Internals of Data Distribution

https://github.com/apple/foundationdb/blob/master/design/data-distributor-internals.md

Basic Understanding of Data Distribution

FoundationDB will only move data for a few reasons:

- to restore replication after a failure
- to split or merge a shard to a size between ~125MB to 500MB (this range is smaller grows and shrinks with database size, these numbers are the max for the largest databases)
- to split or merge a shard because of a lot of recent writes (write hotspot)
- to balance the bytes stored across the storage servers

It does not balance based on high read traffic. When it moves a shard it only considers bytes stored, not write traffic.

This means it is possible to randomly assign the same storage server to multiple high read or write traffic ranges.

Troubleshooting

In the case of a hot storage server, you can look at the `StorageMetrics` TraceEvents in the logs to see if you observe higher read or write traffic compared to other servers.

The TraceEvents `MovingData` and `TotalDataInFlight` can show the type of work currently being done by data distribution.

`RelocateShardStartSplitx100` will give you sampled view of why shards are being split, and `RelocateShardMergeMetrics` will be logged every time two shards are merged.

`TeamHealthChanged` will tell you when a team of servers has become unhealthy.

Observing Data Movement

The cluster can tell you how much data there is to move and how much it is currently moving through `status` on the client.

```
Data:
  Replication health     - Healthy
  Moving data            - 0.043 GB
  Sum of key-value sizes - 88 MB
  Disk space used        - 382 MB
```

There is no ETA published for data movement. It should be noted, though, that data movement can happen constantly, particularly if there are lots of writes happening.

Adjusting Distribution Speeds

There aren't any easily available controls for data distribution. There are some knobs that control its behavior, but in general changing knobs is something that should be done with great caution, and if used they have to be set at the startup of your fdbserver processes. Some of the relevant knobs can be found in `fdbserver/knobs.h`, such as: * `DD_MOVE_KEYS_PARALLELISM` * `MOVE_KEYS_KRM_LIMIT` * `DD_MERGE_COALESCE_DELAY`

It is not recommended to change them without a particular need, and one should be extremely careful when dealing with changes that may have unknown consequences.