Hadoop Interview Questions On HBase In 2018

Recommended by 140 users

Shubham Sinha (https://www.edureka.co/blog/author/shubham-sinha/) | Jul 27,2018

(https://www.facebook.com/sharer.php?u=https://www.edureka.co/blog/interview-questions/hbase-interview-questions/) (https://twitter.com/share?url=https://www.edureka.co/blog/interview-questions/hbase-interview-questions/&text=Hadoop Interview Questions On HBase In 2018) (https://www.linkedin.com/shareArticle?mini=true&url=https://www.edureka.co/blog/interview-questions/hbase-interview-questions/&title=Hadoop Interview Questions On HBase In 2018) (https://plus.google.com/share?url=https://www.edureka.co/blog/interview-questions/hbase-interview-questions/)

Add to Bookmark (https://www.edureka.co/blog/interview-questions/hbase-interview-questions/) Email this Post 14.1K

(https://www.edureka.co/blog/interview-questions/hbase-interview-questions/#comments-wrapper) 0 (https://www.edureka.co/blog/interview-questions/hbase-interview-questions/#disqus_thread)

Apache HBase Interview Questions

Looking out for Apache HBase Interview Questions that are frequently asked by employers? Here is the blog on Apache HBase interview questions in Hadoop Interview Questions series. I hope you must not have missed the earlier blogs of our ***Hadoop Interview Question series*** *(https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/)*.

After going through the HBase interview questions, you will get an in-depth knowledge of questions that are frequently asked by employers in Hadoop interviews related to HBase.

In case you have attended any HBase interview previously, we encourage you to add your questions in the comments tab. We will be happy to answer them, and spread the word to the community of fellow job seekers.

Hadoop Interview Questions and Answers | Edureka

Important points to remember about Apache HBase:

- Apache HBase is a NoSQL column oriented database which is used to store the sparse data sets. It runs on top of the Hadoop distributed file system (HDFS) and it can store any kind of data.
- Clients can access HBase data through either a native Java API, or through a Thrift or REST gateway, making it accessible from any language.

♣ *Tip: Before going through this Apache HBase interview questions, I would suggest you to go through* ***Apache HBase Tutorial*** *(https://www.edureka.co/blog/hbase-tutorial)* and ***HBase Architecture*** *(https://www.edureka.co/blog/hbase-architecture/) to revise your HBase concepts.*

Now moving on, let us look at the Apache HBase interview questions.

1. What are the key components of HBase?

The key components of HBase are Zookeeper, RegionServer and HBase Master.

Key components of HBase

Component	Description
Region Server	A table can be divided into several regions. A group of regions is served to the clients by a Region Server
HMaster	It coordinates and manages the Region Servers (similar as NameNode manages DataNodes in HDFS).
ZooKeeper	Zookeeper acts like as a coordinator inside HBase distributed environment. It helps in maintaining server state inside the cluster by communicating through sessions.

2. When would you use HBase?

- HBase is used in cases where we need random read and write operations and it can perform a number of operations per second on a large data sets.
- HBase gives strong data consistency.
- It can handle very large tables with billions of rows and millions of columns on top of commodity hardware cluster.

3. What is the use of get() method?

get() method is used to read the data from the table.

4. Define the difference between Hive and HBase?

Apache Hive is a data warehousing infrastructure built on top of Hadoop. It helps in querying data stored in HDFS for analysis using Hive Query Language (HQL), which is a SQL-like language, that gets translated into MapReduce jobs. Hive performs batch processing on Hadoop.

Apache HBase is NoSQL key/value store which runs on top of HDFS. Unlike Hive, HBase operations run in real-time on its database rather than MapReduce jobs. HBase

partitions the tables, and the tables are further splitted into column families.

Hive and HBase are two different Hadoop based technologies – Hive is an SQL-like engine that runs MapReduce jobs, and HBase is a NoSQL key/value database of Hadoop. We can use them together. Hive can be used for analytical queries while HBase for real-time querying. Data can even be read and written from HBase to Hive and vice-versa.

5. Explain the data model of HBase.

HBase comprises of:

- Set of tables.
- Each table consists of column families and rows.
- Row key acts as a Primary key in HBase.
- Any access to HBase tables uses this Primary Key.
- Each column qualifier present in HBase denotes attributes corresponding to the object which resides in the cell.

6. Define column families?

Column Family is a collection of columns, whereas row is a collection of column families.

7. Define standalone mode in HBase?

It is a default mode of HBase. In standalone mode, HBase does not use HDFS—it uses the local filesystem instead—and it runs all HBase daemons and a local ZooKeeper in the same JVM process.

8. What is decorating Filters?

It is useful to modify, or extend, the behavior of a filter to gain additional control over the returned data. These types of filters are known as decorating filter. It includes SkipFilter and WhileMatchFilter.

9. What is RegionServer?

A table can be divided into several regions. A group of regions is served to the clients by a Region Server.

10. What are the data manipulation commands of HBase?

Data Manipulation commands of HBase are:

- **put** – Puts a cell value at a specified column in a specified row in a particular table.
- **get** – Fetches the contents of a row or a cell.
- **delete** – Deletes a cell value in a table.
- **deleteall** – Deletes all the cells in a given row.
- **scan** – Scans and returns the table data.
- **count** – Counts and returns the number of rows in a table.
- **truncate** – Disables, drops, and recreates a specified table.

11. Which code is used to open a connection in HBase?

Following code is used to open a HBase connection, here *users* is my HBase table:

```
Configuration myConf = HBaseConfiguration.create();
HTable table = new HTable(myConf, “users”);
```

12. What is the use of truncate command?

It is used to disable, drop and recreate the specified tables.

♣ ***Tip:*** *To delete table first disable it, then delete it.*

13. What happens when you issue a delete command in HBase?

Once you issue a delete command in HBase for cell, column or column family, it is not deleted instantly. A tombstone marker in inserted. Tombstone is a specified data, which is stored along with standard data. This tombstone makes hides all the deleted data.

The actual data is deleted at the time of major compaction. In Major compaction, HBase merges and recommits the smaller HFiles of a region to a new HFile. In this process, the same column families are placed together in the new HFile. It drops deleted and expired cell in this process. All the results from scan and get filters the deleted cells.

14. What are different tombstone markers in HBase?

There are three types of tombstone markers in HBase:

- Version Marker: Marks only one version of a column for deletion.
- Column Marker: Marks the whole column (i.e. all version) for deletion.
- Family Marker: Marks the whole column family (i.e. all the columns in the column family) for deletion

15. HBase blocksize is configured on which level?

The blocksize is configured per column family and the default value is 64 KB. This value can be changed as per requirements.

16. Which command is used to run HBase Shell?

./bin/hbase shell command is used to run the HBase shell. Execute this command in HBase directory.

17. Which command is used to show the current HBase user?

whoami command is used to show HBase user.

Learn Hadoop from Experts

(https://www.edureka.co/big-data-and-hadoop)

18. What is the full form of MSLAB?

MSLAB stands for Memstore-Local Allocation Buffer. Whenever a request thread needs to insert data into a MemStore, it doesn't allocates the space for that data from the heap at large, but rather allocates memory arena dedicated to the target region.

19. Define LZO?

Lempel-Ziv-Oberhumer (LZO) is a lossless data compression algorithm that focuses on decompression speed.

20. What is HBase Fsck?

HBase comes with a tool called hbck which is implemented by the HBaseFsck class. HBaseFsck (hbck) is a tool for checking for region consistency and table integrity problems and repairing a corrupted HBase. It works in two basic modes – a read-only inconsistency identifying mode and a multi-phase read-write repair mode.

21. What is REST?

Rest stands for Representational State Transfer which defines the semantics so that the protocol can be used in a generic way to address remote resources. It also provides support for different message formats, offering many choices for a client application to communicate with the server.

22. What is Thrift?

Apache Thrift is written in C++, but provides schema compilers for many programming languages, including Java, C++, Perl, PHP, Python, Ruby, and more.

23. What is Nagios?

Nagios is a very commonly used support tool for gaining qualitative data regarding cluster status. It polls current metrics on a regular basis and compares them with given thresholds.

24. What is the use of ZooKeeper?

The ZooKeeper is used to maintain the configuration information and communication between region servers and clients. It also provides distributed synchronization. It helps in maintaining server state inside the cluster by communicating through sessions.

Every Region Server along with HMaster Server sends continuous heartbeat at regular interval to Zookeeper and it checks which server is alive and available. It also provides server failure notifications so that, recovery measures can be executed.

25. Define catalog tables in HBase?

Catalog tables are used to maintain the metadata information.

26. Define compaction in HBase?

HBase combines HFiles to reduce the storage and reduce the number of disk seeks needed for a read. This process is called compaction. Compaction chooses some HFiles from a region and combines them. There are two types of compactions.

- **Minor Compaction**: HBase automatically picks smaller HFiles and recommits them to bigger HFiles.
- **Major Compaction**: In Major compaction, HBase merges and recommits the smaller HFiles of a region to a new HFile.

27. What is the use of HColumnDescriptor class?

HColumnDescriptor stores the information about a column family like compression settings, number of versions etc. It is used as input when creating a table or adding a column.

28. Which filter accepts the pagesize as the parameter in hBase?

PageFilter accepts the pagesize as the parameter. Implementation of Filter interface that limits results to a specific page size. It terminates scanning once the number of filter-passed the rows greater than the given page size.

Syntax: PageFilter (<page_size>)

29. How will you design or modify schema in HBase programmatically?

HBase schemas can be created or updated using the Apache HBase Shell or by using Admin in the Java API.

Creating table schema:

```
Configuration config = HBaseConfiguration.create();
HBaseAdmin admin = new HBaseAdmin(conf); // execute command through admin&lt;/span&gt;&lt;/pre&gt;

// Instantiating table descriptor class
HTableDescriptor t1 = new HTableDescriptor(TableName.valueOf("employee"));

// Adding column families to t1
t1.addFamily(new HColumnDescriptor("professional"));
```

```
t1.addFamily(new HColumnDescriptor("professional"));
t1.addFamily(new HColumnDescriptor("personal"));

// Create the table through admin
admin.createTable(t1);
```

♣ *Tip: Tables must be disabled when making ColumnFamily modifications.*

For modification:

```
String table = “myTable”;
admin.disableTable(table);
admin.modifyColumn(table, cf2); // modifying existing ColumnFamily
admin.enableTable(table);
```

30.What are the filters are available in Apache HBase?

The filters that are supported by HBase are:

- **ColumnPrefixFilter**: takes a single argument, a column prefix. It returns only those key-values present in a column that starts with the specified column prefix.
- **TimestampsFilter**: takes a list of timestamps. It returns those key-values whose timestamps match any of the specified timestamps.
- **PageFilter**: takes one argument, a page size. It returns page size, number of rows from the table.
- **MultipleColumnPrefixFilter**: takes a list of column prefixes. It returns key-values that are present in a column that starts with any of the specified column prefixes.
- **ColumnPaginationFilter**: takes two arguments, a limit and an offset. It returns limit number of columns after offset number of columns. It does this for all the rows.
- **SingleColumnValueFilter**: takes a column family, a qualifier, a comparison operator and a comparator. If the specified column is not found, all the columns of that row will be emitted. If the column is found and the comparison with the comparator returns true, all the columns of the row will be emitted.
- **RowFilter**: takes a comparison operator and a comparator. It compares each row key with the comparator using the comparison operator and if the comparison returns true, it returns all the key-values in that row.
- **QualifierFilter**: takes a comparison operator and a comparator. It compares each qualifier name with the comparator using the comparison operator and if the comparison returns true, it returns all the key-values in that column.
- **ColumnRangeFilter**: takes either minColumn, maxColumn, or both. Returns only those keys with columns that are between minColumn and maxColumn. It also takes two boolean variables to indicate whether to include the minColumn and maxColumn or not. If you don’t want to set the minColumn or the maxColumn, you can pass in an empty argument.
- **ValueFilter**: takes a comparison operator and a comparator. It compares each value with the comparator using the compare operator and if the comparison returns true, it returns that key-value.
- **PrefixFilter**: takes a single argument, a prefix of a row key. It returns only those key-values present in a row that start with the specified row prefix.
- **SingleColumnValueExcludeFilter**: takes the same arguments and behaves same as SingleColumnValueFilter. However, if the column is found and the condition passes, all the columns of the row will be omitted except for the tested column value.
- **ColumnCountGetFilter**: takes one argument, a limit. It returns the first limit number of columns in the table.
- **InclusiveStopFilter**: takes one argument, a row key on which to stop scanning. It returns all key-values present in rows up to and including the specified row.
- **DependentColumnFilter**: takes two arguments required arguments, a family and a qualifier. It tries to locate this column in each row and returns all key-values in that row that have the same timestamp.
- **FirstKeyOnlyFilter**: takes no arguments. Returns the key portion of the first key-value pair.
- **KeyOnlyFilter**: takes no arguments. Returns the key portion of each key-value pair.
- **FamilyFilter**: takes a comparison operator and comparator. It compares each family name with the comparator using the comparison operator and if the comparison returns true, it returns all the key-values in that family.
- **CustomFilter**: You can create a custom filter by implementing the Filter class.

31. How do we back up a HBase cluster?

There are two broad strategies for performing HBase backups: backing up with a full cluster shutdown, and backing up on a live cluster. Each approach has benefits and limitation.

Full Shutdown Backup

Some environments can tolerate a periodic full shutdown of their HBase cluster, for example, if it is being used as a back-end process and not serving front-end webpages.

- **Stop HBase**: Stop the HBase services first.
- **Distcp**: Distcp could be used to either copy the contents of the HBase directory in HDFS to either the same cluster in another directory, or to a different cluster.
- **Restore**: The backup of the HBase directory from HDFS is copied onto the ‘real’ HBase directory via distcp. The act of copying these files, creates new HDFS metadata, which is why a restore of the NameNode edits from the time of the HBase backup isn’t required for this kind of restore, because it’s a restore (via distcp) of a specific HDFS directory (i.e., the HBase part) not the entire HDFS file-system.

Live Cluster Backup

The environments which cannot handle downtime uses Live Cluster Backup.

- **CopyTable**: Copy table utility could either be used to copy data from one table to another on the same cluster, or to copy data to another table on another cluster.
- **Export**: Export approach dumps the content of a table to HDFS on the same cluster.

32. How HBase Handles the write failure?

Failures are common in large distributed systems, and HBase is no exception.

If the server hosting a MemStore that has not yet been flushed crashes. The data that was in memory, but not yet persisted are lost. HBase safeguards against that by writing to the WAL before the write completes. Every server that’s part of the.

HBase cluster keeps a WAL to record changes as they happen. The WAL is a file on the underlying file system. A write isn’t considered successful until the new WAL entry is successfully written. This guarantee makes HBase as durable as the file system backing it. Most of the time, HBase is backed by the Hadoop Distributed Filesystem (HDFS). If HBase goes down, the data that were not yet flushed from the MemStore to the HFile can be recovered by replaying the WAL.

33. While reading data from HBase, from which three places data will be reconciled before returning the value?

The read process will go through the following process sequentially:

- For reading the data, the scanner first looks for the Row cell in Block cache. Here all the recently read key value pairs are stored.
- If Scanner fails to find the required result, it moves to the MemStore, as we know this is the write cache memory. There, it searches for the most recently written files, which has not been dumped yet in HFile.
- At last, it will use bloom filters and block cache to load the data from the HFile.

34. Can you explain data versioning?

In addition to being a schema-less database, HBase is also versioned.

Every time you perform an operation on a cell, HBase implicitly stores a new version. Creating, modifying and deleting a cell are all treated identically, they are all new versions. When a cell exceeds the maximum number of versions, the extra records are dropped during the major compaction.

Instead of deleting an entire cell, you can operate on a specific version within that cell. Values within a cell are versioned and it is identified the timestamp. If a version is not mentioned, then the current timestamp is used to retrieve the version. The default number of cell version is three.

35. What is a Bloom filter and how does it help in searching rows?

HBase supports Bloom Filter to improve the overall throughput of the cluster. A HBase Bloom Filter is a space efficient mechanism to test whether a HFile contains a specific row or row-col cell.

Without Bloom Filter, the only way to decide if a row key is present in a HFile is to check the HFile's block index, which stores the start row key of each block in the HFile. There are many rows drops between the two start keys. So, HBase has to load the block and scan the block's keys to figure out if that row key actually exists.

Check out our Hadoop Course

(https://www.edureka.co/big-data-and-hadoop)

Conclusion:

I hope these Apache HBase Interview Questions were helpful for you. This is just a part of our Hadoop Interview Question series. Kindly, refer to the links given below and enjoy the reading:

- ***Top 50 Hadoop Interview Questions*** (https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/)
- ***Hadoop Cluster Interview Questions*** (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hadoop-cluster/)
- ***Hadoop HDFS Interview Questions* (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/)**
- ***Hadoop MapReduce Interview Questions*** (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-mapreduce/)
- ***Pig Interview Questions (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-pig/)***
- ***Hive Interview Questions (https://www.edureka.co/blog/interview-questions/hive-interview-questions/)***

Got a question for us? Mention them in the comments section and we will get back to you.

About Shubham Sinha (26 Posts (https://www.edureka.co/blog/author/shubham-sinha/))

Shubham Sinha is a Big Data and Hadoop expert working as a Research Analyst at Edureka. He is keen to work with Big Data related technologies such as Hadoop, Spark, Flink and Storm and web development technologies including Angular, Node.js & PHP.

Share on (https://www.linkedin.com/shareArticle?mini=true&u=https%3A%2F%2Fwww.edureka.co%2Fblog%2Finterview-questions%2Fhbase-interview-questions%2F&title=HBase%20Interview%20Questions%20and%20Answers%20For%202018%20%7C%20Edureka&summary=&source=https%3A%2F%2Fwww.edureka.co%2Fblog%2Finterview-questions%2Fhbase-interview-questions%2F) (https://twitter.com/intent/tweet?text=HBase%20Interview%20Questions%20and%20Answers%20For%202018%20%7C%20Edureka&via=edurekaIN&url=https%3A%2F%2Fwww.edureka.co%2Fblog%2Finterview-questions%2Fhbase-interview-questions%2F) (https://plus.google.com/share?url=https%3A%2F%2Fwww.edureka.co%2Fblog%2Finterview-questions%2Fhbase-interview-questions%2F)

PREVIOUS NEXT

Got your brain cells running?
Stay tuned to latest technology updates

Enter your Email Address

SUBSCRIBE

Related Posts

HBase Tutorial: HBase Introduction and Facebook Case Study
20.8K
(https://www.edureka.co/blog/hbase-tutorial)

HBase Architecture: HBase Data Model & HBase Read/Write Mechanism
19.6K
(https://www.edureka.co/blog/hbase-architecture/)

Browse Categories

Big Data NoSQL (https://www.edureka.co/blog/category/big-data-nosql/) | Blockchain (https://www.edureka.co/blog/category/blockchain/)
Business Intelligence (https://www.edureka.co/blog/category/business-intelligence/) | Cloud Computing (https://www.edureka.co/blog/category/cloud-computing/)
Cyber Security (https://www.edureka.co/blog/category/cyber-security/) | Deep Learning (https://www.edureka.co/blog/category/deep-learning/)
Finance (https://www.edureka.co/blog/category/finance/) | Frameworks (https://www.edureka.co/blog/category/frameworks/)
Marketing (https://www.edureka.co/blog/category/marketing/) | Mobile Development (https://www.edureka.co/blog/category/mobile-development/)
Operations (https://www.edureka.co/blog/category/operations/) | Programming (https://www.edureka.co/blog/category/programming/)
Project Management (https://www.edureka.co/blog/category/project-management/) | Robotic Process Automation (https://www.edureka.co/blog/category/robotic-process-automation/)
Success Story (https://www.edureka.co/blog/category/success-story/) | Systems & Architecture (https://www.edureka.co/blog/category/systems-architecture/)
Systems Engineering (https://www.edureka.co/blog/category/systems-engineering/) | Testing (https://www.edureka.co/blog/category/testing/)

Comments

0 Comments

0 Comments https://www.edureka.co/blog/ Rajiv Chaudhuri
Recommend Tweet Share Sort by Best
Start the discussion…
Be the first to comment.
ALSO ON HTTPS://WWW.EDUREKA.CO/BLOG/
Top Technical Skills to Secure Jobs of the Future
4 comments • 3 months ago
Gvsm Chaithanya — gvsmc1996@gmail.com
K-Nearest Neighbors Algorithm Using Python
1 comment • 2 months ago
mehru — thank you so much. This article really helped...i want to implement this on mnist …
Top 10 Reasons Why You Should Learn Selenium
1 comment • 6 months ago
Partho — Since there is no Global Certification in Selenium what is the benchmark then.....
Cybersecurity Tools You Must Know – Tools for Cyber Threats
1 comment • 3 months ago
Gvsm Chaithanya — what are the industrial applications for information security
Subscribe Add Disqus to your siteAdd DisqusAdd Disqus' Privacy PolicyPrivacy PolicyPrivacy

Subscribe
to our newsletter
Enter your Email Address
SUBSCRIBE

Related Blogs

HBase Tutorial: HBase Introduction and Facebook Case Study (https://www.edureka.co/blog/hbase-tutorial)
(https://www.edureka.co/blog/hbase-tutorial)

HBase Architecture: HBase Data Model & HBase Read/Write Mechanism (https://www.edureka.co/blog/hbase-architecture/)
(https://www.edureka.co/blog/hbase-architecture/)

Top 50 Hadoop Interview Questions You Must Prepare In 2018 (https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/)
(https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/)

Top Hadoop Interview Questions To Prepare In 2018 – HDFS (https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/)
(https://www.edureka.co/blog/interview-questions/hadoop-interview-questions-hdfs-2/)

Big Data Analytics Courses

Big Data Hadoop Certification Training (/big-data-and-hadoop)
(/big-data-and-hadoop)

Python Spark Certification Training using PySpark (/pyspark-certification-training)
(/pyspark-certification-training)

Hadoop Administration Certification Training (/hadoop-admin)
(/hadoop-admin)

Apache Kafka Certification Training (/apache-kafka)
(/apache-

© 2014 Brain4ce Education Solutions Pvt. Ltd. All rights Reserved.

"PMP®","PMI®", "PMI-ACP®" and "PMBOK®" are registered marks of the Project Management Institute, Inc.
MongoDB®, Mongo and the leaf logo are the registered trademarks of MongoDB, Inc.