Enterprise applications that require scale and availability are being migrated to non-relational cloud databases (such as Apache HBase, Apache Cassandra, DataStax, MongoDB, CouchBase, Amazon DynamoDB, Microsoft DocumentDB, et al.). At the same time, enterprises are onboarding next-generation applications (analytics, e-commerce, IoT, etc.) natively on non-relational databases. Due to the distributed nature of these non-relational databases, traditional backup and recovery solutions are unable to meet new data protection requirements, including: application-consistent and online backups (versioning), granular and orchestrated recovery, restore to different topology for staging and test/dev use cases, and any point-in-time recovery for near-zero RTOs. Enterprises value their applications and data, and are struggling to find a next-generation horizontal scale-out data protection solution to help them recover from data loss. This paper will assess one such cloud database: MongoDB. There will be an assessment of 3 existing backup and recovery solutions for MongoDB, and a comparison of value and deployment costs.

THE 3 SOLUTIONS

Solution #1: Hidden Secondary based Scripted Solution
This is a manually scripted solution, which requires a hidden secondary node for backup purposes. At regular intervals, the node is locked (i.e. quiesced), file system snapshotted and transferred to a secondary storage repository. Alternatively, customers may also use MongoDump tool to dump the data in binary format before transferring to secondary storage. This is the most rudimentary solution, not scalable, and is error prone. Above all, this method doesn’t provide a consistent point-in-time backup of sharded clusters that is required to recover cleanly from data loss.

Solution #2: MongoDB Ops Manager based Backup and Recovery
Customers that use the Enterprise version of MongoDB subscription license get access to a DIY backup solution. However, the solution architecture is extremely complex as mentioned here. This results in massive costs to purchase, deploy, and maintain servers and storage.

Solution #3: Datos IO RecoverX Solution
Datos IO has redefined data protection with its innovative CODRTM architecture that simplifies data protection. In addition, users can achieve continuous data protection at a collection-level granularity.
**COMPARISON OF THE BASICS**

**Recovery Time:** This is the most important metric for any enterprise. A scripted solution (#1) requires a large amount of manual effort, typically several hours to days, to restore the data and apply operational logs on the last backup copy. However, there is more to it, which is the hidden cost of restore. It is not easy to restore granularly at a collection-level and hence, restore operations involve more data than necessary. With Datos IO RecoverX, you simply restore the collection through one-click recovery.

**Data Consistency:** For sharded MongoDB clusters, getting consistent backups is a big challenge. The scripted solution (#1) works on replica set basis and hence does nothing to ensure consistency across shards. RecoverX ensures that consistency is maintained across the shards before taking backups.

**VALUE COMPARISON**

There are multiple elements that make up the cost of a data protection solution—software licensing, infrastructure costs and operational costs. We have used Amazon AWS as a proxy to get infrastructure costs, but the same arguments hold true for any other on-premise or cloud environment. Let’s take a real customer scenario:

**MongoDB Environment**
- A sharded 3x4 MongoDB cluster (3 replica set, 1 primary + 2 secondary + 1 hidden node)
  - Database Size: 3TB (primary on-disk data set)
- A sharded 1x4 MongoDB cluster (1 replica set, 1 primary + 2 secondary + 1 hidden node)
  - Database Size: 1TB (primary on-disk data set)

Daily change rate: 5%
Backup Retention Time: 14 days

**Solution #1: Hidden Secondary Solution**

Assuming that a weekly full and daily incremental strategy is used, below is the total storage required:
- Total data size = \[ 4TB \times 2 + 0.2TB \times 12 \] = 10.4TB
  - Storage Type = AWS EBS
  - Secondary storage costs = $12.5K/year
- MongoDB licensing for hidden secondary ~ $10K x 2 = $20K/year
- Cost of deployment (including developing and maintaining scripts)
  - Assuming 2 weeks to develop and 3 weeks/year to maintain at $1K/day rate
  - Total operational costs = $25K

Total cost of the solution (first year): ~$57.5K/year

**Solution #2: Ops Manager Based Backup Solution**

This solution is hard to understand and even harder to implement as listed here. Looking at the minimum configuration required for this solution, the cost breakdown is listed below:
- Servers = 2x 4-core (backup daemon) + 3x 4-core (backup metadata and oplog stores)
  - Assuming m4.xlarge instances ($1.8K/yr)
  - Total server costs = $9K/yr
- Storage = 2.5 x 4TB (backup daemon) + 3 x 4TB (backup and oplog storage)
  - Assuming EBS storage ($1.2K/yr) for backup daemon and for rest S3 storage ($0.5K/yr)
  - Total storage costs = $17.5K/yr
- MongoDB licensing for data bearing nodes - $10K x 3 = $30K/yr
- Cost of deployment
  - Assuming 3 weeks to deploy and make this solution work at $1K/day rate
  - Total deployment cost = $15K

Total cost of the solution (first year) ~ $71.5K/year

There may potentially be more licensing costs for Head DB and storage that are not included in this analysis. Further, maintaining this complex solution will incur on-going operational costs.

**Solution #3: Datos IO RecoverX**

Datos IO has developed a continuous data protection solution for MongoDB that provides massive cost benefits given its next-generation CODR architecture. Users can recover their data at a collection level granularity up to the last minute. Not only this, deploying RecoverX requires only a single server and 1x the storage. Below are calculations that show cost savings with RecoverX.
- Server = 1x 8-core
  - Assuming m4.2xlarge instances ($3.6K/yr)
  - Total server costs = $3.6K/yr
- Total data size = \[ 4TB + 0.2TB \times 13 \] = 6.6TB
  - Storage Type = AWS S3
  - Secondary storage costs = $3K/yr
- RecoverX licensing cost estimate ~ $6K x 4TB = $24K/yr
- Cost of deployment
  - Assuming 1 day to deploy at $1K/day rate
  - Total deployment cost = $1K

Total cost of the RecoverX solution (TCO) ~ $31.6K/year (~57% less)
**SUMMARY**

To recap, you should be aware of the value you are getting with each solution, and the hidden costs of infrastructure, deployment and operations. Datos IO RecoverX not only provides unique data protection features such as any point-in-time recovery up to the last minute, but also reduces your operational and deployment cost significantly. If you are struggling with high operational and deployment costs, please feel free to reach out to our backup and recovery specialists.

**COMPARISON OF 3 SOLUTIONS FOR BACKUP AND RECOVERY OF MONGODB**

<table>
<thead>
<tr>
<th></th>
<th>Hidden Secondary Backup</th>
<th>Native Database Backup</th>
<th>Datos IO RecoverX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recovery Time</td>
<td>Very large (hours to days)</td>
<td>Low-Mid</td>
<td>Low (minutes to hours)</td>
</tr>
<tr>
<td>Recovery Point Objective (data loss)</td>
<td>High</td>
<td>Mid-High</td>
<td>Low (up to last minute)</td>
</tr>
<tr>
<td>Data Source Failure Handling</td>
<td>Poor</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Operational Efficiency</td>
<td>Poor</td>
<td>Good</td>
<td>Good</td>
</tr>
<tr>
<td>Storage Costs</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Vendor Lock-in</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Overall Solution Cost</td>
<td>High</td>
<td>Very High</td>
<td>Low</td>
</tr>
</tbody>
</table>

**ABOUT DATOS IO**

Datos IO provides cloud-scale data protection software to deliver reliable recovery in minutes (vs. hours), reduction of up to 70 percent in secondary storage costs and increased productivity of application owners, IT Ops and DevOps teams. Datos IO has been recognized by Gartner in 2016 Hype Cycle for storage Technologies for new category of Cloud Backup. Backed by Lightspeed Venture Partners and True Ventures, Datos IO is headquartered in San Jose, California.