

Breakout Exercise: Making Scaling Decisions Based on Hardware

Your company has decided to deploy a new storage platform for a data science project, and your group has been put in charge of scaling and deployment of the system.

The chosen NoSQL solution works based on a *hash-ring*. Put simply, the system is sharded across N servers at any one time (the ring). Incoming I/O is sharded to one of the servers on each request. Data is balanced across the ring, such that I/O hotspots are largely avoided. Adding a node to the hash ring causes data to rebalance, such that adding 1 server requires all N servers send $1/(N+1)$ of their data volume to the new server.

The incoming data is written at a constant rate of 500GB/day. Queries of the data will produce a constant read rate of 10% of the total stored data volume per day.

Your group needs to make two decisions:

1. What is initial hardware configuration? Specifically
 - a. How many servers are in the initial hash ring?
 - b. What is the drive configuration for each server?
 - c. What is the cost of the initial deployment?
2. Determine the size of the hash ring at 6 and 12 months after initial deployment.
 - a. How many rebalances have taken place?
 - b. How much additional hardware cost been incurred, in units of “disk cost”

For designing hardware, your group may choose from a stock servers with **either**

- 12, 4TB Spinning Hard disks
 - Each disk costs 1
 - Each disk can do 200MB/s IO
- 12, **2TB** SSDs
 - Each disk costs 3
 - Each disk can do 500MB/s IOs
- 12 disks, mixed
 - Using SSD and Spinning disk metrics above.

At the end of the breakout period, your group will briefly explain your choices to the rest of the class.