HW1 Twitter Data Generator Analysis

Wesley Jiang

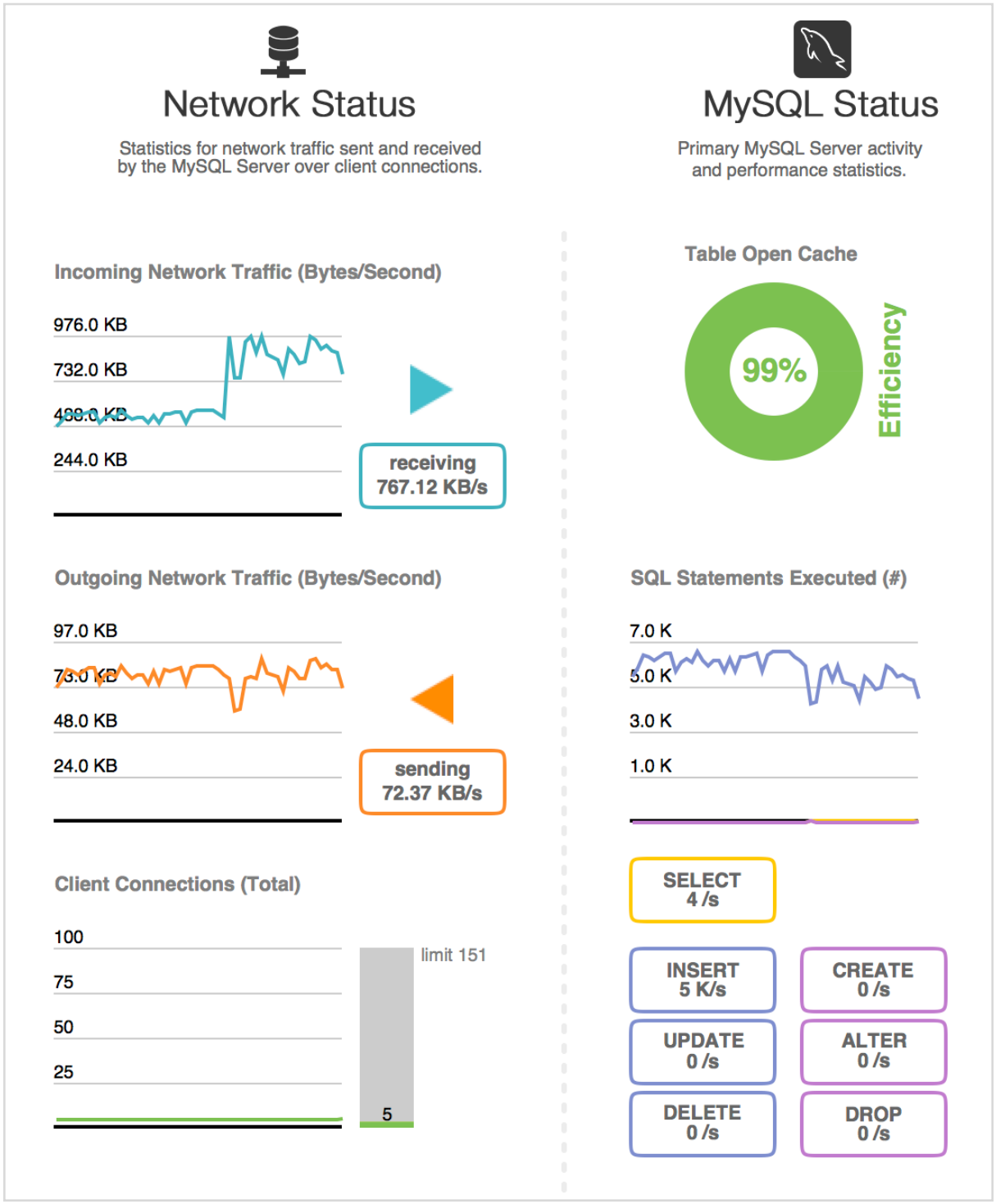
Hardware configuration:
CPU: 2.6 GHz Quad Core
RAM: 16 GB 2133 MHz LPDDR3
Storage: 256GB SSD

Inset Speed

Originally the one million inserts took around 5 minutes with an average of 3k inserts per second using the default innoDB engine. To speed up the process, I moved the commit statement outside of the for loop for inserts, which means all the sql statements execute first and then calls one commit after. However, MySQL auto commits batches of records to optimize performance, but the last batch never commits before the connection is closed so the last inserts get rolled back automatically. The speed did picked up to 7k inserts per second with a total time of 2.56 minutes but the last batch did not get stored into the database.

To fix this problem while still having a fast insert rate, MyISAM storage engine is used instead of innodb. The inserts had a relative faster 7k inserts/second with a total time of 2.36 minutes.

For the timeline requests it was about to perform on average 5 requests per second. And when performed concurrently the both operations slow down. The select speed is 4 / second while the inserts dropped to 5k per second.



Data Distribution

The data assumed that each users has 50 followers and each user has 100 tweets. After trying a random distributions of followers instead of every user has 50 followers, the speed for inserting and selecting both increased to around 6 or 7 per second.

The retrieval slows down just slightly as the records fill up the database, the retrieval rate drops from 3 selects to 2 selects and the insert rates drops from 7k per second to 5k per second.

A billion tweets would obviously not be ideal given the limitaion of hardware such as CPU,

RAM, disk, and also the limitation of MySQL database itself. Storing and querying will be extremely slow as it scales to the size of billions.

Space and Cost

My system could not keep up with the current twitter data generation rate, the max my system achieved was 7k per second 3k short from the ideal speed. My computer has 250G of SSD and the million inserts data were about 92MB for 2.36 minutes, which means that my drive would fill up in about 102 hours , about 4 days. In a year the storage needs is about 22 TB. The cost of hard drive is about $0.05/G, which means that the annual cost is about $1100.

#Class/DS4300