DATOS IO

11 Pitfalls of Native Cassandra Backup Tools

Cassandra (Apache OSS version, DataStax Enterprise Edition / DSE) is a NoSQL non-relational database that is becoming increasingly popular with the emergence of enterprise use cases such as customer 360, IoT, and personalization.

One of the things to keep in mind is that you need an enterprise-grade backup and recovery solution that can effectively and efficiently protect data at scale.

Although native backup tools exist, here are 11 pitfalls to avoid.

1 Node level snapshots are not equal to database backup

The method of backup in native solutions is node-by-node snapshots which are essentially local snapshot backups **(i.e. hand coded scripts).** This is the most rudimentary solution which means it is not scalable and is error prone. Above all, this solution does not provide a point-in-time backup so you cannot recover in the event of catastrophic data loss.

2 Increased backup storage costs

In a native backup solution, **all replica copies are kept in backup storage.** These backups are stored in the Cassandra nodes themselves, as well as in optional secondary storage. In addition, there is no special handling of compacted sstables which means that newly generated compacted sstables are backed up, in addition to the original sstables where the new stables originated from. All of this translates into increased backup storage costs.

3 Longer recovery times with repairs upon recovery

Scripted solutions require a huge amount of manual effort to restore the data, and node-by-node restore from backup storage increases recovery time and network traffic. In addition, you need to restore all replicas during recovery which increases the time it takes for restore. But, it isn't just about the length of time. It is also the hidden cost of restore. After restoring the data, **DBAs** need to run cluster-wide repairs to bring the cluster to a consistent state.

4 Lack of Any Point-in-Time Recovery (APIT)

Enterprises frequently need to refresh their test and development clusters with the latest production data to enable continuous integration and continuous development. However, these clusters have different topologies **(number of nodes)** than production database clusters. It takes hours, if not days to refresh each cluster using native solutions — leading to loss of developer productivity.

5 No data masking option during recovery

Native tools do not give you the option to mask out certain columns during recovery of confidential data, such as personally identifiable information **(PII)** data. This has large implications for enterprises that handle sensitive data, including name, address, phone number, and social security number.

6 Lack of failure handling support during backup / recovery operations

In a native solution, if a source node fails during backup operations, the backups for that node stop. This may result in data loss or a large amount of inconsistency in a backed up data set. This is a significant limitation in any large scale production environment where nodes may fail often. In fact, you need your backup solution to perform most when failures occur.

7 Lack of support of time-to-live handling

There is no ability to adjust **TTL** during restores. Hence, if **TTL** is already expired during recovery, restored data is automatically expired by Cassandra.

8 Limited support for backup storage targets

Native tools are limited to choosing local file system or Amazon S3 as backup storage targets. There is no option to store backups to other **"S3 compatible"** object storage providers. Furthermore, there is no option to store backups to Google Cloud Storage or object storage targets for on-premise deployments.

9 Limited protection granularity level

In native solutions, only keyspace-level backup is available. There is no flexibility to back up using column-family level. This means all column families in a keyspace will be backed up using the same policy **(backup frequency and retention)**. Additionally, column families that are not needed, but in the same keyspace will be backed up as well.

10 Lower performance

Native solutions use sequential data streams for both backup and recovery as opposed to using parallel and distributed data movement.

11 Limited data management use cases

Native solutions do not give you the ability to restore to a different cluster with a different topology. What this means is that you cannot use the same cluster for restore to **QA/Dev/Test** clusters of different topology/capacity.

i About Datos IO

Datos IO provides application-centric data management for next-generation applications and databases, allowing enterprises to have confidence in their data being available always and anywhere.

Backed by Lightspeed Ventures, True Ventures, NetApp, and Cisco Investments, Datos IO is headquartered in San Jose, California.