Hfile

Data is stored in hfile which is in the form of key/values.

When the memstore accumulates enough data the entire sorted keyvalue is written to a new hfile in hdfs this is a sequential write. It is very fast as it avoids moving the disk drive head

Hfile structure

A hfile contains a multi-layered index which allows hbase to seek the data without having to read the whole file. The multi-level index is like a b+ tree. Key value pars are stored in increasing order. Indexes point by row key to key value in 64kb blocks. Each block hs its own leaf index.

The last key of each block is put in the intermediate index. The root index points to the intermediate index. The trailer points t the meta blocks, and is written at the end of persisting the data to the file. The trailer also has information like bloom filters and time rrange info. Bloom filters help to skip files that do not contain a certain row key. The time range info is useful for skipping the file if it is not in the time range the read is looking for.

Hregion server

The region servers have regions that communicate with the cient and handle data-relate operations. Handle read and write requests for all the regions under it. Decide the size of the region by following the region size threshold.

Meta table

The meta table keeps a list of all regions in the system. The meta table structure is as follows

Key: region key format([table],[region start key],[region id])

Values: info: regioninfo,server,serverstartcode

When a table is in the process of splitting two other columns will be created. The values for these columns are also serialized HRegionInfo instances. After the region has been split eventually this row will be deleted.

Zookeeper

Zookeeper is an opensource project thet provides services like maintaining configuration information,naming providing distributed synchronization

Zookeeper has ephemeral nodes representing different region servers. Master servers use these nodes to discover available servers. In addition to availability, the nodes are also used to track server failures or network partitions.

Write

When the client gives a command write, the following steps occur

Instructions to directed to write ahead log and first writes important logs to it. Although it is not the area where the data is stored, it is done for the fault tolerant purpose. So, later if error occurs while writing data, hbase always has wal to look into. Once the log entry is done, the data to be written is forwarded to MemStore which is actually the RAM of data node. All the data is written in the MemStore which is faster than RDBMS. Later the data is dumped n Hfile,where the actual data is stored in hdfs. If the MemCache is full, the data is stored in Hfile directory

Read

Read process starts when a client ends request to Hbase. A request is sent to zookeeper which keeps all the status for the distributed system, when HBase is also present,Refer the figure above.

Zookeper has location for META table which is present in HRegion server. When a client requests zookeeper, it gives the address for the table. The process continues to HRegionServer and gets to META table, where it gets the region address of table where the data is present to be read. Moving forward to a specific HRegion, the process enters the blockcache where data is present from previous read. If a user queries the same records, the client will get the same data in no time. If the table is found the process returns to client with the data as result.

If the table is not found the process starts to search Memstore since data would have been written to Hfile sometime back. If it found the process returns to client with the data as result. If the table is not found the process moves forward in search of data within the HFile. The data wil be located here and once the search is completed the process takes required data and moves forward. The data taken from HFile is thelatest read data and can be read by user again. Hece the data is written in BlockCache, so that the next time,it can be instantly accessed by the client. When the data is written in BlockCache and all the search is completed, the read process wit required data will be returned to client along with ACK