Disaster Recovery Approaches

Authors avatar

Mohsin Banedar (Devops Engineer at Clairvoyant LLC)

Jul 12, 2017

Watch Watch

Disaster Recovery Plan

Hadoop disaster recovery plan is one of the most complicated and expensive elements to protect data.

Some of the best approaches have balanced cost, complexity, and recovery time while planning for disaster recovery.

Having two Hadoop clusters seems to be excellent disaster recovery plan in Hadoop;

we need to make sure that data insertion happens to both Hadoop clusters simultaneously.

We have come across implementations where data written on Hadoop is also inserted to other non-Hadoop locations.

This allows us, if data in Hadoop is lost or corrupted, to load data from this other location into Hadoop.

Problems secenrios where backup is important.

1.DataCenter or Availability Zones is down(Cloud ENV)

2. Hardware crashes(1.node crash, 2.Rack Failure, 3.Data Corruption)

3.Deletion of Data in production environment.

4.Permanent Physical Server damage or etc.

What we need to protect and take backup of

1. HDFS Data

2. Metadata

3.Configurations.

4.Application backup(however Config backup will take the backup of Application Configuration)

Backup Approaches

Approach 1:

Replication of the data in multiple nodes in multiple datcenter.

Approach 2:

Backup the HDFS data, Configuration, Metadata of master roles, Metastore data to S3

Create a pipeline and transfer the data from S3 to HDFS.

usnig distcp we can achive it.

Approach 3:

Transfer the data from one cluster to another cluster using distcp, make sure we have DR cluster

This is the better approach because we can use DR cluster for processing as well.

BDR is best approach to copy the HDFS, metastore and metadata to the another cluster, we can take some directory backup or metastore backup,however it is cost effective because

we need to build DR cluster for it.

Approach 4:

While ingesting data into active cluster we can parallel write the data in DR cluster.

Recovery of data to HDFS

We will be recovering the data,configuration, metadata and metastore from S3 for cloud environment.

Approach 1:Copy the data from S3 to HDFS.

If we are backing up the '/' HDFS data we will copying it in '/' location.

We also copy files in particular folder or particular directory, thats how we can recover the HDFS data

It all depend on the structure of which we are taking the backup.

Example:

for restore we can use below Command (BASH SCRIPT)

```
su $user -c "hadoop distcp s3n://"$s3_access_key":"$s3_secret_key"@"$s3_location""
hdfs://$ActiveNameNode/$HdfsPath
```

This is the best to restore which data and the location it need to get copied.

Approach 2: For metadata(masters roles)

We should keep a different metadata directory, it can be local dir or on S3.

To recover meta data for masters, if we dont have high Avability in our cluster and nodes crashes or any instance goes down.

We can add a new node in the cluster with Namenode service and copy the metadata dir from s3 to the disk where Namenode is pointing

for metadata. And restart the namenode service.

Approach 3:

Using Cloudera BDR, if we are takeing the backup of HDFS and metastore depend on the policy we have mentioned(immediate, scheduled).

We can use the DR cluster for processing in case datacenter or cluster crashes. and can make a new cluster with the DR cluster.

Also if we build the new cluster we can also copy the data vise versa from DR Custer to New active cluster

Approach 4:(physical servers)

Keeping the nodes in different datacneter / different racks, if racks or any datacenter goes down we will be having replica of the data,

after adding the nodes to the cluster we can run balancer command on it. This is suitable for physical servers.

Approach 5: (applcation restore)

If we have configuration backup on S3 with the help of distcp we can overwrite the config on the cluster.

And thats how we can get the application updated for the jobs running on the cluster.