HBASE INTERVIEW QUESTIONS

http://www.tutorialspoint.com/hbase/hbase_interview_questions.htm

Dear readers, these **HBase Interview Questions** have been designed specially to get you acquainted with the nature of questions you may encounter during your interview for the subject of **HBase**. As per my experience good interviewers hardly plan to ask any particular question during your interview, normally questions start with some basic concept of the subject and later they continue based on further discussion and what you answer:

What are the different commands used in Hbase operations?

There are 5 atomic commands which carry out different operations by Hbase.

Get, Put, Delete, Scan and Increment.

How to connect to Hbase?

A connection to Hbase is established through Hbase Shell which is a Java API.

What is the role of Master server in Hbase?

The Master server assigns regions to region servers and handles load balancing in the cluster.

What is the role of Zookeeper in Hbase?

The zookeeper maintains configuration information, provides distributed synchronization, and also maintains the communication between clients and region servers.

When do we need to disable a table in Hbase?

In Hbase a table is disabled to allow it to be modified or change its settings. .When a table is disabled it cannot be accessed through the scan command.

Give a command to check if a table is disabled.

Hbase > is_disabled “table name”

What does the following table do?

hbase > disable_all 'p.*'

The command will disable all the table starting with the letter p

What are the different types of filters used in Hbase?

Filters are used to get specific data form a Hbase table rather than all the records.

They are of the following types.

- Column Value Filter
- Column Value comparators
- KeyValue Metadata filters.
- RowKey filters.

Name three disadvantages Hbase has as compared to RDBMS?

- Hbase does not have in-built authentication/permission mechanism
- The indexes can be created only on a key column, but in RDBMS it can be done in any column.
- With one HMaster node there is a single point of failure.

What are catalog tables in Hbase?

The catalog tables in Hbase maintain the metadata information. They are named as −ROOT− and .META. The −RROT− table stores information about location of .META> table and the .META> table holds information about all regions and their locations.

Is Hbase a scale out or scale up process?

Hbase runs on top of Hadoop which is a distributed system. Haddop can only scale uo as and when required by adding more machines on the fly. So Hbase is a scale out process.

What are the step in writing something into Hbase by a client?

In Hbase the client does not write directly into the HFile. The client first writes to WAL*WriteAccessLog*, which then is accessed by Memdtore. The Memstore Flushes the data into permanent memory from time to time.

What is compaction in Hbase?

As more and more data is written to Hbase, many HFiles get created. Compaction is the process of merging these HFiles to one file and after the merged file is created successfully, discard the old file.

What are the different compaction types in Hbase?

There are two types of compaction. Major and Minor compaction. In minor compaction, the adjacent small HFiles are merged to create a single HFile without removing the deleted HFiles. Files to be merged are chosen randomly.

In Major compaction, all the HFiles of a column are emerged and a single HFiles is created. The delted HFiles are discarded and it is generally triggered manually.

What is the difference between the commands delete column and delete family?

The Delete column command deletes all versions of a column but the delete family deletes all columns of a particular family.

What is a cell in Hbase?

A cell in Hbase is the smallest unit of a Hbase table which holds a piece of data in the form of a tuple{row,column,version}

What is the role of the class HColumnDescriptor in Hbase?

This class is used to store information about a column family such as the number of versions, compression settings, etc. It is used as input when creating a table or adding a column.

What is the lower bound of versions in Hbase?

The lower bound of versions indicates the minimum number of versions to be stored in Hbase for a column. For example If the value is set to 3 then three latest version wil be maintained and the older ones will be removed.

What is TTL *Timetolive* in Hbase?

TTL is a data retention technique using which the version of a cell can be preserved till a specific time period.Once that timestamp is reached the specific version will be removed.

Does Hbase support table joins?

Hbase does not support table jons. But using a mapreduce job we can specify join queries to retrieve data from multiple Hbase tables.

What is a rowkey in Hbase?

Each row in Hbase is identified by a unique byte of array called row key.

What are the two ways in which you can access data from Hbase?

The data in Hbase can be accessed in two ways.

- Using the rowkey and table scan for a range of row key values.
- Using mapreduce in a batch manner.

What are the two types of table design approach in Hbase?

They are − i Short and Wide ii Tall and Thin

In which scenario should we consider creating a short and wide Hbase table?

The short and wide table design is considered when there is

- There is a small number of columns
- There is a large number of rows

In Which scenario should we consider a Tall-thin table design?

The tall and thin table design is considered when there is

- There is a large number of columns
- There is a small number of rows

Give a command to store 4 versions in a table rather than the default 3.

hbase > alter 'tablename', {NAME => 'ColFamily', VERSIONS => 4}

What does the following command do?

```
hbase > alter 'tablename', {NAME => 'colFamily', METHOD => 'delete'}
```

This command deletes the column family form the table.

Give the commands to add a new column family “*newcolfamily*” to a table “*tablename*” which has a existing column family“*oldcolfamily*”.

```

Hbase > disable ‘tablename’
Hbase > alter ‘tablename’ {NAME => ‘oldcolfamily’,NAME=>’newcolfamily’}
Habse > enable ‘tablename’
```

What is the Hbase shell command to only 10 records form a table?

```

scan 'tablename', {LIMIT=>10,
STARTROW=>"start_row",
STOPROW=>"stop_row"}
```

What does the following command do?

major_compact 'tablename'

Run a major compaction on the table.

How does Hbase support Bulk data loading?

There are two main steps to do a data bulk load in Hbase.

- Generate Hbase data file*StoreFile* using a custom mapreduce job) from the data source. The StoreFile is created in Hbase internal format which can be efficiently loaded.
- The prepared file is imported using another tool like comletebulkload to import data into a running cluster. Each file gets loaded to one specific region.

How does Hbase provide high availability?

Hbase uses a feature called region replication. In this feature for each region of a table, there will be multiple replicas that are opened in different RegionServers. The Load Balancer ensures that the region replicas are not co-hosted in the same region servers.

what is HMaster?

The Hmaster is the Master server responsible for monitoring all RegionServer instances in the cluster and it is the interface for all metadata changes. In a distributed cluster, it runs on the Namenode.

What is HRegionServer in Hbase?

HRegionServer is the RegionServer implementation. It is responsible for serving and managing regions. In a distributed cluster, a RegionServer runs on a DataNode.

What are the different Block Caches in Hbase?

HBase provides two different BlockCache implementations: the default on-heap LruBlockCache and the BucketCache, which is *usually* off-heap.

How does WAL help when a RegionServer crashes?

The Write Ahead Log *WAL* records all changes to data in HBase, to file-based storage. if a RegionServer crashes or becomes unavailable before the MemStore is flushed, the WAL ensures that the changes to the data can be replayed.

Why MultiWAL is needed?

With a single WAL per RegionServer, the RegionServer must write to the WAL serially, because HDFS files must be sequential. This causes the WAL to be a performance bottleneck.

In Hbase what is log splitting?

When a region is edited, the edits in the WAL file which belong to that region need to be replayed. Therefore, edits in the WAL file must be grouped by region so that particular sets can be replayed to regenerate the data in a particular region. The process of grouping the WAL edits by region is called log splitting.

How can you disable WAL? What is the benefit?

WAL can be disabled to improve performance bottleneck.

This is done by calling the Hbase client field Mutation.writeToWAL*false*.

When do we do manula Region splitting?

The manual region splitting is done we have an unexpected hotspot in your table because of many clients querying the same table.

What is a Hbase Store?

A Habse Store hosts a MemStore and 0 or more StoreFiles *HFiles*. A Store corresponds to a column family for a table for a given region.

Which file in Hbase is designed after the SSTable file of BigTable?

The HFile in Habse which stores the Actual data*notmetadata* is designed after the SSTable file of BigTable.

Why do we pre-create empty regions?

Tables in HBase are initially created with one region by default. Then for bulk imports, all clients will write to the same region until it is large enough to split and become distributed across the cluster. So empty regions are created to make this process faster.

What is hotspotting in Hbase?

Hotspotting is asituation when a large amount of client traffic is directed at one node, or only a few nodes, of a cluster. This traffic may represent reads, writes, or other operations. This traffic overwhelms the single machine responsible for hosting that region, causing performance degradation and potentially leading to region unavailability.

What are the approaches to avoid hotspotting?

Hotspotting can be avoided or minimized by distributing the rowkeys across multiple regions. The different techniques to do this is salting and Hashing.

Why should we try to minimize the row name and column name sizes in Hbase?

In Hbase values are always freighted with their coordinates; as a cell value passes through the system, it’ll be accompanied by its row, column name, and timestamp. If the rows and column names are large, especially compared to the size of the cell value, then indices that are kept on HBase storefiles $StoreFile(HFile)$ to facilitate random access may end up occupying large chunks of the HBase allotted RAM than the data itself because the cell value coordinates are large.

What is the scope of a rowkey in Habse?

Rowkeys are scoped to ColumnFamilies. The same rowkey could exist in each ColumnFamily that exists in a table without collision.

What is the information stored in hbase:meta table?

The Hbase:meta tables stores details of region in the system in the following format.

info:regioninfo $serializedHRegionInfoinstanceforthisregion$

info:server $server:portoftheRegionServercontainingthisregion$

info:serverstartcode $start - timeoftheRegionServerprocesscontainingthisregion$

What is a Namespace in Hbase?

A Namespace is a logical grouping of tables . It is similar to a database object in a Relational database system.

How do we get the complete list of columns that exist in a column Family?

The complete list of columns in a column family can be obtained only querying all the rows for that column family.

When the records are fetched form a Hbase tables, in which order are the sorted?

The records fetched form Hbase are always sorted in the order of rowkey-> column Family-> column qualifier-> tiestamp.

What is Next ?

Further you can go through your past assignments you have done with the subject and make sure you are able to speak confidently on them. If you are fresher then interviewer does not expect you will answer very complex questions, rather you have to make your basics concepts very strong.

Second it really doesn't matter much if you could not answer few questions but it matters that whatever you answered, you must have answered with confidence. So just feel confident during your interview. We at tutorialspoint wish you best luck to have a good interviewer and all the very best for your future endeavor. Cheers :-)