THE REALITIES OF NOSQL BACKUPS
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>2</td>
</tr>
<tr>
<td>NOSQL DATABASES</td>
<td>2</td>
</tr>
<tr>
<td>PROBLEM: LACK OF COMPREHENSIVE BACKUP AND RECOVERY SOLUTIONS</td>
<td>3</td>
</tr>
<tr>
<td>Scripts &amp; Retrofitting</td>
<td>4</td>
</tr>
<tr>
<td>Persistent Replication</td>
<td>5</td>
</tr>
<tr>
<td>One-off Tools</td>
<td>5</td>
</tr>
<tr>
<td>Nodetool Snapshot</td>
<td>5</td>
</tr>
<tr>
<td>Mongodump for MongoDB</td>
<td>5</td>
</tr>
<tr>
<td>A LOOK AT TRILIO’S VIRTUAL SNAPSHOT TECHNOLOGY (VAST)</td>
<td>7</td>
</tr>
<tr>
<td>Application Awareness</td>
<td>7</td>
</tr>
<tr>
<td>One Click Restore</td>
<td>8</td>
</tr>
<tr>
<td>Selective Restore</td>
<td>8</td>
</tr>
<tr>
<td>Compelling Benefits</td>
<td>8</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>9</td>
</tr>
</tbody>
</table>
INTRODUCTION

According to Gartner, 80% of data in organizations today is unstructured (text, documents, transactions, audio, video, social media, etc.). With the production of global data predicted to be 44 times greater in 2020 than it was in 2009, organizations large and small need to plan for growth.¹

From the most cutting-edge companies to the more traditional fields like insurance and banking, the adoption of next-gen solutions like NoSQL are not only allowing for the ingestion of massive amounts of data but also empowering the “Always On” end user to access and analyze more information than ever before.

This paper discusses the various techniques used to backup the “crown jewels” of organizations that resides on NoSQL databases. The focus is then turned to how trilioData’s Virtual Snapshot Technology (VAST) can easily meet requirements of today resulting in saved time, resources and an organization’s infrastructure.

NoSQL Databases

In recent years NoSQL databases have emerged as an alternative to RDBMS systems where applications require scale and performance and strict ACID properties are not a requirement. Initially they have been incubated and widely deployed in large-scale datacenters like Facebook, Google and Amazon but are gradually finding their way into enterprises. Enterprises are attracted for their scale, performance and availability that traditional RDBMS systems cannot match.

As NoSQL data stores become a larger component of “big data”, the scalability, availability, cost effectiveness and flexibility of NoSQL databases such as MongoDB™, Apache Cassandra™ (or “Cassandra”) and others allow organizations to distribute data between multiple virtual machines (or “VMs” or “Nodes” or “Shards”) and easily grow their environments on an as needed basis.

¹ “Extracting Insight from Unstructured Data”, by Salil Godika, CMSWire, November 21, 2014
A workload is expressed as a collection of VMs, Nodes or Shards, the interconnectivity between the machines and the virtual disks mapped to those machines. Backup of these sets of VMs alone cannot describe the dependencies between VMs of these applications and require in-depth understanding of application configuration to learn the interdependency between these VMs.

**PROBLEM: LACK OF COMPREHENSIVE BACKUP AND RECOVERY SOLUTIONS**

With increasingly complex and critical IT environments, companies are looking for ways to fully protect their business, while at the same time providing easier and faster recovery times. One of the biggest challenges when deploying NoSQL databases in enterprises is the availability of comprehensive backup and recovery solutions. Today the database vendors provide the backup and recovery tools but most enterprises find them inadequate. The scripts and tools that were developed to address and fill the gaps often require additional resources and are hard to maintain. Historically, some enterprises have tried to retrofit existing backup solutions, only to find that file-level backups and VM-level backups cannot completely address their needs either. Part of the issue is that any backup vendor that leverages SAN and storage-based snapshots are unfit solutions, as NoSQL databases exclusively use commodity servers with direct attached storage...not SAN.

Additionally, the dynamic nature of these NoSQL databases also make existing backup solutions insufficient. For example: Adding additional NoSQL compute nodes and retiring existing nodes will cause data to move between nodes. When these topology changes occur, the historical backups become invalid as the data distribution on these nodes has changed.

Unlike the traditional RDBMS ecosystem that is based on SAN storage and backup techniques designed for those ecosystems, the resiliency currently within NoSQL’s distributed architecture allows for a number of benefits. The downside to this resiliency is replicated transactions such as poisoned data, user error and lost data.
To support these types of applications and the complex orchestration of NoSQL’s compute, network, firewall settings, network configurations and storage resources, preparing for point-in-time backup and disaster recovery has never been more important.

- Data loss ranks as number one concern of IT leaders: With 68% probability, the report shows that data loss and privacy breaches are the most prevalent concern for IT leaders over the next 12-18 months. ²

- (EMC) 51% percent of the organizations polled said that when it comes to "emerging workloads" -- hybrid cloud, big data and mobile, essentially -- IT does not have a disaster recovery plan in place. Should trouble befall their storage environments, 71% said that they aren’t fully confident in their ability to recover lost data.³

- Unscheduled downtime has a dramatic financial impact on businesses of all sizes, yet most businesses don’t have adequate recovery technologies in place. Seventy-nine percent reported they have had a major IT failure within the past two years, and only 7 percent were confident that they could recover operations within two hours.⁴

**Scripts & Retrofitting**

From Recovery to Analysis to Migration, point-in-time backups provide a number of benefits. Today, the common method of managing and executing partial backup and recovery is executed by internal scripts. This approach coupled with a recovery process can require a lot of resources: time, staff orchestration, validation, quality controls and maintenance = High Total Cost of Ownership. “Setting it and forgetting it” is not appropriate here. As one more item to manage, this approach can take a toll on IT professionals. Furthermore, as datacenters and environments continue to evolve so must scripting approaches due to lack of automated environmental awareness. The net result are departments executing repair jobs to reconstitute environments. This repair significantly affects SLA’s to bring the application back “online”. Furthermore, due to scripting and the manual nature of current tools, backups are done in most cases once a day. This means that your RPO (Recovery Point Objective) is at least a 24 hour window.

---

³ “Over $1.7 Trillion Lost Per Year from Data Loss and Downtime According to Global IT Study”, EMC, December 2, 2014
⁴ “New Research Reveals Major Challenges with Traditional Backup and Disaster Recovery”, Dimensional Research, November 2014
THE REALITIES OF NOSQL BACKUPS

Persistent Replication

A better but still not complete remedy to this problem is active-active replication of NoSQL workloads in other public/private clouds – known as disaster avoidance. This technique has its Pros and Cons. While there are benefits of continuity, replication technologies would propagate a deletion, misconfiguration or error through all protection tiers if they are real-time. The only way a protection policy can really be successful with replication is to have a schedule to provide a point-in-time recovery within the tools in use. Therefore, a combination of disaster avoidance and achievable point-in-time backup and recovery make for true resiliency and continuity.

One-off Tools

With the advent of NoSQL databases come basic tools to manage these applications. Each one of these tools are specifically tailored for their database environment. While these aid administrators to some degree, they leave many functional gaps required for proper management, maintenance and protection mandatory in today's enterprises. The following reflect a few of these tools:

Node tool Snapshot

Cassandra users without a proper backup and disaster recovery platform use this command to backup data using a snapshot. Depending on how you use the command, the following data is included:

- All keyspaces on a node.
- One or more keyspaces and all tables in the named keyspaces.
- A single table.

Cassandra flushes the node before taking a snapshot, takes the snapshot, and stores the data in the snapshots directory of each keyspace in the data directory. Ultimately, Cassandra users choose to leave the snapshot on the node itself or rsync (aka copy) to a backup store. If a Cassandra user chooses to rsync, the result is management of multiple backup buckets/locations for each of the nodes in the cluster.

Mongodump for MongoDB

mongodump is a utility for creating a binary export of the contents of a database. Consider using this utility as part of an effective backup strategy. Use mongodump in conjunction with mongorestore to

---

5 Apache Cassandra™ 2.0, DataStax Documentation – https://www.datastax.com/documentation/cassandra/2.0/cassandra/tools/toolsSnapShot.html
restore databases. Reading directly from MongoDB data files without an active mongod. The mongodump does not dump the content of the local database.

- The data format used by mongodump from version 2.2 or later is incompatible with earlier versions of mongod. Do not use recent versions of mongodump to back up older data stores.
- mongodump overwrites output files if they exist in the backup data folder. Before running the mongodump command multiple times, either ensure that you no longer need the files in the output folder (the default is the dump/ folder) or rename the folders or files.
A LOOK AT TRILIO’S VIRTUAL SNAPSHOT TECHNOLOGY (VAST)

With Trilio’s one of a kind subscription-based Business Assurance platform, Enterprise IT and Cloud Service Providers can now leverage Backup & Disaster Recovery as a Service for Cloud solutions in both VMware and OpenStack. This multi-tenant, self-service, policy-based solution is designed to protect NoSQL Application Workloads from data corruption or data loss, providing point-in-time snapshots, configuration and change awareness and seamless 1-click recovery.

Unlike traditional vendors that take a snapshot of the database or the application data running on a single compute node alone, Trilio takes a point-in-time snapshot of the entire NoSQL cluster -- that consist of the compute resources, network configurations, and storage data as a whole. The benefits are faster and reliable recovery, easier migration of NoSQL Applications between Cloud platforms and simplified virtual cloning of the cluster in its entirety. That said, like many traditional vendors Trilio’s incremental-forever architecture eliminates large data movements from the capture process. This approach significantly reduces the amount of time required to capture changes made to the production environment, as full database backups are no longer required.

Application Awareness

Integral part of the VAST technology includes application aware workflows. VAST out of box support includes MongoDB, Cassandra and a few other general-purpose workflows. VAST workflow framework is quite extensible and additional workflows can be easily integrated to support new applications.

Application aware workflows are responsible for application and environmental discovery, mapping the application resources to underlying virtualization artifacts including VMs, networks, ports,
security groups and virtual disks. Each application workflow is also responsible for discovering application topology. For example, a Cassandra specific workflow discovers the datacenters, all keyspaces the replication strategy and options for each keyspace. Workflows are also responsible for consistent backup of the application wide resources, restore points and application reconfiguration.

**One Click Restore**

Any backup copy, irrespective of its complexity can be restored with one click. One click process evaluates the target platform and restores the copy once the target platform passes the validation successfully.

**Selective Restore**

The Selective Restore process provides enormous flexibility with the restore process. A user is not limited to same virtualization platform and hypervisors. Selective restore process discovers the target platforms and provides various possible options to map backup image resources to new platform. This includes new hypervisor clusters, availability zones, networks, storage volumes etc.

---

![Diagram of single click process](image)

**KEY FEATURES**

| Time Machine | Policy Based Protection |
| Multi-tenancy | Custom Workflow |
| Single Pane of Glass | Automated Discovery |
| REST API |

**Compelling Benefits**

| TCO Savings | Data Protection |
| Ease of Capture | Cloud Agility |
| Environmental Intelligence/Awareness | Set-in and Forget-it |
| Single Click Recovery | |
CONCLUSION

KEY FINDING 1:
- Choosing a particular method of data protection usually depends on the budget, as well as what RTO (Recovery Time Objective) and RPO (Recovery Point Objective) required by a company and defined in its BC (Business Continuity) or DR (Disaster Recovery) plan. Backup should be the cornerstone of that plan.

KEY FINDING 2:
- NoSQL and other big data class of application architectures are vastly different than traditional applications and require new methods to provide data protection. With application aware workflows, trilioVault is intimately aware of each application data protection needs, vastly enhancing the reliability of your data protection plans.

KEY FINDING 3:
- With Trilio, a single solution can now be deployed to replace one or more tools, including backup software, disaster recovery, business continuity or test and development tools resulting in resource savings and speed to recovery.

About trilioData (“Trilio”)

Trilio is the most important piece of software that you do not know you need. With our best-in-class Business Assurance Platform, Trilio provides Backup & Disaster Recovery as a Service for Enterprise IT organizations and Cloud Service Providers deploying solutions on VMware and OpenStack. The platform is designed to protect NoSQL workloads such as Cassandra and MongoDB from data corruption or data loss. Through the solution’s one click policy-driven point-in-time backup and recovery, customers can rest assured that regardless of environmental or topographical changes their NoSQL applications and data are secure….recovery is just one click away.
THANK YOU.

FOR MORE INFORMATION CONTACT:
INFO@TRILIODATA.COM