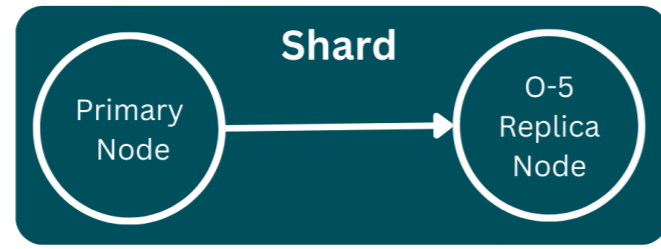


1

Terminology to Understand ElastiCache Cluster Modes

1



Shard

- 1 A shard is a logical partition of your dataset
- 2 It consists of
 - 1 One Primary Node Handles all write operations
 - 2 Zero to Five Replica Nodes Provide redundancy and read scalability
- 3 Each shard is responsible for managing a portion of the data in the cluster
- 4 In a cluster with 3 shards, the data is divided into 3 parts, and each shard handles one part

2

Primary Node

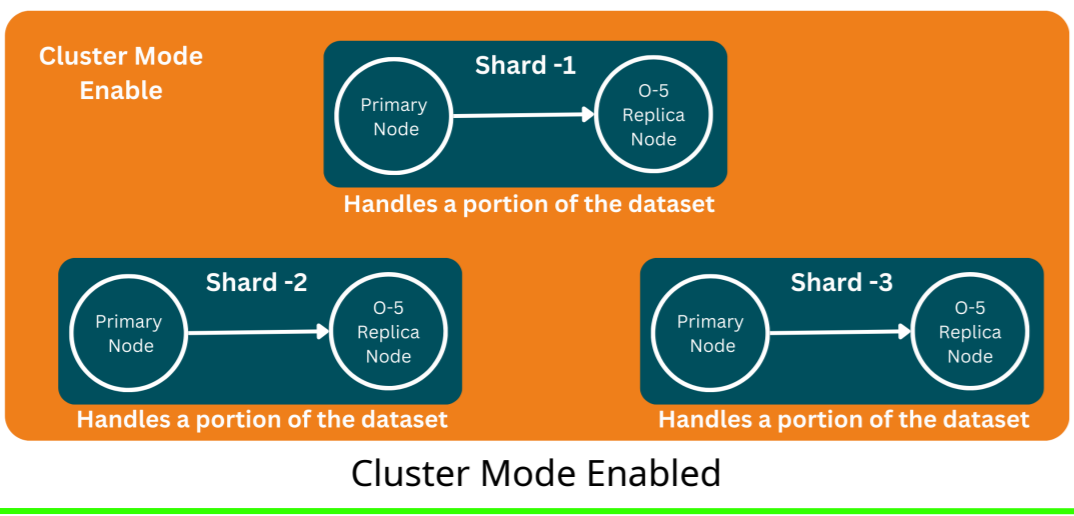
- 1 The primary node is the main node in a shard that Stores the shard's data
- 2 Handles all write operations and provides the source for replication
- 3 Each shard has exactly one primary node

3

Replica Node

- 1 A replica node is an exact copy of the primary node within a shard
- 2 By offloading read traffic from the primary node, replicas improve performance for read-heavy workloads
- 4 Replicas ensure high availability by taking over if the primary node fails

2



Cluster Mode Enabled

1

Introduction

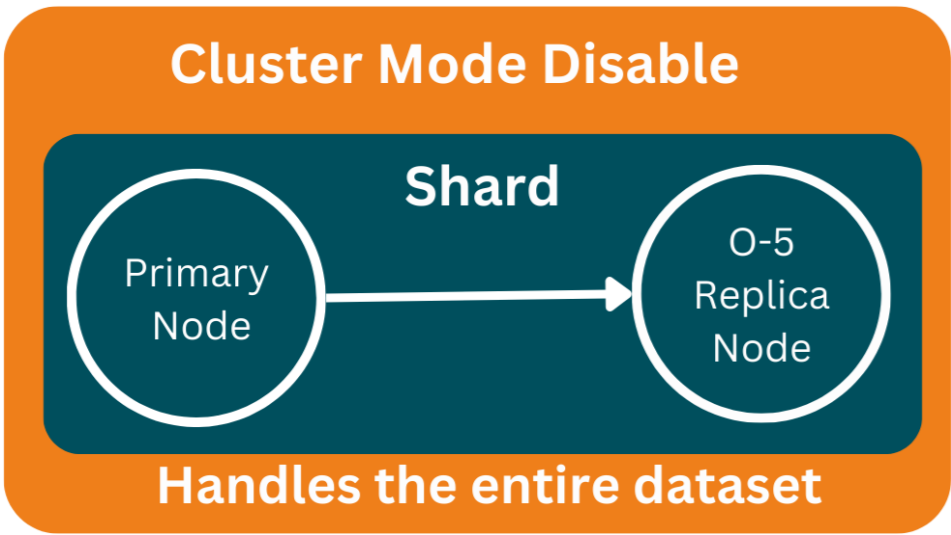
- 1 In Cluster Mode Enabled, your data is partitioned across multiple shards, enabling horizontal scaling and high performance.
- 3 Each shard stores a subset of your data and operates independently, allowing the cluster to handle larger datasets and higher traffic efficiently
- 4 This mode is only supported by Redis in Amazon ElastiCache, as Memcached does not support sharding or replicas

2

Key Features

- 1 Data Partitioning Across Shards
 - 1 The dataset is divided into multiple shards using a hashing mechanism
 - 2 Each shard handles a portion of the data, allowing the cluster to scale horizontally
- 2 High Performance
 - 1 By distributing data across shards, the cluster can handle a higher volume of read and write requests simultaneously
 - 2 Optional replicas further improve read scalability
- 3 High Availability
 - 1 Each shard can have up to 5 replicas for redundancy
 - 2 If a primary node in a shard fails, one of its replicas within the same shard is promoted to primary automatically (when automatic failover is enabled)

3



Cluster Mode Disabled

1

Introduction

- 1 In Cluster Mode Disabled, all data is stored within a single primary node, and optional replica nodes can be added for high availability and read scaling.
- 2 This mode is supported by both Redis and Memcached, but they function slightly differently

2

Key Features

- Single Node Group
 - 1 The entire dataset is managed by one primary node
 - 2 Additional replica nodes (up to 5 for Redis) can be added
 - 3 No Data Partitioning All data resides in one node group, making it simpler to set up and manage
 - 4 Scalability Limited to the memory and capacity of the single primary node

3

Key Difference Between Redis and Memcached Replicas

- 1 When configuring replicas for Redis in Cluster Mode Disabled
 - 1 Replicas are exact copies of the primary node
 - 2 They support read scaling and provide high availability in case the primary fails
 - 3 Maximum replicas: 5 per primary node
- 2 Memcached allows you to configure multiple replica but
 - 1 In Memcached, when you add replica nodes, these are independent nodes, not true replicas
 - 2 The dataset is distributed across these nodes using client-side hashing, and each node holds a portion of the data
 - 3 Advantages of Adding Replica
 - 1 Improves performance by distributing the load across nodes
 - 2 Not for High Availability If one Memcached node fails, the data stored on that node is lost, as there is no failover or redundancy