

HBase Explanation

A chat system such as this one will need the ability to efficiently perform random reads and writes because there will be many users sending messages to one another on a daily basis. Like other NoSQL systems, HBase is not completely ACID compliant; however, HBase does guarantee atomicity within a particular row. The ability to find a user and update their list of messages coupled with the guarantee that the entire write will be complete makes HBase the optimal choice for a chat system with this sort of structure. HBase also has the capability to perform both partial and complete scans of its data. This would allow someone to quickly look at the information about a range of users. For example, in this assignment we need to be able to display all the messages that have been sent to a particular user. This task is easily completed using the partial scan capability provided by HBase. By providing the first message and the last message received by a user we can perform a scan over the data in order to display only the messages we are looking for.

The Row key we chose for our project was the userID + the message number. We chose this key because it would be unique for each row because each message would have a different message number and since it is concatenated with the userID it would be easy to scan the messages for each user. At first we wanted to structure the row key with the userID+ the message Title; however, we realized that message Titles do not necessarily need to be unique.

For the Column Families we chose “message” and “user” We decided that this was the most logical way to store the data as the “message” column family would contain the message information (title of the message, data/time, and the message contents) while the “user” column family would contain the actual address of the user. While the row key does contain the userID it does not contain the actual email address of the particular user, so we decided to include it in the “user” column family.

Columns inside “user” column family:

- user’s email address

Columns inside “message” column family:

- message title
- message body
- date and time