NoSQL Interview Questions And Answers Guide.

Global Guideline.

https://globalguideline.com/

NoSQL Job Interview Preparation Guide.

Question # 1

Please tell me what is impedance mismatch in Database terminology?

Answer:-

It is the difference between the relational model and the in-memory data structures. The relational data model organizes data into a structure of tables and rows, or more properly, relations and tuples. In the relational model, a tuple is a set of name-value pairs and a relation is a set of tuples. All operations in SQL consume and return relations, which leads to the mathematically elegant relational algebra.
This foundation on relations provides a certain elegance and simplicity, but it also introduces limitations. In particular, the values in a relational tuple have to be simple-they cannot contain any structure, such as a nested record or a list. This limitation isn't true for in-memory data structures, which can take on much richer structures than relations. As a result, if you want to use a richer in-memory data structure, you have to translate it to a relational representation to store it on disk. Hence the impedance mismatch-two different representations that require translation

Read More Answers.

Question # 2

Do you have any idea about Aggregate-oriented databases?

Answer:-

An aggregate is a collection of data that we interact with as a unit. Aggregates form the boundaries for ACID operations with the database. Key-value, document, and column-family databases can all be seen as forms of aggregate-oriented database. Aggregates make it easier for the database to manage data storage over clusters. Aggregate-oriented databases work best when most data interaction is done with the same aggregate; aggregate-ignorant databases are better when interactions use data organized in many different formations.
Aggregate-oriented databases make inter-aggregate relationships more difficult to handle than intra-aggregate relationships. They often compute materialized views to provide data organized differently from their primary aggregates. This is often done with map-reduce computations.

Read More Answers.

Question # 3

What is BigSQL?

Answer:-

Big Data is a culmination of numerous research and development projects at IBM. So IBM has taken the work from these various projects and released it as a technology preview called Big SQL.IBM claims that Big SQL provides robust SQL support for the Hadoop ecosystem:
- it has a scalable architecture
- it supports SQL and data types available in SQL '92, plus it has some additional capabilities
- it supports JDBC and ODBC client drivers
- it has efficient handling of "point queries"
Big SQL is based on a multi-threaded architecture, so it's good for performance and the scalability in a Big SQL environment essentially depends on the Hadoop cluster itself that is its size and scheduling
policies.

Read More Answers.

Question # 4

Explain the features of BigSQL?

Answer:-

IBM claims that Big SQL provides robust SQL support for the Hadoop ecosystem:
- it has a scalable architecture;
- it supports SQL and data types available in SQL '92, plus it has some additional capabilities;
- it supports JDBC and ODBC client drivers;
- it has efficient handling of "point queries" (and we'll get to what that means);
- there are a wide variety of data sources and file formats for HDFS and HBase that it supports;
- And it, although is not open source, it does interoperate well with the open source ecosystem within Hadoop.

Read More Answers.

Question # 5

What is JAQL?

Answer:-

Jaql is a JSON-based query language that translates into Hadoop MapReduce jobs. JSON is the data interchange standard that is humanreadable like XML but is designed to be lighter-weight.
Jaql programs are run using the Jaql shell. We start the Jaql shell using the jaqlshell command. If we pass no arguments, we start it in interactive mode. If we pass the -b argument and the path to a file, we will execute the contents of that file as a Jaql script. Finally, if we pass the -e argument, the Jaql shell will execute the Jaql statement that follows the -e.
There are two modes that the Jaql shell can run in: The first is cluster mode, specified with a -c argument. It uses the Hadoop cluster if we have one configured. The other option is minicluster mode, which starts a minicluster that is useful for quick tests. The Jaql query language is a data-flow language.

Read More Answers.

Question # 6

What is Hive?

Answer:-

Hive can be thought of as a data warehouse infrastructure for providing summarization, query and analysis of data that is managed by Hadoop.Hive provides a SQL interface for data that is stored in Hadoop.And, it implicitly converts queries into MapReduce jobs so that the programmer can work at a higher level than he or she would when writing MapReduce jobs in Java. Hive is an integral part of the Hadoop ecosystem that was initially developed at Facebook and is now an active Apache open source project.

Read More Answers.

Question # 7

Explain some benefits of Impala?

Answer:-

1) One of the key ones is low latency for executing SQL queries on top of Hadoop. And part of this has to do with bypassing the MapReduce infrastructure which involves significant overhead, especially when starting and stopping JBMs.
2) Cloudera also claims several magnitudes of improvement in performance compared to executing the same SQL queries using Hive.
3) Another benefit is that if we really wanted to look under the hood at what Cloudera has provided in Impala or if we wanted to tinker with the code, the source code is available for you to access and download.

Read More Answers.

Question # 8

Explain Flume?

Answer:-

Flume is an open source software program developed by Cloudera that acts as a
service for aggregating and moving large amounts of data around a Hadoop cluster
as the data is produced or shortly thereafter. Its primary use case is the gathering of log files from all the machines in a cluster to persist them in a centralized store such as HDFS.
In Flume, we create data flows by building up chains of logical nodes and
connecting them to sources and sinks. For example, say we wish to move data from
an Apache access log into HDFS. You create a source by tailing access log and use a logical node to route this to an HDFS sink.

Read More Answers.

Question # 9

What is data wizard?

Answer:-

A Data Wizard is someone who can consistently derive money out of data, e.g. working as an employee, consultant or in an other capacity, by providing value to clients or extracting value for himself, out of data. Even a guy who design statistical models for sport bets, and use his strategies for himself alone, is a data wizard.Rather than knowlege, what makes a data wizard successful is craftsmanship, intuition and vision, to compete with peers who share the same knowledge but lack these other skills.

Read More Answers.

Question # 10

What is Not Only SQL (NoSQL)?

Answer:-

A NoSQL or Not Only SQL database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Motivations for this approach include simplicity of design, horizontal scaling and finer control over availability. The data structure differs from the RDBMS, and therefore some operations are faster in NoSQL and some in RDBMS.

Read More Answers.

Question # 11

Do you know what is the key difference between Replication and Sharding?

Answer:-

a) Replication takes the same data and copies it over multiple nodes. Sharding puts different data on different nodes
b) Sharding is particularly valuable for performance because it can improve both read and write performance. Using replication, particularly with caching, can greatly improve read performance but does little for applications that have a lot of writes. Sharding provides a way to horizontally scale writes.

Read More Answers.

Question # 12

What is Pros?

Answer:-

a) Graph databases seem to be tailor-made for networking applications. The
prototypical example is a social network, where nodes represent users who have various kinds of relationships to each other. Modeling this kind of data using any of the other styles is often a tough fit, but a graph database would accept it with relish.
b) They are also perfect matches for an object-oriented system.

Read More Answers.

Question # 13

Explain the drawbacks and limitations associated with Hive?

Answer:-

1) The SQL syntax that Hive supports is quite restrictive. So for example, we are not
allowed to do sub-queries, which is very very common in the SQL world. There is no windowed aggregates, and also ANSI joins are not allowed. And in the SQL world there are a lot of other joins that the developers are used to which we cannot use with Hive.
2) The other restriction that is quite limiting is the data types that are supported, for example when it comes to Varchar support or Decimal support, Hive lacks quite severely
3) When it comes to client support the JDBC and the ODBC drivers are quite limited and there are concurrency problems when accessing Hive using these client drivers.

Read More Answers.

Question # 14

Can you please explain the difference between NoSql vs Relational database?

Answer:-

The history seem to look like this:
Google needs a storage layer for their inverted search index. They figure a traditional RDBMS is not going to cut it. So they implement a NoSQL data store, BigTable on top of their GFS file system. The major part is that thousands of cheap commodity hardware machines provides the speed and the redundancy.
Everyone else realizes what Google just did.
Brewers CAP theorem is proven. All RDBMS systems of use are CA systems. People begin playing with CP and AP systems as well. K/V stores are vastly simpler, so they are the primary vehicle for the research.
Software-as-a-service systems in general do not provide an SQL-like store. Hence, people get more interested in the NoSQL type stores.
I think much of the take-off can be related to this history. Scaling Google took some new ideas at Google and everyone else follows suit because this is the only solution they know to the scaling problem right now. Hence, you are willing to rework everything around the distributed database idea of Google because it is the only way to scale beyond a certain size.

Read More Answers.

Question # 15

Do you know how Cassandra writes?

Answer:-

Cassandra writes first to a commit log on disk for durability then commits to an in-memory structure called a memtable. A write is successful once both commits are complete. Writes are batched in memory and written to disk in a table structure called an SSTable (sorted string table). Memtables and SSTables are created per column family. With this design Cassandra has minimal disk I/O and offers high speed write performance because the commit log is append-only and Cassandra doesn't seek on writes. In the event of a fault when writing to the SSTable Cassandra can simply replay the commit log

Read More Answers.

Question # 16

What is Cassandra?

Answer:-

Cassandra is an open source scalable and highly available â€œNoSQLâ€• distributed database management system from Apache. Cassandra claims to offer fault tolerant linear scalability with no single point of failure. Cassandra sits in the Column-Family NoSQL camp. The Cassandra data model is designed for large scale distributed data and trades ACID compliant data practices for performance and availability. Cassandra is optimized for very fast and highly available writes. Cassandra is written in Java and can run on a vast array of operating systems and platform.

Read More Answers.

Question # 17

What is Cons?

Answer:-

a) Because of the high degree of interconnectedness between nodes, graph
databases are generally not suitable for network partitioning.
b) Graph databases don't scale out well.

Read More Answers.

Question # 18

What is Cassandra Data Model?

Answer:-

The Cassandra data model has 4 main concepts which are cluster, keyspace, column, column family.
Clusters contain many nodes (machines) and can contain multiple keyspaces.

A keyspace is a namespace to group multiple column families, typically one per application.
A column contains a name, value and timestamp .
A column family contains multiple columns referenced by a row keys.

Read More Answers.

Question # 19

Tell me what are the pros and cons of Graph database?

Answer:-

Pros
a) Graph databases seem to be tailor-made for networking applications. The
prototypical example is a social network, where nodes represent users who have various kinds of relationships to each other. Modeling this kind of data using any of the other styles is often a tough fit, but a graph database would accept it with relish.
b) They are also perfect matches for an object-oriented system.
Cons
a) Because of the high degree of interconnectedness between nodes, graph
databases are generally not suitable for network partitioning.
b) Graph databases don't scale out well.

Read More Answers.

Question # 20

What is Apache HBase?

Answer:-

Apache HBase is an open source columnar database built to run on top of the Hadoop Distributed File System (HDFS). Hadoop is a framework for handling large datasets in a distributed computing environment.HBase is designed to support high table-update rates and to scale out horizontally in distributed compute clusters. Its focus on scale enables it to support very large database tables e.g. ones containing billions of rows and millions of columns.

Read More Answers.

Question # 21

Tell me how Big SQL works?

Answer:-

The Big SQL engine analyzes incoming queries.It separates portions to execute at the server versus the portions to be executed by the cluster. It rewrites queries if necessary for improved performance; determines the appropriate storage handle for data; produces the execution plan and executes and coordinates the query.
IBM architected Big SQL with the goal that existing queries should run with no or few modifications and that queries should be executed as efficiently as the chosen storage mechanisms allow. And rather than build a separate query execution infrastructure they made Big SQL rely much on Hive, so much of the data manipulation language, the data definition language syntax, and the general concepts of Big SQL are similar to Hive. And Big SQL shares catalogues with Hive via the Hive metastore.Hence each can query each other's tables.

Read More Answers.

Question # 22

What is Impala?

Answer:-

Impala is a SQL query system for Hadoop from Cloudera. It is currently in beta; and it has been opensource and it's source can be downloaded from Gitap. It supports the same SQL syntax, the ODBC driver and the user interface (which is Beeswax) as Apache Hive.We can use it to query data, whether it is stored in ADFS or Apache H-base. And we can do selects, joins and aggregate functions.

Read More Answers.

Question # 23

Explain the drawbacks of Impala?

Answer:-

1)Impala isn't a GA offering yet.So as a beta offering, it has several limitations in terms of functionality and capability; for example, several of the data sources and file formats aren't yet supported.
2)Also ODBC is currently the only client driver that's available, so if we have JDBC applications we are not able to use them directly yet.
3)Another Impala drawback is that it's only available for use with Cloudera's distribution of Hadoop; that is CDH 4.1.

Read More Answers.

Question # 24

What do you know about Impala?

Answer:-

Impala is a SQL query system for Hadoop from Cloudera. The Cloudera positions Impala as a "real-time" query engine for Hadoop and by "real-time" they imply that rather than running batch oriented jobs like with MapReduce, we can get much faster query results for a certain types of queries using Impala over an SQL based front-end.
It does not rely on the MapReduce infrastructure of Hadoop, instead Impala implements a completely separate engine for processing queries. So this engine is a specialized distributed query engine that is similar to what you can find in some of the commercial pattern related databases. So in essence it bypasses MapReduce.

Read More Answers.

Question # 25

Explain "Polyglot Persistence" in NoSQL?

Answer:-

In 2006, Neal Ford coined the term polyglot programming, to express the idea that applications should be written in a mix of languages to take advantage of the fact that different languages are suitable for tackling different problems. Complex applications combine different types of problems, so picking the right language for each job may be more productive than trying to fit all aspects into a single language.
Similarly, when working on an e-commerce business problem, using a data store for the shopping cart which is highly available and can scale is important, but the same data store cannot help you find products bought by the customers' friends-which is a totally different question. We use the term polyglot persistence to define this hybrid approach to persistence.

Read More Answers.

Question # 26

Explain the modes of operation that Flume supports?

Answer:-

Flume supports three modes of operation: single node, pseudodistributed,
and fully distributed.
Single node is useful for basic testing and getting up and running quickly
Pseudo-distributed is a more production like environment that lets us build more complicated flows while testing on a single physical machine
Fully distributed is the mode that run in for production. The fully distributed
mode offers two further sub-modes: a standalone mode with a single
master and a distributed mode with multiple masters.

Read More Answers.

Databases Programming Most Popular Interview Topics.

1 : SQL and PL/SQL Frequently Asked Interview Questions and Answers Guide.

2 : MySQL Programming Frequently Asked Interview Questions and Answers Guide.

3 : MS SQL Server Frequently Asked Interview Questions and Answers Guide.

4 : Database Administrator (DBA) Frequently Asked Interview Questions and Answers Guide.

5 : Data Structures Frequently Asked Interview Questions and Answers Guide.

6 : SQL Frequently Asked Interview Questions and Answers Guide.

7 : Data Modeling Frequently Asked Interview Questions and Answers Guide.

8 : RDBMS Frequently Asked Interview Questions and Answers Guide.

9 : Stored Procedure Frequently Asked Interview Questions and Answers Guide.

10 : PostgreSQL Frequently Asked Interview Questions and Answers Guide.

About Global Guideline.

Global Guideline is a platform to develop your own skills with thousands of job interview questions and web tutorials for fresher's and experienced candidates. These interview questions and web tutorials will help you strengthen your technical skills, prepare for the interviews and quickly revise the concepts. Global Guideline invite you to unlock your potentials with thousands of **Interview Questions with Answers** and much more. Learn the most common technologies at Global Guideline. We will help you to explore the resources of the World Wide Web and develop your own skills from the basics to the advanced. Here you will learn anything quite easily and you will really enjoy while learning. Global Guideline will help you to become a professional and Expert, well prepared for the future.

* This PDF was generated from https://GlobalGuideline.com at **November 29th, 2023**

* If any answer or question is incorrect or inappropriate or you have correct answer or you found any problem in this document then don't hesitate feel free and e-mail us we will fix it.

You can follow us on FaceBook for latest Jobs, Updates and other interviews material.
www.facebook.com/InterviewQuestionsAnswers

Follow us on Twitter for latest Jobs and interview preparation guides
https://twitter.com/InterviewGuide

Best Of Luck.

Global Guideline Team
https://GlobalGuideline.com
Info@globalguideline.com