Data Quantity Requirements

Your cluster should manage a large quantity of data if sharding is to have an effect. The default *chunk* size is 64 megabytes. And the *balancer* (page 684) will not begin moving data across shards until the imbalance of chunks among the shards exceeds the *migration threshold* (page 685). In practical terms, unless your cluster has many hundreds of megabytes of data, your data will remain on a single shard.

In some situations, you may need to shard a small collection of data. But most of the time, sharding a small collection is not worth the added complexity and overhead unless you need additional write capacity. If you have a small data set, a properly configured single MongoDB instance or a replica set will usually be enough for your persistence layer needs.

Chunk size is `user configurable`. For most deployments, the default value is of 64 megabytes is ideal. See *Chunk Size* (page 688) for more information.

Production Cluster Architecture

In a production cluster, you must ensure that data is redundant and that your systems are highly available. To that end, a production cluster must have the following components:

- **Three Config Servers** Each *config server* (page 670) must be on separate machines. A single *sharded cluster* must have exclusive use of its *config servers* (page 670). If you have multiple sharded clusters, you will need to have a group of config servers for each cluster.
- **Two or More Replica Sets As Shards** These replica sets are the *shards*. For information on replica sets, see *Replication* (page 559).
- **One or More Query Routers (`mongos`)** The `mongos` instances are the routers for the cluster. Typically, deployments have one `mongos` instance on each application server.

  You may also deploy a group of `mongos` instances and use a proxy/load balancer between the application and the `mongos`. In these deployments, you *must* configure the load balancer for *client affinity* so that every connection from a single client reaches the same `mongos`.

  Because cursors and other resources are specific to an single `mongos` instance, each client must interact with only one `mongos` instance.

See also:

Deploy a Sharded Cluster (page 691)

Sharded Cluster Test Architecture

Warning: Use the test cluster architecture for testing and development only.

For testing and development, you can deploy a minimal sharded clusters cluster. These **non-production** clusters have the following components:

- One *config server* (page 670).
- At least one shard. Shards are either *replica sets* or a standalone `mongod` instances.
- One `mongos` instance.

See

Production Cluster Architecture (page 672)