Anatomy of an Append Operation

Let’s learn how GFS handles an append operation.

Record append operation is optimized in a unique way that distinguishes GFS from other distributed file systems. In a normal write, the client specifies the offset at which data is to be written. Concurrent writes to the same region can experience race conditions, and the region may end up containing data fragments from multiple clients. In a record append, however, the client specifies only the data. GFS appends it to the file at least once atomically (i.e., as one continuous sequence of bytes) at an offset of GFS’s choosing and returns that offset to the client. This process is similar to the append operation on a file opened with O_APPEND (https://man7.org/linux/man-pages/man2/open.2.html) mode on a POSIX-compliant file system but without the race conditions when multiple writers do so concurrently.

Record Append is a kind of mutation that changes the contents of the metadata of a chunk. When an application tries to append data on a chunk by sending a request to the client, the client pushes the data to all replicas of the last chunk of the file just like the write operation. When the client forwards the request to the primary, the primary checks whether appending the record to the existing chunk will increase the chunk’s size more than its limit (maximum size of a chunk is 64MB). If this happens, it pads the chunk to the maximum limit, commands the secondary to do the same, and requests the clients to try to append to the next chunk. If the record fits within the maximum size, the primary appends the data to its replica, tells the secondary to write the data at the exact offset where it has, and finally replies success to the client.

If an append operation fails at any replica, the client retries the operation. Due to this reason, replicas of the same chunk may contain different data,

possibly including duplicates of the same record in whole or in part.

Hence, GFS does not guarantee that all replicas are byte-wise identical; instead, it only ensures that the data is written at-least-once as an atomic unit.

← Back

Anatomy of a Write Operation

Next →

GFS Consistency Model and Snapshot...

Mark as Completed

Report an Issue