Name – Aditya Somani Roll No. – T1851061 PRN No. : 71901204L

Assignment 1

Aim: Study and design a database with suitable example using following database systems: SQL, Redis, Hbase, MongoDB and Neo4J

1. Relational Database: SQL

SQL is Structured Query Language, which is a computer language for storing, manipulating and retrieving data stored in a relational database.

SQL is the standard language for Relational Database System. All the Relational Database Management Systems (RDMS) like MySQL, MS Access, Oracle, Sybase, Informix, Postgres and SQL Server use SQL as their standard database language. Also, they are using different dialects, such as −

- MS SQL Server usingT-SQL,
- Oracle usingPL/SQL,
- MS Access version of SQL is called JET SQL (native format)etc.

Why SQL?

SQL is widely popular because it offers the following advantages −

- Allows users to access data in the relational database managementsystems.
- Allows users to describe thedata.
- Allows users to define the data in a database and manipulate thatdata.
- Allows to embed within other languages using SQL modules, libraries &pre-compilers.
- Allows users to create and drop databases andtables.
- Allows users to create view, stored procedure, functions in adatabase.
- Allows users to set permissions on tables, procedures andviews.

A Brief History of SQL

- **1970** − Dr. Edgar F. "Ted" Codd of IBM is known as the father of relationaldatabases. He described a relational model fordatabases.
- **1974** − Structured Query Languageappeared.
- **1978** − IBM worked to develop Codd's ideas and released a product named System/R.
- **1986** − IBM developed the first prototype of relational database and standardized by ANSI. The first relational database was released by Relational Software which later came to be known asOracle.

SQL Commands

The standard SQL commands to interact with relational databases are CREATE, SELECT, INSERT, UPDATE, DELETE and DROP. These commands can be classified into the following groups based on their nature −

DDL - Data Definition Language

Sr.No.	Command &Description
1	**CREATE** Creates a new table, a view of a table, or other object in the database.
2	**ALTER** Modifies an existing database object, such as a table.
3	**DROP** Deletes an entire table, a view of a table or other objects in the database.

DML - Data Manipulation Language

Sr.No.	Command & Description
1	**SELECT** Retrieves certain records from one or more tables.
2	**INSERT** Creates a record.
3	**UPDATE** Modifies records.
4	**DELETE** Deletes records.

DCL - Data Control Language

Sr.No.	Command & Description
1	**GRANT** Gives a privilege to user.
2	**REVOKE** Takes back privileges granted from user.

Characteristics of SQL

1. SQL is an ANSI and ISO standard computer language for creating andmanipulating databases.
2. SQL allows the user to create, update, delete, and retrieve data from adatabase.
3. SQL is very simple and easy tolearn.
4. SQL works with database programs like DB2, Oracle, MS Access, Sybase, MS SQL Severetc.

Advantages of SQL:

1. High Speed: SQL Queries can be used to retrieve large amounts of records froma database quickly andefficiently.
2. Well Defined Standards Exist: SQL databases use long-established standard, which is being adopted by ANSI & ISO. Non-SQL databases do not adhere to any clearstandard.
3. No Coding Required: Using standard SQL it is easier to manage databasesystems without having to write substantial amount ofcode.
4. Emergence of ORDBMS: Previously SQL databases were synonymous with relational database. With the emergence of Object Oriented DBMS, object storage capabilities are extended to relationaldatabases.

Disadvantages of SQL:

1. Difficulty in Interfacing: Interfacing an SQL database is more complex than adding afew lines ofcode.
2. More Features Implemented in Proprietary way: Although SQL databases conform to ANSI & ISO standards, some databases go for proprietary extensions to standard SQL to ensure vendorlock-in.

2. Key Value:Redis

Redis is an open source, advanced key-value store and an apt solution for building high performance, scalable web applications. Redis is an in-memory database open-source software project implementing a networked, in-memory key-value store with optional durability. Redis supports different kinds of abstract data structures, such as strings, lists, maps, sets, sorted sets, hyperloglogs, bitmaps and spatial indexes. The project is mainly developed by Salvatore Sanfilippo and is currently sponsored by Redis Labs.

Redis has three main peculiarities that sets it apart.

- Redis holds its database entirely in the memory, using the disk only forpersistence.
- Redis has a relatively rich set of data types when compared to many key-value data stores.
- Redis can replicate data to any number ofslaves.

History

The name Redis means REmote DIctionary Server. Salvatore Sanfilippo, the original developer of Redis, was hired by VMware in March, 2010. In May, 2013, Redis was sponsored by Pivotal Software (a VMware spin-off). In June 2015, development became sponsored by Redis Labs.

According to monthly rankings by DB-Engines.com, Redis is often ranked the most popular key-value database. Redis has also been ranked the #4 NoSQL database in user satisfaction and market presence based on user reviews, the most popular NoSQL database in containers, and the #1 NoSQL database among Top 50 Developer Tools & Services.

Redis Data Structures

Redis supports these data structures:

- Binary-safe strings - lists or collections of string elements are sorted according tothe order ofinsertion.
- Sets and sorted sets - collections of unique, unsorted string elements and collections in which every string element is associated with a floating number value called ascore.
- Hashes - maps composed of fields associated with values. Both the field and the valueare strings.
- Bit arrays (bitmaps) - use special commands to handle string values like an array ofbits.
- HyperLogLogs - a data structure that can estimate the number of items in aset.
- Geospatial indexes - data that is stored as coordinatepairs.

Redis Advantages

Following are certain advantages of Redis.

1. Exceptionally fast − Redis is very fast and can perform about 110000 SETs persecond, about 81000 GETs persecond.
2. Supports rich data types − Redis natively supports most of the datatypes that developers already know such as list, set, sorted set, and hashes. This makes it easy to solve a variety of problems as we know which problem can be handled better by which datatype.
3. Operations are atomic − All Redis operations are atomic, which ensures that if twoclients concurrently access, Redis server will receive the updatedvalue.
4. Multi-utility tool − Redis is a multi-utility tool and can be used in a number of use cases such as caching, messaging-queues (Redis natively supports Publish/Subscribe), any short-lived data in your application, such as web application sessions, web page hit counts,etc.

Redis Versus Other Key-value Stores

- Redis is a different evolution path in the key-value DBs, where values can contain more complex data types, with atomic operations defined on those datatypes.
- Redis is an in-memory database but persistent on disk database, hence it represents a different trade off where very high write and read speed is achieved with the limitation of data sets that can't be larger than thememory.
- Another advantage of in-memory databases is that the memory representation of complex data structures is much simpler to manipulate compared to the same data structure on disk. Thus, Redis can do a lot with little internalcomplexity.

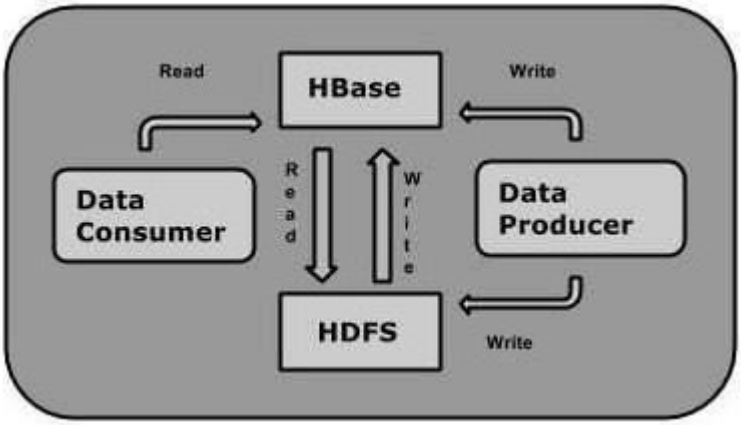
3. Column Oriented:Hbase

HBase is a distributed column-oriented database built on top of the Hadoop file system. It is an open-source project and is horizontally scalable.

HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data. It leverages the fault tolerance provided by the Hadoop File System (HDFS).

It is a part of the Hadoop ecosystem that provides random real-time read/write access to data in the Hadoop File System.

One can store the data in HDFS either directly or through HBase. Data consumer reads/accesses the data in HDFS randomly using HBase. HBase sits on top of the Hadoop File System and provides read and write access.



Storage Mechanism in HBase

HBase is a **column-oriented database** and the tables in it are sorted by row. The table schema defines only column families, which are the key value pairs. A table have multiple column families and each column family can have any number of columns. Subsequent column values are stored contiguously on the disk. Each cell value of the table has a timestamp. In short, in an HBase:

- Table is a collection ofrows.
- Row is a collection of columnfamilies.
- Column family is a collection ofcolumns.
- Column is a collection of key value pairs.Given

below is an example schema of table inHBase.

Rowid	Column Family			Column Family			Column Family			Column Family		
	col1	col2	col3	col1	col2	col3	col1	col2	col3	col1	col2	col3
1												

2												
3												

Column Oriented and Row Oriented

Column-oriented databases are those that store data tables as sections of columns of data, rather than as rows of data. Shortly, they will have column families.

Row-Oriented Database	Column-Oriented Database
It is suitable for Online Transaction Process (OLTP).	It is suitable for Online Analytical Processing (OLAP).
Such databases are designed for small number of rows and columns.	Column-oriented databases are designed for huge tables.

Features of HBase

- HBase is linearlyscalable.
- It has automatic failuresupport.
- It provides consistent read andwrites.
- It integrates with Hadoop, both as a source and adestination.
- It has easy java API forclient.
- It provides data replication acrossclusters.

Where to Use HBase

- Apache HBase is used to have random, real-time read/write access to BigData.
- It hosts very large tables on top of clusters of commodityhardware.
- Apache HBase is a non-relational database modeled after Google's Bigtable. Bigtable acts up on Google File System, likewise Apache HBase works on top of Hadoop andHDFS.

Applications of HBase

- It is used whenever there is a need to write heavyapplications.
- HBase is used whenever we need to provide fast random access to availabledata.
- Companies such as Facebook, Twitter, Yahoo, and Adobe use HBaseinternally.

HBase History

Year	Event
Nov 2006	Google released the paper on BigTable.

Feb 2007	Initial HBase prototype was created as a Hadoop contribution.
Oct 2007	The first usable HBase along with Hadoop 0.15.0 was released.
Jan 2008	HBase became the sub project of Hadoop.
Oct 2008	HBase 0.18.1 was released.
Jan 2009	HBase 0.19.0 was released.
Sept 2009	HBase 0.20.0 was released.
May 2010	HBase became Apache top-level project.

4. Document Oriented:MongoDB

MongoDB is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document.

Database

Database is a physical container for collections. Each database gets its own set of files on the file system. A single MongoDB server typically has multiple databases.

Collection

Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database. Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or relatedpurpose.

Document

A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data. The following table shows the relationship of RDBMS terminology with MongoDB.

RDBMS	MongoDB
Database	Database
Table	Collection

Tuple/Row	Document
column	Field
Table Join	Embedded Documents
Primary Key	Primary Key (Default key _id provided by mongodb itself)
Database Server and Client	
Mysqld/Oracle	mongod
mysql/sqlplus	mongo

MongoDB Features

- **General purpose database**, almost as fast as the key:value NoSQLtype.
- **Highavailability**.
- **Scalability** (from a standalone server to distributed architectures of huge clusters). This allows us to shard our database transparently across all our shards. This increases the performance of our dataprocessing.
- **Aggregation**: batch data processing and aggregate calculations using nativeMongoDB operations.
- **Load Balancing**: automatic data movement across different shards for load balancing. The balancer decides when to migrate the data and the destination Shard, so they are evenly distributed among all servers in the cluster. Each shard stores the data for a selected range of our collection according to a partitionkey.
- **Native Replication**: syncing data across all the servers at the replicaset.
- **Security**: authentication, authorization,etc.
- **Advanced usersmanagement**.
- **Automatic failover**: automatic election of a new primary when it has gonedown.

Advantages of MongoDB

- Load Balancing and Sharding: If you have huge amounts of data or want to distribute the traffic of your database among various machines to balance the load, MongoDB carries a number of advantages over traditional databases. Moreover, Sharding, which is MongoDB's unique approach for fulfilling the requirements of growth in data, makes use of horizontal scaling and allows you to multiple machines for the purpose of supporting the growth ofdata.
- Flexibility: It doesn't need data structures that are unified in nature across all the objects in use. This makes using MongoDB much simpler than RDBMS. On the other hand, data consistency is very important at times and is generally a very good thing, thus, it is advised that you should make use of unified datastructure.

- Speed: As all the data is generally at a single location, MongoDB are extremely quick. However, this only stands true when the data you are working on is actually a document. If you are working on a data that emulates relational model, your code will be required to carry out multiple independent queries for the purpose of retrieving single document and this will make it slower than aRDBMS.

Disadvantages of MongoDB

- Usage of Memory: As MongoDB stores the key name along with every document, it naturally consumes more memory. Moreover, as slow queries and joins are not possible because join within the code are required to be performed, you often have to deal with duplicatedata.
- No Joins: Like a relational database, joins are simply not possible in MongoDB. Thus, if you ever need the functionality of joins, you will be required to create multiple queries, which you'll have to join manually in thecode.
- Still Under Development: While SQL was developed in the 1980s, MongoDB entered into the market in 2009. As a result, MongoDB is not that extensively documented or tested and also lacks the availability of support andexperts.

5. Graph Database:Neo4J

Neo4j is the world's leading open source Graph Database which is developed using Java technology. It is highly scalable and schema free (NoSQL).

What is a Graph Database?

A graph is a pictorial representation of a set of objects where some pairs of objects are connected by links. It is composed of two elements - nodes (vertices) and relationships(edges).

Graph database is a database used to model the data in the form of graph. In here, the nodes of a graph depict the entities while the relationships depict the association of thesenodes.

Popular Graph Databases

Neo4j is a popular Graph Database. Other Graph Databases are Oracle NoSQL Database, OrientDB, HypherGraphDB, GraphBase, InfiniteGraph, and AllegroGraph.

Why Graph Databases?

Nowadays, most of the data exists in the form of the relationship between different objects and more often, the relationship between the data is more valuable than the data itself.

Relational databases store highly structured data which have several records storing the same type of data so they can be used to store structured data and, they do not store the relationships between the data.

Unlike other databases, graph databases store relationships and connections as first-class entities.

The data model for graph databases is simpler compared to other databases and, they can be used with OLTP systems. They provide features like transactional integrity and operational availability.

Advantages of Neo4j

Following are the advantages of Neo4j.

- **Flexible data model** − Neo4j provides a flexible simple and yet powerful datamodel, which can be easily changed according to the applications andindustries.
- **Real-time insights** − Neo4j provides results based on real-timedata.
- **High availability** − Neo4j is highly available for large enterprise real-time applications with transactionalguarantees.
- **Connected and semi structures data** − Using Neo4j, you can easily representconnected and semi-structureddata.
- **Easy retrieval** − Using Neo4j, you can not only represent but also easilyretrieve (traverse/navigate) connected data faster when compared to otherdatabases.
- **Cypher query language** − Neo4j provides a declarative query language to represent the graph visually, using an ascii-art syntax. The commands of this language are in human readable format and very easy tolearn.
- **No joins** − Using Neo4j, it does NOT require complex joins to retrieve connected/related data as it is very easy to retrieve its adjacent node or relationship details without joins or indexes.

Features of Neo4j

Following are the notable features of Neo4j −

- **Data model (flexible schema)** − Neo4j follows a data model named native property graph model. Here, the graph contains nodes (entities) and these nodes are connected with each other (depicted by relationships). Nodes and relationships store data in key-value pairs known as properties.
  In Neo4j, there is no need to follow a fixed schema. You can add or remove properties as per requirement. It also provides schema constraints.
- **ACID properties** − Neo4j supports full ACID (Atomicity, Consistency, Isolation,and Durability) rules.
- **Scalability and reliability** − You can scale the database by increasing the number of reads/writes, and the volume without effecting the query processing speed and data integrity. Neo4j also provides support for replication for data safety and reliability.
  Cypher Query Language − Neo4j provides a powerful declarative query language known as Cypher. It uses ASCII-art for depicting graphs. Cypher is easy to learn and can be used to create and retrieve relations between data without using the complex queries likeJoins.
- **Built-in web application** − Neo4j provides a built-in Neo4j Browser web application. Using this, you can create and query your graph data.
- **Drivers** − Neo4j can work with −
  - REST API to work with programming languages such as Java, Spring, Scalaetc.
  - Java Script to work with UI MVC frameworks such as NodeJS.

- It supports two kinds of Java API: Cypher API and Native Java API to develop Java applications. In addition to these, you can also work with other databases such as MongoDB, Cassandra,etc.

- **Indexing** − Neo4j supports Indexes by using ApacheLucence.

Conclusion:-

We have studied a database with suitable example using following database systems: SQL, Redis, Hbase, MongoDB and Neo4J.