - Monitoring.

28. what ~~the~~ are som commen challenges in managing large scale database?
scalability - performance slowdown, high

storage, data security, backup, and ensuring data quality.

29 popular database platforms and their use cases.

- MySQL - web apps
- Snowflake - analytics and cloud data warehousing
- postgreSQL : advanced relational workload.
oracle - enterprise level systems

30 main data storage formats used in analytics.

~~SS~~

CSV, JSON, Avro