BIG DATA
Assignment-2

SUBMITTED BY: **SUBMITTED TO:**

AYUSH KUMAR JHA   SHELLY MA'AM
500086400
B.C.A–IOT (2020–2023)

Q2.Discuss the concept of Polyglot Persistence with the help of an example .(5 marks)

Polyglot persistence is an enterprise storage term used to describe choosing different data storage/data stores technologies to support the various data types and their storage needs. Polyglot persistence is essentially the idea that an application can use more than one core database (DB)/storage technology.

Polyglot persistence suggests that database engineers/architects first figure out how they want to manipulate data and then choose the database technology that best fits their needs. This approach is used to solve data storage efficiency issues, simplify operations and eliminate fragmentation.

How to use polyglot persistence

Unfortunately, polyglot persistence cannot be downloaded; it must be designed for the specific data architecture used by an enterprise. However, it has the flexibility to be used with SQL, NoSQL or hybrid database systems.

Various factors should be taken into consideration when deciding to move to a polyglot persistence storage system. For example, if big data is used in the promotion of a business or product, then changes will have to be made to the system to support this practice. The most ideal environment is one where only one persistence technology exists; however, this is not suitable for big data problem solving.

Data storage types

Polyglot persistence uses data storage types such as:

- External hard drive
- Network-attached storage
- Cloud storage
- Solid-state drive
- USB flash drive

Q3.Highlight the reasons for preferring NoSQL database for Big Data? Elaborate the various types of NoSQL database. (5 marks)

NoSQL databases were created in response to the limitations of traditional relational database technology. When compared to relational databases, NoSQL databases are often more scalable and provide superior performance. In addition, the flexibility and ease of use of their data models can speed development in comparison to the relational model, especially in the cloud computing environment.

Each specific type of NoSQL database has different strengths, but all share fundamental characteristics that allow them to:

Handle large volumes of data at high speed with a scale-out architecture

Store unstructured, semi-structured, or structured data

Enable easy updates to schemas and fields

Be developer-friendly

Take full advantage of the cloud to deliver zero downtime

Types of NoSQL databases

Document databases store data in documents similar to JSON (JavaScript Object Notation) objects. Each document contains pairs of fields and values. The values can typically be a variety of types including things like strings, numbers, booleans, arrays, or objects.

Key-value databases are a simpler type of database where each item contains keys and values.

Wide-column stores store data in tables, rows, and dynamic columns.

Graph databases store data in nodes and edges. Nodes typically store information about people, places, and things, while edges store information about the relationships between the nodes.

Q4. Illustrate the various components of Data Lake.(5 marks)

FIVE KEY COMPONENTS OF A DATA LAKE ARCHITECTURE

1.   Data ingestion

A highly scalable ingestion-layer system that extracts data from various sources, such as websites, mobile apps, social media, IoT devices, and existing Data Management systems, is required. It should be flexible to run in batch, one-time, or real-time modes, and it should support all types of data along with new data sources.

2.   Data Storage

A highly scalable data storage system should be able to store and process raw data and support encryption and compression while remaining cost-effective.

3.   Data Security

Regardless of the type of data processed, data lakes should be highly secure from the use of multi-factor authentication, authorization, role-based access, data protection, etc.

4.   Data Analytics

After data is ingested, it should be quickly and efficiently analyzed using data analytics and machine learning tools to derive valuable insights and move vetted data into a data warehouse.

5.   Data Governance

The entire process of data ingestion, preparation, cataloging, integration, and query acceleration should be streamlined to produce enterprise-level Data Quality. It is also important to track the changes to key data elements for a data audit.