**What is NoSQL and Why NoSQL is used**

Q1. Let’s consider an RDBMS table that has the following schema:

|  |  |
| --- | --- |
| **Column Name** | **Data Type** |
| Roll No | NUMBER (Used to represent numbers) |
| NAME | VARCHAR2 (Same as string in Java) |
| Age | NUMBER (Used to represent numbers) |
| Year | NUMBER (Used to represent numbers) |
| Department | VARCHAR2 (Same as string in Java) |
| Sex | VARCHAR2 (Same as string in Java) |

In SQL, the INSERT command is used for inserting data into RDBMS tables. Consider the syntax of the INSERT command given below and select the correct Insert command that will execute without giving any errors.

INSERT INTO <TABLE\_NAME> VALUES (‘STRING’, NUMBER ….).

Anything enclosed within single quotes is considered as VARCHAR2. Let’s assume that there are no implicit type conversions when inserting data into a relational database. Select the correct INSERT statement.

1. INSERT INTO <TABLE\_NAME> VALUES (‘20’, ‘Siddharth’, 28, 2, ‘CSE’, ‘M’ )
2. **INSERT INTO <TABLE\_NAME> VALUES (20, ‘Siddharth’, 28, 2, ‘3’, ‘M’ )**
3. INSERT INTO <TABLE\_NAME> VALUES (20, ‘Siddharth’, 28, 2, ‘CSE’, ‘M’, 50.5 )
4. INSERT INTO <TABLE\_NAME> VALUES (‘20’, ‘Siddharth’, ‘28’, ‘2’, ‘3’, ‘M’ )

Ans b

* Option a is incorrect. The values to be inserted into the table do not adhere to the schema. The first field is ‘Roll No’ which is of type ‘Number’, whereas, in this option, we are treating the field as a String as it is enclosed within single quotes.
* **Option b is correct. All the values adhere to the schema of the table.**
* Option c is incorrect. The values to be inserted into the table do not adhere to the schema. This command is trying to insert seven values, whereas the table has only six columns.
* Option d is incorrect. The values to be inserted into the table do not adhere to the schema. This command is trying to insert six values, all enclosed within single quotes. Numeric data values should not be enclosed within quotes.

Q2. Which data format can be stored and processed efficiently using RDBMSs?

1. Image
2. Video
3. **Data arranged in rows and columns**
4. Log files

Ans c.

* Option a is incorrect. Image files contain unstructured data, and RDBMS is not suited for storing images and processing them.
* Option b is incorrect. Video files contain unstructured data, and RDBMS is not suited for storing videos and processing them.
* **Option c is correct. Data arranged in rows and columns represent structure, and RDBMS is well-suited for storing data in a tabular format and processing them.**
* Option d is incorrect. Text files contain unstructured data, and RDBMS is not suited for storing text files in their original format and then processing them.

Q3) State whether the following statement is true or false:  
“NoSQL was designed because SQL systems were incompetent in processing transactional data.”

1. True
2. **False**

The answer is False because SQL systems are designed to process transactional data. But, NoSQL databases were designed to accommodate high volumes of unstructured data in distributed storage.

Q4) Select the most suitable use case where NoSQL should be used:

1. When the data security is of the utmost priority
2. When data is arranged in rows and columns
3. **When a data storage in the range of TB is required**
4. When the retrieval speed of data is not critical

Ans c

* Option a is incorrect. NoSQL databases are not used specifically to provide high security to the stored data.
* Option b is incorrect. NoSQL databases are not specifically designed to store structured data.
* **Option c is correct. NoSQL databases are designed to store high volumes of data in the range of TB distributed across clusters of machines.**
* Option d is incorrect. NoSQL databases are designed to provide quick access to data.

Q5) Select the option that is not valid:

1. SQL databases make use of a fixed schema, whereas NoSQL datastores do not make use of a fixed schema
2. **The query language used for querying NoSQL datastores adds a layer of abstraction on top of MapReduce. They convert the query into a series of MapReduce jobs.**
3. NoSQL refers to “Not Only SQL”
4. MapReduce scans the entire dataset for any query, whereas NoSQL has the ability to fetch a single record without scanning the entire dataset

* Option a is incorrect. It is a valid statement. SQL databases indeed make use of a fixed schema. On the other hand, NoSQL datastores make use of dynamic schemas, which can be modified on the fly.
* **Option b is correct. This statement is False. The queries are not converted to MapReduce because NoSQL datastores are designed to perform low latency data lookups.**
* Option c is incorrect. It is a valid statement. NoSQL datastores are indeed called as “Not only SQL”.
* Option d is incorrect. It is a valid statement. MapReduce scans the entire dataset. NoSQL datastores are designed to perform low latency data lookups.

Q6) What are the benefits of having a dynamic schema?

**Video2: CAP Theorem**

Q1) Which of the following is not a reason why NoSQL has become a popular solution among some organisations?

1. Better scalability
2. Faster data lookup
3. Allows data to be stored across multiple nodes
4. **Ability to store consistent data eternally**

* Option a is incorrect. NoSQL datastores provide scalability by various mechanisms, such as horizontal scalability and sharding.
* Option b is incorrect. Some NoSQL datastores such as HBase provide low latency data lookup.
* Option c is incorrect. NoSQL datastores are designed to store data across multiple nodes.
* **Option d is correct. NoSQL datastores do not guarantee consistency all the time. In case of a network partition, some datastores may not return the recent value.**

Q2) Which of the following use cases will give priority to consistency over availability?

1. Prices of air tickets on a travel portal
2. Notifications on Facebook
3. **A customer’s bank account details that includes all latest transactions**
4. None of these

* Option a is incorrect. The prices of the air tickets on a travel portal may not be the latest at the time when we are checking. Sometimes, the prices are updated just before the payment. Here, availability is given a higher priority over consistency.
* Option b is incorrect. Delaying of notifications is completely fine. But it is not acceptable if facebook.com goes down because of the delay in notifications. Here, availability is given a higher priority over consistency.
* **Option c is correct. Bank account details are sensitive information, and they have to be consistent and correct. So, if the latest accurate information is not available, then it’s better to declare an outage than showing wrong information to the customer.**
* Option d is incorrect. Go through the use cases again. One of the remaining options is the correct answer.

Q3) Pick the statement that is true about the CAP theorem?

1. In CAP, ‘A’ stands for Accuracy
2. **If partition tolerance is not considered, then the datastore will be an RDBMS system**
3. Consistency guarantees that every node in the distributed system has an identical configuration
4. None of the above

* Option a is incorrect. In CAP, ‘A’ stands for availability.
* **Option b is correct. RDBMS systems are not used to process data using distributed computing. Hence, partition tolerance is out of context for RDBMS systems.**
* Option c is incorrect. Consistency guarantees that every node in the distributed system returns the same, most successful, recent write.
* Option d is incorrect. Go through the use cases again; one of the remaining options is the correct answer.

Q4) There is a need to store the transactional data generated by a bank’s ATM. The data is to be stored in a tabular format. According to the CAP theorem, which type of datastore should be used for this dataset?

1. **CA**
2. CP
3. AP
4. None of these

* **Option a is correct. An RDBMS system will be required to store the transactional data in a structured format. As per the CAP theorem, RDBMS systems support consistency and availability together.**
* Option b is incorrect. An RDBMS system will be required to store the transactional data in a structured format. As per the CAP theorem, RDBMS systems do not support consistency and partition tolerance together.
* Option c is incorrect. An RDBMS system will be required to store the transactional data in a structured format. As per the CAP theorem, RDBMS systems do not support availability and partition tolerance together.
* Option d is incorrect. Go through the use cases again; one of the remaining options is the correct answer.

Q5) For a social networking website, there is a need to store data about users and their friends’ in a datastore. Let’s assume that the data would be stored in a distributed storage system. Here, due to a system outage, if the latest total count of the friends for a user is not available, the portal should still be available and reflect older stats. According to the CAP theorem, which type of datastore should be used?

1. CA
2. CP
3. **AP**
4. None of these

* Option a is incorrect. This answer is incorrect as partition tolerance has not been considered.
* Option b is incorrect. This answer is incorrect as consistency is given priority over availability.
* **Option c is correct. This answer is correct because the system needs to be highly available in case of a partition tolerance requirement.**
* Option d is incorrect. Please go through the use cases again, one of the remaining options is the correct answer.

**Video3: Introduction to HBase.**

**Q1)** Which one of the following advantages is not provided by HBase because of the HDFS?

1. **Low latency retrieval of data**
2. Distributed storage
3. Horizontal scalability
4. Fault tolerance

* **This option is correct. HDFSs do not provide low latency retrieval of data. MapReduce is used for reading data directly from the HDFS in a sequential manner.**
* This option is incorrect. HDFSs provide distributed storage.
* This option is incorrect. HDFSs are horizontally scalable. The capacity of an HDFS can be increased by adding extra nodes to the cluster.
* This option is incorrect. HDFSs are fault tolerant and ensure that there is minimum data loss by means of data replication.

Q2) Pick the correct tabular representation of HBase from the following tables. (Here, RowID is the RowKey.)

|  |  |  |
| --- | --- | --- |
| **RowID** | **Field1** | **Field2** |
| **1** | A | 10 |
| **3** | B | 20 |
| **2** | C | 30 |
| **5** | D | 40 |
| **9** | E | 50 |

B.

|  |  |
| --- | --- |
| **Field1** | **Field2** |
| A | 10 |
| B | 20 |
| C | 30 |
| D | 40 |
| E | 50 |

C.

|  |  |  |
| --- | --- | --- |
| **RowID** | **Field1** | **Field2** |
| **1** | A | 10 |
| **2** | B | 20 |
| **3** | C | 30 |
| **5** | D | 40 |
| **9** | E | 50 |

D. None of the above

* Option ‘a’ is incorrect. The rows are not sorted according to the row IDs.
* Option ‘b’ is incorrect. The table does not have the row ID column, which is impossible in HBase.
* **Option ‘c’ is correct. The rows in the given table are sorted as per the RowKey, i.e. rowID.**
* Option ‘d’ is incorrect. Check the options again. One of the options is correct.

**Video 4. Data Models in HBase**

Q1) Pick the option which is not true with respect to column families?

1. Column families comprise more than one columns
2. **The number of columns present in a column family is fixed**
3. Column families allow the corresponding table to scale linearly
4. None of the above

* Option ‘a’ is incorrect. This statement is true. Column families are collections of columns in HBase tables.
* **Option ‘b’ is correct. Therefore, this statement is incorrect. The number of columns in a column family is not fixed. This gives flexibility to the HBase table.**
* Option ‘c’ is incorrect. This statement is true. Column families allow the tables to scale linearly. This feature ensures that HBase does not
* Option ‘d’ is incorrect. Check the options again. One of the options is correct.

Q2) Pick a valid difference between HBase tables and RDBMS tables?

1. In HBase, rows are sorted by RowKeys, whereas in RDBMS, rows are not sorted by any keys.
2. HBase tables have column families, which comprise multiple columns, whereas RDBMS tables do not have the concept of column families.
3. **HBase tables follow a strict schema, whereas RDBMS tables do not follow any strict schema and are flexible.**
4. HBase tables are known as NoSQL databases, whereas RDBMS tables are known as SQL databases.

* Option ‘a’ is incorrect. In HBase, rows are sorted by RowKeys, whereas in RDBMS, rows are not sorted in any particular order.
* Option ‘b’ is incorrect. HBase tables comprise column families, which are collections of columns. On the other hand, RDBMS tables do not have any feature called column families.
* **Option ‘c’ is correct. HBase tables do not follow a strict schema, whereas RDBMS tables follow a well-defined schema.**
* Option ‘d’ is incorrect. HBase tables are known as NoSQL databases, whereas RDBMS tables are known as SQL databases.

Let’s assume the following syntax for retrieving the data present in a single cell of an HBase table:

get ‘<table name>’, ‘rowkey’, ‘<column name>’, ‘version’

The column name is represented as ‘column family: column name’.

In the ‘get’ command, table name, row key, and column name are the mandatory parameters. The version field is optional. If you don’t provide any version in the command then, by default, the latest value will be picked.

Let’s consider the sample table given below with the name ‘Employee’. This table can record a maximum of two versions for each column. Version ‘V1’ refers to the latest information. Here, the column families are ‘Personal Details’ and ‘Professional Details’. The columns in the column family ‘Personal Details’ are ‘Name’ and ‘Age’. The columns in the column family ‘Professional Details’ are ‘Company Name’ and ‘Designation’:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ROWID** | **Personal Details** | | **Professional Details** | | | |
|  | **Name** | **Age** | **Company Name** | | **Designation** | |
|  | **V1** | **V1** | **V1** | **V2** | **V1** | **V2** |
| **1** | Siddharth | 27 | Wipro | Infosys | Software Engineer | Systems Engineer |
| **2** | Sandeep | 27 | ITC Infotech | Accenture | Senior Business Analyst | Business Analyst |
| **3** | Jayadeep | 22 | Symantec |  | Intern |  |
| **4** | Onkar | 28 | Wipro | ITC Infotech | Senior Project Manager | Project Manager |

Note: The blank cells indicate that there are no entries for those cells. It means that for ROWID 3, column ‘Professional Details: Company Name’ and version ‘V2’ have no recorded value yet.

Answer the questions based on the information provided above:

Q3)

What would be the correct syntax for the ‘get’ command to retrieve the data in version ‘V2’ of ‘company name’ and ROWID 2?

1. get ‘Employee’, ‘2’, ‘Company Name’, ‘V2’
2. get ‘Employee’, ‘Company Name’, ‘2’, ‘V2’
3. **get ‘Employee’, ‘2’, ‘Professional Details:Company Name’, ‘V2’**
4. get ‘Employee’, ‘2’, ‘Professional Details:Company Name’, ‘V1’

* Option ‘a’ is incorrect. The column name is not preceded by the qualifier name.
* Option ‘b’ is incorrect. The column name is not preceded by the qualifier name, and the RowKey is incorrectly written after the column name. Refer to the syntax for more clarity.
* **Option ‘c’ is correct. The command is in accordance with the syntax mentioned above.**
* Option ‘d’ is incorrect. In the question, you are asked to retrieve the data for V2 and not V1.

Q4)

What would be the output of the command mentioned below?

get ‘Employee’, ‘4’, ‘Professional Details:Designation’

1. Project Manager
2. **Senior Project Manager**
3. Command will not run and flag an error
4. Senior Business Analyst

* Option ‘a’ is incorrect. If the version is not mentioned in the query then, by default, the latest value is picked. In this table, the latest value is denoted by V1.
* **Option ‘b’ is correct. If the version is not mentioned in the query then, by default, the latest value is picked. In this table, the latest value is denoted by V1.**
* Option ‘c’ is incorrect. The command is correct. If the version is not mentioned in the query then, by default, the latest value is picked. In this table, the latest value is denoted by V1.
* Option ‘d’ is incorrect. ‘Senior Business Analyst’ is the latest designation for Row ID 2.

Q5) What will be the output if the command mentioned below is executed on the table given earlier?

get ‘Employee’, ‘3’

1. Jayadeep
2. 22
3. **Command will not run and flag an error**
4. None of these

* Option ‘a’ is incorrect. This is not the correct answer. Read the entire problem description again to answer this question correctly.
* Option ‘b’ is incorrect. This is not the correct answer. Read the entire problem description again to answer this question correctly.
* **Option ‘c’ is correct. This command will not run as the mandatory field, i.e. the column name followed by the column family is missing.**
* Option ‘d’ is incorrect. This is not the correct answer. Read the entire problem description again to answer this question correctly.

**Video no: 5 Common Operations performed on an HBase table**

Q1) Pick the command which is not used in HBase?

1. get
2. put
3. **insert**
4. scan

* Option ‘a’ is incorrect. ‘Get’ is a valid HBase command and is used for retrieving data from a single cell in an HBase table.
* Option ‘b’ is incorrect. ‘Put is a valid HBase command and is used for entering data into an HBase table.
* **Option ‘c’ is correct. Insert is an invalid command.**
* Option ‘d’ is incorrect. ‘Scan’ is a valid HBase command and is used for displaying all the records present in an HBase table.

Q2) Pick the correct command to create the table ‘Employee’ with the structure shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Personal Details** | | **Professional Details** | | | |
| **Name** | **Age** | **Company Name** | | **Designation** | |

1. create ‘Employee’, ‘Personal Details:Name’, ‘Personal Details:Age’, ‘Professional Details:Company Name’, ‘Professional Details:Designation’
2. **create ‘Employee’, ‘Personal Details’, ‘Professional Details’**
3. create ‘Employee’, ‘RowKey’, ‘Personal Details’, ‘Professional Details’
4. None of these

* Option ‘a’ is incorrect. While creating a table, it is not mandatory to specify the columns along with their column families.
* **Option ‘b’ is correct. While creating a table, only the column families are specified in the create command.**
* Option ‘c’ is incorrect. This command will run without giving any error, but an extra column family will be created with the name ‘RowKey’. As per the table structure provided above, the column family named ‘RowKey’ is not required.
* Option ‘d’ is incorrect. Check the options again. One of the options is correct.

Q3) What will be the output if the following command is executed after creating the ‘Employee’ table before inserting any data?

**scan ‘Employee’**

1. The command won’t run and will flag an error
2. Default values will be returned
3. **The command runs without flagging any error**
4. None of the above

* Option ‘a’ is incorrect. The command will run without flagging any error.
* Option ‘b’ is incorrect. HBase does not store default values in empty tables.
* **Option ‘c’ is correct. The command will run without flagging any error. However, zero rows will be returned as the table is empty.**
* Option ‘d’ is incorrect. Check the options again. One of the options is correct.

Q4) Which one of the following statements is NOT true about the ‘Put’ statement?

1. The Put command works similar to the put method of HashMap in Java
2. The Put command is used for adding data to HBase tables
3. The Put command can be used for updating the existing values
4. **A single Put command can add multiple cell values**

* Option ‘a’ is incorrect. This is a valid statement. The Put command works similar to the put method in HashMap. In HashMap, the put method is used for adding a value to a given key. Similarly, the put command is used for adding a single data value to the given key. Here, the key is a combination of the RowKey and the column.
* Option ‘b’ is incorrect. This is a valid statement. The Put command is used for adding data to the HBase table.
* Option ‘c’ is incorrect. This is a valid statement. The Put command is used for updating the existing values in the HBase table
* **Option ‘d’ is correct. A single Put command cannot add multiple cell values.**

Q5) Which one of the following options is not a mandatory parameter in a delete command?

1. Table name
2. **Timestamp version**
3. Row key
4. Column Name

* Option ‘a’ is incorrect. In every delete command, the table name has to be there.
* **Option ‘b’ is correct. The version parameter in the delete command is optional. Every delete command must have the table name, row key, and column name.**
* Option ‘c’ is incorrect. In every delete command, the row key is mandatory.
* Option ‘d’ is incorrect. In every delete command, the column name is mandatory.

Q6) Let’s assume that the current state of the ‘Employee’ table is as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ROWID** | **Personal Details** | | **Professional Details** | | | |
|  | **Name** | **Age** | **Company Name** | | **Designation** | |
|  | **V1** | **V1** | **V1** | **V2** | **V1** | **V2** |
| **1** | Siddharth | 27 | Wipro | Infosys | Software Engineer | Systems Engineer |
| **2** | Sandeep | 27 | ITC Infotech | Accenture | Senior Business Analyst | Business Analyst |

Which one of the following commands can delete the V2 company name for the employee ‘Siddharth’?

1. delete ‘Employee’, ‘Siddharth’,’Professional Details:Company Name’, V2
2. delete ‘Employee’, ‘Siddharth’,Personal Details:Company Name’, V2
3. **delete ‘Employee’, ’1’, ‘Professional Details:Company Name’, V2**
4. delete ‘Employee’, ‘1’, ’Siddharth’, ‘Professional Details:Company Name’, V2

* Option ‘a’ is incorrect. The second parameter in the delete command has to be a row ID.
* Option ‘b’ is incorrect. The second parameter in the delete command has to be a row ID.
* **Option ‘c’ is correct. The command is syntactically correct. This will run and produce the desired result.**
* Option ‘d’ is incorrect. An additional field ‘Siddharth’ is included, which is wrong.

**Optional Project**

Create a table named ‘Mobile\_Subscribers’ with two column families, namely, ‘Personal\_Details’ and ‘Plan\_Details’. Maintain the VERSION counts for the column family Personal\_Details as 2 and Plan\_Details as 5. Phone numbers of the subscribers are to be used as row keys.

Q1) Add the data mentioned below into the table:

1. 7007989009(Row Key),

Personal\_Details:Name = ‘Rahul’,

Personal\_Details:Document\_Type=’Aadhar’, Personal\_Details:Document\_Number=’100934561234’

Plan\_Details:Call\_Plan=’149’

Plan\_Details:Data\_Plan=’399’

b. 8965969019(Row Key),

Personal\_Details:Name = ‘Sonia’,

Personal\_Details:Document\_Type=’Aadhar’, Personal\_Details:Document\_Number=’232934568901’

Plan\_Details:Data\_Plan=’499’

c. 98745654312(Row Key),

Personal\_Details:Name = ‘Hemraj’,

Personal\_Details:Document\_Type=’Passport’, Personal\_Details:Document\_Number=’HJ128MN’

Plan\_Details:Call\_Plan=’81’

d. 98745654312(Row Key),

Personal\_Details:Name = ‘Shalini’,

Personal\_Details:Document\_Type=’Passport’, Personal\_Details:Document\_Number=’HJ128MN’

Plan\_Details:Call\_Plan=’169’

Plan\_Details:Data\_Plan=’399’

Q2)

1. Let’s assume that after a month Rahul’s call plan got expired and he did not recharge. Update the new call plan to 0. Here, 0 or non-existence of the call\_plan entry represents that the user is not in possession of any active call plan.
2. Sonia opts for a new data plan of 599 INR. Update this in the table.
3. Hemraj opts for a new data plan of value 1099. Update this in the table.
4. Shalini updates her submitted document to Aadhaar, and her Aadhaar number is 134267541235. Update this in the table.

Q3)

1. Fetch the rows for which the document type is Aadhaar
2. Fetch the rows for which the document type is Passport
3. Fetch all data present in the column Name
4. Fetch all data present in the column family Plan\_Details

Q4)

1. Fetch two versions of the column Data\_Plan
2. Fetch two versions of the column Call\_Plan
3. Pick the earliest Call\_Plan for Rahul
4. Pick the earliest data\_plan for Sonia
5. Pick the earliest data\_plan for Hemraj
6. Pick the earliest Document\_Type for Shalini

Q5)

1. Delete the earliest Call\_Plan for Rahul
2. Delete the earliest data\_plan for Sonia
3. Delete the earliest data\_plan for Hemraj
4. Delete the earliest Document\_Type for Shalini

Q6)

Create a table named ‘Mobile\_Subscribers1’ using the details given above for the Mobile\_Subscribers table using Java.

Q7)

Insert all the data mentioned in Q1 into the HBase table Mobile\_Subscribers1 using Java.

Q8)

Carry out all the aforementioned changes in Q2 in the HBase table Mobile\_Subscribers2 using Java.

Q9)

Perform the delete tasks mentioned in Q5 using Java.