

COHESITY

Version 4.3.0
March 2020

Cohesity Imanis Data User Guide

Table of Contents

1	Introduction: Imanis Data software.....	8
2	Overview of Imanis Data Software	14
2.1	Imanis Data Deployment.....	14
2.2	Version Compatibility Matrix.....	17
2.3	Data Repository Usage Matrix.....	20
3	Getting Started with Imanis Data Software	21
3.1	Logging in Imanis Data Software	21
3.2	QuickNav.....	23
3.3	HTTPS Support	23
3.4	Main Menu	24
3.5	Dashboard	25
3.6	Imanis Data Licensing.....	32
3.7	Smart System Notifications	34
3.7.1	<i>Recovery Alert Notification Band</i>	34
3.8	System Notification Band	34
4	Users & Roles	36
4.1	SMTP Config.....	36
4.1.1	<i>Adding SMTP</i>	36
4.1.2	<i>Editing SMTP server</i>	38
4.1.3	<i>Deleting SMTP server</i>	38
4.2	Domains.....	39
4.3	Adding a Domain.....	39
4.4	Editing a Domain	42
4.5	Deleting a Domain.....	42
4.6	Users.....	43
4.6.1	<i>Adding a Local User & Assigning a Role</i>	43

4.6.2 <i>Editing a Local User</i>	44
4.6.3 <i>Deleting a Local User</i>	44
4.6.4 <i>Resetting Password</i>	45
4.6.5 <i>Configuring a Domain user</i>	46
4.6.6 <i>Editing a Domain user</i>	47
4.6.7 <i>Deleting a Domain user</i>	47
4.7 Roles	48
4.7.1 <i>Creating a role</i>	49
4.7.2 <i>Editing a role</i>	50
4.7.3 <i>Deleting a role</i>	50
5 Policies	52
5.1 Creating Policies.....	52
5.1.1 <i>Backup Policy</i>	52
5.1.2 <i>DLM Policy</i>	57
5.1.3 <i>Recovery Policy</i>	60
5.1.4 <i>Global Cloud Policy</i>	62
5.1.5 <i>Direct Replication Policy</i>	63
5.2 Managing Policies	64
5.2.1 <i>Editing a policy</i>	64
5.2.2 <i>Deleting a policy</i>	66
6 Data Repositories	68
6.1 Prerequisites of Adding Data Repositories.....	68
6.1.1 <i>Hadoop (HDFS, Hive & HBase)</i>	68
6.1.2 <i>Cassandra</i>	73
6.1.3 <i>Couchbase</i>	80
6.1.4 <i>MongoDB</i>	80
6.1.5 <i>Amazon Glacier</i>	87
6.1.6 <i>Amazon S3</i>	88
6.1.7 <i>Azure Blob Storage</i>	89
6.1.8 <i>Cloudian</i>	90

6.2 Adding Data Repositories	90
6.2.1 <i>Auto Discovery</i>	91
6.3 Manual Configuration	104
6.3.1 <i>Hadoop (HDFS, Hive, HBase)</i>	104
6.3.2 <i>Couchbase</i>	108
6.3.3 <i>MongoDB</i>	111
6.3.4 <i>Amazon Glacier</i>	113
6.3.5 <i>Amazon S3</i>	116
6.3.6 <i>Cloudian</i>	121
6.3.7 <i>Azure Blob Storage</i>	125
6.4 Managing Data Repositories	128
6.4.1 <i>Setting a Blackout Window</i>	128
6.4.2 <i>Removing a Blackout Window</i>	128
6.4.3 <i>Editing a Data Repository</i>	129
6.4.4 <i>Deleting a Data Repository</i>	129
7 Data Backup	131
7.1 Getting Started with Data Backup	131
7.1.1 <i>Backing Up Data for HDFS or Hive</i>	131
7.1.2 <i>Backing Up Data for HBase</i>	135
7.1.3 <i>Backing Up Data for Cassandra</i>	139
7.1.4 <i>Backing Up Data for Couchbase</i>	144
7.1.5 <i>Backing Up Data for MongoDB</i>	149
8 Data Lifecycle Management	154
8.1 Overview.....	154
8.1.1 <i>Data Lifecycle Management for HDFS & Hive</i>	154
9 Data Recovery	159
9.1 Overview.....	159
9.2 Getting Started with Data Recovery	159
9.2.1 <i>Recovering Data for HDFS</i>	159

9.2.2 Recovering Data for Hive.....	166
9.2.3 Recovering Data for HBase.....	176
9.2.4 Recovering Data for Cassandra	186
9.2.5 Recovering Data for Couchbase.....	208
9.2.6 Recovering Data for MongoDB.....	243
9.3 Recovery Sandbox.....	254
9.3.1 Couchbase.....	255
9.3.2 Cassandra	263
9.3.3 Hadoop.....	273
10 Data Mirroring	282
10.1 Overview.....	282
10.1.1 Getting Started with Data Mirroring	282
10.1.2 Data Mirroring for Hive	282
10.1.3 Data Mirroring for HDFS.....	294
10.1.4 Data Mirroring for HBase	298
10.1.5 Data Mirroring for Cassandra	303
10.1.6 Data Mirroring for Couchbase	310
10.1.7 Data Mirroring for MongoDB.....	324
10.2 Direct Replication for Hadoop	329
10.2.1 Direct Replication for HDFS and HBase	330
10.2.2 Direct Replication for HIVE.....	334
11 Data Pipeline	338
11.1 Overview.....	338
11.2 Getting Started with Data Pipeline	338
11.2.1 Data Pipeline for HDFS or Hive.....	338
11.2.2 Data Pipeline for HBase	355
11.2.3 Data Pipeline for Cassandra.....	360
11.2.4 Data Pipeline for Couchbase	369
11.2.5 Data Pipeline for MongoDB	384
12 Managing Jobs	390

12.1 Rerunning ‘Run Now’ jobs	390
12.2 Cloning a job	391
12.3 Stopping a job	391
12.4 Editing a job	392
12.5 Deactivating a job	393
12.6 Activating & Deactivating multiple jobs.....	394
12.7 Deleting a job.....	394
12.8 Monitoring Jobs	395
12.8.1 <i>Viewing Workflow Details</i>	395
12.8.2 <i>Viewing Job Run Statistics</i>	397
12.8.3 <i>Viewing Advanced Job Run Statistics</i>	399
13 Appendix A: Rules for Data Inclusions and Exclusions.....	402
13.1 HDFS Regular Expressions	402
13.1.1 <i>Inclusion for HDFS</i>	403
13.1.2 <i>Exclusions for HDFS</i>	403
13.2 Hive Regular Expressions.....	403
13.2.1 <i>Inclusions for Hive</i>	404
13.2.2 <i>Exclusions for Hive</i>	405
14 Appendix B: EC2 Installation Cheat Sheet	406
14.1 Setting Up Multi-DC Configuration	406
14.2 Verifying the Data Source of EC2 Cassandra Cluster	407
15 Appendix C: Quick Troubleshooting	408
15.1 Job Log Collection Tool.....	408
15.2 Masking dialog box does not respond	409
15.3 Valid revisions for objects not available	410
15.4 Email not received after job is complete or failed.....	410
15.5 Hive row count does not match after recovery	411
16 Appendix D: Mirroring Workflow in TDE/Encrypted Environment in Cassandra	412

17 Appendix D: Mirroring Workflow in TDE/Encrypted Environment in Hadoop	415
17.1 HDFS	415
17.2 HBase.....	415
17.3 HIVE	415
18 Appendix E: Couchbase Point-in-Time (PIT) Recovery Limitation.....	417

1 Introduction: Imanis Data software

Imanis Data software is an enterprise-grade big data lifecycle management solution. Imanis Data software functions as your secondary cluster, referred to as the Imanis Data cluster, to seamlessly blend with your current big data infrastructure. Supported data repositories include Hadoop (HDFS, Hive, and HBase), Cassandra, Couchbase, Amazon Glacier, Amazon S3, Azure Blob Storage, and MongoDB.

Imanis Data software is installed on commodity hardware running Linux and provides policy-based data management capabilities for backup, archive, data pipeline and data mirroring between Hadoop clusters. The scale-out architecture of Imanis Data software enables enterprises to operate multiple nodes as a single cluster and grow the cluster non-disruptively to an Exabyte scale level at a low price point.

The software archives old data to commodity hardware and Amazon Glacier, using 80% less storage. Using compression and de-duplication, Imanis Data software typically provides 5X or better optimization. Imanis Data software provides an easy way to set up automated policies to do backup, recovery, and archive data on a consistent basis. The software provides integrated network throttling and the ability to pre-set “blackout” times to reduce impact on the source. This feature frees up the network and reduces impact on the production systems, freeing data administrators from having to manually throttle back systems during peak usage on production systems.

Imanis Data Software Solves Big Data Challenges

Imanis Data software can be used in a big data environment for a variety of data management purposes:

- Backup & Data Recovery
- Archiving
- Data Pipeline
- Data Mirroring
- Direct Replication

Backup & Data Recovery

Imanis Data software provides robust and comprehensive protection for your primary data store. With its data backup and restore functionality, Imanis Data software helps protect your data in case of hardware failure, site failures, and user or application errors. For disaster recovery (DR), Imanis Data software provides automatic, efficient, and frequent data replication to the DR site. Enterprises can quickly restore should there be a disaster at a production site as data is fetched from the storage tier where it resides.

There are different types of failures that can occur more often than expected, such as user errors, software bugs, ransomware, application corruptions, hardware failures, and site failures. Not having a solution in place can have catastrophic consequences.

Imanis Data software solves these challenges and gives confidence to enterprises. For backup, Imanis Data software provides automatic, efficient, and frequent backups to the Imanis Data cluster. Furthermore, the Imanis Data storage optimization can store de-duplicated data on S3/Blob Storage thus reducing your storage footprint even further. Enterprises can restore from the most current backup or to a specific point in time with the Imanis Data FastFind technology.

Key Benefits:

- Rapid Restore: Leverages Imanis Data FastFind so enterprises can find the backed-up data and restore it quickly
- Erasure Coding: Ensures data durability so enterprises can withstand unexpected hardware failures on the Imanis Data cluster
- Storage Optimization: Typically, 5X or greater (compared to standard Hadoop) with compression, de-duplication, and erasure coding thus saving significant money and resources on CAPEX and OPEX
- App-Aware Integration: Ensures that the data and metadata (database schema, file attributes, etc.) are backed up and restored
- Incremental Forever: Ensures incremental changes on the production systems are tracked and captured on the target, which leads to faster transfers and lower network traffic
- Scale-Out Architecture: Enables enterprises to start small and grow non-disruptively to exabyte-scale at a lower price point thanks to the flexible Imanis Data deployment model that runs on bare metal or in virtualized environments, and on-premises or in the Cloud
- Archiving

With data growing exponentially, enterprises must likewise grow their Hadoop cluster, which requires adding more nodes, resulting in higher hardware and software costs, as well as higher operations cost of managing the cluster.

Consider the following scenario: Take a 100-node cluster with an estimated annual data growth rate of 30% for three years. After the third year, the Hadoop cluster consists of 220 nodes, with 120 nodes as the incremental difference. With hardware, licensing, and other expenditures, the cost of storing all of that data could easily top \$1M. That's a significant expense! Imanis Data software provides the ideal solution to prevent Hadoop cluster sprawl. Imanis Data software enables enterprises to set up automated user-defined policies for archiving and retention, using significantly less storage.

With Imanis Data software, enterprises can automate the process. Let's take an example of an enterprise that analyzes three months (90 days) worth of click-stream data on their production system. At month four, the data becomes less relevant. Instead of keeping the data on the production cluster, the enterprise can archive it to the Imanis Data cluster. After one year, the data can be moved to Amazon Glacier.

Key Benefits:

- Rapid Restore: Leverages Imanis Data FastFind so enterprises can find the backed-up data and restore it quickly
- User-Defined Policies: Eliminates manual scripting and unnecessary waiting
- App-Aware Integration: Ensures that the data as it appeared in the production source is maintained on the destination cluster
- Cloud Integration: Enables enterprise to seamlessly archive and retain data on to Imanis Data Cluster or Cloud (Amazon Glacier)

Data Pipeline

The benefits of adopting Big Data technologies like Hadoop and Cassandra are significant: the ability to analyze, derive insights from large quantities of data. But implementing a Big Data approach takes a team: administrators, engineers, scientists, architects, and others. Realizing the benefits of these new platforms requires not only managing production environments, but also non-production environments for test and development purposes. And therein lies the challenge. The development and data science teams need relevant samples of the production data for analysis and application development.

The greatest challenge lies in being able to sync the data between production and non-production sandboxes. The ‘development’ and ‘data science’ teams require current data but must wait for administrators to set up the pipeline, which can take days or weeks. To compensate, engineers and data scientists manually write data pipeline scripts, which wastes precious engineering resources and takes significant time away from development and analytics work.

Enterprises must ask: Should engineers and data scientist manage the data or analyze the data? The solution is Imanis Data software that intelligently and automatically manages the data flow from production to multiple test and development sandboxes, enabling engineers and data scientists to spend their time analyzing data, not managing it.

Key Benefits:

- User-Defined Policies: Eliminates manual scripting and unnecessary waiting
- App-Aware Integration: Ensures that the data and metadata as it appeared in the production source is maintained on the destination cluster
- Retention Operations: Enables enterprises to manage sandbox environment longevity with user-defined policies

Data Masking

To meet compliance guidelines, enterprises are required to conceal private and sensitive data such as credit card or social security numbers in testing or development environments. In this case, a technique referred to as

data masking is used to conceal private and sensitive data so that unauthorized users do not have access to the actual data.

Using Imanis Data software, you can mask various types of data, such as id, name, address, credit card number, social security number, and so on. For example, you can mask a 16-digit credit card number with a randomly generated 16-digit number before copying the data to the test cluster.

Data Masking enables enterprises to comply with data privacy and protection mandates, such as Sarbanes-Oxley, Payment Card Industry (PCI), Data Security Standard (DSS) and Health Insurance Portability and Accountability Act (HIPAA) which restrict the use of actual customer data.

Key Benefits:

- Meeting Compliance Guidelines: Enterprises can meet strict compliance guidelines easily
- Zero Scripting: Imanis Data software provides an automated and intelligent data syncing solution between production and non-production sandboxes thus eliminating scripting completely

Data Sampling

At times, enterprises have a need to retrieve a subset of their production data for testing, application development and training environments. In this case, a technique referred to as data sampling is used to select a smaller, targeted subset of the production without needing the entire production data.

For example, a table has 1 million rows on the primary cluster and you want to copy a certain number of rows to a test cluster. Imanis Data software enables you to execute this task.

Key Benefit:

- Retrieving Subset of Production Data Made Easy: Ensures that smaller, targeted subsets of production are made available for testing environments without needing the entire production data

Data Mirroring

Data Mirroring is defined as the process of copying data from one cluster to a different cluster in one single, unified step. In Data Mirroring, the data recovery process starts as soon as the data backup process is completed, unlike the Data Pipeline process where the data backup process and the data recovery process can be configured to have their own respective operating frequency and schedule.

Key Benefits:

- App-Aware Integration: Ensures that the data as it appeared in the production source is maintained on the target
- Efficient Disaster Recovery (DR): Provides automatic replication to the data recovery site, thus allowing enterprises to quickly and efficiently restore the data should there be a disaster at a production site.
- Retention Operations: Enables enterprises to manage sandbox environment longevity with user-defined policies

Direct Replication

You can replicate the exact copy of your data directly from your primary cluster onto to the destination cluster.

Key Benefits:

- Fast and Secure: Provides data migration in a secure environment
- Simple and Efficient: Clean and simple interface allows you to setup the direct replication job quickly and execute the workflow

Features & Benefits of Imanis Data software

- The Imanis Data platform is based on innovative technology that enables the intelligent orchestration and storage of data throughout its lifecycle.
- Single Pane of Glass is a single dashboard location to manage data on multiple repositories: Enables administrators the ability to manage all the data workflows, such as archiving, backup, restore, and test/dev management from a single location, saving precious time and resources
- Storage Optimization provides significant storage reduction using compression, de-duplication, and erasure coding, typically realizing 5X or better optimization: With 5X reduction, enterprises save significant money and resources on CAPEX and OPEX
- Scale Out Architecture using commodity hardware is the design model of Imanis Data software, instead of a scale up architecture and proprietary hardware: Enables enterprises to start small and grow non-disruptively to exabyte-scale at a lower price point thanks to the flexible Imanis Data deployment model that runs on bare metal or in virtualized environments, and on-premises or in the Cloud
- Data Durability provides erasure coding capabilities to mitigate the multiple drive and server failures that can lead to possible data loss: Enables enterprises to withstand failures because erasure coding offers a greater failure threshold than what Hadoop and other big data solutions offer
- Incremental Forever means just the incremental changes in the production systems are tracked and captured: Reduces the overall data movement because only the incremental changes are tracked and moved, which leads to faster transfers and lower network traffic
- User-Defined Policies provide automation and control over what, how, and when to move data: Eliminates having to manually write and execute scripts for data pipelines, archiving, backup, restore, and other data lifecycle management (DLM) tasks
- Seamless Cloud Integration means moving data to and from Glacier, S3, Azure Blob Storage or Azure Data Lake is as easy as moving data to local data stores: Imanis Data optimizes data management for the cloud enabling enterprises to integrate Cloud as part of their overall big data solution thus reducing their storage footprint even further
- Rapid Restore gives users a way to rapidly restore from an event because Imanis Data software keeps a catalog of all the data: Eliminates having to remember where the data was archived, or backed up by using the Imanis Data FastFind tool, which enables rapid restoration of the data

- Application Aware means that the data schema from the source application is maintained all the way through to the destination cluster: Provides users with the ability to restore data at an application granular level
- Retention Operations allow enterprises to easily implement their data retention policies, setting detailed parameters throughout the data lifecycle: Enables enterprises to quickly and easily comply with any data retention policy
- Point in Time (PIT) Revisions: Provides continuous and ultra-granular backups thus enabling enterprises to quickly and easily restore the database to the state that it was at the specific date and time you identified when restoring your database

2 Overview of Imanis Data Software

This section discusses the architecture and technology behind Imanis Data software. You can also view the version compatibility matrix to understand the various applications and their respective versions that are supported by Imanis Data.

2.1 Imanis Data Deployment

The following graphic describes a typical Imanis Data deployment in an infrastructure:

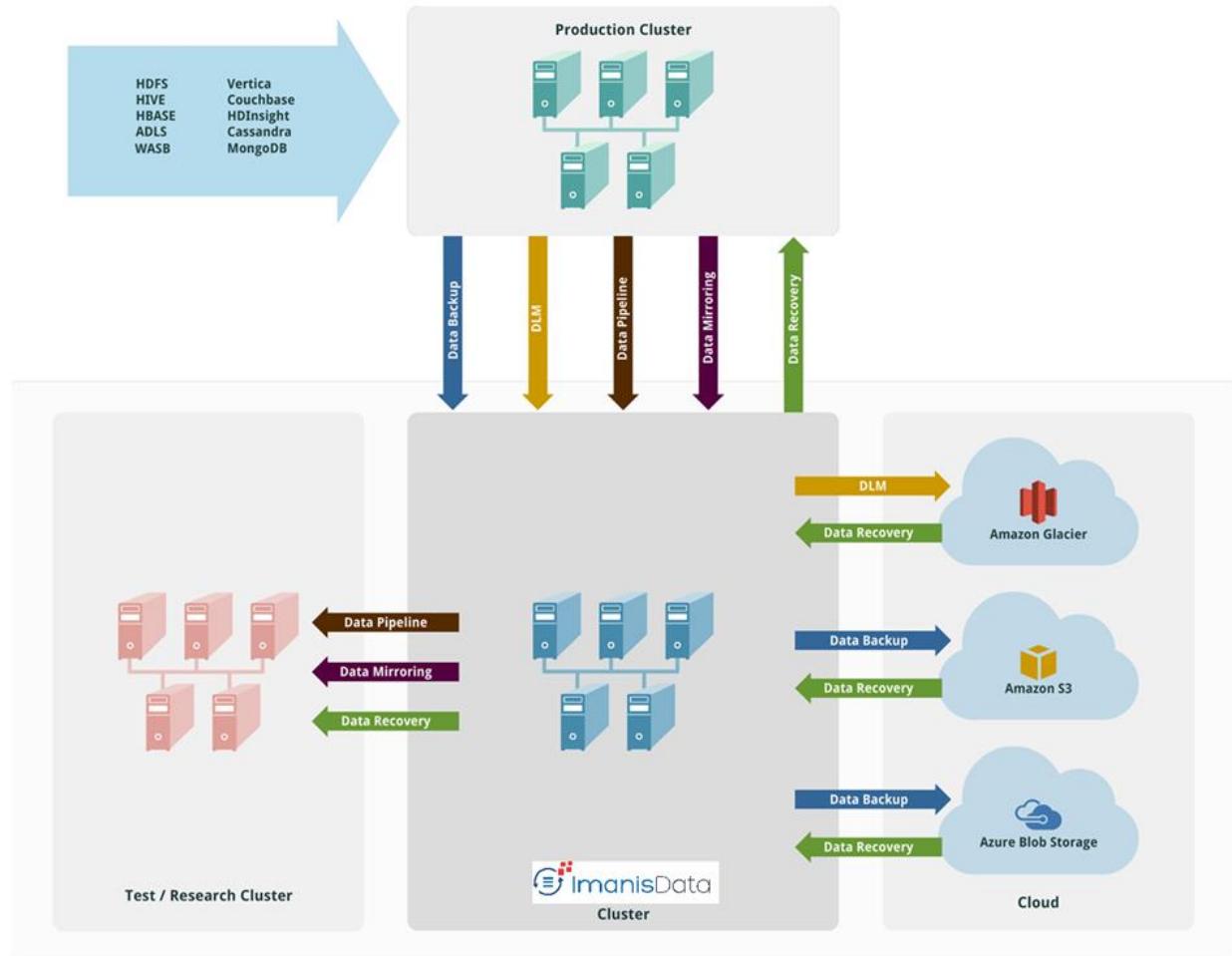


Figure 1: Cohesity Imanis Data Deployment

1. Intelligent Storage Layer

The Intelligent Storage Layer stores archive and backup data in a distributed file system. The storage layer uses techniques such as global block level de-duplication to find duplicate patterns within the incoming data and then stores it only once on the cluster. Data is then compressed using aggressive compression algorithms. Erasure coding improves data redundancy in a fault-tolerant manner by distributing data fragments across multiple nodes of the Imanis Data cluster.

- a. Block-level De-duplication: Imanis Data software incorporates specialized data reduction techniques that identifies and remove duplicate blocks within data without compromising its fidelity, and then stores only the unique data segments each time data is ingested into the Imanis Data cluster. Imanis Data software uses “post-process de-duplication” to identify redundant data. In this procedure, data is stored in the Imanis Data cluster upon ingest and later de-duplicated offline
- b. Compression: Once the data has been de-duplicated, the data chunks are aggressively compressed, further reducing the storage footprint
- c. Erasure Coding: Once data has been de-duplicated and compressed, data redundancy is improved by applying a technique called erasure coding. Imanis Data software erasure coding breaks data into six fragments, creates three additional redundant fragments, and stores these nine fragments across different Imanis Data nodes. This protects data from up to 3 node failures and only requires 50% more storage overhead. In contrast, a 3-replica redundancy model protects data from up to 2 node failures and requires 200% more storage

The following example explains how the Intelligent Storage Layer works:

Scenario: An organization has 30 TB (10 TB * 3 replicas) of old data in a production Hadoop cluster. To free up storage space on the production cluster, data is moved to the Imanis Data cluster.

2. Data Movers

Data movers are used to move data from the application to the Imanis Data cluster while maintaining application-level data consistency. Data movers bring data into the Imanis Data cluster through the following workflows: Data Backup, Data Recovery, Data Pipeline, and Data Mirroring

- a. Data Backup and Recovery: The Imanis Data file system is built on a storage tiering model which optimizes data management for the cloud. Imanis Data can federate data over multiple tiers of storage transparently based on user defined policies. For example, Imanis Data file system transparently migrates data between different tiers when a user defines a backup policy to retain the data on Imanis Data cluster for 10 days and in the cold tier for six months. Imanis Data software also performs periodic, incremental backups of a primary data set. Once the incremental changes have been stored on the Imanis Data cluster, Imanis Data software creates a restore point by taking a snapshot of the file system. To restore data, a user invokes the Imanis Data UI, browses through the timeline of restore points, and selects the relevant point and the appropriate data set that needs to be recovered. Imanis Data file system presents a unified namespace across different storage tiers so if a user performs a data migration data from local storage to cloud storage using a policy, the data movement happens

asynchronously without the user needing to be aware of the underlying migration process thus making the process completely transparent

- b. Data Lifecycle Management (DLM): Using Imanis Data software, you can significantly reduce the size of the primary cluster by managing the number of replicas or archiving older data to the Imanis Data cluster based on user-defined policies. The DLM policy starts with full replica state in which data objects remain in the state for a specific number of days (user-defined) after which the number of replicas of a file on the primary cluster is reduced and a new copy is created for the Imanis Data cluster. In the reduced replica state, Imanis Data software periodically reduces the replica count by 1 on the primary cluster and increases the replica count by 1 on the Imanis Data cluster. This process is repeated automatically until the replica count on the data repository is zero, after which the data are archived to Imanis Data. The data objects stay on the Imanis Data cluster for a specific number of days (user-defined) after which the data objects are archived to Amazon Glacier and then deleted after a defined time frame.
- c. Data Pipeline: Imanis Data software enables you to set up data pipelines that sync data from production cluster to multiple test and development sandboxes by:
 - Masking sensitive customer data so that unauthorized users do not have access to the actual data
 - Sampling specific cross sections to create smaller, more targeted databases while being able to retain the original database's referential

Data Pipeline allows users to set up intelligent and automated data sync policies, instead of having to develop custom scripts or manual procedures. For example, you can configure a Data Pipeline workflow to take daily backups from primary cluster onto Imanis Data cluster and weekly recovery from Imanis Data cluster to the destination data repository. Thus, the Data Pipeline empowers engineers and data scientists to spend their time analyzing data, not managing it.

- d. Data Mirroring: Data Mirroring is defined as the process of copying data from one data repository to another in one single, unified flow. The data recovery process starts as soon as the data backup process is completed, unlike the Data Pipeline process where the data backup process and the data recovery process can be configured to have their own respective operating frequency.

2.2 Version Compatibility Matrix

Imanis Data software supports the following in RELEASE 4.3.0:

APPLICATION	SUPPORTED ENTERPRISE VERSIONS	
	CLOUDERA	HORTONWORKS
HADOOP	6.2.0	3.1.0
	6.1.1	3.0.1
	6.0.1	2.6.5
	5.16.1	2.6.4
	5.14.1	2.6.0
	5.12.2	2.5.4
	5.8.2	2.5.0
	5.7.6	2.3.6
	-	2.2.0

APPLICATION	SUPPORTED ENTERPRISE VERSIONS	
	APACHE	DATASTAX ENTERPRISES
CASSANDRA	3.11	6.7**
	3.10	5.1.10
	3.0	5.1.9
	2.1.11	5.1.8
	2.1.8	5.1.7
	2.1	5.1.2
	-	5.1.0
	-	5.0
	-	4.8

** Refer to the Cassandra limitations section In Release Notes for more information.

APPLICATION	SUPPORTED ENTERPRISE VERSIONS	
	COUCHBASE INC.	
COUCHBASE		6.0
		5.5.2
		5.1.0

APPLICATION	SUPPORTED ENTERPRISE VERSIONS
	COUCHBASE INC.
	5.0.1
	4.6.3
	4.6.2
	4.5.0
	4.1.1

APPLICATION	SUPPORTED ENTERPRISE VERSIONS
	MONGODB INC.
MONGODB	4.2**
	4.0
	3.6
	3.4

Table 1: Version Compatibility Matrix

** Refer to the MongoDB limitations section in Release Notes for limitations of MongoDB 4.2 support.

2.3 Data Repository Usage Matrix

Imanis Data software supports the following:

APPLICATIONS	DATA REPOSITORIES		CLOUD STORAGE		DESCRIPTION
	SOURCE	DESTINATION	GLOBAL	REGULAR	
Amazon Glacier	Not Available	Not Available	Not Available	Available	Available through DLM workflow only
Amazon S3	Not Available	Not Available	Available	Available	Available through Backup, Mirroring and Data Pipeline workflows only
Cloudian	Not Available	Not Available	Available	Available	Available through Backup, Mirroring and Data Pipeline workflows only
Azure Blob Storage	Not Available	Not Available	Available	Available	Available through Backup, Mirroring and Data Pipeline workflows only

Table 1: Data Repository Usage Matrix

Refer to the sections on Global Cloud Data Repository and Regular Cloud Data Repository for more information.

3 Getting Started with Imanis Data Software

This section describes the first-time logging in procedure and key GUI elements of Imanis Data software.

3.1 Logging in Imanis Data Software

You need the URL and login credentials to access the Imanis Data GUI.

When you are logging in for the first time, you will be asked to change the default password. This is a mandatory, one-time activity only, without which the Imanis Data GUI cannot be accessed.

If you are using the Evaluation License of Imanis Data, you can continue to log in and evaluate Imanis Data software for your purpose until the license gets expired. Once the Evaluation License gets expired, you must contact the Imanis Data Software Support Team for the renewal of the license.

To log in Imanis Data software, do the following:

1. Open Firefox browser on your local computer and type the URL, that you received, in the address bar. Cohesity Imanis Data software login page appears (see screenshot below):

The screenshot shows a login interface for the Imanis Data Software. It consists of a light gray rectangular form with a thin green border. Inside the form, there are three input fields: a 'Username' field containing 'admin', a 'Password' field containing several redacted dots, and a 'Domain' dropdown menu set to 'Imanis Data'. Below these fields is a large blue rectangular button labeled 'Sign In' in white text. At the very bottom of the page, outside the main form, is a footer bar with the text 'Imanis Data version4.1' and '© 2018 Cohesity, Inc.'

Figure 2: Login screen

2. Type the login credentials in the Username and Password fields and then click the **Domain** dropdown menu to select a domain.

3. Click the **Sign In** button to login. The **Password Reset** dialog box will be displayed. This dialog box is displayed only when you are logging into Cohesity Imanis Data for the first time.
4. In the **Reset Password** page, do the following:
 - a. Type the password that you received in the email in the **Current password** field.
 - b. Type a new password in the **New password** field and then type the new password again in the **Re-enter new password** field. Make sure that passwords match in the **New Password** and **Re-enter new password** fields.
 - c. Click the **Reset Password** button.

The dialog box has a light gray background and a thin green border. It contains the following elements:

- A title "Reset Password" centered at the top.
- A "Current password" input field with a horizontal line below it.
- A "New password" input field with a horizontal line below it.
- A "Re-enter new password" input field with a horizontal line below it.
- A blue rectangular button labeled "Reset Password" at the bottom right.

Figure 3: Password reset dialog box

5. Click the **Continue** button. The QuickNav page is displayed if you do not have data repositories, policies, or workflows defined in the system else the Dashboard page is displayed.

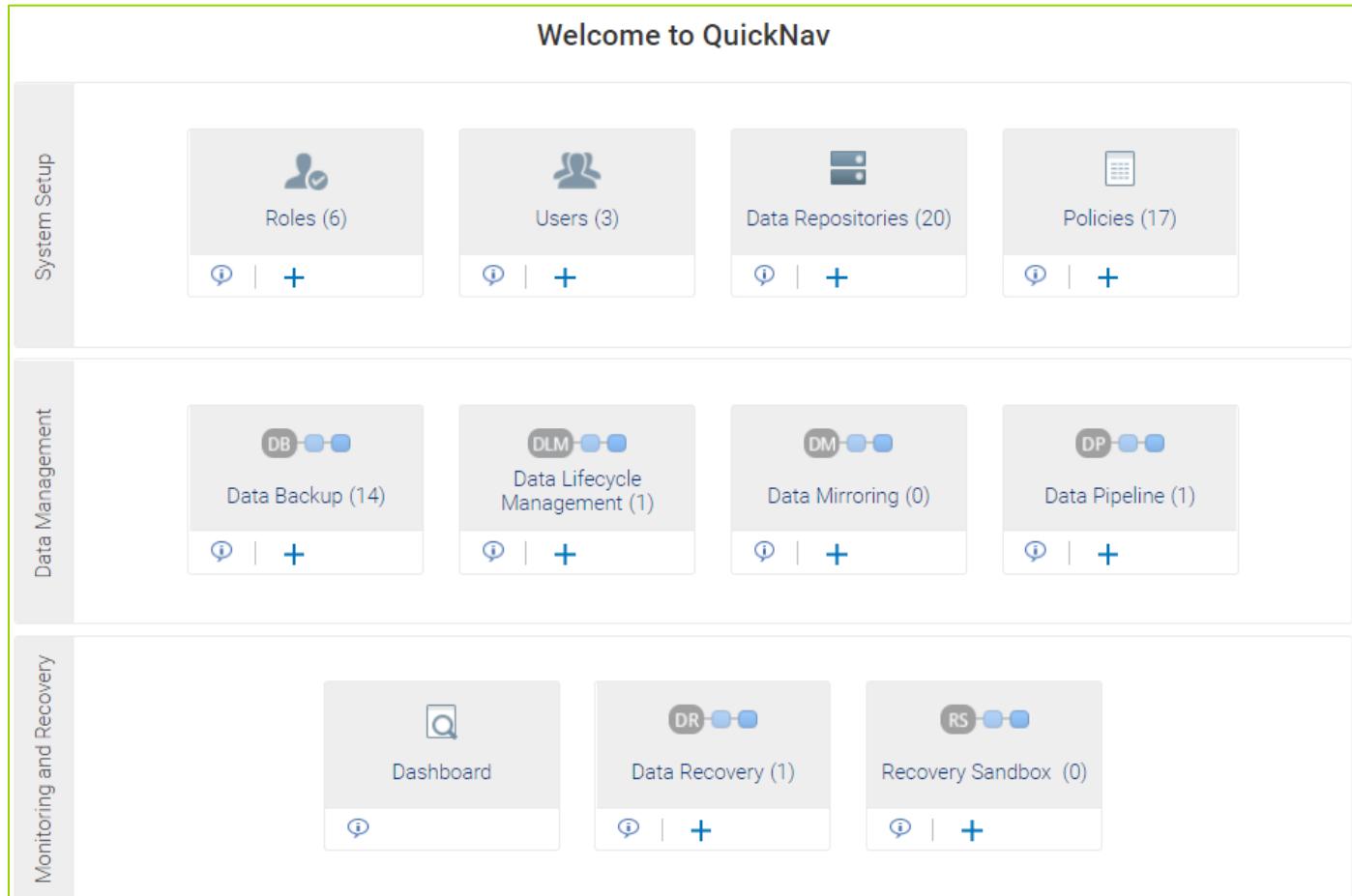
The dialog box has a light gray background and a thin green border. It contains the following elements:

- An icon of a green circle with a white checkmark.
- The word "Success" next to the icon.
- A message "Password successfully reset." below the icon.
- A blue rectangular button labeled "Continue" at the bottom right.

Figure 4: Password Successfully Reset dialog box

3.2 QuickNav

The QuickNav feature provides a user-friendly gateway to quickly set up and navigate Imanis Data software. You can quickly create roles and users, various workflows by clicking and add new data repositories and policies by the  icon. Also, view informative tooltips for each function by simply hovering the mouse over the  icon.



The screenshot shows the 'Welcome to QuickNav' interface. On the left, there are three vertical navigation panels: 'System Setup' (selected), 'Data Management', and 'Monitoring and Recovery'. The main area is divided into a 3x4 grid of cards. Each card has a title, a small icon, and two buttons at the bottom: a tooltip icon () and a plus icon ().

Welcome to QuickNav				
System Setup	 Roles (6)	 Users (3)	 Data Repositories (20)	
	 	 	 	
	 Data Backup (14)	 Data Lifecycle Management (1)	 Data Mirroring (0)	 Data Pipeline (1)
	 	 	 	 
Monitoring and Recovery	 Dashboard	 Data Recovery (1)	 Recovery Sandbox (0)	
		 	 	

Figure 5: QuickNav screen

3.3 HTTPS Support

Imanis Data UI supports HTTPS. Look for the padlock icon on your internet browser indicating that the session is running in a secure mode.

3.4 Main Menu

You can navigate through all the sub-menus in Imanis Data software through the Main Menu.

You may not have access to all the modules; the capabilities and privileges are assigned by role, and your role provides access to the modules you need to do your work. For more information, refer to the section **Users and Roles**.

Main Menu	
Monitoring and Recovery	<p>The Monitoring and Recovery menu enables you to:</p> <ul style="list-style-type: none">• Launch the Dashboard• Manage Data Recovery Workflows• Run Recovery Workflows in a Sandbox
Data Management	<p>The Data Management menu enables you to:</p> <ul style="list-style-type: none">• Manage Data Backup Workflows• Manage Data Lifecycle Management Workflows• Manage Data Mirroring Workflows• Manage Data Pipeline Workflows
System Setup	<p>The System Setup menu enables you to:</p> <ul style="list-style-type: none">• Set up users (local and domain users)• Manage roles• Set up domains• Configure SMTP• Manage data repositories• Manage policies
Audit Log	<p>The Audit Log menu enables you to:</p> <ul style="list-style-type: none">• View Audit Logs of Imanis Data cluster
About	<p>The About menu enables you to:</p> <ul style="list-style-type: none">• View License Type: (Evaluation or Production)• View License Expiry Date• View Days Remaining:

- View Licensee Company name

3.5 Dashboard

The dashboard captures and reports important information, in the form of interactive pie-charts, line graphs, and bubbles, so that administrators can effectively operate the Imanis Data cluster. Using the dashboard, administrators can track metrics like Data Movement, Storage Savings, Storage Consumption, Workflows, and System Health.

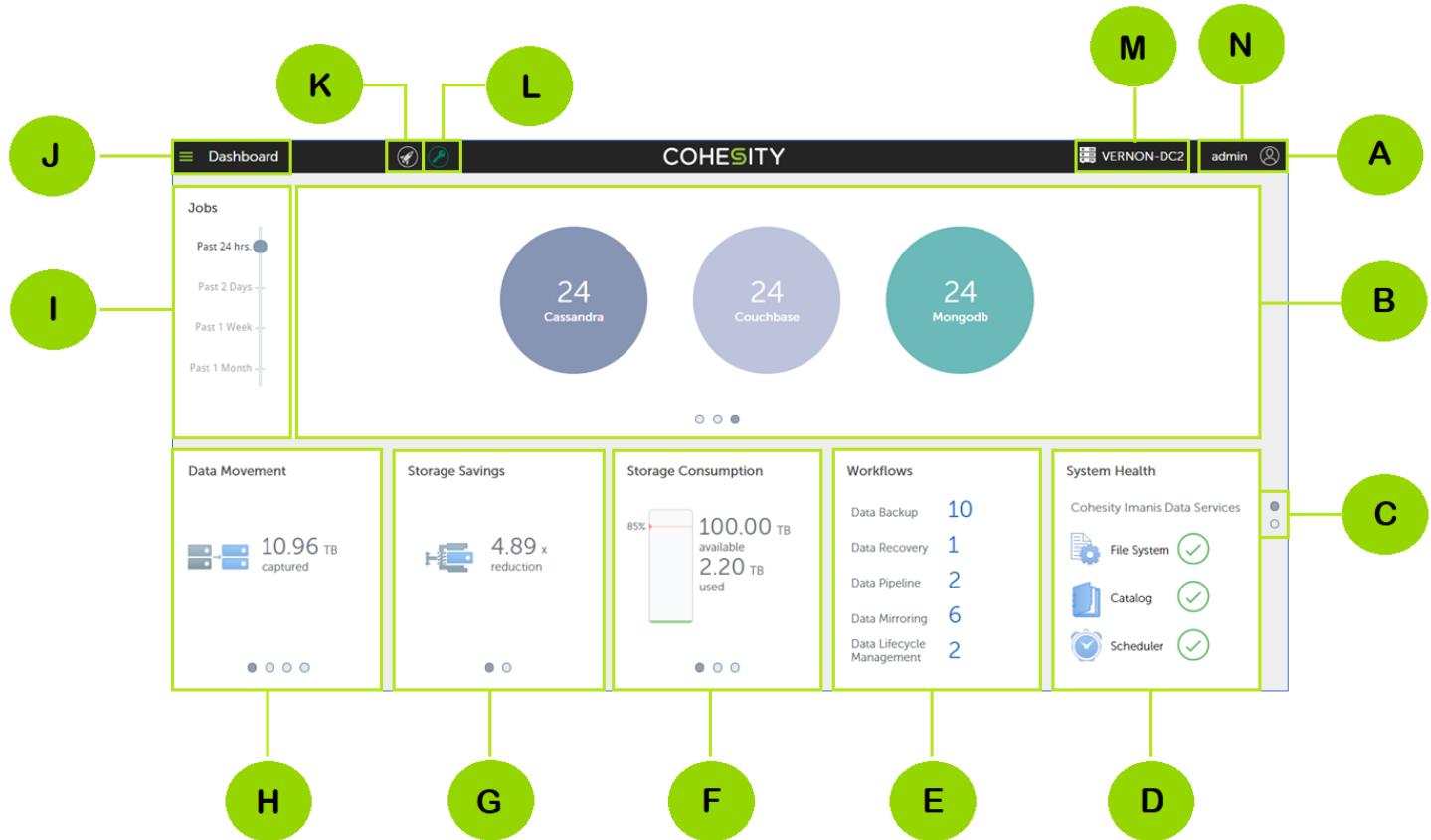


Figure 6: Dashboard screen

NOTE: The following screenshot is meant for illustration purposes only. If no jobs are configured, the Jobs area will appear blank, that is, job bubbles will not appear.

SYMBOLS	DESCRIPTION	ACTION TO TAKE
A	<ul style="list-style-type: none"> ▪ Profile tab 	<ul style="list-style-type: none"> ▪ Click the Profile tab to change the password and log out of Imanis Data software
B	<ul style="list-style-type: none"> ▪ Job Bubbles indicate the number of jobs that are in progress, failed, and completed 	<ul style="list-style-type: none"> ▪ Click on the individual job bubble to view detailed information such as job name, type, application, data repository, status, job start and end date, and the progress bar. See an example screenshot here. ▪ Click on the small dots right under the bubbles to see a granular view of the high-level information which is grouped accordingly: <ul style="list-style-type: none"> ▪ Grouped by stats (the default view) ▪ Grouped by job type (number of backup, restore, and DLM jobs) ▪ Grouped by application type (number of HDFS, Hive, Cassandra jobs and so on)
C	<ul style="list-style-type: none"> ▪ Imanis Data All Jobs Details 	<ul style="list-style-type: none"> ▪ Click on the small dot right under the graphic to see a granular view of each “completed” and “failed” job (see the screenshot and description table below)
D	<ul style="list-style-type: none"> ▪ System Health indicates the health (state) of the Imanis Data services, such as File System, Catalog, and Scheduler 	<ul style="list-style-type: none"> ▪ The  icon indicates the service is up and running so NO action is required ▪ The  icon indicates the service is down. In this case, you must contact the Imanis Data Administrator

SYMBOLS	DESCRIPTION	ACTION TO TAKE
	<ul style="list-style-type: none"> ▪ Workflows indicate the total number of workflows 	<ul style="list-style-type: none"> ▪ Click the number corresponding to the workflow to view details. For example, if you click the number, say 8, against the Data Recovery workflow, the Data Recovery workflow page is displayed where you can view all the eight workflows
	<ul style="list-style-type: none"> ▪ Storage Consumption indicates the available (free) space and used space in the Imanis Data cluster 	<ul style="list-style-type: none"> ▪ Click on the small dot right under the graphic to view a line graph that indicates the storage consumption (Y axis) made over a period of time (X axis) on the Imanis Data cluster
	<ul style="list-style-type: none"> ▪ Storage Savings indicates the total size of data reduction achieved on the Imanis Data cluster. For example, the size of the data moved from primary cluster to the Imanis Data cluster is 1 TB; however, Imanis Data software uses its proprietary technology to reduce the size to 500 GB, so your storage savings show a 2x reduction 	<ul style="list-style-type: none"> ▪ Click on the small dot right under the graphic to view a line graph that indicates the storage savings (Y axis) made over a period of time (X axis) on the Imanis Data cluster ▪
	<ul style="list-style-type: none"> ▪ Data Movement indicates the size of the data moved from primary clusters to the Imanis Data cluster, for example, 1000 MB backed up 	<ul style="list-style-type: none"> ▪ Click on the small dots right under the graphic to view more granular, drilled down information about the backed up data in the form of interactive pie-charts
	<ul style="list-style-type: none"> ▪ Imanis Data Job History 	<ul style="list-style-type: none"> ▪ Move the slider to view Imanis Data job history from the past 24 hours to 1 month. ▪ For example, if you move the slider to the 'Past 24 hrs' option, the Job Bubbles dynamically change to indicate the number of jobs that are in progress, completed, and failed from the past 24 hours

SYMBOLS	DESCRIPTION	ACTION TO TAKE
J	<ul style="list-style-type: none"> Main Menu 	<ul style="list-style-type: none"> Click the Main Menu to access all the sub-menus like Monitoring, Recovering, Backup, Data Management, System Setup and so on. Refer to the section Imanis Data License Types for more information
K	<ul style="list-style-type: none"> QuickNav 	<ul style="list-style-type: none"> The default landing page when there are no workflows are configured in the Imanis Data cluster. However, using QuickNav users can quickly create roles and users, add new data repositories and policies, create data backup, data lifecycle, data mirroring, and data pipeline workflows
L	<ul style="list-style-type: none"> License Information 	<ul style="list-style-type: none"> Click the License Key Icon to view license validity information for each of the applications in your system. Refer to the section Imanis Data Licensing for more information.
M	<ul style="list-style-type: none"> Cluster Name 	<ul style="list-style-type: none"> Name of the cluster where you have installed or set up Imanis Data

Table 2: Dashboard Table

Cassandra Jobs (24)							
Job Name	Type	Application	Data Repository	Status	Started On	Ended On	Progress
Cass_Hourly2	Backup	Cassandra	Cassandra Production	Completed	2019-09-03 11:00:25 AM	2019-09-03 11:05:22 AM	<div style="width: 90%;">00:04:57</div>
Cass_Hourly2	Backup	Cassandra	Cassandra Production	Completed	2019-09-03 10:00:25 AM	2019-09-03 10:05:45 AM	<div style="width: 95%;">00:05:20</div>
Cass_Hourly2	Backup	Cassandra	Cassandra Production	Completed	2019-09-03 09:00:25 AM	2019-09-03 09:04:57 AM	<div style="width: 93%;">00:04:32</div>
Cass_Hourly2	Backup	Cassandra	Cassandra Production	Completed	2019-09-03 08:00:24 AM	2019-09-03 08:05:49 AM	<div style="width: 98%;">00:05:24</div>

Figure 7: Jobs screen

FIELDS & BUTTONS	DESCRIPTION	ACTION TO TAKE
▪ Job Name	▪ Displays the name of the job	▪ Click on the job name to view historical stats of that job in the Stats tab. ▪ The  icon indicates that some of the data objects were skipped due to errors during the job run. Refer to the screenshot here.
▪ Type	▪ Displays the job type such as backup, restore, DLM, and so on.	▪ NA
▪ Application	▪ Type of application on which the job is being executed. Possible values: HDFS, Hive, HBase, Cassandra, Couchbase, MongoDB, Azure.	▪ NA
▪ Data Repository	▪ Displays the name of the data repository against which the job is running.	▪ NA
▪ Status	▪ Status of the job. Possible Values: Completed, Failed, and Running.	▪ Click the  icon to view the Imanis Data Logs and Yarn Logs. You can also see the Start Time - End Time of the job and the Job Run ID. ▪ Refer to the screenshot here.
▪ Started On	▪ Job start date and time.	▪ NA
▪ Ended On	▪ Job end date and time. The end date and time appear only for completed and failed jobs.	▪ NA

FIELDS & BUTTONS	DESCRIPTION	ACTION TO TAKE
▪ Progress	▪ Displays the elapsed time of a job.	▪ Hover the mouse on the progress bar to see percent complete for running jobs ▪ Click the Stop icon  to stop the job. Refer to the section Stopping a Job for more information

Table 3: Jobs screen table

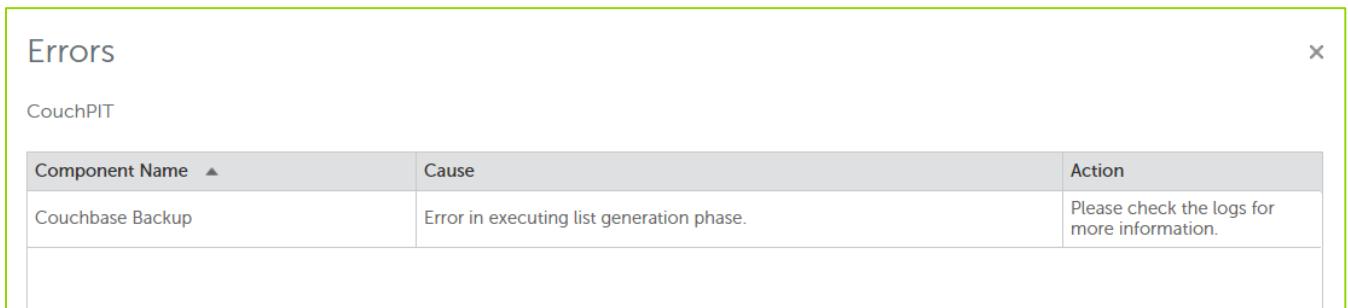


Figure 8: Viewing Job Errors Screen

Logs X

Cass_Hourly2

● Completed
 Start time: 2019-09-03 11:00:25 AM
 End time: 2019-09-03 11:05:22 AM
 Job Run ID: 1567508425488

<div style="border-bottom: 1px solid #ccc; padding: 5px;">▼ Cohesity Imanis Data Logs</div> <div style="border-bottom: 1px solid #ccc; padding: 5px; background-color: #e0f2fd;">■ Data Mover</div> <div style="border-bottom: 1px solid #ccc; padding: 5px;">▼ Yarn Logs</div> <div style="border-bottom: 1px solid #ccc; padding: 5px;">▶ node0</div> <div style="padding: 5px;">node0_YC_0</div> <div style="padding: 5px;">node0_YC_1</div> <div style="padding: 5px;">node0_YC_2</div> <div style="border-bottom: 1px solid #ccc; padding: 5px;">▶ node1</div> <div style="border-bottom: 1px solid #ccc; padding: 5px;">▶ node2</div>	<pre> 2015-11-27 00:14:11 INFO mapreduce.Job:Running job: job_1448610330496_0030 2015-11-27 00:14:18 INFO mapreduce.Job:Job job_1448610330496_0030 running in uber mode : false 2015-11-27 00:14:18 INFO mapreduce.Job: map 0% reduce 0% 2015-11-27 00:14:27 INFO mapreduce.Job: map 100% reduce 0% 2015-11-27 00:14:27 INFO mapreduce.Job: Job job_1448610330496_0030 completed successfully 2015-11-27 00:14:28 INFO mapreduce.Job: Counters: 30 File System Counters FILE: Number of bytes read=0 FILE: Number of bytes written=139769 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=7476 HDFS: Number of bytes written=0 HDFS: Number of read operations=3 HDFS: Number of large read operations=0 HDFS: Number of write operations=0 Job Counters Launched map tasks=1 Data-local map tasks=1 Total time spent by all maps in occupied slots (ms)=6560 Total time spent by all reduces in occupied slots (ms)=0 Total time spent by all map tasks (ms)=6560 Total vcore-seconds taken by all map tasks=6560 Total megabyte-seconds taken by all map tasks=26869760 Map-Reduce Framework Map input records=0 Map output records=0 Input split bytes=242 Spilled Records=0 Failed shuffles=0 Merged Map outputs=0 GC time elapsed (ms)=0 CPU time spent (ms)=2690 Physical memory (bytes) snapshot=476782592 virtual memory (bytes) snapshot=4331425792 Total committed heap usage (bytes)=2026373120 File Input Format Counters Bytes Read=7234 File Output Format Counters Bytes Written=0 2015-11-27 00:14:28 INFO zookeeper.ZooKeeper: Initiating client connection, connectstring=talena- 33:2181,talena-41:2181,talena-42:2181 sessiontimeout=60000 watcher=org.kiji.schema.layout.impl.ZooKeeperClient\$sessionwatcher@4c380929 2015-11-27 00:14:28 INFO zookeeper.ClientCnxn: opening socket connection to server talena- 33/10.1.10.23:2181.will not attempt to authenticate using SASL (unknown error) 2015-11-27 00:14:28 INFO zookeeper.ClientCnxn: socket connection established to talena-33/10.1.10.23:2181, initiating session </pre>
--	--

Figure 9: Viewing Job Logs screen

3.6 Imanis Data Licensing

Imanis Data supports component-based licensing. There are two types of Imanis Data licenses: Evaluation and Production.

Evaluation License: The Evaluation license is valid for a limited period. Typically, 30 days.

Users will NOT be permitted to login to Imanis Data software if none of the license are active. To extend the trial period of the evaluation license or upgrade the evaluation license to a production license, contact Imanis Data Technical Support.

Production License: The Production license is valid for a cluster for a specified period as per the licensed duration. Once the production license expires, the user can still login to Imanis Data software.

The following section illustrates what a user can do and cannot do when the production licenses expires:

When the Production license expires, a user CAN ...

- Use the Recovery and Recovery Sandbox workflows
- Access cloud repositories like Azure, Glacier, and S3
- Edit workflows or jobs related to databases whose licenses have expired

When the Production license expires, a user CANNOT...

- Create new data repositories or create new workflows of that database type. For example, if the Cassandra database type license expires then user will not be permitted to add new source data repository or create new backup workflow of Cassandra database type.
- Rerun an existing one-time job or resume previously suspended workflows related to any databases

The following dialog box is displayed when the Key icon present on the Dashboard is clicked:



Figure 10: Viewing Licence Information dialog box

3.7 Smart System Notifications

Smart system notifications related to various events are displayed on the Imanis Data GUI from time to time. You can click the notification band to view detailed information about the corresponding event and take appropriate action to resolve it.

3.7.1 Recovery Alert Notification Band

The recovery alert notifications feature displays a list of backup workflows for which corresponding workflows were not run in a while. The Recovery Alert Notification Band is displayed on the page in the following scenarios:

- Recovery workflow for a backup was never run
- Recovery workflow for a backup was unsuccessful
- Recovery workflow for a backup was run 60 days ago

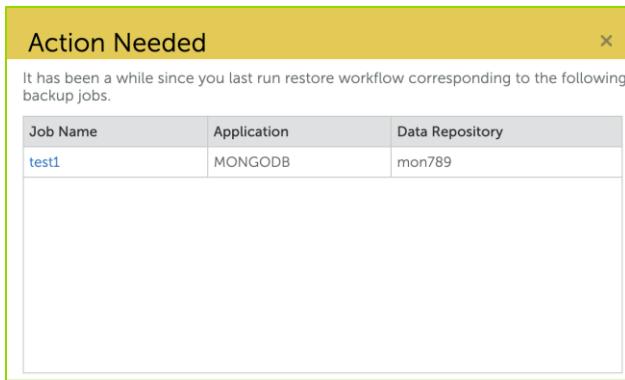
To view recovery alert information, do the following:

1. Click the **Check Details** button on the Recovery Alert Notification Band (see the following screenshot).



⚠ It's been a while since you last run restore workflow. [Check Details](#)

2. In the **Action Needed** dialog box, view the list of corresponding backup workflows for which recovery workflows must be run.



3.8 System Notification Band

You can click on the band to view information regarding the failed component(s).

Contact the Imanis Data Software Support Team to resolve any issues with Imanis Data services.

To view system notification information, do the following:

1. Click the **Check Details** button on the System Notification Band (see the following screenshot).


System requires immediate attention. [Check Details](#)
2. In the **Action Needed** dialog box, view detailed information regarding the corresponding event or events such as severity, component, status, and condition of the specific service or services.

4 Users & Roles

This chapter describes the process of configuring Imanis Data software once it has been installed in your infrastructure. Once the setup is complete, you can access the System Setup menu in Imanis Data software to manage data in your big data environment. The System Setup menu is a gateway for managing users, roles, data repositories, and policies. Data Repositories and Policies are explained in a different chapter.

To access the Users and Roles menu, do the following:

- Click the **Main Menu**  > **System Setup**.

Imanis Data software user management supports local users and LDAP/active directory users for authentication. The following steps will help you to get started with managing users:

- Adding SMTP
- Adding a domain
- Creating a role
- Adding a user and assigning a role

4.1 SMTP Config

SMTP stands for Simple Mail Transfer Protocol. Configure SMTP in Imanis Data software to enable email notification for user management, job completion, job failures, and daily digests of system activity.

4.1.1 Adding SMTP

This section describes the process of adding SMTP server.

NOTE: Imanis Data software permits you to configure only one SMTP server.

To add an SMTP server, do the following:

1. Click the **Main Menu** > **System Setup** > **User** > **SMTP Config** tab. The following page appears:

Figure 11: SMTP Configuration Default Screen

2. In the **SMTP Config** page, click the **+ Configure** button or the **+** icon to configure an SMTP server.
3. Type a user name in the **User Name** field and other relevant information about your SMTP server. For example, see the following screenshot:

Figure 12: SMTP Configuration Screen

4. Click **Save**. A confirmation message will be displayed on the page indicating SMTP Configuration is saved successfully.

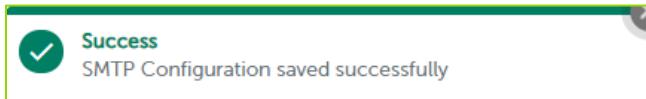


Figure 13: SMTP Configuration Successfully Saved Configuration Dialog Box

4.1.2 Editing SMTP server

Imanis Data software enables you can edit the parameters of the SMTP server.

To edit an SMTP server, do the following:

1. Click the **Main Menu** > **System Setup** > **User** > **SMTP Config** tab.
2. In the left pane, identify the SMTP server that you want to edit and select it.
3. Click the edit icon to activate the editing mode.
4. Make the appropriate changes and click **Save**.

4.1.3 Deleting SMTP server

Imanis Data software enables you to delete the parameters of the SMTP server.

To delete an SMTP server, do the following:

1. Click the **Main Menu** > **System Setup** > **User** > **SMTP Config** tab.
2. In the left pane, identify the SMTP server that you want to delete and select it.
3. Click the delete icon to remove the identified SMTP server. The following dialog box is displayed:



Figure 14: SMTP Delete Confirmation Dialog Box

4. On the dialog box, type the name of the SMTP server to confirm your decision to delete the SMTP configuration. Once you type the name, then click the **I understand the consequences, delete this SMTP Configuration** button is activated.

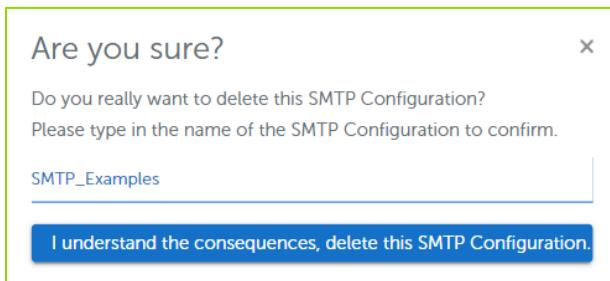


Figure 15: SMTP Delete Confirmation Dialog Box 2

5. Click the "**I understand the consequences, delete this SMTP Configuration**" button to delete your SMTP configuration permanently. A confirmation message will be displayed on the page indicating SMTP is successfully deleted.



Figure 16: SMTP Successfully Deleted Confirmation Dialog Box

4.2 Domains

Imanis Data software supports LDAP for user authentication. Currently only LDAP client authentication in anonymous and simple authentication modes is supported.

4.3 Adding a Domain

Imanis Data software enables you to add a domain through the User menu option.

To add a domain, do the following:

1. Click the **Main Menu**  > **System Setup** > **User** > **Domains** tab. The following page appears:

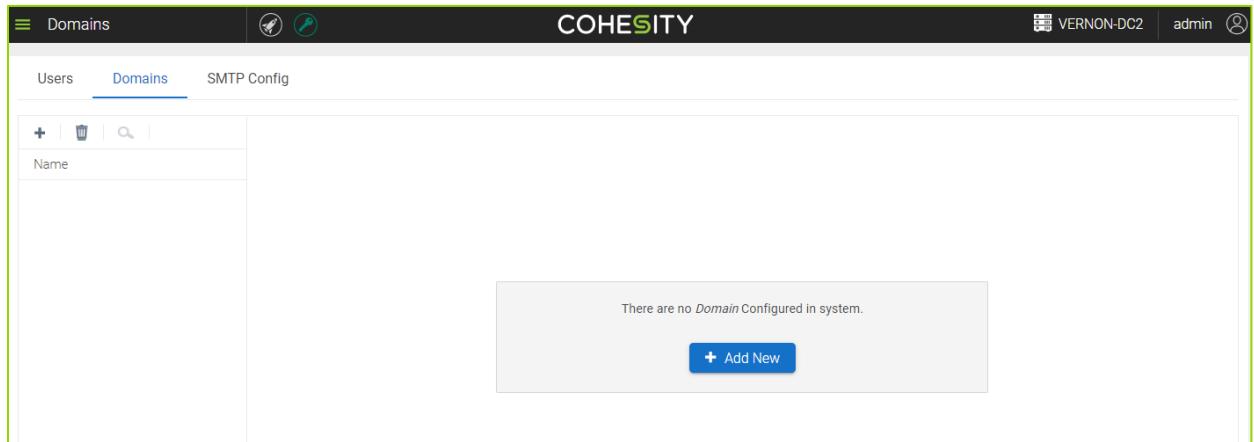


Figure 17: Domains Default Page

2. On the **Domains** page, click the **+ Add New** button or the **+** icon to add a domain. The following page is displayed: Type a domain name, LDAP address, port number, and Base DN in the respective fields:

Domain Name	LDAP	Save	Cancel
LDAP Address	10.1.10.63		
Port	389		
Base DN	dc=imanisdata,dc=com		
Unique Attribute Name	uid		
Mail Attribute Name	mail		

Figure 18: Configuring Domains

3. Type a domain name, LDAP address, port number, and Base DN in the respective fields:

BUTTONS & FIELDS	DESCRIPTION
Domain Name	<ul style="list-style-type: none"> Name of the domain
LDAP Address	<ul style="list-style-type: none"> Type the name of the server where LDAP is hosted.
Port	<ul style="list-style-type: none"> Type the LDAP port (the default port is 389)
Base DN	<ul style="list-style-type: none"> Type the top level DN of your LDAP directory tree
Unique Attribute Name	<ul style="list-style-type: none"> Type LDAP attribute name to uniquely search for user's DN (Distinguished Name). This attribute will map to the username field in Imanis Data. The default value is uid. Possible values: uid, sAMAccountName, etc.
Mail Attribute Name	<ul style="list-style-type: none"> Type LDAP attribute name that stores the email id of the user. <p>NOTE: If the specified mail attribute does not exist on the LDAP user account, then it won't be auto populated on a successful LDAP user search. In that case, the mail attribute needs to be entered manually. (Note: Email is needed for LDAP user addition into Imanis Data)</p>

Table 4: Domains Configuration Table

- Select the **Yes** option in the **Enable LDAP Authentication** field to connect to LDAP server using simple authentication mode:

Enable LDAP Authentication
<input checked="" type="radio"/> Yes <input type="radio"/> No

Figure 19: Enabling LDAP Authentication

- Type the LDAP administrator's username and password in the respective **Username** and **Password** fields to allow LDAP client simple authentication. Specify the complete DN of the admin user in the username field. For example, see the following screenshot:

Username
admin
Password
.....

Figure 20: LDAP Administrator Username & Password

6. Click **Save**. A confirmation message will be displayed on the page indicating Domain configuration is saved successfully.

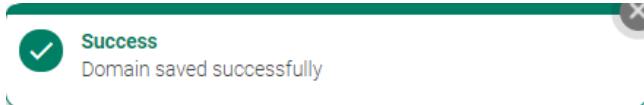


Figure 21: Domain Successfully Saved Confirmation Dialog Box

4.4 Editing a Domain

Imanis Data software allows you to change the LDAP address, port number, and Base DN of the domain.

To edit a domain, do the following:

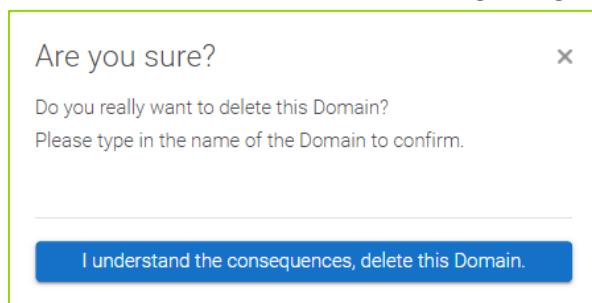
1. Click the **Main Menu** > **System Setup** > **User** > **Domains** tab.
2. On the **Domains** page, in the left-hand pane, identify the domain that you want to edit and select it.
3. Click the edit icon and make the required changes.
4. Click **Save**.

4.5 Deleting a Domain

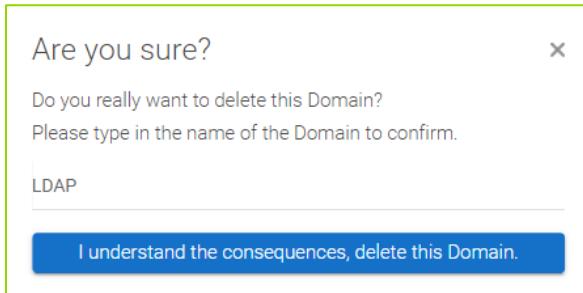
Imanis Data software enables you to delete a domain permanently from the user management console. Once a deleted, a domain cannot be retrieved.

To delete a domain, do the following:

1. Click the **Main Menu** > **System Setup** > **User** > **Domains** tab.
2. On the **Domains** page, in the left-hand pane identify the domain user that you want to delete, select it and then click the icon. the following dialog box is displayed:



3. On the dialog box, type the name of the domain to confirm your decision to delete it. Once you type the name, then the **I understand the consequences, delete this user** button is activated.



4. Click the **I understand the consequences, delete this domain** button to delete the role permanently. A confirmation message will be displayed on the page indicating LDAP is successfully deleted.



4.6 Users

A user account represents an entity who will use Imanis Data software. You can configure two types of users: A Local User, which is also called as Imanis Data User, and a Domain User.

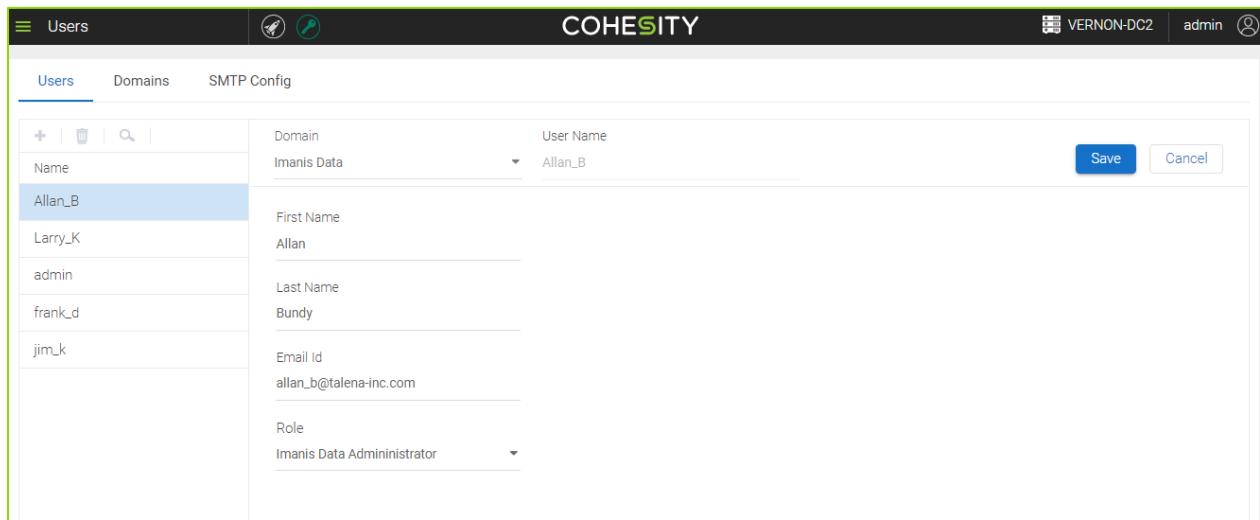
NOTE: If you add a user before configuring SMTP, Imanis Data software displays the following message: No SMTP configurations found. Password of the new user will be same as user name.

4.6.1 Adding a Local User & Assigning a Role

Imanis Data software enables you to add a local user through the Users option.

To add a local user and assign a role, do the following:

1. Click the **Main Menu**  > **System Setup** > **Users**.
2. On the **Users** page, click the  icon to configure a local user.
3. Select **Imanis Data** from the **Domain** drop-down menu and then type the user name in the **User Name** field.
4. Proceed to type the **First Name**, **Last Name**, and **Email ID**.



5. Assign a role to this user by selecting a role from the **Role** drop-down menu. For example, Data Scientist.
6. Click **Save**. An email containing Imanis Data software login password is sent to the email id specified in the above procedure.

NOTE: An Admin user with super-user privileges is already created in the user management console. You are not permitted to edit or delete this super user.

4.6.2 Editing a Local User

Imanis Data software allows you to change the first name, last name, email id, and role of a user.

To edit a local user, do the following:

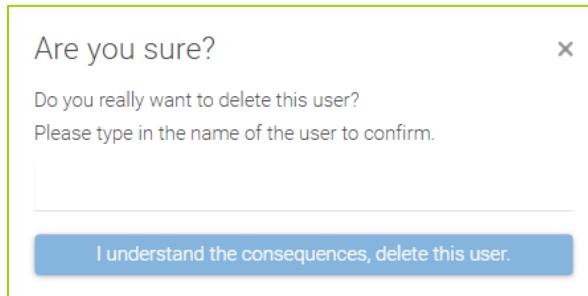
1. Click the **Main Menu** > **System Setup** > **Users**.
2. On the **Users** page, in the left pane identify the local user that you want to edit and select it.
3. Click the edit icon and make the required changes. For example, changing the first name, last name, email id, and role. However, you cannot change the Domain and User Name.
4. Click **Save**.

4.6.3 Deleting a Local User

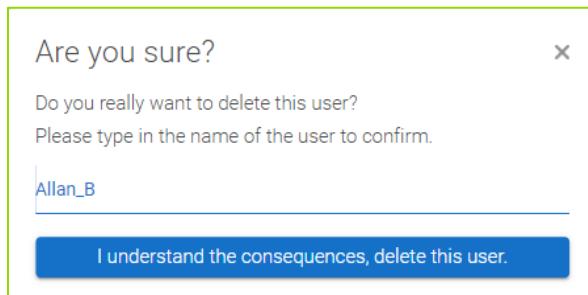
Imanis Data software enables you to delete a local user permanently from the user management console. Once a deleted, a user cannot log on to Imanis Data software.

To delete a role, do the following:

1. Click the **Main Menu**  > **System Setup** > **Users**.
2. In the left-hand pane, identify the user that you want to delete, select it and then click the  icon. The following dialog box is displayed:



3. In the dialog box, type the name of the user to confirm your decision to delete it. Once you type the name, then the **I understand the consequences, delete this user** button is activated.



7. Click the **I understand the consequences, delete this user** button to delete the role permanently. A confirmation message will be displayed on the page indicating user is successfully deleted.



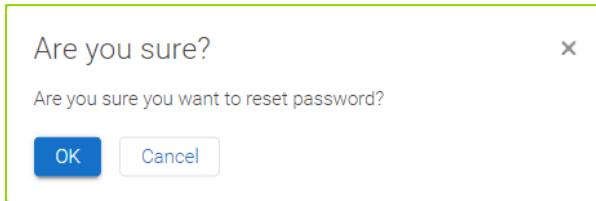
4.6.4 Resetting Password

The Reset Password link is displayed only for the local Imanis Data users.

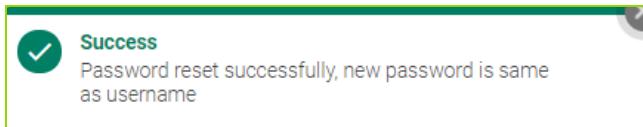
To reset a password, do the following:

1. Click the **Main Menu**  > **System Setup** > **Users**.
2. On the **Users** page, from the left-hand pane identify the name of the local user for whom you want to reset the password and select it.
3. Click the **Reset Password** link.

4. Click **Yes** on the confirmation message **Are you sure you want to reset the password?**



5. When you click **Yes**, Imanis Data software reset the password and set the new password as the username.



4.6.5 Configuring a Domain user

It is recommended that you first create an LDP domain and role before configuring a domain user.

Prerequisite: A valid LDAP domain and a user role is a must to map to the domain user.

To add a domain user, do the following:

1. Click the **Main Menu** > **System Setup** > **Users**.
2. On the **Users** page, click the icon to configure a local user.
3. Select the LDAP domain from the **Domain** drop-down menu and then type the unique name of the user

in the Username field and then click the search icon . Imanis Data software will search for this value under the unique attribute name (specified at the time of domain creation) of LDAP accounts. Imanis Data software performs an exact match. It is this value that gets stored as the Imanis Data username once the user is saved.

The following fields will be auto-populated by Imanis Data *only* if they're available in matched LDAP user account: First Name, Last Name, Email ID.



4. Assign **role** from the Role drop-down menu after Imanis Data software auto-populates the user information like First Name, Last Name, and Email id.
5. Click **Save**. An email containing Imanis Data software login password is sent to the email id specified in the above procedure. The Imanis Data username for this newly added user is same as the value of the unique attribute.

4.6.6 Editing a Domain user

Imanis Data software allows you to change only the role of the domain user.

To edit a domain user, do the following:

1. Click the **Main Menu** > **System Setup** > **Users**.
2. On the **Users** page, in the left-hand pane identify the domain user that you want to edit and select it.
3. Click the edit icon and make the required changes. For example, changing the role. You are not permitted to change the first name, last name, and email id.
4. Click **Save**.

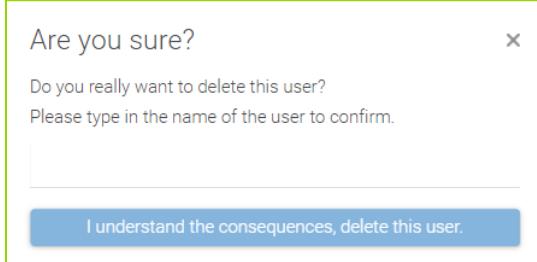
4.6.7 Deleting a Domain user

Imanis Data software enables you to delete a domain local user permanently from the user management console. Once a deleted, a domain user cannot be retrieved.

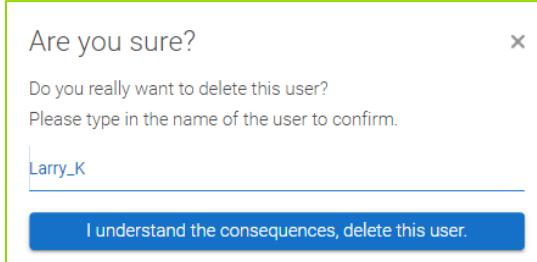
To delete a domain user, do the following:

1. Click the **Main Menu** > **System Setup** > **Users**.

2. In the left-hand pane, identify the domain user that you want to delete, select it and then click the  icon.



3. On the confirmation dialog box, type the name of the domain user to confirm your decision to delete it. Once you type the name, then the **I understand the consequences, delete this user** button is activated.



8. Click the **I understand the consequences, delete this user** button to delete the role permanently. A confirmation message will be displayed on the page indicating domain user is successfully deleted.



4.7 Roles

Roles have a set of capabilities and users are assigned to these roles. For example, a backup operator role could be given the following capabilities: create, view, edit, run and delete data backup workflows.

Accordingly, all users assigned to that role will inherit those capabilities. On the **Roles** page, you will see the following tabs:

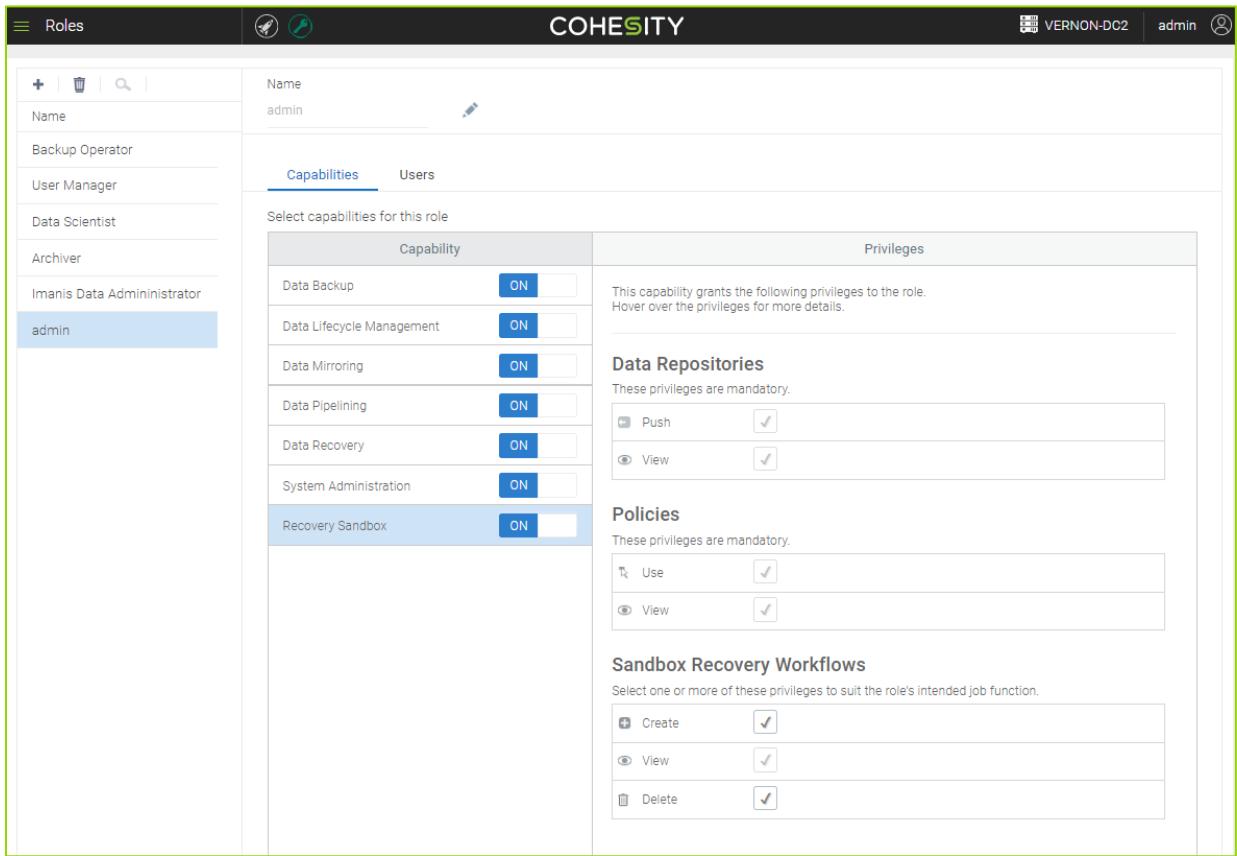
- **Capabilities:** Select a role to view the capabilities associated with a role. The **Capabilities** tab is selected by default
- **Users:** Select a role and then click the **Users** tab to see all the users assigned to a role. This tab is visible only if you assign a role to a user

4.7.1 Creating a role

You can create a role from the left-hand pane of Imanis Data software.

To create a role, do the following:

1. Click the **Main Menu**  > **System Setup** > **Roles**. The following page is displayed:



Capability	ON	Privileges
Data Backup	<input checked="" type="checkbox"/>	This capability grants the following privileges to the role. Hover over the privileges for more details.
Data Lifecycle Management	<input checked="" type="checkbox"/>	
Data Mirroring	<input checked="" type="checkbox"/>	
Data Pipelining	<input checked="" type="checkbox"/>	
Data Recovery	<input checked="" type="checkbox"/>	
System Administration	<input checked="" type="checkbox"/>	
Recovery Sandbox	<input checked="" type="checkbox"/>	

2. Edit the default admin profile by clicking the  edit icon and make the required changes in the **Capability** and **Privileges** section.
3. Click the  icon to configure a user and then type the role name. For example, Data Scientist or Admin.
4. In the **Capability** section, identify the capability that you want this particular role to perform, click the toggle on-off slider to turn on that particular capability and then select one or more privileges from the **Privileges** section to assign it to the role. For example, you turn on **Data Recovery** and then assign **Create**, **View**, and **Delete** privileges to the role by clicking the respective check-boxes.

NOTE: Mandatory privileges are pre-selected, and you are not allowed to edit the mandatory privileges of capabilities.

5. Click **Save**.

NOTE: An Admin user with super-user privileges is already created in the user management console. You are not permitted to delete this super user.

4.7.2 Editing a role

Imanis Data software allows you to add or remove capability and privileges assigned to a role.

To edit a role, do the following:

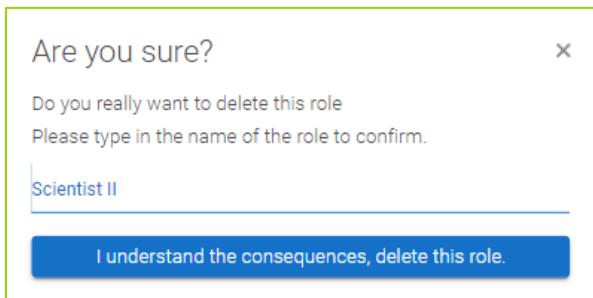
1. Click the **Main Menu**  > **System Setup** > **Roles**.
2. On the **Roles** page, identify the role that you want to edit, select it, and then click the edit icon.
3. Choose to add or remove capability and the privileges assigned to it.
4. For example, you choose to turn on the **System Administration** capability and assign the Users and Roles and Data Repositories privileges to the role. This action will allow the role to create, edit, and create users, roles, domains, and SMTP protocol in the Users and Roles and create, edit and delete data repositories too:
5. Click **Save**.

4.7.3 Deleting a role

Imanis Data software enables you to delete a role permanently from the user management console. If a role, previously assigned to multiple users, is deleted, the users can still log on to Imanis Data software, however, the users cannot perform any tasks because they do not have the capability or privileges. For example, a role with Data Backup capability and privileges to create, edit, and delete data back workflows is created and assigned to a user. If the role is deleted, the user can login to Imanis Data software, however, he/she cannot access the Data Backup menu or perform any tasks related to data backup.

To delete a role, do the following:

1. Click the **Main Menu**  > **System Setup** > **Roles**.
2. In the left-hand pane, identify the role that you want to delete, select the role, and then click the  icon.
3. On the confirmation page, type the name of the role to confirm your decision to delete it. Once you type the name, then the **I understand the consequences, delete this role** button is activated.

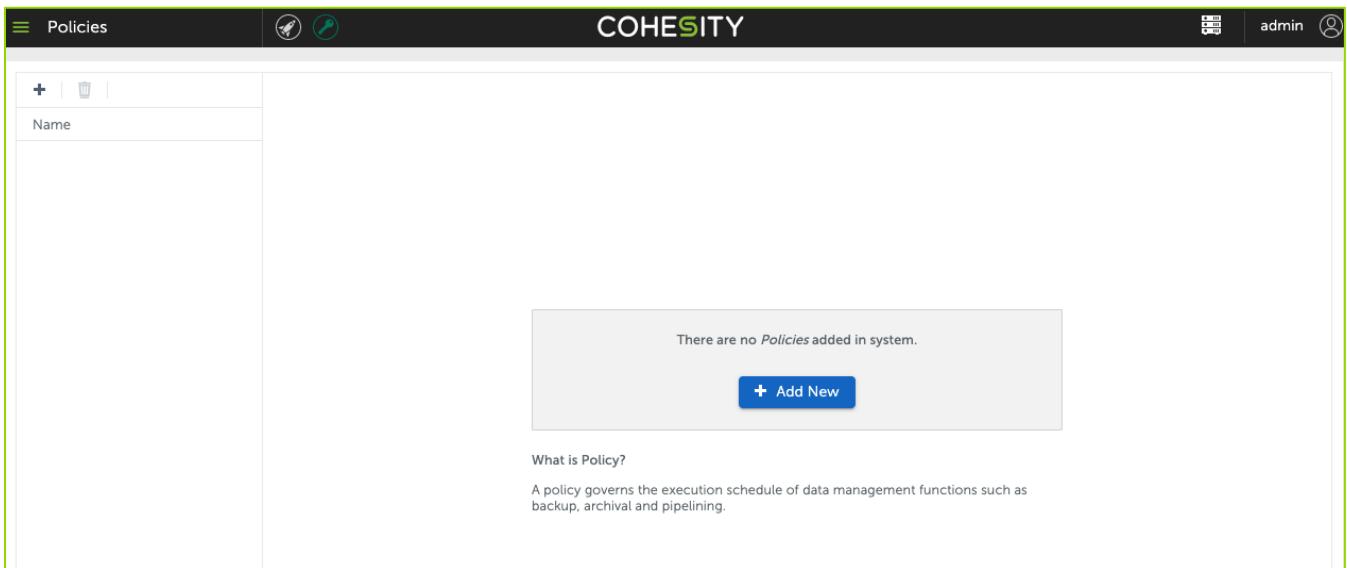


4. Click the **I understand the consequences, delete this role** button to delete the role permanently. A confirmation message will be displayed on the page indicating role is successfully deleted.



5 Policies

A policy governs the execution priority and schedule of backup, archival, mirroring, pipelining, and recovery workflows. You can access the Policies under the System Setup option.



This section describes the process of creating, editing, and deleting a policy in Imanis Data software.

5.1 Creating Policies

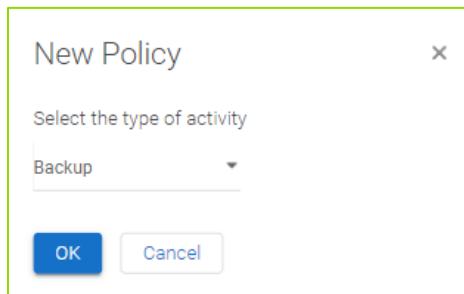
This section describes the process of creating a backup policy, DLM policy, and Recovery policy in Imanis Data software.

5.1.1 Backup Policy

Backups are probably the simplest form of disaster recovery. The purpose of backups is to make identical copies of objects on separate storage with a user-defined policy. Imanis Data software does a full backup first and then incremental forever. You can also choose to retain data on the cloud with Amazon S3 or Azure Blob Storage.

To create a backup policy, do the following:

1. Click the **Main Menu** > **System Setup** > **Policies** and then click the icon. The following dialog box appears:



2. In the **New Policy** dialog box, under the **Select the type of activity** drop-down menu, select the **Backup** option, and then click **OK**.
3. Type a policy name in the **Policy Name** field.
The policy name can include alphanumeric characters, numbers and/or special characters. A good practice is to create a policy with intuitive names and brief descriptions. This practice helps users, other than the policy creator, to use the policy appropriately. For example, backup_even_hours or backup_recurring.
4. In the **Retention** area, do one of the following:
 - Set the retention to cloud feature to '**Yes**', if you want to backup data in native format to cloud storage such as Amazon S3 or Azure Blob storage

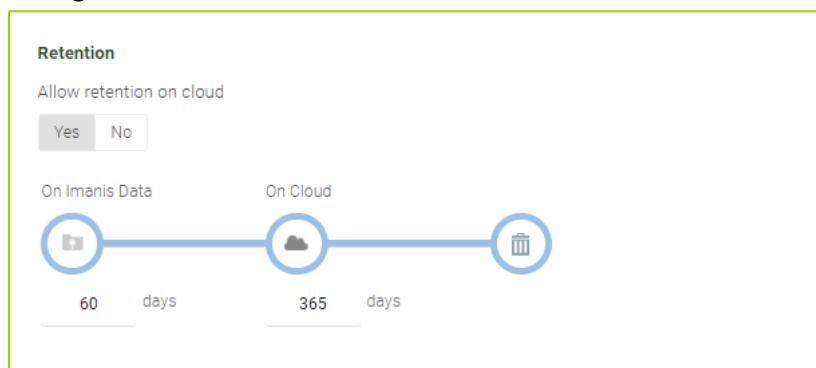
The screenshot shows the Cohesity UI for managing policies. On the left, there's a sidebar with 'Policies' and a search bar. The main area shows a table with a single row for 'Backup_Policy'. To the right of the table, there's a 'Retention' configuration panel. It includes a 'Retention' section with a 'Allow retention on cloud' toggle (set to 'Yes'), and a timeline diagram showing data moving from 'On Immanis Data' to 'On Cloud' over 60 days and 365 days. The 'Save' and 'Cancel' buttons are at the top right of the panel.

- Set the retention to cloud feature to '**No**', if you want to backup data in de-duplicated format to Amazon S3 or Azure Blob Storage. However, first ensure that you have setup Global S3 or Global Azure data repository for movement of de-duplicated data. If you DO not setup Global S3 or Global Azure data repository, data will be retained on Imanis Data cluster only



Here is the procedure in brief:

- Click **Yes** in the **Allow retention on cloud** option and then type a number in the **days** field in the **On Imanis Data** option and the **On Cloud** option. This action retains data objects, first on the Imanis Data cluster and later on the Cloud (Amazon S3) after which the data objects will be deleted automatically. Imanis Data enables you to set a schedule to retain data on the cloud through the Cloud Retention feature



For example, if you want to retain your backup data for 2 months on the Imanis Data cluster and for a year on the Cloud, then type 60 in the days field under the **On Imanis Data** option and 365 in the days field under **On Cloud** option. In addition, you can set a schedule to retain data on the cloud through the Cloud Retention feature.

The process executed in three simple steps:

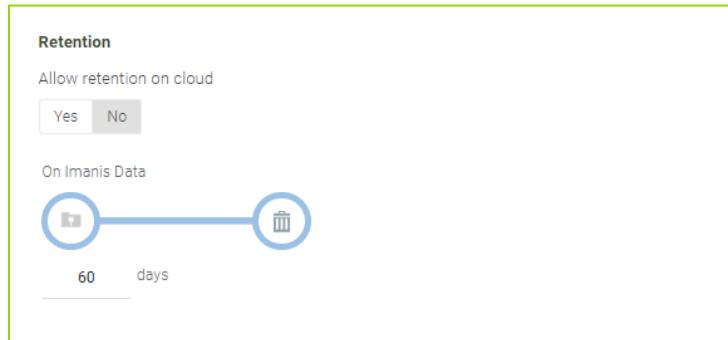
- a. Imanis Data software will retain your backup data on Imanis Data cluster for 60 days.
- b. When 60 days conclude, Imanis Data software will move the backup data onto the Cloud (S3 or Azure) and retain it for another 365 days.
- c. After 365 days conclude, Imanis Data software will automatically delete the data from the cloud.

IMPORTANT: You must type a value in the days field under the On Imanis Data and On Cloud option, without which Imanis Data software will not allow you to save the policy.

- Click **No** in the Allow retention on cloud option and then type a number in the days field under the On Imanis Data option to retain data objects on the Imanis Data cluster (only until a S3 or Azure data repository is enabled to function as a Global Cloud data repository). After the 'number of days' are concluded, the data objects will be deleted automatically

IMPORTANT: In case a S3 or Azure data repository is not enabled for movement of de-duplicated data, the data will be retained on Imanis Data cluster. However, the data is automatically moved to S3 or Azure, when Yes is selected for the Use this for deduplication option in S3 or Azure data repository for enabling the movement of de-duplicated data.

For example, if you want to retain your backup data for 1 year for regulatory compliance, then type 365 in the days field. This action retains your backup data on the Imanis Data cluster and then gradually deletes the data as it completes a year.



5. In the **Backup Schedule** area, do one of the following:

- Click the **One time, immediately** option to activate the policy right away
- Click the **At specified times** option to select hours and days of the week

For example, if you select 09 and 21 hours in the hours of the day option and the Weekdays option, then a job will at 9 A.M. and 9 P.M. on Monday, Tuesday, Wednesday, Thursday, and Friday

Backup Schedule

One time, immediately	At specified times	At a specified frequency
-----------------------	--------------------	--------------------------

Backup at these hours of the day

All	08	09	10	11	12	13	14	15	16	17	18	19
	07	06	05	04	03	02	01	00	23	22	21	20

Repeat on these days of the week

None	All	Weekdays	Mo	Tu	We	Th	Fr	Sa	Su
------	-----	----------	----	----	----	----	----	----	----

- Click the **At a specified frequency** option to select minutes and days of the week. For example, if you select the 20 minutes option and the **Weekdays** option, then a job will run three times in an hour on Monday, Tuesday, Wednesday, Thursday, and Friday

Backup Schedule

One time, immediately	At specified times	At a specified frequency
-----------------------	--------------------	--------------------------

⚠ Use this option sparingly as it may put considerable load on system resources.

Every 20 minutes

15	20	30
----	----	----

Repeat on these days of the week

None	All	Weekdays	Mo	Tu	We	Th	Fr	Sa	Su
------	-----	----------	----	----	----	----	----	----	----

- In the **Cloud Schedule** area, select hours and days of the week to migrate data onto the S3 cloud repository (see screenshot below):

Cloud Schedule

Migrate at these hours of the day

All	08	09	10	11	12	13	14	15	16	17	18	19
	07	06	05	04	03	02	01	00	23	22	21	20

Migrate on these days of the week

None	All	Weekdays	Mo	Tu	We	Th	Fr	Sa	Su
------	-----	----------	----	----	----	----	----	----	----

- In the **Priority** area, click **High**, **Medium**, or **Low**. This option allows you to select the kind of priority that you want to set for your policy.

Priority
<input type="radio"/> High <input type="radio"/> Medium <input type="radio"/> Low

- Click **Save**. A confirmation message will be displayed on the page indicating policy is successfully saved.



5.1.2 DLM Policy

Data Life Cycle Management (DLM) is a popular policy-based approach to manage the flow of data throughout its life cycle — from the creation of the data (hot data) and initial storage (warm data) to the time when it moves to a long-term archive to the cloud (cold). You can create DLM policies to manage the flow of your data throughout its lifecycle. Archive your data to the cloud using Amazon S3 or Amazon Glacier using this policy.

To create a data lifecycle management policy, do the following:

- Click the **Main Menu** > **System Setup** > **Policies** and then click the icon. The following dialog box appears:

New Policy	X
Select the type of activity	
Lifecycle Management	▼
<input type="button"/> OK <input type="button"/> Cancel	

- In the **New Policy** dialog box, under the **Select the type of activity** drop-down menu, select the **Lifecycle Management** option, and then click **OK**.
- Type a policy name in the **Policy Name** field. The policy name can include alphanumeric characters, numbers and/or special characters. A good practice is to have intuitive policy names so that users, other than the policy creator, can read the name and use it appropriately. For example, recovery_now or recovery_recurring.
- In the **Lifecycle** area, type values for the DLM options. Refer to the **Data Lifecycle Management** Table below to type values in the fields.

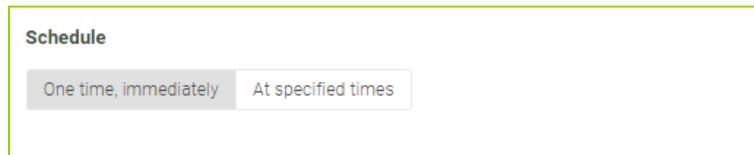


BUTTONS & FIELDS	DESCRIPTION	ACTION TO TAKE
Full Replica 	<ul style="list-style-type: none"> Number of days for which data objects remain in the Full Replica state of the primary cluster. Imanis Data software does not take any action on the data objects in this period. For example, Full Replica is set to 30 days. On Day 31, the number of replicas of a file on the primary cluster is reduced and a new copy is created on the Imanis Data cluster. 	<ul style="list-style-type: none"> Type a number (of days) for which Imanis Data software will not execute any task on the data objects in primary cluster The Forever and Skip option is not available for Full Replica.
Reduced Replica 	<ul style="list-style-type: none"> Number of days for which data objects remain in the Reduced Replica state of the primary cluster. For example, Reduced Replica is set to 15 days. On day 16, Imanis Data software reduces the replica count by 1 on the primary cluster and increases the replica count by 1 on the Imanis Data cluster. This process is repeated automatically until the replica 	<p>Do one of the following:</p> <ul style="list-style-type: none"> Select or Type Skip to ignore or bypass reduced replica step in the DLM process Select or Type a number (of days) after which copy of a file on the primary cluster is reduced further and a new copy is added to the Imanis Data cluster In Reduced Replica, only the Skip option is available. Even if you type Forever, Imanis Data

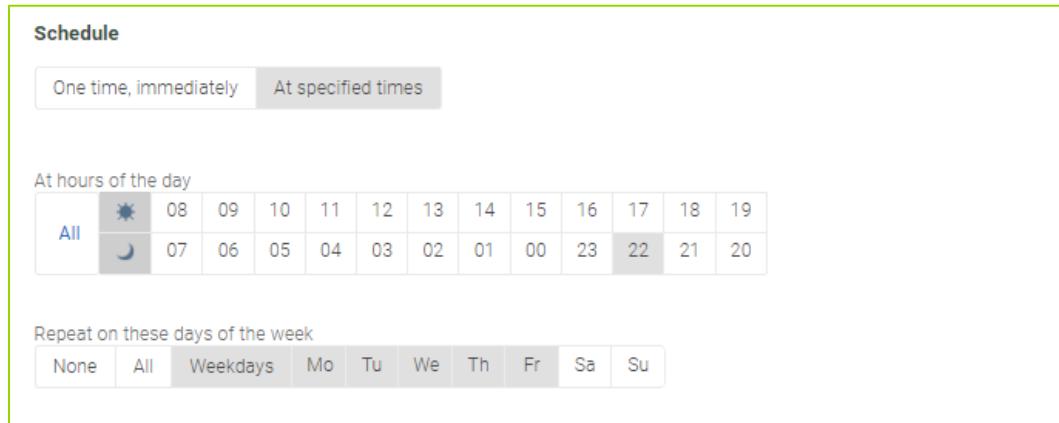
	<p>count on the data repository is zero, after which the data is archived.</p>	<p>software will not allow you to save the policy.</p>
Archived 	<ul style="list-style-type: none"> Number of days for which data objects are stored in the Imanis Data cluster. For example, Archived is set to 100 days. On day 101, the data objects are either deleted or moved to Amazon Glacier. However, if you type the Forever value, data will be archived perpetually on the Imanis Data cluster until you decide to delete it. 	<p>Do one of the following:</p> <ul style="list-style-type: none"> Select Forever to perpetually archive data on the Imanis Data cluster until you decide to delete it Select or Type number (of days) for which copy of a file must be archived on the Imanis Data cluster. After the number of days elapse, the file is deleted or moved to Glacier. <p>Even if you type Skip, Imanis Data software will not allow you to save the policy.</p>
On Cloud 	<ul style="list-style-type: none"> Number of days for which data objects can be stored on the Cloud (Amazon Glacier) 	<p>Do one of the following:</p> <ul style="list-style-type: none"> Select Skip to ignore or bypass the Cloud option step in the DLM process Select Forever to perpetually archive data on the Cloud until you decide to delete it Select or Type number (of days). After the days elapse Imanis Data software deletes copy of file from the Cloud
	<ul style="list-style-type: none"> This is the last stage of the DLM process. 	<ul style="list-style-type: none"> No action to take

5. Select the **Days or Hours** option. This option is the lifecycle unit (in terms of time) of the DLM policy.
6. In the **Schedule** area, click **One time, immediately** or **At specified times**.

- If you select **One time, immediately** the policy is activated right away



- If you select **At specified times**, you need to select the hours and days of the week



7. In the **Priority** area, click **High**, **Medium**, or **Low**. This option allows you to select the kind of priority that you want to set for your policy.
8. Click **Save**. A confirmation message will be displayed on the page indicating policy is successfully saved.



IMPORTANT: Imanis Data software allows you to automate deletion of data associated with a Data Lifecycle Management (DLM) workflow from the Imanis Data cluster except when the associated policy has the 'One time, immediately' option turned on in its Schedule section.

5.1.3 Recovery Policy

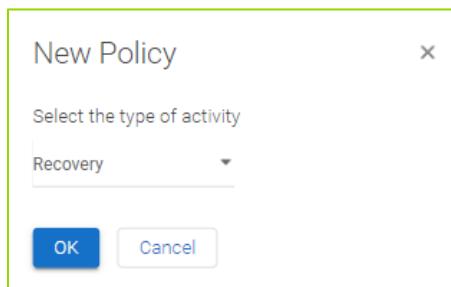
Imanis Data software allows you to create recovery policies to do a full or incremental restore of your backed up data. For Cassandra, you can do recovery at the keyspace and the table level for Cassandra and Solr-enabled DataStax Enterprise (DSE) Cassandra.

IMPORTANT: The process of creating a recovery policy is optional. Recovery policy is needed only if you want to schedule a recovery of your data. For ad hoc or on-demand recoveries, policies are not required. You can just simply create a recovery workflow, identify the data, specify recovery locations and options, and click **Submit**.

To create a recovery policy, do the following:

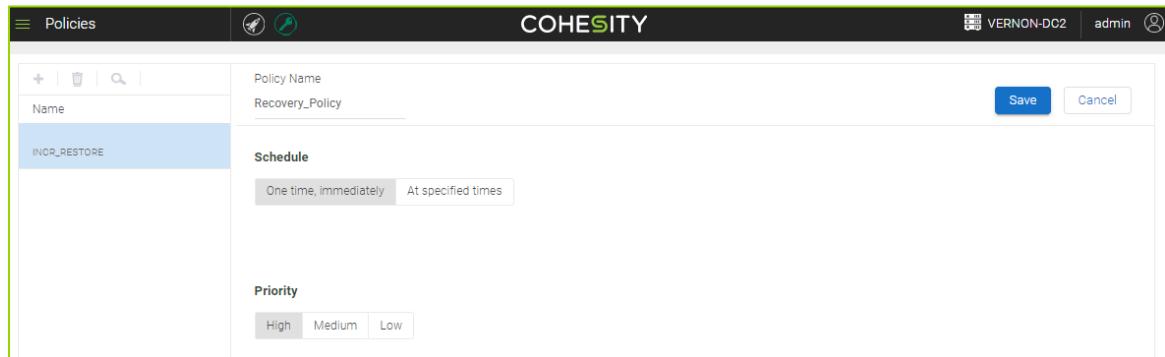
1. Click the **Main Menu**  > **System Setup > Policies** and then click the  icon.

The following dialog box appears:

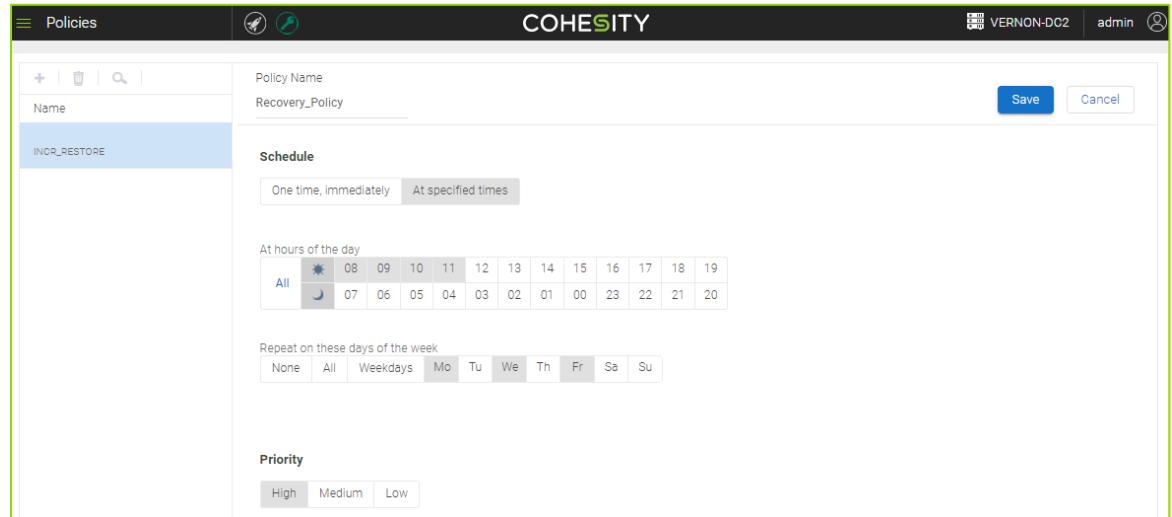


2. In the **New Policy** dialog box, under the **Select the type of activity** drop-down menu, select the **Recovery** option, and then click **OK**.
3. Type a policy name in the **Policy Name** field.
4. In the **Schedule** area, do one of the following:

- Select the **One time, immediately** option to activate the policy immediately



- Select the **At specified times** option to select the hours and days of the week (see screenshot below). In other words, activate the recurring policy



- In the **Priority** area, click **High**, **Medium**, or **Low**. This option allows you to select the kind of priority that you want to set for your policy.
- Click **Save**. A confirmation message will be displayed on the page indicating policy is successfully saved.



5.1.4 Global Cloud Policy

The Global Cloud policy is pre-configured policy in Imanis Data software.

The Global Cloud policy governs the schedule and priority of moving de-duplicated backup data from Imanis Data cluster to Global Cloud Data Repository: An Amazon S3, an Azure Blob storage, or Cloudian setup for long term storage of de-duplicated data.

By default, the Global Cloud policy is set migration of data at 0000 hours on all days of the week. However, you can set your own data migration schedule by editing the default schedule associated with the Global Cloud policy.

You are not permitted to change the priority, rename, or delete the Global Cloud policy.

IMPORTANT: The policy is relevant only if you have created a cloud data repository (S3, Azure, or Cloudian) that is enabled for movement of de-duplicated data. For more information, refer to the section on S3 data repository, Azure Blob Storage or Cloudian data repository.

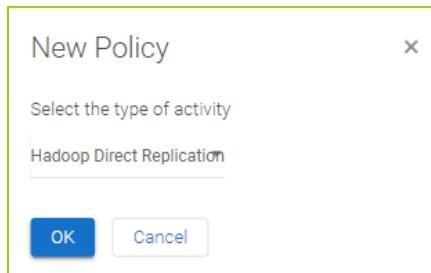
5.1.5 Direct Replication Policy

You can create a direct replication policy to replicate an exact copy of data from the source cluster to destination cluster. This policy can be created for Hadoop only.

To create a direct replication policy, do the following:

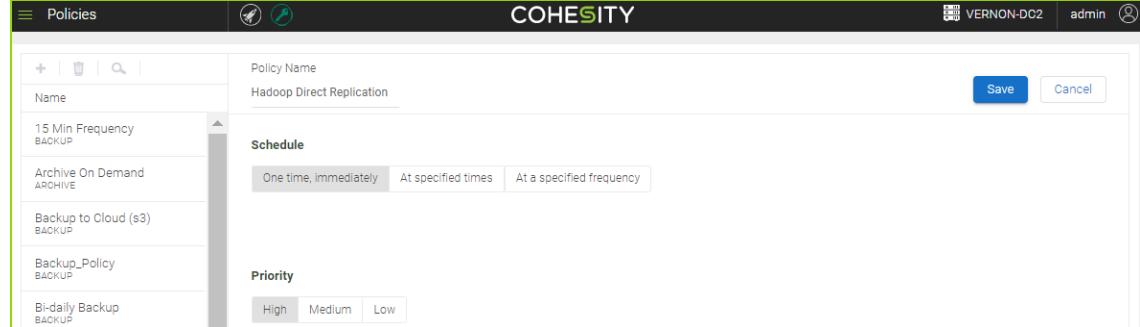
1. Click the **Main Menu**  > **System Setup** > **Policies** and then click the  icon.

The following dialog box appears:

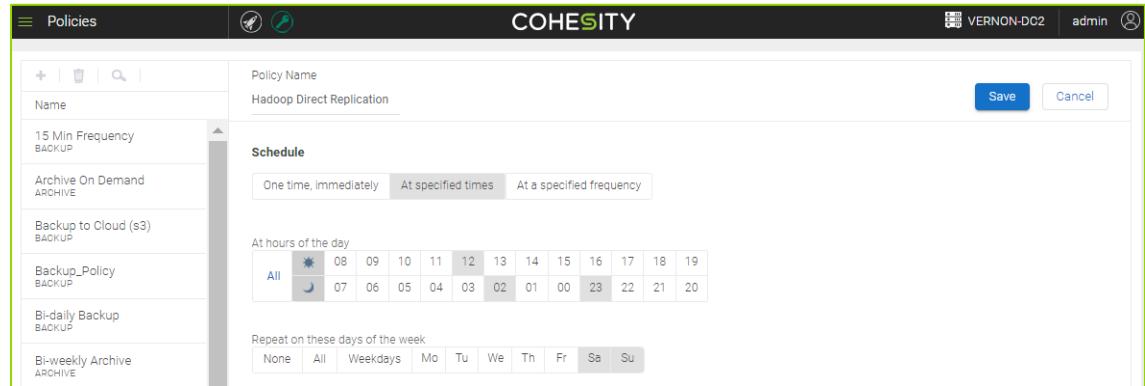


2. In the **New Policy** dialog box, under the **Select the type of activity** drop-down menu, select the **Hadoop Direct Replication** option, and then click **OK**.
3. Type a policy name in the **Policy Name** field.
4. In the **Schedule** area, do one of the following:

- Select the **One time, immediately** option to activate the policy immediately



- Select the **At specified times** option to select the hours and days of the week (see screenshot below). In other words, activate the recurring policy



- Select the **At a specified frequency** option to select frequency of running job at every 15, 20 and 30 minutes apart on all days, weekdays, or specific days of the week (see screenshot below).



- In the **Priority** area, click **High**, **Medium**, or **Low**.
- This option allows you to select the kind of priority that you want to set for your policy.
- Click **Save**.

5.2 Managing Policies

Imanis Data software allows you to edit the backup, recovery, and data lifecycle management policies, even if the policies are being used in a workflow.

5.2.1 Editing a policy

This section describes the procedure of editing policies and the do's and don'ts of editing the policies.

5.2.1.1 Backup Policy

To get started with editing a Backup policy, all you have to do is identify the policy that you want to edit, select it, and then click the  icon.

You can edit the following in the backup policy:

- If the backup policy has **Schedule** value set as **At specified times**, you can edit the hours in the **Migrate at these hours of the day** option and select specific days, all the days, or weekdays in the **Migrate on these days of the week** option
- If you have selected **No** in the **Allow retention on cloud** option to disable retention of data on the Cloud then you can edit to **Yes**. Once you make this change, make sure you set data migration schedule to the cloud in the Cloud Schedule area in the backup policy itself. Also, ensure that you edit each job using the policy and specify the cloud data repository details
- If the backup policy has **Priority** value set as **High**, you can change it **Medium** or **Low** as per your need However, you cannot change or edit the following once you create a backup policy and use it in a workflow:
 - If the **Schedule** value is set as **One time, immediately** it cannot be edited to **At specified times** or **At a specified frequency**
 - If the **Schedule** value is set as **At specified times** it cannot be edited to **One time, immediately** or **At a specified frequency**
 - If the **Schedule** value is set as **At a specified frequency** it cannot be edited to **One time, immediately** or **At specified times**
 - If you have selected **Yes** in the **Allow retention on cloud** option to enable retention of data on the Cloud then you cannot edit the option to **No**

5.2.1.2 DLM Policy

Once you create a DLM policy and use it in a workflow, Imanis Data software does not allow you to do the following:

- If the DLM policy has **On Cloud** values set as **Forever** or **Number of Days**, Imanis Data software does not allow you to change the value to **Skip**. Similarly, if the **On Cloud** value is set as **Skip**, Imanis Data software does not allow you to edit the values to **Forever** or **Number of Days**
- If the DLM policy has **Schedule** value set as **One time, immediately** it cannot be edited to **At specified times**. Similarly, if the **Schedule** value is set as **At specified times** it cannot be edited to **One time, immediately**.

You can, however, edit the following in the DLM policy:

- If the DLM policy has **Lifecycle** value (time unit) set as **Hours**, you can change it to **Days** and vice versa.

- If the DLM policy has **Schedule** value set as **At specified times**, you can edit the hours in the **At hours of the day** option and select specific days, all days, or weekdays in the **Repeat on days of the week** option.
- If the DLM policy has **Priority** value set as **High**, you can change it **Medium** or **Low** as per your need.

5.2.1.3 Recovery Policy

To get started with editing a Recovery policy, all you have to do is identify the policy that you want to edit, select it, and then click the  icon.

You can edit the following in the recovery policy:

- If the recovery policy has **Schedule** value set as **At specified times**, you can edit the hours in the **At hours of the day** option and select specific days, all days, or weekdays in the **Repeat on days of the week** option
 - If the recovery policy has **Priority** value set as **High**, you can change it **Medium** or **Low** as per your need
- However, you cannot change or edit the following once you create a recovery policy and use it in a workflow:
- If the recovery policy has **Schedule** value set as **One time, immediately** it cannot be edited to **At specified times**. Similarly, if the **Schedule** value is set as **At specified times** it cannot be edited to **One time, immediately**.

5.2.1.4 Global Cloud Policy

To get started with editing a Global Cloud policy, select the policy, and then click the  icon.

The Global S3 policy is an out-of-the-box policy in Imanis Data software that governs the schedule and priority of moving de-duped backup data from Imanis Data to S3, Azure Blob Storage, or Cloudian. By default, the Global Cloud policy is set to migrate data at 0000 hours on all days of the week, however, you can edit the default schedule in the Global Cloud policy by selecting:

- The hours in the **Migrate at these hours of the day** option
- The specific days, all the days, or weekdays in the **Migrate on these days of the week** option

You are not permitted to change the priority, rename, or delete the Global Cloud policy.

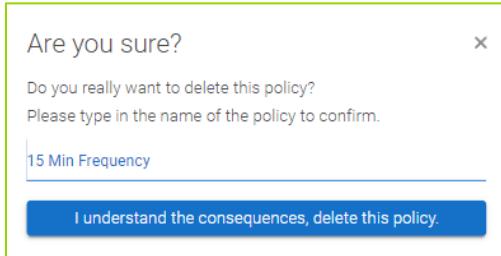
5.2.2 Deleting a policy

Imanis Data software allows you to delete a policy permanently. However, you cannot delete a policy if it is being actively used by a workflow.

NOTE: Once a policy is deleted it cannot be recovered.

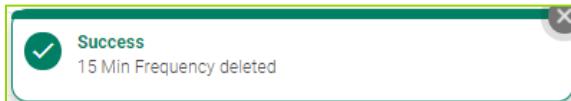
To delete a policy, do the following:

1. Identify the policy that you want to remove and click the  icon.
2. On the confirmation page, type the name of the policy to confirm your decision to delete the policy. Once you type the name, then the **I stand the consequences, delete this policy** button is activated.



NOTE: Once a policy is deleted it cannot be recovered.

3. Click the **I understand the consequences, delete this policy** button to delete your policy permanently. A confirmation message will be displayed on the page indicating policy is successfully deleted.



6 Data Repositories

A data repository is a production, test, or development cluster whose data you want Imanis Data to manage and protect. You must manually add data repositories Imanis Data software which communicates with, and writes data to and from, the data repository.

6.1 Prerequisites of Adding Data Repositories

This section describes the respective prerequisites for each of the data repositories that you need to execute or perform before adding the data repository.

6.1.1 Hadoop (HDFS, Hive & HBase)

This section describes the prerequisites for HDFS, Hive & HBase.

6.1.1.1 Kerberos-enabled Hadoop

The following section describes the steps you need to perform for adding Kerberized Hadoop data repository with non-default principals.

By default, Imanis Hadoop agents (hdfs and hive) connect to Hadoop data repository using Imanis cluster's service principal (hdfs@TALENA.REALM). For hbase, "hbase.super" principal is used.

From Imanis Data 3.4.0 release, Hadoop agents can be configured to connect to Hadoop data repository with alternative principals however before specifying such principals for connection do the following:

Adding Kerberized Hadoop data repository with non-default principals:

1. Add this principal realm's information in \${INSTALL_DIR}/conf/krb5.conf on each of Imanis nodes (use script \${INSTALL_DIR}/bin/talena-kerberos.sh of Imanis software with argument addrealm).
2. Append this principal's keytab entry in Keytab file of Imanis cluster (\$INSTALL_DIR/conf/talena.keytab) on all nodes.
3. Make sure that these principals are super user in their respective Hadoop application.
4. Restart UI.

6.1.1.2 SSL-enabled Hadoop

The following section describes the tasks you need to perform before adding SSL-enabled Hadoop as a source data repository in Imanis Data cluster.

Before adding SSL-enabled Hadoop, do the following:

1. Consult your Hadoop system admin to get the SSL certificates of all nodes in your production Hadoop system.
2. Copy all SSL certificate (*.cer) files to Imanis Data nodes where hdfs user has access. The scp command uses SSH to copy files and directories between remote hosts without starting an FTP session.

```
scp ImanisData13new.cer  
root@ImanisData1.qa.com:/home/hdfs/ImanisData13and14cer  
  
scp ImanisData14new.cer  
root@ImanisData1.qa.com:/home/hdfs/ImanisData13and14cer
```

3. Import the SSL certificates using the following command. The command must be run as user hdfs and use password used while creating certificate.

```
[hdfs@ImanisData1 ImanisData13and14cer]$ $INSTALL_DIR/bin/talena-ssl.sh  
import ImanisData13 /home/hdfs/ImanisData13and14cer/ImanisData13new.cer  
  
Importing SSL certificate into Imanis Data...  
  
Enter the Truststore password: Certificate was added to keystore  
[hdfs@ImanisData1 ImanisData13and14cer]$  
[hdfs@ImanisData1 ImanisData13and14cer]$ $INSTALL_DIR/bin/talena-ssl.sh  
import ImanisData14 /home/hdfs/ImanisData13and14cer/ImanisData14new.cer  
  
Importing SSL certificate into Imanis Data...  
  
Enter the Truststore password:  
  
Certificate was added to keystore
```

4. Run the above command for all the certificate files.
5. Run the following command from any one node of Imanis Data cluster to verify if your production is accessible from Imanis Data node:

```
hdfs dfs -ls swebhdfs://<namenode>:50470/
```

6. Restart the Imanis Data GUI with the following command:

```
[hdfs@ImanisData1 ImanisData13and14cer]$ talena-services.sh restart UI
```

7. Go ahead and add this SSL-enabled Hadoop as a data source repository in Imanis Data software.

6.1.1.3 Kerberos + SSL-enabled Hadoop

The following section describes the tasks you need to perform before adding Kerberos + SSL-enabled Hadoop as a source data repository in Imanis Data cluster.

Before adding Kerberos + SSL-enabled Hadoop, do the following:

1. First, follow the steps mentioned in the Kerberos-enabled Hadoop section for Kerberos-related configuration required for adding Kerberized Hadoop primary.
2. Then, consult your Hadoop system admin to get the SSL certificates of all nodes in your production Hadoop system.
3. Copy all SSL certificate (*.cer) files to Imanis Data nodes where hdfs user has access. The scp command uses SSH to copy files and directories between remote hosts without starting an FTP session.

```
scp ImanisData13new.cer  
root@ImanisData1.qa.com:/home/hdfs/ImanisData13and14cer  
  
scp ImanisData14new.cer  
root@ImanisData1.qa.com:/home/hdfs/ImanisData13and14cer
```

4. Import the SSL certificates using the following command. The command must be executed as **hdfs** user and a password must be set when creating the truststore for the first time. All future certificate imports to the talena truststore must use the same password.

```
[hdfs@ImanisData1 ImanisData13and14cer]$ $INSTALL_DIR/bin/talena-ssl.sh  
import ImanisData13 /home/hdfs/ImanisData13and14cer/ImanisData13new.cer  
  
Importing SSL certificate into Imanis Data...  
  
Enter the Truststore password: Certificate was added to keystore  
[hdfs@ImanisData1 ImanisData13and14cer]$  
[hdfs@ImanisData1 ImanisData13and14cer]$ $INSTALL_DIR/bin/talena-ssl.sh  
import Imanis Data14 /home/hdfs/ImanisData13and14cer/ImanisData14new.cer  
  
Importing SSL certificate into Imanis Data...  
  
Enter the Truststore password:  
  
Certificate was added to keystore
```

5. Run the following command to login with user **hdfs**:

```
kinit -kvT <path_to_talena_keytab_file> hdfs@PRIMARYKDC.REALM
```

6. Run the List command (**ls**) as the following:

```
hdfs dfs -ls swebhdfs://<namenode>:50470/
```

7. Restart the Imanis Data GUI with the following command:

```
[hdifs@ImanisData1 ImanisData13and14cer]$ talena-services.sh restart UI
```

8. Go ahead and add this Kerberos + SSL-enabled Hadoop as a data source repository in Imanis Data software.

6.1.1.4 Hive

The following section describes the tasks you need to perform to work with Hive and Kerberized Hive.

6.1.1.4.1 Configuring core-site.xml for backing up Kerberized Hive

This is case applicable for backing up Kerberized Hive and Hive Masking Sampling.

You have to enable the following in your core-site.xml. This action can be done from Cloudera manager or Ambari UI.

```
<property><name>hadoop.proxyuser.hdfs.groups</name><value>*</value></property>
<property><name>hadoop.proxyuser.hdfs.hosts</name><value>*</value></property>
<property><name>hadoop.proxyuser.hive.groups</name><value>*</value></property>
<property><name>hadoop.proxyuser.hive.hosts</name><value>*</value></property>
```

Also, ensure value of min.user.id (Yarn configuration) is less than uid of Linux Hive user of primary cluster:

```
[root@qa conf]# grep min.user.id -R /etc/hadoop/conf/*
/etc/hadoop/conf/container-executor.cfg:min.user.id=500
[root@qa conf]# id hive
uid=500(hive) gid=500(hadoop) groups=500(hadoop)
```

6.1.1.5 HBase

The following section describes the tasks you need to perform to work with HBase and Kerberos HBase.

6.1.1.5.1 Maintaining the delete markers on HBase primary cluster

This case is applicable only if you are using Data Mirroring in HBase.

Let's assume a row was deleted on the primary cluster and a delete marker was created masking the deleted value. These delete markers are purged only during major compactions which may happen immediately after the row was deleted. In this case, the next incremental backup will not be able to capture these delete markers.

As the incremental backup fails to capture these delete propagations, they will not be propagated to the destination cluster.

Thus, it is recommended to maintain delete propagations a little longer on the primary cluster.

To maintain delete propagations on HBase primary, do the following:

1. Log on to the HBase primary cluster.
2. Access the hbase-site.xml file in the primary cluster.
3. Set the value of parameter hbase.hstore.time.to.purge.deletes as greater than twice that of the backup frequency. For example, if a backup is scheduled for daily run, then the value of the parameter must be set to 2 days, that is 172800000 in milliseconds.

```
<property>
<name>hbase.hstore.time.to.purge.deletes</name>
<value>172800000</value>
</property>
```

6.1.1.5.2 Configuration Entry of hbase.superuser

This case is applicable only if you are using Kerberos HBase.

Imanis Data supports only one value of hbase.superuser in hbase-site.xml is supported. If your production cluster configuration file has multiple values, make sure that you keep only one whose keytab you have imported into keytab file of Imanis Data cluster. Also, keytab of user hbase.superuser must be imported to Imanis Data keytab.

To import keytab of user hbase.superuser to Imanis Data keytab, do the following:

1. Export the keytab of hbase.superuser to /tmp/hbase.keytab.
2. Copy this keytab file to one of the nodes in the Imanis Data Cluster at some temporary location, for example /tmp and merge this keytab with Imanis keytab installed in conf directory.

```
# ktutil
ktutil: read_kt /tmp/hbase.keytab
ktutil: read_kt $INSTALL_DIR/conf/talena.keytab
ktutil: write_kt /tmp/talena.keytab
ktutil: quit
```

Where INSTALL_DIR is your installation directory. This command will generate talena.keytab in your current directory.

3. Copy talena.keytab to all Imanis Data nodes at the location \$INSTALL_DIR/conf/. Ensure that the keytab is owned and readable only by user \$SERVICE_USER:\$SERVICE_USER.
4. Verify if the keytab is working fine from each node in the cluster.

```
# kinit -kVt $INSTALL_DIR/conf/talena.keytab hbase@PRIMARYKDC.REALM  
should succeed
```

6.1.2 Cassandra

The following sections describe the tasks you need to perform before adding a Cassandra data repository. The following section describes prerequisites of adding a Kerberized Cassandra.

6.1.2.1 Creating Private Key on Linux

For adding a Cassandra data repository, either a Linux password or a private key is needed. SSH keys can be used to establish a secure connection. This section describes the procedure of generating an SSH key on Linux.

1. Open a terminal and type the following:

```
ssh-keygen
```

You'll see a response like the following:

```
Enter file in which to save the key (/home/cassandrauser/.ssh/id_rsa):  
Created directory '/home/cassandrauser/.ssh'.  
Enter passphrase (empty for no passphrase):  
Enter same passphrase again:  
Your identification has been saved in /home/cassandrauser/.ssh/id_rsa.  
Your public key has been saved in /home/cassandrauser/.ssh/id_rsa.pub.  
The key fingerprint is:  
58:3a:80:a5:df:17:b0:af:4f:90:07:c5:3c:01:50:c2 cassandrauser@ImanisData-  
inc
```

2. Press <Enter> to accept the default location and file name. If the .ssh directory doesn't exist, the system creates one for you. The default location is the .ssh folder in your Home directory.
3. Enter, and re-enter, a passphrase when prompted and if required. The passphrase is optional.
4. Your public and private SSH key should now be generated. Open the file manager and navigate to the .ssh directory. You should see two files: `id_rsa` and `id_rsa.pub`.

6.1.2.2 Creating SSL certificates in Cassandra nodes and importing them in Imanis Data cluster

This section describes the procedure of creating SSL certificates in Cassandra nodes and importing them in Imanis Data cluster. However, if you already have SSL certificate (*.cer) files, you can start referring to Step #4 onwards.

1. Check the location of Truststore in Cassandra Primary and Password. You must execute the following command from the directory which contains the `cassandra.yaml` file.

```
[root@talena13 conf]# grep "truststore:" cassandra.yaml
    truststore: /home/cassandra/certs/cassandra.truststore
    truststore: /home/cassandra/certs/cassandra.truststore
[root@talena13 conf]#
[root@ talena13 certs]# grep password /root/apache-cassandra-
2.1.11/conf/cassandra.yaml
# - PasswordAuthenticator relies on username/password pairs to authenticate
users. It keeps usernames and hashed passwords in system_auth.credentials
table.

# The passwords used in these options must match the passwords used when
generating
    keystore_password: talena2016
    truststore_password: talena2016
    keystore_password: talena2016

# Set truststore and truststore_password if require_client_auth is true
truststore_password: talena2016
```

2. Go to location `/home/cassandra/certs/` and create SSL certificates using the preceding password, that is, `talena2016`. The "`/root/jdk1.7.0_45`" is the Java home directory in Cassandra nodes

```
[root@talena13 certs]# /root/jdk1.7.0_45/jre/bin/keytool -export -alias
talena13 -file talena13new.cer -keystore cassandra.keystore

Enter keystore password:

Certificate stored in file <talena13new.cer>
```

3. Repeat Step #2 in all the Cassandra nodes:

```
[root@talena14 certs]# /root/jdk1.7.0_45/jre/bin/keytool -export -alias
talena14 -file talena14new.cer -keystore cassandra.keystore

Enter keystore password:

Certificate stored in file <talena14new.cer>

[root@talena14 certs]#
```

Where `/root/jdk1.7.0_45` is the Java home directory in Cassandra nodes.

4. Copy all SSL certificate (*.cer) files to Imanis Data nodes where `hdfs` user has access. The `scp` command uses SSH to copy files and directories between remote hosts without starting an FTP session.

```
scp talena13new.cer root@talena1.qa.com:/home/hdfs/talena13and14cer
```

```
scp talena14new.cer root@talenal.qa.com:/home/hdfs/talena13and14cer
```

- Import the SSL certificates using the following command. The command must be executed as **hdfs** user and a password must be set when creating the truststore for the first time. All future certificate imports to the talena truststore must use the same password.

```
[hdfs@talenal talena13and14cer]$ $INSTALL_DIR/bin/talena-ssl.sh import  
talena13 /home/hdfs/talena13and14cer/talena13new.cer
```

```
Importing SSL certificate into Talena...
```

```
Enter the Truststore password: Certificate was added to keystore  
[hdfs@talenal talena13and14cer]$
```

```
[hdfs@talenal talena13and14cer]$ $INSTALL_DIR/bin/talena-ssl.sh import  
talena14 /home/hdfs/talena13and14cer/talena14new.cer
```

```
Importing SSL certificate into Talena...
```

```
Enter the Truststore password:
```

```
Certificate was added to keystore
```

- Restart the Imanis Data GUI

```
[hdfs@talenal talena13and14cer]$ talena-services.sh restart UI
```

6.1.2.3 Creating SSL certificates in Imanis Data cluster and importing them in all Cassandra nodes

This section describes the process of generating certificates in Imanis Data cluster and importing them in all Cassandra nodes.

- Go to the directory where Imanis Data TrustStore directory located and do the following:

```
cd $INSTALL_DIR/conf  
  
[hdfs@talenal conf]$ ll  
total 20  
-rw-r--r-- 1 hdfs QA 7620 May 19 12:16 monitor-conf.xml  
-rw----- 1 hdfs QA 2188 May 10 14:19 talenal.keystore  
-rw----- 1 hdfs QA 1880 May 19 12:26 talena.truststore
```

- Create certificates for Imanis Data nodes. In the current release, Imanis data keystore password is a random string that can be found in \$INSTALL_DIR/hadoop-2.3.0/etc/hadoop/ssl-client.xml. The keystore password is stored in the value field of below property in this file:

```
<property>
  <name>ssl.client.keystore.password</name>
  <value>YTuFONHyxh</value>
  <description>Optional. Default value is "".
  </description>
</property>
```

```
[hdfs@talena1 conf]$ keytool -export -alias talena1 -file talena1.cer -
keystore talena1.keystore

Enter keystore password:
Certificate stored in file <talena1.cer>

[hdfs@talena2 conf]$ keytool -export -alias talena2 -file talena2.cer -
keystore talena2.keystore

Enter keystore password:
Certificate stored in file <talena2.cer>

[hdfs@talena3 conf]$ keytool -export -alias talena3 -file talena3.cer -
keystore talena3.keystore

Enter keystore password:
Certificate stored in file <talena3.cer>
```

3. Copy the preceding Imanis Data certificate (*.cer) files to all nodes in Cassandra. Ensure that the Imanis Data certificate (*.cer) files have all the necessary permissions on Cassandra node for Cassandra service user. For example, after copying Imanis Data certificate (*.cer) files to all nodes in Cassandra, you must change the owner of that certificate to Cassandra on all the Cassandra node.

```
[hdfs@talena3 conf]$ scp talena3.cer root@talena13:/home/cassandra/certs/
[hdfs@talena3 conf]$ scp talena3.cer root@talena14:/home/cassandra/certs/
[hdfs@talena2 conf]$ scp talena2.cer root@talena13:/home/cassandra/certs/
[hdfs@talena2 conf]$ scp talena2.cer root@talena14:/home/cassandra/certs/
[hdfs@talena1 conf]$ scp talena1.cer root@talena13:/home/cassandra/certs/
[hdfs@talena1 conf]$ scp talena1.cer root@talena14:/home/cassandra/certs/
```

4. Import Imanis Data certificates on each of the Cassandra nodes. Ensure that this step is repeated for all the remaining Cassandra nodes.

```
[root@talena13 certs]# keytool -import -v -trustcacerts -alias talena1 -
file talena1.cer -keystore .truststore
Enter keystore password:
Owner: CN=talena1, OU=, O=, L=, ST=, C=
```

```
Issuer: CN=talena1, OU=, O=, L=, ST=, C=
Serial number: 203a2356
Valid from: Tue Jul 11 19:14:33 IST 2017 until: Mon Oct 09 19:14:33 IST
2017
Certificate fingerprints:
MD5: 17:4B:40:00:DA:A0:38:82:3A:88:37:4D:BE:8F:92:C7
SHA1: A7:C1:5E:B8:A3:7F:49:C3:C8:66:76:51:7B:E3:28:D4:F5:D7:62:C9
SHA256:
FC:C3:99:69:9F:EC:40:EC:8B:E3:DD:FD:68:3E:30:01:66:F7:E8:DA:97:6F:71:13:19:
1B:35:6D:38:4B:30:77
Signature algorithm name: SHA256withRSA
Version: 3
Extensions:
#1: ObjectId: 2.5.29.14 Criticality=false
SubjectKeyIdentifier [
KeyIdentifier [
0000: BA 16 CB 9A CE DD 71 8B 98 4F 5F 75 39 CB F3 65 .....q..O_u9..e
0010: BF 8A B7 FF ....
]
]
Trust this certificate? [no]: yes
Certificate was added to keystore
[Storing .truststore]
[root@talena13 certs]# keytool -import -v -trustcacerts -alias talena2 -
file talena2.cer -keystore .truststore
Enter keystore password:
Owner: CN=talena2, OU=, O=, L=, ST=, C=
Issuer: CN=talena2, OU=, O=, L=, ST=, C=
Serial number: 456a0a54
Valid from: Tue Jul 11 19:14:31 IST 2017 until: Mon Oct 09 19:14:31 IST
2017
Certificate fingerprints:
MD5: 0E:36:CD:50:86:EE:47:BB:17:EA:E3:B9:60:9F:A7:C2
SHA1: 37:FD:A0:BF:B1:E3:15:1F:99:6B:AC:7D:91:66:43:65:85:0C:41:AC
SHA256:
8E:1F:11:71:1C:EC:BA:52:F1:42:CE:53:3C:63:C4:47:09:71:23:13:21:B0:E0:4B:AC:
CE:73:44:D5:B8:77:7D
Signature algorithm name: SHA256withRSA
Version: 3
Extensions:
#1: ObjectId: 2.5.29.14 Criticality=false
SubjectKeyIdentifier [
KeyIdentifier [
0000: 41 ED 3E 7D 10 56 62 01 A7 A1 21 5F F6 D9 7C D0 A.>..Vb...!.....
0010: D0 2E 03 34 ...4
]
]
Trust this certificate? [no]: yes
```

```
Certificate was added to keystore  
[Storing .truststore]
```

NOTE: Ensure that permissions of Cassandra's trust store file are given to the Cassandra user.

6.1.2.4 Getting to know Cassandra SSL requirements

The `cassandra.yaml` file is the main configuration file for Cassandra. In this file, you must change the following properties and restart the node for the changes to take effect.

- `client_encryption_options` (**Client-to-node encryption**)
 - `enabled`: By default, the `enabled` option is set to False. If the `enabled` option is set to True, then you must copy Cassandra certificates into the Imanis Data Truststore using `$INSTALL_DIR/bin/talena-ssl.sh`
 - `require_client_auth`: By default, the `require_client_auth` is set to False. If the `require_client_auth` is set to True, then you must copy Imanis Data SSL certificates into every Cassandra nodes TrustStore and restart Cassandra services for the changes to take effect
- `server_encryption_options` (**Node-to-node encryption**)
 - `internode_encryption`: If this option is set to all, dc or rack, then you must copy Cassandra certificates into Imanis Data TrustStore using `$INSTALL_DIR/bin/talena-ssl.sh`
 - `require_client_auth`: By default, the `require_client_auth` is set to False. If set to true, then you must copy Imanis Data SSL certificates into every Cassandra nodes TrustStore and restart Cassandra node for the changes to take effect

6.1.2.5 Setting the number of open files limit for Cassandra

The default setting for the number of open files in Cassandra primary is 1024. However, to be able to increase the number of open files limit in Cassandra, you must set `ulimit -n 100000` in Cassandra primary. This task is important without which you may face an error on the Primary cluster while executing Cassandra restore.

6.1.2.6 Importing Kerberos configuration files to Imanis Data

This section discusses the prerequisites of adding Kerberized Cassandra in Imanis Data. Refer to the following section for adding Kerberized Cassandra.

Prior to using Kerberized Cassandra cluster, you must import keytab and `krb5.conf` to Imanis Data cluster. However, in case of multiple KDC Server, you must import only the merged `krb5.conf` file which contains configuration for all KDC setups.

Following are the examples of merging the krb5.conf file: When Kerberized Imanis Data and Kerberized Cassandra are not sharing the Kerberos setups or when multiple Kerberized Cassandra clusters have their own Kerberos setups.

To import the configuration files, do the following:

1. Copy the krb5.conf file and user keytab file in a temporary location on any one node on the Imanis Data cluster. Ensure that the hdfs user has read permissions for the files.
2. Run the following command:

```
$INSTALL_DIR/bin/talena-jaasConfig.sh -k <keytab-path> -p <user-principal>
```
3. Enter the path of the krb5.conf file. The utility generates a Java Authentication and Authorization Service (JAAS) configuration file which is used by the Imanis Data GUI for user authentication purposes. If you have already added a few Kerberos Cassandra repositories and now want to add new Kerberos Cassandra Primary, then you must first merge the existing Imanis Data krb5.conf and the new KDC server krb5.conf.
4. Copy the path of the JAAS configuration file on a Notepad file and use it later during Imanis Data GUI verification for Cassandra user authentication.

6.1.2.7 Primary Directory Permission & Space Requirement for PIT Recovery

This section discusses the permission required for the Archive Log Directory and Commitlog Restore Path Directory which is required before using the Point-in-Time (PIT) Recovery feature.

6.1.2.7.1 Archive Log Directory

The Archive log directory is specified while enabling Point-in-Time (PIT) Recovery feature. The following must be specified for the Archive log directory:

- The directory should be present on all Cassandra nodes with same path
- The directory should have read, write, and execute permission for the backup user

6.1.2.7.2 Commitlog Restore Path Directory

The Commitlog Restore Path Directory is specified during creation of Cassandra PITR recovery job. You must create this directory on all Cassandra primary nodes. Imanis Data PIT recovery job will copy commitlog files to this directory.

As a good practice, the user must ensure that the Commitlog Restore Path Directory is empty.

The following must be specified for the Commitlog Restore Path Directory:

- The directory should be present on all Cassandra nodes of destination cluster with same path.
- The directory should have read, write and execute permission for Imanis Data recovery user

- The partition hosting this directory should have sufficient free disk space so that it can accommodate archived commitlog files

6.1.3 Couchbase

The following sections describe the actions you need to perform before adding a Couchbase data repository.

6.1.3.1 Creating SSL certificates in Couchbase and importing them in Imanis Data cluster

Couchbase Server client libraries support client-side encryption using the SSL protocol by encrypting data in-flight between the client and the server. This section describes the procedure of creating SSL certificates in Couchbase and importing them in Imanis Data cluster.

If you already have SSL certificate (*.cer) files, you can refer to Step #4. However, if you do not have the certificate, follow these steps:

1. Access the Couchbase Web Console and copy the certificate file.
2. On Imanis Data cluster, run command **cat > cluster.cert** to copy-paste the certificate, and then press **Ctrl+D** to save the certificate.
3. Import the SSL certificates using a `talena-ssl.sh` script. Run the script as hdfs user and set a password as per your preference. The SSL certificate will now be successfully added to the keystore.
4. Certificate Authority (CA) certificates can be configured on Couchbase version 4.5 and above. If you have Certificate Authority (CA) signed certificates, all chain pem files should be imported too. For more information, refer to the Couchbase documentation [here](#) and [here](#).

For more information, refer to the [Couchbase documentation on SSL based client-server communication](#).

6.1.4 MongoDB

This section describes the prerequisites of adding MongoDB:

6.1.4.1 Setting up mongod and mongos with TLS/SSL Certificate and Key

This section describes the steps you need to perform to set up mongod and mongos with TLS/SSL Certificate and Key.

To configure mongod or mongos to use TSL/SSL connections, do the following:

1. Modify MongoDB config to do the following:
 - set **net.ssl.mode** to **requireSSL**

- set **net.ssl.PEMKeyFile** to the .pem file that contains the TLS/SSL certificate and key in the MongoDB configuration file.

2. If the key is encrypted then specify the passphrase to decrypt using **net.ssl.PEMKeyPassword**.

```
net:  
  ssl:  
    mode:requireSSL  
    PEMKeyFile: /etc/ssl/mongodb.pem  
    PEMKeyPassword: passphrase (optional)
```

3. Restart mongod and mongos daemons by specifying above configuration files. For example:

```
$ mongod --config <pathToConfigFile>  
$ mongos --config <pathToConfigFile>
```

4. Optionally, you can also configure SSL for mongod and mongos using command-line options instead of configuration file:

```
$ mongod --sslMode_requireSSL --sslPEMKeyFile <filename>[--sslPEMKeyPassword  
<value>]  
$ mongos --sslMode_requireSSL --sslPEMKeyFile <filename>[--sslPEMKeyPassword  
<value>]
```

6.1.4.2 Importing MongoDB cluster certificates to Imanisdata cluster

For MongoDB backup/recovery using Imanisdata, it requires all the mongod and mongos SSL certificates of primary cluster to be imported to Imanisdata cluster truststore.

- Import SSL Certificates to truststore using following command:

```
$ talena-ssl.sh import <aliasname><pathToCertificateFile>
```

6.1.4.3 Creating Talena Data Repository for SSL Enabled MongoDB Cluster

While creating data repository “SSL Required” option must be selected to connect to SSL enabled MongoDB cluster.

6.1.4.4 MongoDB SSL with Single CA (Certificate Authority)

This section explains how to configure MongoDB cluster with single CA (Certificate Authority), so that Imanisdata cluster just requires single CA certificate to be imported to access the Mongodb cluster.

6.1.4.5 Generating CA certificate (self signed)

This certificate is typically provided by third party.

```
openssl genrsa -out ca.key 2048          // Generate a certificate key
# openssl genrsa -out ca.key 2048
Generating RSA private key, 2048 bit long modulus
.....+++
.....++
e is 65537 (0x10001)
```

```
openssl req -new -x509 -key ca.key -out ca.crt // Generate certificate
# openssl req -new -x509 -key ca.key -out ca.crt
You are about to be asked to enter information that will be incorporated
into your certificate request.
What you are about to enter is what is called a Distinguished Name or a DN.
There are quite a few fields but you can leave some blank
For some fields there will be a default value,
If you enter '.', the field will be left blank.
-----
Country Name (2 letter code) [XX]:IN
State or Province Name (full name) []:MH
Locality Name (eg, city) [Default City]:PUNE
Organization Name (eg, company) [Default Company Ltd]:TALENA
Organizational Unit Name (eg, section) []:DEV
Common Name (eg, your name or your server's hostname) []:talenal.qa.com
Email Address []:
```

6.1.4.6 Generating private and public key for MongoDB Server

```
openssl genrsa -out private.key 2048          // Generate a private Key
# openssl genrsa -out private.key 2048
Generating RSA private key, 2048 bit long modulus
.....++++
.....++++
e is 65537 (0x10001)
[root@talenal vikram]# ls
mongod.conf  private.key
#
```

```
openssl rsa -in private.key -noout -text          // Print key
# openssl rsa -in private.key -noout -text
Private-Key: (2048 bit)
modulus:
00:af:71:a8:1c:75:e7:d7:76:1f:95:32:64:57:96:
58:be:9c:0d:06:91:41:b4:94:b2:c1:b8:60:c5:da: ...
publicExponent: 65537 (0x10001)
privateExponent:
1f:52:37:24:fd:97:aa:4b:98:4f:d6:73:3b:7f:c7:
D0:1f:a3:e3:ac:43:02:ae:19:4b:a4:53:02:7f:7b: ...
prime1:
00:e2:32:ce:ae:5c:1d:5d:8b:23:c8:06:27:48:43:
55:d7:fb:98:9d:3c:43:7e:cc:71:0d:e2:46:44:f7: ...
prime2:
00:c6:8f:02:1f:69:d0:bf:10:6e:fd:ec:6c:80:2f:
A2:8a:6b:0a:ff:2d:5a:2f:52:10:29:8b:c4:a1:5e: ...
exponent1:
0a:c7:67:07:f2:05:c8:36:60:2f:20:f3:f0:42:9d:
2d:9c:a4:aa:21:7c:09:e7:ce:1f:5e:40:00:fb:52: ...
exponent2:
09:b7:5f:cc:37:ba:5e:4f:28:8f:46:6d:7c:cc:57:
F0:dc:12:1a:b0:96:74:30:58:d8:1f:9d:cc:a3:b4: ...
coefficient:
3f:3f:25:2d:26:e0:90:e1:e0:3d:9f:b0:43:76:8d:
A7:8e:14:f6:a1:25:9a:df:a5:08:84:24:3a:1e:ed: ...
```

```
openssl rsa -in private.key -pubout -out public.key // Get public key
# openssl rsa -in private.key -pubout -out public.key
writing RSA key
```

```
openssl rsa -in public.key -pubin -noout -text // Print public Key
# openssl rsa -in public.key -pubin -noout -text
Public-Key: (2048 bit)
Modulus:
00:af:71:a8:1c:75:e7:d7:76:1f:95:32:64:57:96:
58:be:9c:0d:06:91:41:b4:94:b2:c1:b8:60:c5:da:
```

6.1.4.7 Generate CSR (Certificate Signing Request) from private key

```
openssl req -new -key private.key -out server.csr

# openssl req -new -key private.key -out server.csr
You are about to be asked to enter information that will be incorporated
into your certificate request.
What you are about to enter is what is called a Distinguished Name or a DN.
There are quite a few fields but you can leave some blank
For some fields there will be a default value,
If you enter '.', the field will be left blank.
-----
Country Name (2 letter code) [XX]:IN
State or Province Name (full name) []:MH
Locality Name (eg, city) [Default City]:PUNE
Organization Name (eg, company) [Default Company Ltd]:TALENA
Organizational Unit Name (eg, section) []:DEV
Common Name (eg, your name or your server's hostname) []:talenal.qa.com
Email Address []:sdfdf

Please enter the following 'extra' attributes
to be sent with your certificate request
A challenge password []:
An optional company name []:
#
```

```
openssl req -in server.csr -noout -text

# openssl req -in server.csr -noout -text
Certificate Request:
    Data:
    Version: 0 (0x0)
    Subject: C=IN, ST=MH, L=PUNE, O=TALENA, OU=DEV, CN=talenal.qa.com/emailAddress=sdfdf
    Subject Public Key Info:
        Public Key Algorithm: rsaEncryption
        Public-Key: (2048 bit)
        Modulus:
            00:af:71:a8:1c:75:e7:d7:76:1f:95:32:64:57:96:
            58:be:9c:0d:06:91:41:b4:94:b2:c1:b8:60:c5:da:
```

6.1.4.8 Generate server PEMkeyFile to be supplied to MongoDB Server

```
openssl x509 -req -in server.csr -CA ca.crt -CAkey ca.key -CAcreateserial -out
server.crt
# openssl x509 -req -in server.csr -CA ca.crt -CAkey ca.key -CAcreateserial -out server.crt
Signature ok
subject=/C=IN/ST=MH/L=PUNE/O=TALENA/OU=DEV/CN=talenal.qa.com/emailAddress=sdfdf
Getting CA Private Key
```

```
openssl x509 -in server.crt -noout -text          // Print server.crt
# openssl x509 -in server.crt -noout -text
Certificate:
    Data:
        Version: 1 (0x0)
        Serial Number: 17667182684445226008 (0xf52e710bf84a3418)
        Signature Algorithm: sha1WithRSAEncryption
        Issuer: C=IN, ST=MH, L=PUNE, O=TALENA, OU=DEV, CN=talenal.qa.com
        Validity
            Not Before: Aug 20 10:16:17 2019 GMT
            Not After : Sep 19 10:16:17 2019 GMT
        Subject: C=IN, ST=MH, L=PUNE, O=TALENA, OU=DEV, CN=talenal.qa.com/emailAddress=sdfd
        Subject Public Key Info:
            Public Key Algorithm: rsaEncryption
            Public-Key: (2048 bit)
                Modulus:
                    00:af:71:a8:1c:75:e7:d7:76:1f:95:32:64:57:96:
                    58:be:9c:0d:06:91:41:b4:94:b2:c1:b8:60:c5:da:
```

```
cat server.crt private.key > server.pem          // Generate server.pem
# cat server.crt private.key > server.pem
# cat server.pem
-----BEGIN CERTIFICATE-----
MIIDUzCCAjsSCCQD1LnEL+Eo0GDANBgkqhkiG9w0BAQUFADBhMQswCQYDVQQGEwJJ
TjELMAkGA1UECAwCTUgxDTALBgNVBAcMBFBVTkUxDzANBgNVBAoMB1RBTEVOQTEM...
-----END CERTIFICATE-----
-----BEGIN RSA PRIVATE KEY-----
MIEogIBAAKCAQEAr3GoHHXn13YflTJKv5ZYvpwNBpFBtJSywbhgxdpQ06uv/t3+
4f+6sptXrs2WYbDX7tyRfPuVYO4BwfHcfVi5Xm3pA2vkXpcPdlGFWyROYaK865v...
-----END RSA PRIVATE KEY-----
```

6.1.4.9 Modifying MongoDB config to use server.pem and ca.crt

```
mode: requireSSL
PEMKeyFile: /path/to/server.pem
CAFfile: /path/to/ca.crt
weakCertificateValidation: true
```

6.1.4.10 Connecting to MongoDB Cluster using CA

Connect to Cluster using CA certificate (ca.crt file) with following mong shell command

```
mongo --ssl --sslCAFile ./ca.crt --host talenal.qa.com --port 27201
```

6.1.4.11 Importing CA certificate to Imanisdata

For Imanisdata cluster to get access to MongoDB cluster, just import ca.crt to talena-truststore using following command:

```
talena.ssh import ca-aliasname ./ca.crt
```

6.1.4.12 Ensuring all nodes nodes in MongoDB cluster are reachable

Make sure that all nodes in the MongoDB cluster must be reachable through hostname and IP address from all nodes in the Imanis Data cluster.

6.1.4.13 Creating a User & Assigning Roles

Prior to adding MongoDB data repository in Imanis Data, you must create a MongoDB user and then assign specific roles to the user. This MongoDB user will be used in Imanis Data software at the time of adding the MongoDB data repository.

You must ensure that the MongoDB user must have clusterAdmin and readWriteAnyDatabase roles. For example:

```
db.createUser({  
    user: "username",  
    pwd: "password",  
    roles : [  
        { role: "clusterAdmin", db: "admin" },  
        { role: "readWriteAnyDatabase", db: "admin" }  
    ]  
})
```

For more information on the preceding roles, refer to the MongoDB documentation on Built-In Roles.

NOTE: This is applicable for LDAP authentication as well.

6.1.4.14 Syncing time on MongoDB cluster and Imanis Data cluster

Ensure that the time on MongoDB cluster and Imanis Data cluster is in sync. This is a mandatory requirement as Imanis Data accesses MongoDB Oplogs for incremental backups.

6.1.4.15 Adding a new shard to an existing MongoDB cluster

The following point must be considered if you are adding a new shard to an existing MongoDB cluster:

- If a shard is added to a MongoDB cluster, the shard cannot be removed before the next run of the backup

NOTE: Adding or removing a shard from the MongoDB cluster will result into a full back up on the workflows that are protecting data for the cluster.

6.1.4.16 Ensuring unique value of _id in a collection

The field name `_id` is reserved for use as a primary key for a document and its value must be unique in the collection. Refer to the following MongoDB documentation for more information:

<https://docs.mongodb.com/manual/core/document/#field-names>

<https://docs.mongodb.com/manual/core/sharding-shard-key/#sharding-shard-key-unique>

6.1.4.17 Ensuring minimum storage for backups

The minimum storage recommended for storing MongoDB backup is 4x backup size including storage space for daily changes.

For example, for 1 TB of backup data, the user must make provision for 4 TB disk space + additional space based on change rate and the retention requirements.

6.1.4.18 Standalone MongoDB clusters are not supported

Standalone MongoDB clusters do not have an Oplog. Oplog is mandatory as Imanis Data accesses MongoDB Oplogs for incremental backups.

Standalone cluster must be converted to replica sets using the process outlined in this reference link

- <https://docs.mongodb.com/manual/tutorial/convert-standalone-to-replica-set/>

6.1.5 Amazon Glacier

In Imanis Data software, you can configure Amazon Glacier to work as a regular cloud data repository only. Adding or configuring the data repository to work as source data repository or a global cloud data repository is not permitted.

To add Amazon Glacier, do the following:

1. Create an account on Amazon Glacier at <http://aws.amazon.com/> and provide “Amazon Glacier Full Access” or “root access” to the user.
2. Create a user and generate the access key and secret key. Every user must have his or her own access key and secret key.
3. Create a region-specific vault.

Amazon Glacier pricing is based on several factors such as the (geographical) region where your data would reside, amount of data, duration of data archival, amount of data to be retrieved, number of requests and so on. To better understand the pricing, visit this unofficial Amazon Glacier cost estimator website.

6.1.6 Amazon S3

In Imanis Data software, you can configure Amazon S3 to work as a Global Cloud Data Repository and a Regular Cloud Data Repository only. You are not permitted to add or configure it to work as Source data repository.

To add Amazon S3, do the following:

1. Create an account on Amazon S3 at <http://aws.amazon.com/s3/>
2. Create a bucket in a region that is geographically close to you to optimize latency, minimize costs, or address regulatory requirements. Refer to the Amazon S3 documentation on using Buckets [here](#).
3. Create an IAM user and generate the access key and secret key. Every IAM user must have his or her own access key and secret key.
4. Grant access to the IAM user to list buckets and give access to the bucket created in Step#2. To give access to the bucket (for example named as ‘TexasWarehouse’), add the following policy using ‘Add Policy Inline’ option in the Amazon S3 Web Console.

```
{  
    "Statement": [  
        {  
            "Effect": "Allow",  
            "Action": "s3>ListAllMyBuckets",  
            "Resource": "arn:aws:s3:::*"  
        },  
        {  
            "Effect": "Allow",  
            "Action": [  
                "s3>ListBucket",  
                "s3:GetBucketLocation"  
            ],  
            "Resource": [  
                "arn:aws:s3::: TexasWarehouse"  
            ]  
        },  
        {  
            "Effect": "Allow",  
            "Action": [  
                "s3>PutObject",  
                "s3>GetObject",  
                "s3>DeleteObject"  
            ],  
            "Resource": [  
                "arn:aws:s3::: TexasWarehouse/*"  
            ]  
        }  
    ]  
}
```

With Amazon S3, you pay only for the storage you actually use. There is no minimum fee and no setup cost. Refer to the Amazon S3 Pricing for more information.

Dos & Don'ts of using Amazon S3 Buckets

- Use a dedicated bucket with all permissions
- Do not use the bucket for any other miscellaneous data except for Imanis Data data
- Data saved in S3 buckets is available in a proprietary format and cannot be used directly. Contact Imanis Data Software Support for reconstruction of the data which may or may not be possible depending on the original data

WARNING: Do not use the same bucket for two different Imanis Data clusters.

6.1.7 Azure Blob Storage

In Imanis Data software, you can configure Azure data repository as Global Cloud Data Repository and Regular Cloud Data Repository only. You are not permitted to add or configure it to work as a Source data repository.

Before you can create a blob storage account, you must have an Azure subscription, which is a plan that gives you access to a variety of Azure services including blob storage account.

To add Azure Blob Storage, do the following:

1. Go to <https://azure.microsoft.com> to open a free Azure account. All you need is a phone number, a credit card, and a Microsoft Account username (formerly Windows Live ID).
2. Add an Azure subscription plan as per your need.
3. A Windows Azure subscription grants you access to Windows Azure services and to the Windows Azure Platform Management Portal.
4. Create an account on Microsoft Azure at <https://portal.azure.com>.
5. It is recommended to set the Primary cluster, Imanis Data cluster, and Azure storage account in the same region to optimize latency, minimize costs, or address regulatory requirements.
6. Create a user and generate the shared access key. Every user must have his or her own access key. You can use the account name and key to create Azure Data repository in Imanis Data.
7. Create a container. Refer to the Microsoft Azure documentation on using Containers.
8. Grant access of the container to the user that was created in step #4.

NOTE: For fine grain access you may create Shared Access Signatures (SAS) and use that instead account name and key. Please create the SAS with sufficient permissions such as:

- Allowed services: Blob
- Allowed resource types: Object, Service

- Allowed permissions: Read, Write, Delete, List, Create

Dos & Don'ts of using Azure Blob Storage

- Use a dedicated container with all permissions
- Do not use the container for any other miscellaneous data except data of Imanis Data
- Data saved in Azure containers is available in a proprietary format and cannot be used directly. Contact Imanis Data Software Support for reconstruction of the data which may or may not be possible depending on the original data

WARNING: Do not use the same container for two different Imanis Data clusters.

6.1.8 Cloudian

In Imanis Data software, you can configure Cloudian to work as a Global Cloud Data Repository and a Regular Cloud Data Repository only. However, you are not permitted to add or configure it to work as Source data repository. Cloudian uses S3 API which is compatible with S3-enabled applications.

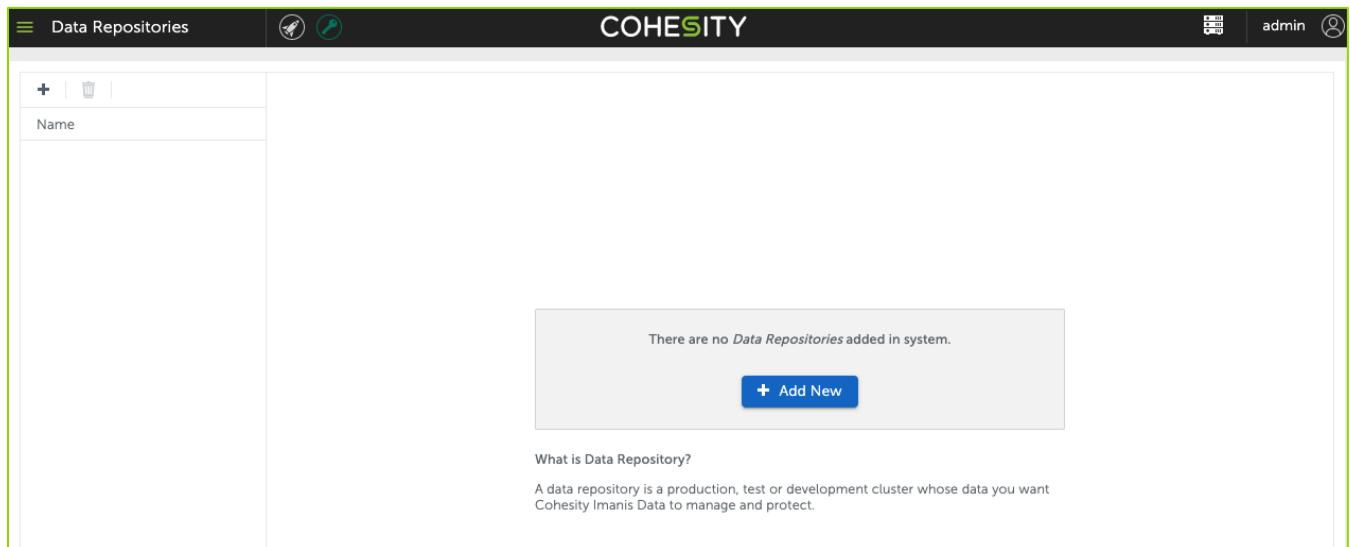
Refer to the following information to ensure you have the Cloudian versions supported by Imanis Data:

- cloudian-redis-2.8.23-1.el7.x86_64
- cloudian-cassandra-2.2.9-1.el7.x86_64
- cloudian-s3-7.0.5-1.el7.x86_64
- cloudian-agent-7.0.5-1.el7.x86_64
- cloudian-tomcat-7.0.85-1.el7.x86_64

6.2 Adding Data Repositories

Imanis Data software supports Hadoop (HDFS, Hive, and HBase), Cassandra, Couchbase, MongoDB, Amazon Glacier, Amazon S3, and Azure Blob Storage data repositories.

However, if license of a database expires then the user will not be permitted to create new data repositories or create backup workflows pertaining to the particular database.



6.2.1 Auto Discovery

Imanis Data software enables administrators to auto discover settings of Hadoop and Cassandra applications in the system. If the primary Hadoop cluster is Namenode HA, you must do the ‘auto discovery of applications’ task only.

IMPORTANT: Before you get started with auto discovery of applications, you must enable WebHDFS on all datanodes and namenodes. If you do not enable WebHDFS, backup may not function as expected.

NOTE: The Auto Discovery feature is available only for Hadoop and Cassandra. It is not available for Glacier, S3, and Azure.

6.2.1.1 Hadoop (HDFS, Hive, HBase)

The following Hadoop applications are supported by Imanis Data software:

- HDFS, a distributed, scalable, and portable file-system
- Hive, a data warehouse software for Hadoop
- HBase, an open-source, distributed, versioned, non-relational database
- Microsoft's Hadoop on Azure service supports Hadoop, HBase, Spark, and Storm.

Imanis Data software automatically discovers the settings in xml configuration files for HDFS, Hive and HBase on your primary Hadoop cluster. Refer to the section Version Compatibility Matrix for information on versions supported by Imanis Data software.

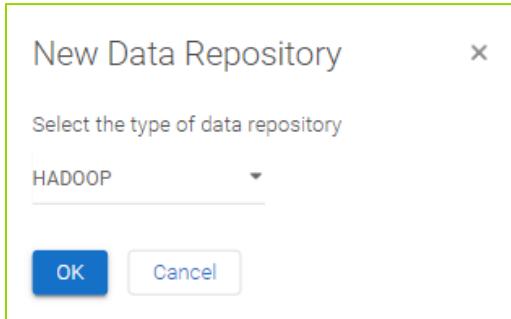
NOTE: Imanis Data software provides an option to connect to your Hive data repository with a custom user other than default user Hive.

This option is valid for non-kerberized Hadoop setup and can be enabled at Imanis Data-site level, that is, all hive data repositories that are added in Imanis Data system can be connected using this user only. You have to ensure that this alternate superuser is able to access all the hive objects and their data. This user also needs all the read, write, and delete permissions.

To enable this option, contact Imanis Data Customer Support on support@imanisdata.com.

To do auto-discovery of Hadoop applications, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the  icon. The following dialog box appears:



3. In the **New Data Repository** dialog box, select **Hadoop** from the **Select the type of data repository** drop-down menu, and then click **OK**.
4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - b. **Description:** A meaningful description of the data repository to help others identify it
5. In the **Applications** section, type the following:
 - a. **Primary Host:** The name or IP address of the machine on which the configuration directories related to the applications exist.
 - b. **Configuration Directory:** A list of one or more paths where configuration files related to the application are present. If any of HDFS, Hive, and/or HBase are Kerberized, then a new field Kerberos Principal is displayed on the screen along with the buttons and fields.

Refer to the following information on configuration files:

- For HDFS: the path must have `core-site` and `hdfs-site.xml`
- For Hive: the path must have `hive-site.xml`
- For Hbase: the path must have `hbase-site.xml`

6. Click **Discover**. This action displays the **Credentials** dialog box that prompts you to type your login credentials.

7. In the **Credentials** dialog box, under **Authentication Mechanism** do one of the following:
 - Click the **Password** option button, type the login and password, and click **OK**
 - Click the **Private Key** option button, type the login id, copy-paste the private key, type the passphrase (if you have set the passphrase), and then click **OK**

This action displays a list of all applications discovered in the data repository by Imanis Data software. For example, HDFS, HBase or Hive.

a. Refer to the following table, if HDFS file system is discovered:

BUTTONS & FIELDS	DESCRIPTION	ACTION TO TAKE	EXAMPLE
Namenode Address	Stores the directory tree of all files in the file system and tracks where the file data is kept across the cluster. It does not store the data of these files itself.	NA	t35vm1
Webhdfs Port	Provides “read and write” access. Supports all HDFS operations (like granting permissions, configuring replication factor and accessing block location)	NA	50070
Kerberos Principal	A unique identity in a Kerberos system to which Kerberos can assign tickets to access Kerberos-aware services. Make sure the kerberos principal is a superuser. Refer to Hadoop documentation for creating alternative superuser.	1. Type the Kerberos Principal. 2. Append the principal's keytab entry to Imanis Data keytab file.	prodhdfs@t25vm8.realm

IMPORTANT: HDFS Connection and Data movement will be done through HDFS Kerberos principal.

b. Refer to the following table, if Hive is auto-discovered:

BUTTONS & FIELDS	DESCRIPTION	ACTION TO TAKE	EXAMPLE
Metastore Address	Stores the metadata for Hive tables and partitions in a relational database	NA	t35vm1
Metastore Port	Service for accessing metadata about Hive tables and partitions	NA	9083
Kerberos Principal	A unique identity in a Kerberos system to which Kerberos can assign tickets to access Kerberos-aware services. Make sure the kerberos principal is a superuser. Refer to Hadoop documentation for creating alternative superuser.	1. Type the Kerberos Principal. 2. Append the principal's keytab entry to Imanis Data keytab file.	prodhive@t25vm8.realm

IMPORTANT: During auto-discovery of Hive, two metastore addresses could be displayed on the Imanis Data UI if the Hive is configured in HA. Make sure that you remove one of the metastore address from the Metastore Address field without which Hive discovery will not succeed.

IMPORTANT: Imanis Data does not support Hive HA.

IMPORTANT: Hive Metastore connection will be executed through Hive Kerberos Principal. However, Data Movement will be executed through HDFS Kerberos Principal. This framework is designed to handle external tables when accessing data outside Hive warehouse directory and thus requires super-user access.

In addition, **Hive Server Address** and **Hive Server Port** fields are also automatically populated depending upon the configuration properties in the `hive-site.xml` file.

c. Refer to the following table, If HBase is discovered:

BUTTONS & FIELDS	DESCRIPTION	ACTION TO TAKE	EXAMPLE
Hadoop Distribution & Version	Available options in the Hadoop Distribution drop-down menu are Cloudera and Hortonworks	Select a Hadoop Distribution type from the Select Hadoop Distribution drop-down menu and then type the version number in the Version field.	If you have Cloudera 5.14.1 as the primary cluster version, then you must type 5.8.2 in the Version field.
Zookeeper Quorum	Comma separated full list of ZooKeeper quorum servers which are set to localhost for local and pseudo-distributed modes of operation. For example, "host1.mydomain.com,host2.mydomain.com,host3.mydomain.com"	NA	t14vm9.qa.com, t14vm8.qa.com
HBase Data Root Directory	The directory shared by region servers and into which HBase persists. The URL should be 'fully-qualified' to include the filesystem scheme. For example, to specify the HDFS directory '/hbase' where the HDFS instance's namenode is running at namenode.example.org on port 9000, set this value to: hdfs://namenode.example.org:9000/hbase.	NA	hdfs://t14vm8.qa.com:8020/apps/hbase/data
Kerberos Principal	A unique identity in a Kerberos system to which Kerberos can assign tickets to access Kerberos-aware services. Make sure the kerberos principal is a superuser. Refer to Hadoop documentation for creating alternative superuser.	1. Type the Kerberos Principal. 2. Append the principal's keytab entry to Imanis Data keytab file.	prodhbase@t25vm8.realm

IMPORTANT: For Kerberized-HBase, ensure that the principal specified during the data repository verification process can traverse the hbase data directory. HBase backup agent traverses this data directory during list generation after taking snapshot of tables.

IMPORTANT: Similar to Hive, the HBase operations (snapshot, listing) will be executed through HBase super user kerberos principal and the Data Movement will be executed through HDFS Kerberos Principal. This framework is designed to copy data to temp directory and then bulk load from that directory. Attributes are also changed during data-copy phase. This process may require super user access.

NOTE: Ensure that the following property is present on primary cluster for setting up external KDC and cross realm trust for HBase:

```
<property>
  <name>hbase.superuser</name>
  <value>hbase</value>
</property>
```

NOTE: In the current release, Imanis Data software does not support different namenode ports in case of HA setup because it is assumed that all namenodes are listening on the same port. For example, if there are two namenodes in a Hadoop cluster namely namenode1, namenode2 having two different http ports: 50070 and 50071 respectively, Imanis Data software does not support it. Let's have a look at the properties listed in the hdfs-site.xml file, where examplenamespace-10-1-10-62 is the name of the nameservice. Imanis Data software does not support the following scenario:

```
<property>
  <name>dfs.namenode.http-address.examplenamespace-10-1-10-
62.namenode1</name>
  <value>t34vm2:50070</value>
</property>

<property>
  <name>dfs.namenode.http-address.examplenamespace-10-1-10-
62.namenode2</name>
  <value>t34vm5:50071</value>
</property>
```

8. Click **Verify** to confirm the selection of your data repository.
9. Set a data backup window and data recovery window, and then click **Save**. Learn more.

6.2.1.2 Cassandra

You can add Cassandra application through the auto discovery option only. Imanis Data software also supports Kerberized Cassandra. Refer to the section Version Compatibility Matrix for information on versions supported by Imanis Data software. Before you add Cassandra data repository make sure perform the following actions:

1. Enable the Remote JMX connection: For Cassandra 2.0.14 (2.0 branch) and Cassandra 2.1.4 (2.1 branch) onwards.
2. Access the `cassandra-env.sh` from the Cassandra configuration directory on all the Cassandra nodes and set `["$LOCAL_JMX" = "yes"]` entry to "no".
3. For the default DSE installation, the preceding file is available in `/etc/dse/cassandra`.
4. For a successful data source discovery in DSE 6.7.x, the following parameter must be set in `/etc/dse/cassandra/cassandra-env.sh`.
`"JVM_OPTS"="$JVM_OPTS -Djava.rmi.server.hostname"`

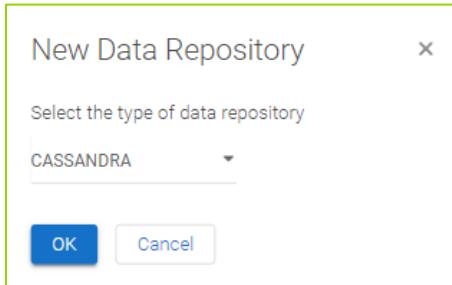
Refer to the section Setting up SSL on the Imanis Data cluster if you are trying to add a Cassandra data repository from an SSL-enabled Primary cluster:

NOTE: It is mandatory to rediscover and re-verify the Cassandra data repository in Imanis Data software whenever Cassandra cluster is upgraded from DSE4.x to DSE5.x. Also, any major version upgrade (2.0 to 2.1 or 2.1 to 3.0) would also require a rediscovery and re-verification.

IMPORTANT: Ensure that Cassandra service is up and running on the "Primary host" that is pointed for application discovery.

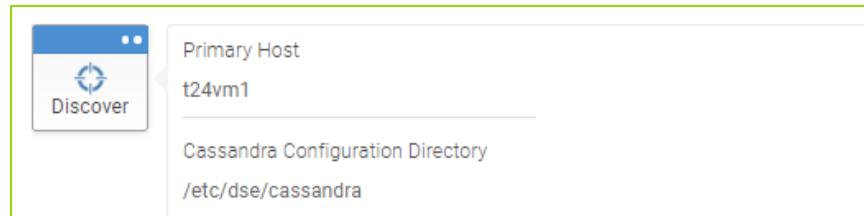
To do auto discovery of Cassandra application, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the  icon. The following dialog box appears:

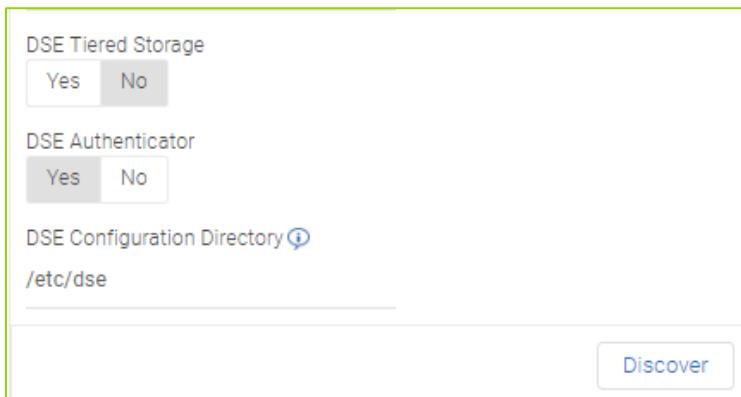


3. In the **New Data Repository** dialog box, select **Cassandra** option from the **Select the type of data repository** drop-down menu, and then click **OK**.
4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.

- b. **Description:** A meaningful description of the data repository to help others identify it.
5. In the **Discover** section, do the following:
- Type the name or IP address of the machine on which the configuration directories related to the applications exist in the **Primary Host** field.
 - Type the directory path where cassandra.yaml exists on the Cassandra primary in **Cassandra Configuration Directory** field. This path has to be same on all nodes.

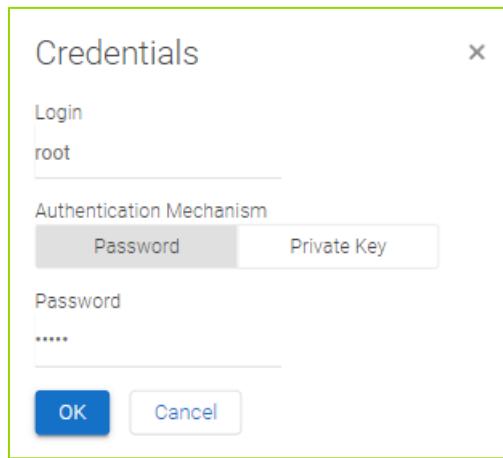


6. In the **DSE tiered storage** option, select the **Yes** option and then type DSE configuration directory details in the **DSE Configuration Directory** field. Ensure that the configuration directory that you mention is same on all the nodes. The DSE tiered storage option is applicable from DSE 5.0 onwards only:

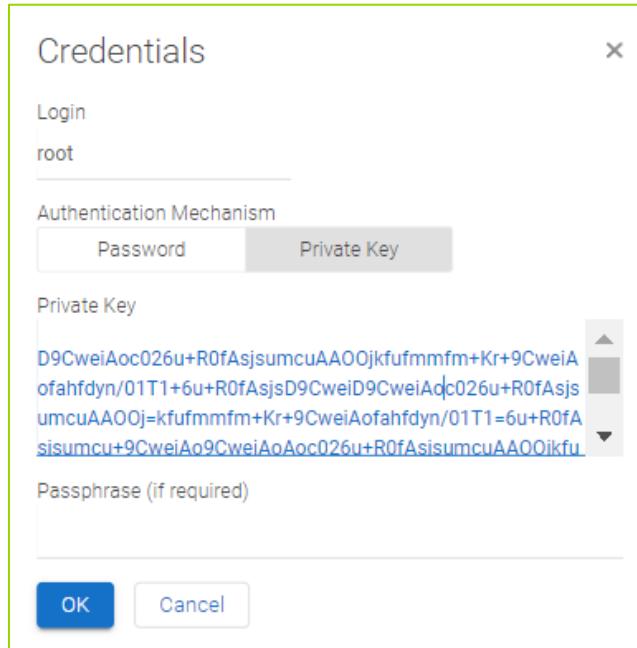


NOTE: Ensure that read and write permissions are granted for SCP (Secure Copy Protocol) based restore for the SSH user on the Cassandra configuration and data directories on all nodes. Also, make sure that the same SSH user (with similar permissions) should be present on all Cassandra nodes.

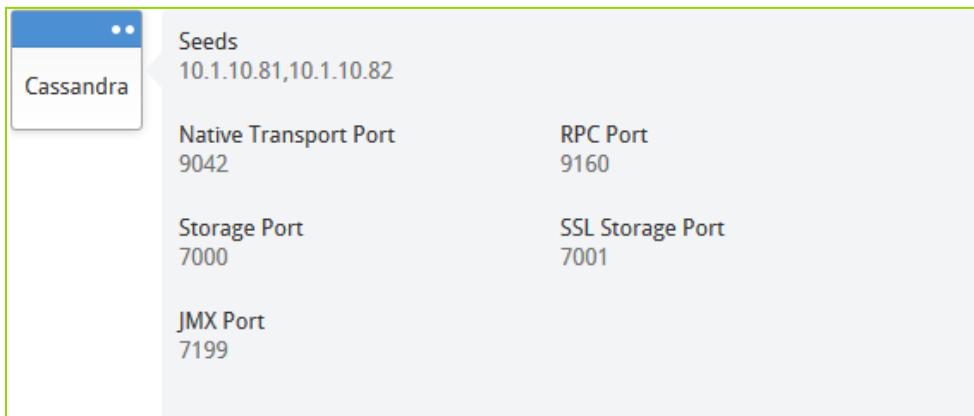
- Click **Discover**. The **Credentials** dialog box will appear.
- In the **Credentials** dialog box, do one of the following:
 - In the **Password** option, type the login and password, and then click **OK**:



- In the **Private Key** option button, type the login id, copy-paste the **private key**, type the passphrase (if you have set the passphrase), and then click OK. For more information on Private Key, refer to the section Creating Private Key on Linux:



- The preceding action auto-populates information related to the Cassandra application that you are trying to add. Refer to the following table for information about the fields:



BUTTONS & FIELDS	DESCRIPTION
Seeds	Cassandra nodes that are mentioned in the <code>cassandra.yaml</code> file
Native Transport Port	Port for the CQL native transport
RPC Port	Remote Procedure Call (RPC) port for general mechanism for client-server applications
Storage Port	TCP port for commands and data.
SSL Storage Port	SSL port for encrypted communication. Internally used by the Cassandra bulk loader
JMX Port	Cassandra management port

10. In the Data Centers field, type all the data centers names. Data Centers is a collection of related nodes. A data center can be a physical data center or virtual data center.

The screenshot shows a text input field with the placeholder "Data Centres". Inside the field, the text "DC1,DC2" is typed. The entire input field is highlighted with a light gray box.

IMPORTANT: In case, it is a Multi-DC environment, all the DC names must be typed manually, if not listed, in the Data Centers field. This is a mandatory step if backups or restores need to be supported at the DC level.

If you leave the Data Centers field as blank, it indicates that DC specific data movement is NOT required. You will NOT be able to select data centers when creating backup or restore jobs.
Hover the mouse on the Data Centers field to refer to the tooltip for more information.

NOTE: Ensure that all data centers are listed in the textbox. In most cases, the data centers will be listed after discovery operation however, if they are not listed, you must manually type the data center names. Leaving the field blank will disallow data center-specific backup or restore. Entering a subset of all data centers may cause problems in data movement.

11. In the **DSE Authenticator** option, select **Yes** and then do one of the following:

- If the DSE Authenticator type is Kerberos, then type the JAAS configuration path
- If the DSE Authenticator type is Internal, then enter the Username and Password

12. In the **Cassandra Authentication** area, do one of the following depending on DSE Authenticator type in the preceding procedure step:

- Copy-paste the JAAS configuration path in the JAAS Config Path field. Refer to the steps mentioned in the Cassandra prerequisites section to get the JAAS configuration path

Cassandra Authentication
JAAS Config Path on Imanis Data
`/opt/imanisdata/conf/jaas-configurati`

- Type the Cassandra Username and Password in the respective fields

Cassandra Authentication
Username Password
`cassandra` `*****`

13. Cassandra has many options in the Authenticator and Authorization options. The following tables displays options that are supported by Imanis Data software:

For Cassandra 2.x, the following options are supported in backup and recovery workflows:

Authenticator

- AllowAllAuthenticator
- PasswordAuthenticator

Authorizer

- AllowAllAuthorizer
- AllowAllAuthorizer

NOTE: KerberosAuthenticator and LdapAuthenticator are not supported for Cassandra 2.x and 3.x

For Cassandra 3.x, the following options are supported options by Imanis Data for backup and recovery workflows:

Authenticator	Authorizer
• AllowAllAuthenticator	• AllowAllAuthorizer
• AllowAllAuthenticator	• CassandraAuthorizer
• DseAuthenticator	• DseAuthorizer
○ internal	

14. In the JMX Authentication area, type the JMX Username and JMX Password in the respective fields.

JMX Authentication

Username: cassandra

Password: [REDACTED]

NOTE: The file containing JMX username and password can be found in `cassandra-env.sh`.

NOTE: In the current release, special characters such as \$, `, and \ are not supported in the following fields: SSH username, Cassandra Username, and JMX Username.

15. In the **DSE Search/Solr** option, click the **Yes** option and then type the host or ip of the nodes (separated by a comma) where Solr is enabled in the **Solr Nodes** field. Specify the configured Solr port number if Solr services are not running on the default 8983 port.

DSE Search/Solr

Yes No

Solr Nodes ⓘ
10.1.10.81,10.1.10.82

Solr Port
8983

16. In the **Point-time (PIT) Recovery** area, do the following:

- In the Enable PIT Recovery option, click Yes if you want to enable the Point in Time recovery feature for this data repository

Point-in-Time (PIT) RecoveryEnable PIT Recovery   Use this option sparingly as it may put considerable load on system resources.

IMPORTANT: If you disable the Point-in-Time (PIT) recovery feature by clicking No, then all the PITR commitlog data backup will be deleted. However, in rare cases, if you decide to delete the PITR-enabled data repository, then you first need to disable the PITR feature, wait for 15 minutes, and then delete the Cassandra data repository.

- In the **Archive log directory** field, type the path of the directory used to archive commit logs on Cassandra primary:

Archive log directory
/backup/archivelogdir

- In **Data Centers** field, select all or one of the data centers:

Data Centers
 Select All
 DC1
 DC2

- In the **Retain PIT metadata for __ days** field, type a number denoting number of days you want to retain the Point in Time metadata in Imanis Data cluster:

Retain PIT metadata for days

The default Point-in-Time (PIT) job run frequency is set to 15 minutes, that is, if the PIT job runs at 1300 hrs then the next PIT job run will be automatically executed at 1315 hrs.

IMPORTANT: If a Cassandra primary cluster is added twice as two different data repositories, then the Point in Time (PIT) Recovery feature must be enabled on only one data repository.

17. Click **Verify** to confirm the addition of the Cassandra application.
18. Set a data backup window and data recovery window, and then click **Save**. Learn more.

IMPORTANT: The `native_transport_port_ssl` option' is not supported for Cassandra 3.0.

6.3 Manual Configuration

If you may want to manually configure data repositories, refer to the following section that describes the process of manually adding HDFS and Hive.

NOTE: Imanis Data software provides an option to connect to your Hive data repository with a custom user other than default user Hive.

This option is valid for non-kerberized Hadoop setup and can be enabled at Imanis Data-site level, that is, all hive data repositories that are added in Imanis Data system can be connected using this user only. You must ensure that this alternate superuser is able to access all the hive objects and their data. This user also needs all the read, write, and delete permissions.

Contact Imanis Data Customer Support for enabling this option.

IMPORTANT: If Kerberos authentication is enabled on primary Hadoop cluster, then the manual discovery feature will not work for Hadoop applications, that is, HDFS, Hive, and HBase.

6.3.1 Hadoop (HDFS, Hive, HBase)

The following Hadoop applications are supported by Imanis Data software:

- HDFS, a distributed, scalable, and portable file-system
- Hive, a data warehouse software for Hadoop
- HBase, an open-source, distributed, versioned, non-relational database

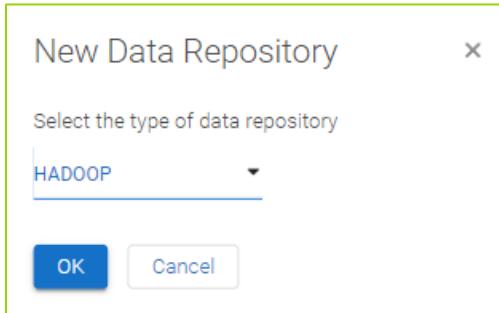
Imanis Data software automatically discovers the xml configuration files for HDFS, Hive and HBase on your primary Hadoop cluster. Refer to the section Version Compatibility Matrix for information on versions supported by Imanis Data software.

6.3.1.1 HDFS

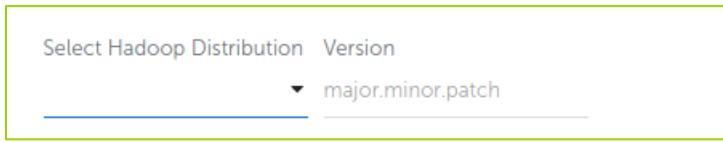
To add an HDFS data repository manually, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.

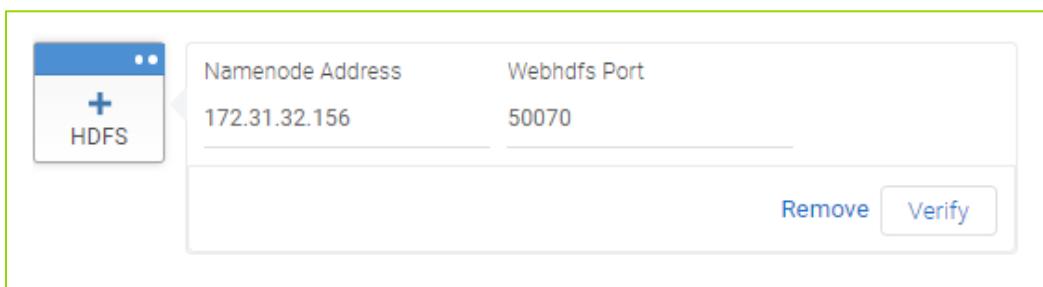
2. On the **Data Repositories** page, click the  icon.
3. In the **New Data Repository** dialog box, select the Hadoop from the Select the type of data repository drop-down menu, and then click **OK**.



4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - b. **Description:** A meaningful description of the data repository to help others identify it.
5. Select a Hadoop Distribution type from the **Select Hadoop Distribution** drop-down menu and then type the version number in the **Version** field.



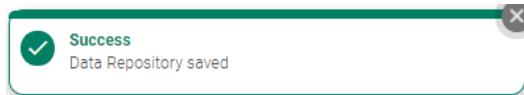
6. To manually add an HDFS data repository, click the  icon and type the details for Namenode Address. The default value of the **Webhdfs Port** field is 50070.



7. Click **Verify** to confirm the selection of your data repository. A confirmation message will be displayed on the page indicating the data repository is successfully verified.



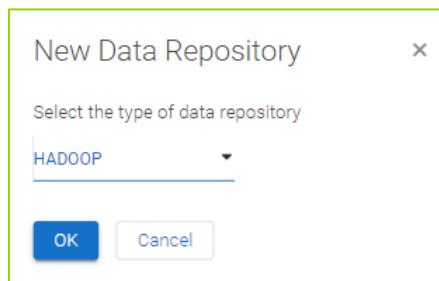
8. Set a data backup window and data recovery window and then click **Save**. A confirmation message will be displayed on the page indicating the data repository is successfully saved.



6.3.1.2 Hive

To add a Hive data store manually, do the following:

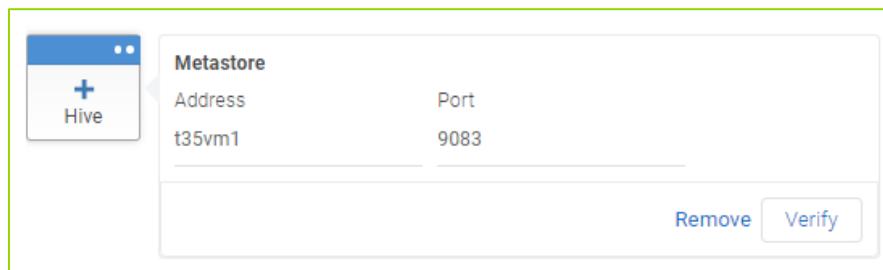
1. Click the **Main Menu** > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the **+** icon. The following dialog box appears:
3. In the **New Data Repository** dialog box, select the Hadoop from the **Select the type of data repository** drop-down menu, and then click **OK**.



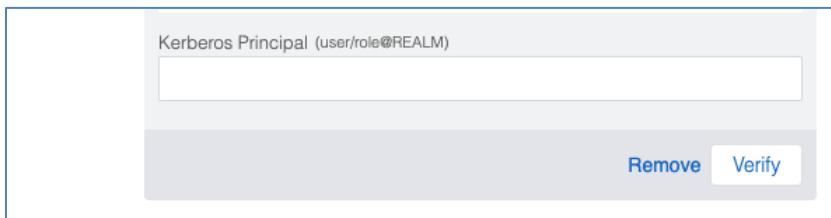
4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - b. **Description:** A meaningful description of the data repository to help others identify it.
5. Select a Hadoop Distribution type from the **Select Hadoop Distribution** drop-down menu and then type the version number in the **Version** field.



6. To manually add Hive data store, click the **Hive** icon, type the Metastore Address. The default value of Metastore Port is 9083.



7. Type the **Kerberos Principal** in the respective field. The Kerberos Principal option is displayed only if the Hive data repository is Kerberos-enabled.



8. Click **Verify** to confirm the selection of your data repository. A confirmation message will be displayed on the page indicating the data repository is successfully verified.



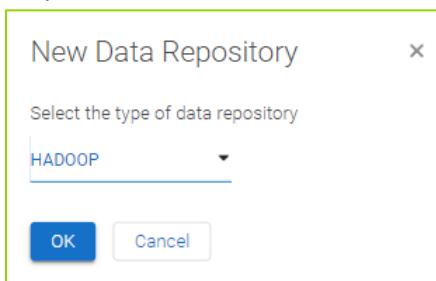
9. Set a data backup window and data recovery window and then click **Save**. A confirmation message will be displayed on the page indicating the data repository is successfully saved.



6.3.1.3 HBase

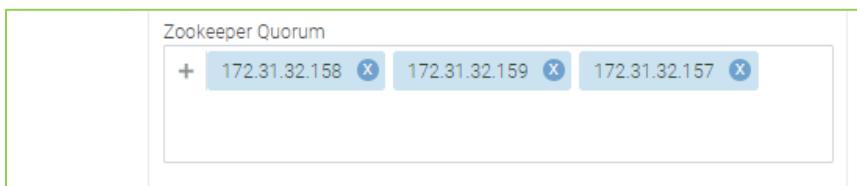
To add an HBase data repository manually, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the  icon. The following dialog box appears:
3. In the **New Data Repository** dialog box, select the Hadoop from the Select the type of data repository drop-down menu, and then click **OK**.

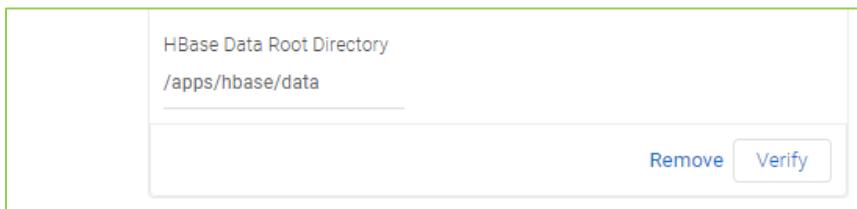


4. Type the following details:
 - Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - Description:** A meaningful description of the data repository to help others identify it.
5. Select a Hadoop Distribution type from the **Select Hadoop Distribution** drop-down menu and then type the version number in the **Version** field.

6. To manually add an HBase data repository, click the  icon and then do the following steps.
7. Click the **+** icon and type the list of servers in the **ZooKeeper Quorum** field. For example, "host1.mydomain.com,host2.mydomain.com,host3.mydomain.com".



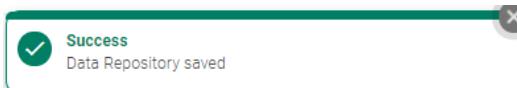
8. Type the path of the root directory in the **HBase Data Root Directory** field.



9. Click **Verify** to confirm the selection of your data repository. A confirmation message will be displayed on the page indicating the data repository is successfully verified.



10. Set a data backup window and data recovery window and then click **Save**. A confirmation message will be displayed on the page indicating the data repository is successfully saved.

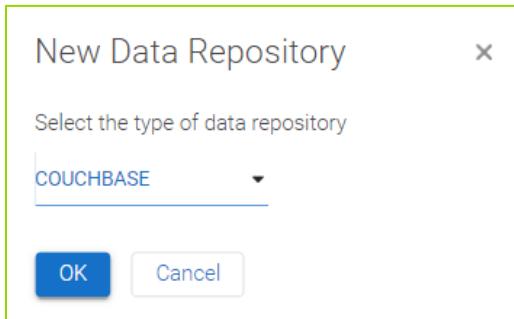


6.3.2 Couchbase

Couchbase Server is an open source, distributed database engineered for scalability, performance, and availability. Refer to the section Version Compatibility Matrix for information on versions supported by Imanis Data software.

To add a Couchbase data repository manually, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the  icon.
3. In the **New Data Repository Workflow** dialog, select the **Couchbase** option from the **Select the type of data repository** drop-down menu, and then click **OK**.



4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - b. **Description:** A meaningful description of the data repository to help others identify it
5. In the **Nodes** field, type comma separated IP addresses of the nodes where Couchbase is enabled.
6. In the **Authentication** field, type the **Username** and **Password**.
7. Do one of the following:
 - Click the **SSL Required** button if Couchbase primary cluster is SSL-enabled. In this case, both the fields HTTP direct port and Carrier direct port will be auto-populated by the default parameters:

- Click the **SSL Not Required** button if Couchbase primary cluster is not SSL-enabled. In this case too, both the fields HTTP direct port and Carrier direct port will be auto-populated by the default parameters however, you can customize by typing the port numbers of your choice in the preceding fields:

<input type="radio"/> SSL Required	<input type="radio"/> SSL Not Required
HTTP direct port	Carrier direct port
8091	11210

8. In the **Point-in-Time (PIT) Recovery** section, do the following:

- a. Click **Yes** to enable the **Point in Time Recovery** option. By default, the option is set to **No**.

Point-in-Time (PIT) Recovery		
Enable PIT Recovery 	<input checked="" type="radio"/> Yes	<input type="radio"/> No
 Use this option sparingly as it may put considerable load on system resources.		

- b. Enter a numerical value to set the retention period in the **In the Retain PIT metadata for__days** field. For example, if you set 35 days, Imanis Data can restore your data to any point in time during the last 35 days.

Retain PIT metadata for 35	days
----------------------------	------

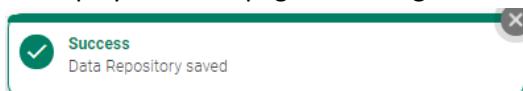
- c. Set the PIT job run frequency in minutes in the **Run PIT job every__minutes** field. For example, if you set the PIT job run frequency as 5 minutes then Imanis Data will run the first PIT job at say 1300 hrs, the second PIT job run at 1305, the third PIT job run at 1310 and so on.

Run PIT job every	5	10	15	minutes
-------------------	---	----	----	---------

11. Click **Verify**. A confirmation message will be displayed on the page indicating the data repository is successfully verified.



12. Set a data backup window and data recovery window and then click **Save**. A confirmation message will be displayed on the page indicating the data repository is successfully saved.

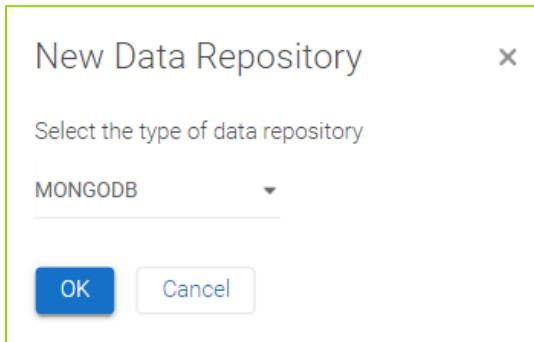


6.3.3 MongoDB

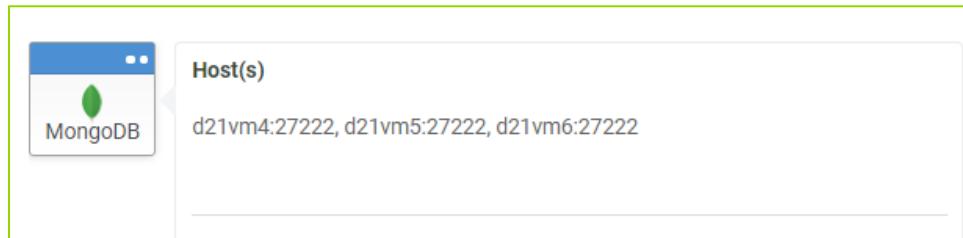
MongoDB is an open source database that uses a document-oriented data model. Refer to the section Version Compatibility Matrix for information on versions supported by Imanis Data software.

To add a MongoDB data repository manually, do the following:

1. Click the **Main Menu** > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the icon. The **New Data Repository** dialog box appears.
3. In the **New Data Repository** dialog box, select the **MongoDB** option from the **Select the type of data repository** drop-down menu, and then click **OK**.



4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - b. **Description:** A meaningful description of the data repository to help others identify it.
5. In the **Nodes** field, enter the node names on the primary cluster where MongoDB is installed:
 - For a sharded cluster enter a comma-separated list of hostname:port of MongoS daemons
 - For a non-sharded cluster enter a comma-separated list of hostname:port MongoD daemons



6. In the **Authentication** area, do one of the following:
 - Click **SCRAM**, if authentication is enabled for the particular MongoDB cluster that you are adding, type the username and password in the Username and Password fields, and then type the database name in the Authenticating Database field:

Authentication**SCRAM** **LDAP** **None**

Username

Administrator

Password

Authenticating Database

admin

- Click **LDAP** for LDAP authentication, and then enter the LDAP Username and LDAP Password.

Authentication**SCRAM** **LDAP** **None**

Username

Password

Authenticating Database

\$external

NOTE: For LDAP, the default Authenticating Database is **\$external**.

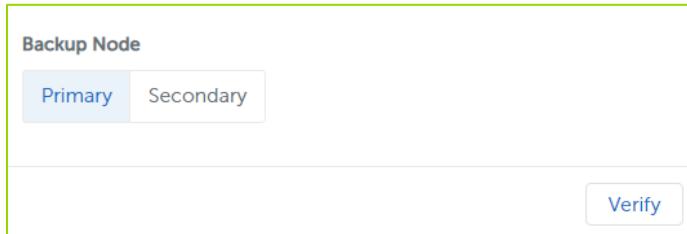
- Click **None**, if authentication is disabled for the MongoDB cluster that you are adding

7. In the **SSL Requirement** option, click **Yes** if the MongoDB is SSL-enabled.

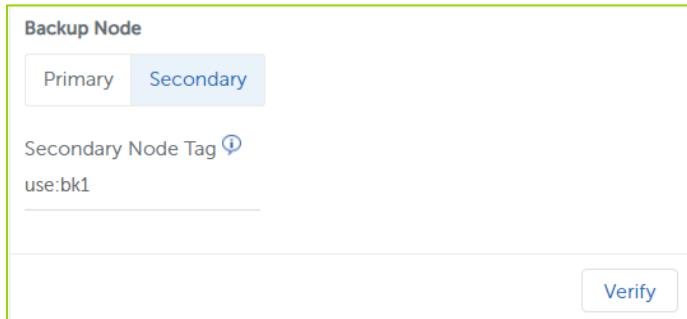
SSL Requirement**Yes** **No**

8. In the **Backup Node** option, do one of the following:

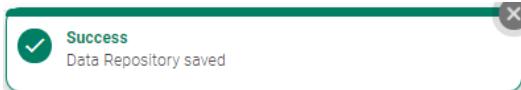
- Click **Primary** to select your primary node as one of the backup nodes



- Click **Secondary** and then specify the secondary node tag associated with a specific secondary node, in case of multiple secondary back nodes, from which backups should be performed In key:value pair.



9. Click **Verify** to add the MongoDB data source in Imanis Data.
10. Set a data backup window and data recovery window, and then click **Save**. A confirmation message will be displayed on the page indicating the data repository is successfully saved.



6.3.4 Amazon Glacier

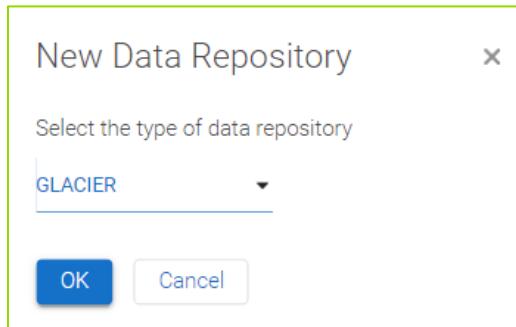
Imanis Data software enables you to reduce OPEX and CAPEX by moving inactive or cold data to a lower-cost storage tier like Amazon Glacier. You can configure Amazon Glacier to work as a regular cloud data repository only. You cannot add or configure it to work as Source data repository or a Global Cloud Data Repository.

Refer to the section Data Repository Usage Matrix for information on using data repositories as source and/or global and regular cloud repository.

To add an Amazon Glacier, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the  icon. The **New Data Repository** dialog box appears.

3. In the **New Data Repository** dialog box, select the **Glacier** option from the **Select the type of data repository** drop-down menu, and then click **OK**.



4. Type the following details:
- Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - Description:** Type a meaningful description of the data repository to help others identify it.
5. In the **Glacier** section, type **the Amazon Glacier Region URL, Vault ID, Access Key, Secret Key, and Bandwidth** in the respective fields. Refer to the following table for more information:

BUTTONS & FIELDS	DESCRIPTION	EXAMPLE
Amazon Glacier Region URL	URL of region-specific folder that is used to	https://glacier.us-west-2.amazonaws.com/example-test

BUTTONS & FIELDS	DESCRIPTION	EXAMPLE
	store objects in Amazon Glacier.	"glacier.us-west-2.amazonaws.com" is region-specific information, that is, West Coast of US.
Vault ID	Region-specific vault name	https://glacier.us-west-2.amazonaws.com/example-test "example-test" is the vault name.
Access Key	A string that uniquely identifies your Amazon Glacier account as well as individual IAM users in your account.	AFIAI9YP1USWBSO6QGXX
Secret Key	A string that uniquely identifies your Amazon Glacier account. To be used in conjunction with the access key ID to cryptographically sign programmatic Glacier requests.	vxRQyG0imAxVYF3kVDclx39K2rlQp49YmDaMzS/0
Bandwidth	Limitation of the data that you want to download in Mbytes per day	

6. Click **Verify** to confirm the selection of your data repository. A confirmation message will be displayed on the page indicating the data repository is successfully verified.



7. Set a data archival window and data retrieval window, and then click **Save**. A confirmation message will be displayed on the page indicating the data repository is successfully saved.



6.3.5 Amazon S3

S3 is a highly scalable, reliable, fast, inexpensive data storage infrastructure.

In Imanis Data software, you can configure Amazon S3 to work as a Global Cloud Data Repository and a Regular Cloud Data Repository only. You cannot add or configure it to work as Source data repository.

Refer to the section Data Repository Usage Matrix for information on using data repositories as source and/or global and regular cloud repository.

6.3.5.1 Global Cloud Data Repository

The Global Cloud data repository should be created if you want to move de-duplicated backup data to the cloud.

To create a Global Cloud data repository, add a S3 data repository, type the mandatory parameters, and click Yes for the Use this for deduplication option. The Yes option enables the movement of de-duplicated data from Imanis Data cluster to the Global Cloud data repository. For more information, refer to the section S3 Data Repository.

Once you create the Global Cloud data repository, all the de-duplicated backup data will be automatically migrated to this S3 data repository based on the schedule and priority defined in the Global Cloud Policy. Please note that at any given point in time, you may only have a single Global Cloud data repository.

Data (in de-duplicated format) is backed up from your Primary cluster to Global Cloud data repository in two simple steps:

1. Data is backed up from the Primary cluster to Imanis Data cluster based on the schedule set in the backup policy associated with the workflow.
2. All the de-duplicated data is moved from Imanis Data cluster to Global Cloud data repository based on the schedule set in the Global Cloud policy. Refer to the section on Global Cloud policy for more information.

6.3.5.2 Regular S3 Data Repository

The S3 data repository should be created if you want to move backup data in native format, instead of de-duplicated format, to S3 data repository.

To create a S3 data repository, add a S3 data repository, type the mandatory parameters, and click No for the Use this for deduplication option. For more information, refer to the section Adding S3 Data Repository.

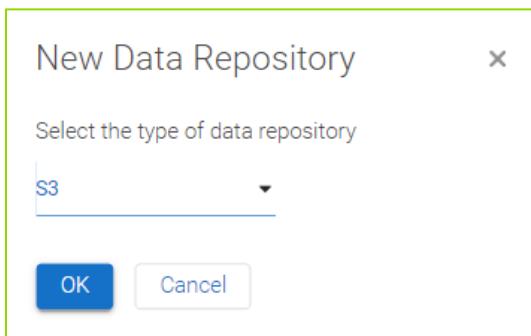
Unlike Global S3 data repository, you can add multiple S3 data repositories in Imanis Data software.

Data (in native format) is backed up from your Primary cluster to S3 data repository in two simple steps:

1. Data is backed up from the Primary cluster to Imanis Data cluster based on the schedule in the backup policy associated with the workflow.
2. Data is moved from Imanis Data cluster to S3 data repository based on the retention parameters of Imanis Data cluster and Cloud Schedule set in the backup policy. Once the retention on Imanis Data cluster elapses, data is automatically backed up from the Imanis Data cluster to S3 data repository based on the Cloud Schedule set in the backup policy.

To add an Amazon S3, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the  icon. The **New Data Repository** dialog box appears.
3. In the **New Data Repository** dialog box, select the **S3** option from the **Select the type of data repository** drop-down menu, and then click **OK**.



4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - b. **Description:** Type a meaningful description of the data repository to help others identify it.
5. In the **S3** section, type the **Amazon S3 URL**, **Access Key**, and **Secret Key**. Then, select the type of encryption from the drop-down menu. Refer to the following table for more information:



Amazon S3 URL	<input type="text" value="https://s3-us-west-2.amazonaws.com"/>
Access Key	<input type="text" value="AGUAJ5P77SYSDLCWNW8E"/>
Secret Key	<input type="text" value="....."/>
Encryption Type	<input type="text" value="No Encryption"/>

BUTTONS & FIELDS	DESCRIPTION	EXAMPLE
Amazon S3 URL	URL of region-specific bucket that is used to store objects in Amazon S3. Ensure that the S3 URL matches the region of the S3 bucket. For more information on Amazon S3 URLs, click here .	https://s3-us-west-2.amazonaws.com “s3.us-west-2.amazonaws.com” is region-specific information, that is, West Coast of US.
Access Key	Alphanumeric string that uniquely identifies your Amazon S3 account as well as individual IAM user in your account. No two accounts can have the same access key.	AGUAJ5P77SYSDLCWNW8E
Secret Key	Alphanumeric string that plays the role of a	vxRQyG0imAxVYF3kVDclx40K3rlQp49YmDaMzS/0

BUTTONS & FIELDS	DESCRIPTION	EXAMPLE
	password which is known to the user only. To be used in conjunction with the access key ID to cryptographically sign programmatic S3 requests.	
Encryption Type	The type of encryption that you can select to store and encrypt your sensitive data for regulatory or compliance reasons. You can choose to have No Encryption, Server-Side Encryption, or Client-Side Encryption. For more information on Encryption Type, refer to the Amazon S3 documentation here .	No Encryption Server-side encryption Client-side encryption S3 Managed Key KMS Managed Key Customer Provided Key KMS Managed Key Symmetric Master Key Asymmetric Master Key

6. Enter the respective key depending on the type of encryption that you select in the preceding step.

The KMS Managed Key will be provided by AWS. Depending on the security policy of your company, refer to the key management tool to generate keys for these types of encryption: Customer Provided Key under Sever Side Encryption, and Symmetric and Asymmetric Master Key under Client-Side Encryption.

7. To be able to access Amazon S3 URLs needs to go through a proxy server, then the S3 repository creation page has the option to specify proxy settings. Do the following to access Amazon S3 though a proxy server:

- Click **Yes** on the **Network proxy** option. The following options are displayed on the screen.
- Set **Proxy server hostname** to the host name or IP address of the proxy server and then set the **Proxy server port**.

- Optionally, you can set the **Proxy server username** and **Proxy server password** if the proxy server access requires authentication. For no authentication, do not populate these fields. Currently BASIC and DIGEST authentication methods are supported.

Network proxy
Yes No

Proxy server hostname
proxyhost.domain.com

Proxy server port
3128

Proxy server username
username

Proxy server password

8. In the **Use as deduplication target** area, do one of the following:

- Click **Yes** to use this S3 data repository for global deduplication purposes, and then click **Discover** to display a list of buckets that you can access in the S3 data repository
- Click **No** to use this S3 data repository for regular backup purposes, click **Verify** to confirm the selection of your data repository

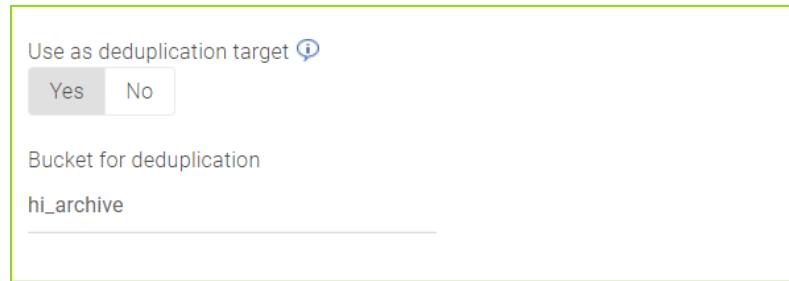
9. If you have clicked **Yes** in the preceding step, then do one of the following:

- Select a bucket from the **Bucket for deduplication** drop-down menu and then click **Verify** to authenticate the selection. If the user does not have permission to list buckets, the bucket name can be entered manually

Use as deduplication target ⓘ
Yes No

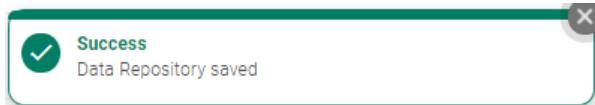
Bucket for deduplication
hi_backup

- Type name of the bucket in the **Bucket for deduplication** field for which you have been granted access by the AWS Account Admin of your organization, and then click **Verify**



For example, as a user you aware that you have been given access to the ArizonaMaps bucket. All you have to do is type the name of the bucket ArizonaMaps in the **Bucket for deduplication** field and then click **Verify**.

10. Click **Save**. A confirmation message will be displayed on the page indicating the data repository is successfully saved.



NOTE: In the current release, Imanis Data software does not support archiving Amazon S3 data to Amazon Glacier. For more information, refer to the Amazon S3 documentation [here](#).

6.3.6 Cloudian

Cloudian has limitless scalability with bucket-level granularity for all storage policies. Cloudian HyperStore offers more capabilities that boost interoperability, data durability, and operational efficiency.

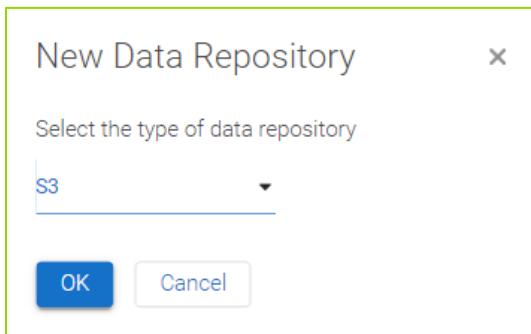
In Imanis Data software, you can configure Cloudian to work as a **Global Cloud Data Repository** and a **Regular Cloud Data Repository** only. You are not permitted to add or configure it as Source data repository.

Refer to the section Data Repository Usage Matrix for information on using data repositories as source and/or global and regular cloud repository.

You can add Cloudian data repository in Imanis Data through the S3 data repository GUI as Cloudian's S3 API is compatible with S3-enabled applications.

To add a Cloudian, do the following:

1. Click the **Main Menu** > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the icon. The following dialog box appears:



3. In the **New Data Repository** dialog box, select the **S3** option from the **Select the type of data repository** drop-down menu, and then click **OK**.
4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - b. **Description:** Type a meaningful description of the data repository to help others identify it.
5. In the **S3** section, type the Amazon S3 URL, Access Key, and Secret Key. Then, select the type of encryption from the drop-down menu. Refer to the following table for more information:

A screenshot of the 'Amazon S3' configuration section. On the left is a small icon of a blue square with a white 3D cube and the letters 'S3' below it. To its right are four input fields: 'Amazon S3 URL' containing 'https://s3-us-west-2.amazonaws.com', 'Access Key' containing 'AGUAJ5P77SYSDLCWNW8E', 'Secret Key' (which is redacted with dots), and 'Encryption Type' set to 'No Encryption'. There is also a small dropdown arrow next to the 'Encryption Type' field.

VERSION	DATE	DOCUMENT HISTORY
Amazon S3 URL	URL of region-specific bucket that is used to store objects in Amazon S3. Ensure that the S3 URL	<p>https://s3-us-west-2.amazonaws.com</p> <p>“s3.us-west-2.amazonaws.com” is region-specific information, that is, West Coast of US.</p>

VERSION	DATE	DOCUMENT HISTORY	
	matches the region of the S3 bucket.		
Access Key	Alphanumeric string that uniquely identifies your Amazon S3 account as well as individual IAM user in your account. No two accounts can have the same access key.	AGUAJ5P77SYS DLCWNW8E	
Secret Key	Alphanumeric string that plays the role of a password which is known to the user only. To be used in conjunction with the access key ID to cryptographically sign programmatic S3 requests.	vxRQyG0imAxVYF3kVDclx40K3rlQp49YmDaMzS/0	
Encryption Type	The type of encryption that you can select to store and encrypt your sensitive data for regulatory or compliance reasons. You can choose to have No Encryption, Server-Side Encryption, or Client Side Encryption.	No Encryption	
		Server-side encryption	S3 Managed Key

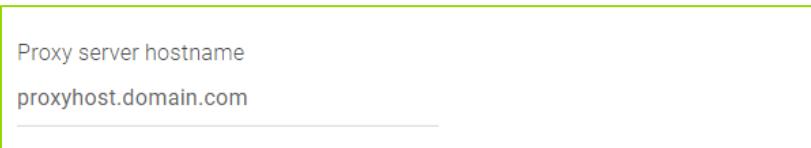
6. In the **Network proxy** option, click **Yes** to access Amazon S3 though a proxy server. The following options will be displayed on the screen:



Network proxy

Yes No

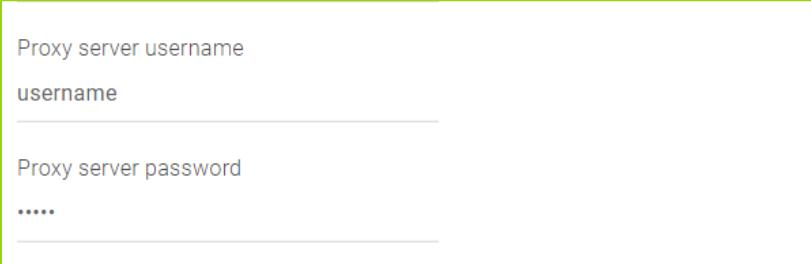
7. Set **Proxy server hostname** to the host name or IP address of the proxy server and then set the **Proxy server port**.



Proxy server hostname

proxyhost.domain.com

8. Optionally, you can set the **Proxy server username** and **Proxy server password** if the proxy server access requires authentication. For no authentication, do not populate these fields. Currently BASIC and DIGEST authentication methods are supported.



Proxy server username

username

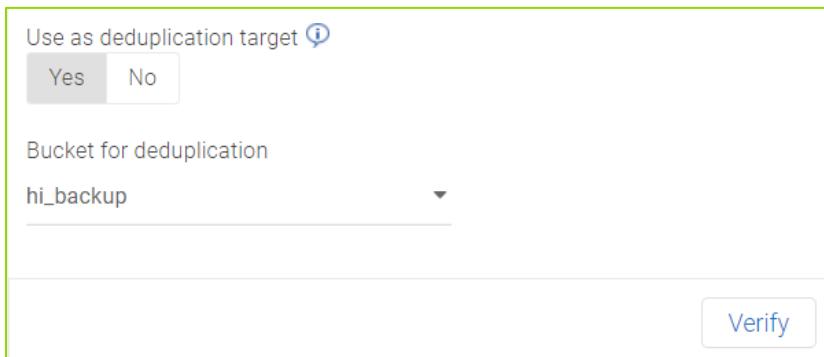
Proxy server password

.....

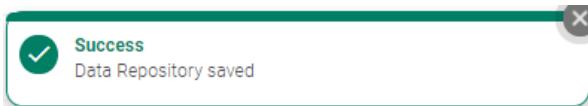
9. In the **Use as deduplication target** area, do one of the following:

- Click **Yes** to use this S3 data repository for global deduplication purposes, and then click **Discover** to display a list of buckets that you can access in the S3 data repository
- Click **No** to use this S3 data repository for regular backup purposes, click **Verify** to confirm the selection of your data repository

10. If you have clicked **Yes** in the preceding step, select a bucket from the **Bucket for deduplication** drop-down menu and then click **Verify** to authenticate the selection. If the user does not have permission to list buckets, the bucket name can be entered manually.



11. Click **Save**. A confirmation message will be displayed on the page indicating the data repository is successfully saved.



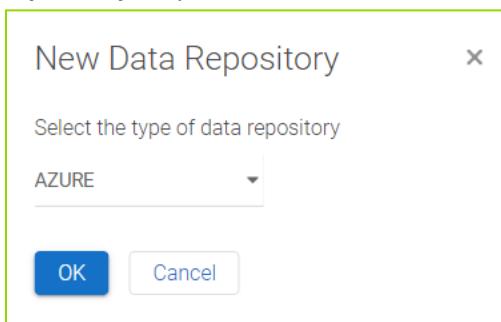
NOTE: In the current release, Imanis Data software does not support archiving Amazon S3 data to Amazon Glacier. For more information, refer to the Amazon S3 documentation [here](#).

6.3.7 Azure Blob Storage

In Imanis Data software, you can configure Azure Blob Storage as Global Cloud Data Repository and Regular Blob Storage Data Repository. Refer to the section Data Repository Usage Matrix for information on using data repositories as source and/or global and regular cloud repository.

To add Azure Blob Storage, do the following:

1. Click the **Main Menu** > **System Setup** > **Data Repositories**.
2. On the **Data Repositories** page, click the icon. The **New Data Repository** dialog box appears.
3. In the **New Data Repository** dialog box, select the **Azure** option from the **Select the type of data repository** drop-down menu, and then click **OK**.



4. Type the following details:
 - a. **Data Repository Name:** Type a name for the data repository. The data repository name can include alphanumeric characters, numbers and/or special characters.
 - b. **Description:** Type a meaningful description of the data repository to help others identify it.
5. In the **Applications** section, type the following:
 - a. **Primary Host:** The name or IP address of the machine on which the configuration directories related to the applications exist.
 - b. **Configuration Directory:** A list of one or more paths where configuration files related to the apps exist.
6. In the **Azure** section, type or copy-paste the connection string in the **Connection String** field, and select the type of encryption from the **Encryption Type** drop-down menu. Refer to the following table for more information:

Data Repositories		COHESITY		VERNON-DC2	admin
<input type="button" value="New"/> <input type="button" value="Edit"/> <input type="button" value="Delete"/> <input type="button" value="Search"/>		Name	Data Repository Name Georgia Maps	Description (optional) Azure Blob Data Repo	<input type="button" value="Save"/> <input type="button" value="Cancel"/>
		AZURE	 <input type="button" value="..."/>	Connection String <pre>DefaultEndpointsProtocol=https;AccountName=ImanisDataqa;AccountKey=xSOUZLzDj5DuEi0Als+5IDzDcgHmfIOgbs+UzAEVzzijcG+ji6KlgSwb/+vSknai22ClhX+tpbzAT+GzXwBbg==</pre>	
				Encryption Type No Encryption	

BUTTONS & FIELDS	DESCRIPTION	EXAMPLE	
Connection String	A connection string includes the authentication information required for your application to access data in an Azure Storage account at runtime.	DefaultEndpointsProtocol=https;AccountName=ImanisDataqa;AccountKey=xSOUZLzDj5DuEi0Als+5IDzDcgHmfIOgbs+UzAEVzzijcG+ji6KlgSwb/+vSknai22ClhX+tpbzAT+GzXwBbg==	
Encryption Type	The type of encryption that you can select to store and encrypt your sensitive data for	Client-side encryption	Symmetric Master Key

BUTTONS & FIELDS	DESCRIPTION	EXAMPLE
	regulatory or compliance reasons. You can choose to have No Encryption or Client-Side Encryption.	Asymmetric Master Key

7. Do one of the following:

- Click **Yes** under the **Use as deduplication target** option to use this Azure data repository for global deduplication purposes, click **Discover** to display a drop-down menu of containers available in the Azure Blob Storage data repository, select a container from the **Container for deduplication** drop-down menu, and then click **Verify** to authenticate the selection

Use as deduplication target i

Yes No

Container for deduplication

adls_stage_backup

Verify

- Click **No** under the **Use as deduplication target** option and then click **Verify** to confirm the selection of your data repository

Use as deduplication target i

Yes No

Verify

8. Click **Save**. A confirmation is displayed that the data repository is successfully saved.



NOTE: In the current release, Imanis Data software does not support changing storage accounts across different runs of the same workflow.

6.4 Managing Data Repositories

This section describes how you can manage data repositories in Imanis Data software. Get to know how you can edit or delete data repositories. Find out how you can set bandwidth and concurrency throttling, set a blackout window, and so on.

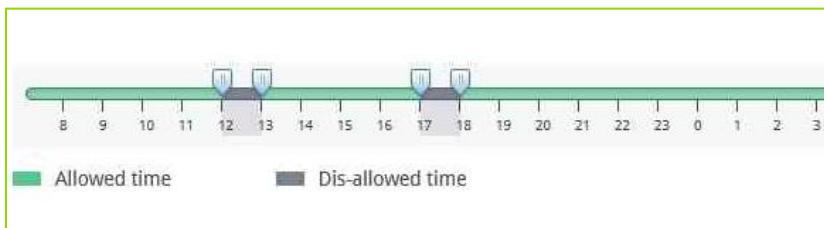
6.4.1 Setting a Blackout Window

You can set blackout windows (in terms of time slots) that will *prevent the initiation* of data backup, data recovery, data archival, and data retrieval during key business periods for selected data repositories.

The following example describes how you can set blackout window in a Hadoop data repository.

To set blackout window, do the following:

1. Identify a time slot on the timeline to set a blackout window.
2. Double-click on the timeline to set a blackout window.



3. Adjust the slider as per your requirement.
4. Repeat the preceding step to set multiple blackout windows.

NOTE: Drag the slider to set the start and stop time for a blackout window.

6.4.2 Removing a Blackout Window

You can manually remove the blackout window from data backup, data recovery, data archival, and data retrieval process timeline.

The following example describes how you can remove blackout window in a Hadoop data repository.

To remove a blackout window, do the following:

1. Identify the blackout window that you want to remove.

2. Click, hold and move the blackout window away from the timeline till you see a dotted rectangle.
3. Drop the dotted rectangle to remove the blackout window (see the following screenshot):



6.4.3 Editing a Data Repository

Administrators may modify the name and description of the data repository as well as the applications that were configured earlier. If an HDFS data repository is available, administrators can add a Hive data repository; however, the HDFS data repository cannot be removed. The Data Backup and Data Recovery Window can also be edited.

Imanis Data software also allows editing of the Address and Port of HDFS data repository and Namenode Address and Webhdfs Port of the Hive data repository. While editing the Address (HDFS) and Namenode Address (Hive), administrators must ensure that the new address or namenode address logically refers to the same cluster that was referred earlier. If not, jobs will fail.

To edit a data repository or cloud storage, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.
2. Identify the data repository or cloud storage that you want to edit and select it.
3. On the right pane, click the edit  icon.
4. Make the appropriate changes; click **Verify** to confirm the selection.
5. Click **Save**.

6.4.4 Deleting a Data Repository

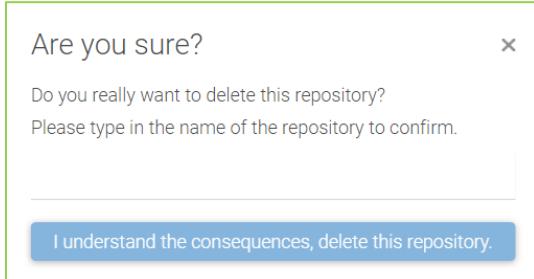
Deleting a data repository, automatically removes the corresponding data residing in the Imanis Data cluster. The data and data repository are deleted permanently and cannot be undone.

Imanis Data software does not allow you to delete a data repository if there are any active or inactive workflows are scheduled for that repository. Unless you delete all the workflows, Imanis Data software does not allow you to delete a data repository. Refer to the section on Dashboard for more information.

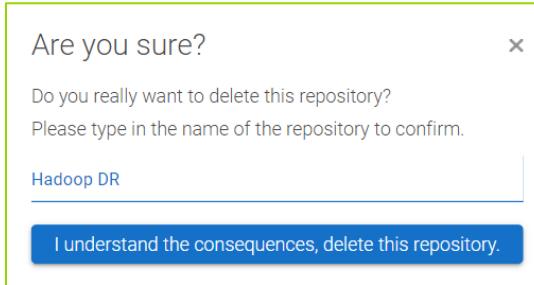
To delete a data repository or cloud storage, do the following:

1. Click the **Main Menu**  > **System Setup** > **Data Repositories**.

2. Identify the data repository or cloud storage that you want, select it, and then click the delete icon . A confirmation message appears.



3. Type or copy-paste the name of the data repository that you want to delete in the confirmation message. When the name of the data repository is entered, the **I understand the consequences, delete this repository** button is activated.



4. Click the **I understand the consequences, delete this repository** button. Data repository is permanently deleted from the system.

7 Data Backup

This section describes the features of the Data Backup menu of Imanis Data software.

IMPORTANT: Before you get started, you must create a data repository and a data backup policy through the System Setup menu.

7.1 Getting Started with Data Backup

In the data backup workflow, you can identify a data repository and select the data objects that you want to back up to the Imanis Data cluster based on a pre-defined backup policy. You can also retain data onto the Cloud with Amazon S3.

IMPORTANT: In the backup policy, if you have earlier selected **No** in the **Allow retention on cloud** option to disable retention of data on the Cloud then you can edit it to **Yes**. However, once you make this change, ensure that you edit each job using the policy and specify the cloud data repository details.

7.1.1 Backing Up Data for HDFS or Hive

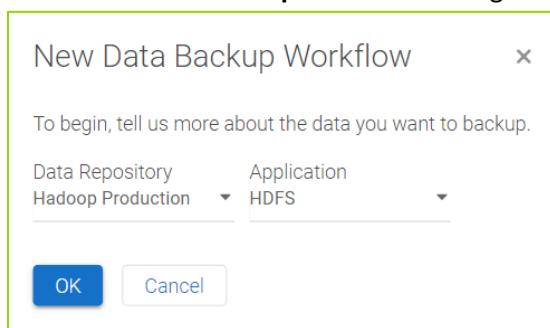
Imanis Data software enables you to back up data for HDFS or Hive.

To backup data in HDFS or Hive, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Backup**.

The following page appears only if you are creating data backup workflows for the first time. You can proceed to click the  button or the  icon to create data backup workflows.

2. On the **Data Backup** page, click the  icon. The **New Data Backup Workflow** dialog box appears.
3. In the **New Data Backup Workflow** dialog box, do the following:



- a. Select a data repository (from where you want to backup the data) from the **Data Repository** drop-down menu.
 - b. Select **HDFS** or **Hive** from the Application drop-down menu.
 - c. Click **OK**.
4. All the files, tables, directories, databases, and partitions available in the data repository that you selected earlier appear in the next page. This is entirely dependent on whether you select Hive or HDFS.
5. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.

Note:

- In Imanis Data software terminology, a job tag is a unique identifier of a job in the system. This job tag is saved in the Imanis Data catalog and can be used during a data recovery to search a job using FastFind.
- Backup/restore of transactional tables for Hive 3.x is not supported.

6. In the **Identify Data** section, do the following:

Objects	Modified Time	Owner	Size
system	2019-09-06 05:01:05 ...	talena	29.7 GB
human resource	2019-09-06 05:01:05 ...	talena	937.6 MB
sales	2019-09-06 05:01:05 ...	talena	2 GB
payroll	2019-09-06 05:01:05 ...	talena	5.7 GB
legal	2019-09-06 05:01:05 ...	talena	971.7 MB
purchasing	2019-09-06 05:01:05 ...	talena	2.3 GB

- a. In the **HDFS** tab, identify the directories and files. that you want to backup by selecting the corresponding check boxes. Optionally, you can also search for data objects by typing in a term in the search box. For example, type A* in the search box to find files, tables, directories, databases, and partitions starting with character ‘A’. However, the search term is case sensitive and can only be executed at the current directory or database level and it does not search lower level directories or databases to look for sub-directories, databases, tables, files, or partitions that match the search term.
- b. In the **Selected Data** tab, verify your selection or click the **X** icons to remove unwanted items.

- c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section Appendix A: Rules for Data Inclusions and Exclusions.

IMPORTANT: Imanis Data software does not support archiving of partitions in Hive. This is a Imanis Data software limitation which may be fixed in future releases.

7. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention. The following screenshot illustrates the selection of backup policy with cloud retention option.

The screenshot shows the 'Specify Policy' section of a backup configuration interface. At the top, it says 'Select a data backup policy' with 'Backup to Cloud (s3)' selected. Below that is a 'Retention' section with the option 'Allow retention on cloud'. Under this, there are two radio buttons: 'Yes' (selected) and 'No'. A timeline diagram shows data moving from 'On Imanis Data' (file icon) through 'On Cloud' (cloud icon) to a trash bin icon. It indicates a retention period of 1 day on Imanis Data and 365 days on Cloud.

8. In the **Specify Options** section under **Cloud options**, do one of the following:

- Select an S3 cloud repository from the **Data Repository** drop-down menu and select a bucket (S3) from the **Buckets** drop-down menu to retain data in the cloud:

The screenshot shows the 'Specify Options' section. Under 'Cloud options', it lists 'Data Repository' as 'S3 Cloud Storage' and 'Buckets' as 'hadoop_production_backup'.

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure) from the **Containers** drop-down menu to retain data in the cloud:

Cloud options ⓘ

Data Repository	Containers
Azure Cloud Storage	adls_stage_backup

As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field.

IMPORTANT: The S3 data repository enabled for global deduplication will not be available for selection in the Cloud options under the Specify Options section. For more information, refer to the section [Adding S3 Data Repository](#).

- Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:
 - In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:

Advanced Options

Bandwidth Throttling ⓘ

Yes No

50 MB/s per Mapper

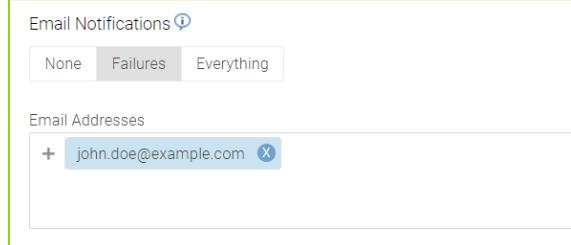
- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:

Concurrency Throttling ⓘ

Yes No

- 6 +

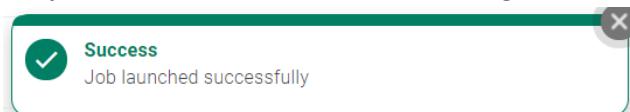
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the Email Addresses field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. Imanis Data software will then start executing your workflow. A confirmation message will be displayed on the page indicating the job is successfully launched. You can monitor the progression of the job on the Dashboard while its running.



IMPORTANT: In Hive, when data is manually added/deleted to/from existing tables, it is recommended that the user issues the ANALYZE command.

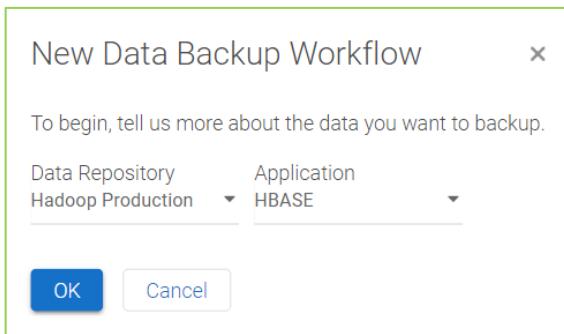
NOTE: Imanis Data software does not back up Hive Index Tables and Hive Views. As a result, you may have to recreate these indexes after restoring the data.

7.1.2 Backing Up Data for HBase

Imanis Data software enables you to back up data at the Namespace and Table level.

To start a backup workflow for HBase, do the following:

- Click the **Main Menu**  > **Data Management** > **Data Backup**.
- On the **Data Backup** page, click the  icon to create a Data Backup workflow.
- In the **New Data Backup Workflow** dialog, select a **HBase** source data repository from the **Data Repository** drop-down menu.



4. Click **OK**. All the Namespaces available in the HBase data repository that you selected earlier appear on the following page.
5. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
6. In the **Identify Data** section, do the following:
 - a. In the **HBase** tab, identify the Namespace from which you want to backup data and select the corresponding check boxes. To select tables within namespaces, double-click a namespace to view tables and select the ones you want to backup.
 - b. In the **Selected Data** tab, verify your selection or click the **X** icons to remove unwanted items.
 - c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section Appendix A: Rules for Data Inclusions and Exclusions.

The screenshot shows the Cohesity Data Protection interface. The top navigation bar includes "Data Backup" and "admin". The main area is titled "Workflow Definition". A card titled "1 Identify Data" is open, showing the "HBASE" tab selected. Below it is a table with columns "Objects" and "Size". The objects listed are "payroll", "system", "human resource", and "sales", each with a checkbox next to it. Other tabs include "Selected Data" and "Rules". At the top right of the card are "Submit" and "Cancel" buttons.

NOTE: Running two or more backup jobs on a particular table at the same time must be avoided as this operation is not supported by HBase.

7. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention. The following screenshot illustrates the selection of backup policy with cloud retention option.

The screenshot shows the 'Specify Policy' step of a backup configuration. Under 'Select a data backup policy', 'Backup to Cloud (s3)' is selected. In the 'Retention' section, 'Allow retention on cloud' is checked, and 'Yes' is selected. A timeline diagram shows data moving from 'On Imanis Data' (1 day) to 'On Cloud' (365 days), and finally being deleted from the cloud. The 'On Cloud' node contains a trash can icon.

8. In the **Specify Options** section under **Cloud options**, do one of the following:
 - Select an S3 cloud repository from the **Data Repository** drop-down menu and select a bucket (S3) from the **Buckets** drop-down menu to retain data in the cloud:

The screenshot shows the 'Specify Options' step. Under 'Cloud options', 'Data Repository' is set to 'S3 Cloud Storage' and 'Buckets' is set to 'hadoop_production_backup'. Both dropdown menus have a downward arrow indicating they are dropdowns.

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure) from the **Containers** drop-down menu to retain data in the cloud:

The screenshot shows the 'Specify Options' pane with a header '3 Specify Options'. Under 'Cloud options', there are two dropdown menus: 'Data Repository' set to 'Azure Cloud Storage' and 'Containers' set to 'adls_stage_backup'.

As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field.

9. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

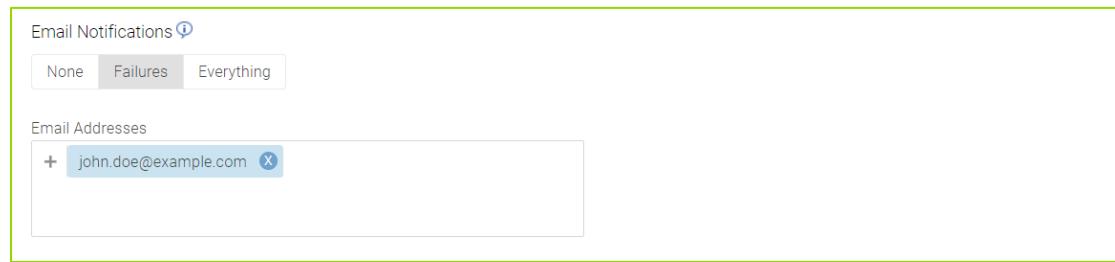
- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:

The screenshot shows the 'Advanced Options' pane expanded. Under 'Bandwidth Throttling', there is a 'Yes' button, a 'No' button, and a slider with a lock icon set to 50 MB/s per Mapper.

- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:

The screenshot shows the 'Advanced Options' pane expanded. Under 'Concurrency Throttling', there is a 'Yes' button, a 'No' button, and a numeric input field with a minus sign (-), a value of 6, and a plus sign (+).

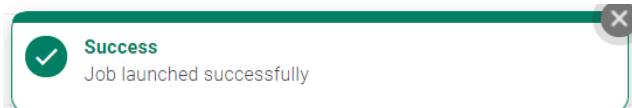
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the Email Addresses field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

10. Click **Submit**. Imanis Data software will then start executing your workflow. A confirmation message will be displayed on the page indicating the job is successfully launched. You can monitor the progression of the job on the Dashboard while its running.



7.1.3 Backing Up Data for Cassandra

Imanis Data software enables you to backup keyspace and table level data for Cassandra. During the data backup process, Imanis Data software also identifies if a table is related to Solr Core and backs up its Solr configuration as well.

IMPORTANT: If you have made any changes to the `cassandra.yaml` configuration file after adding a Cassandra data repository in Imanis Data software, then you must auto-discover the Cassandra data repository again. For example, if authentication or authorization type is changed in the `cassandra.yaml` file, then auto-discovering the Cassandra data repository is mandatory.

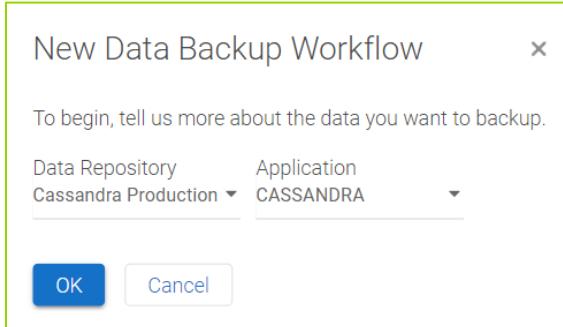
IMPORTANT: Backup is not supported for the following system keyspaces:

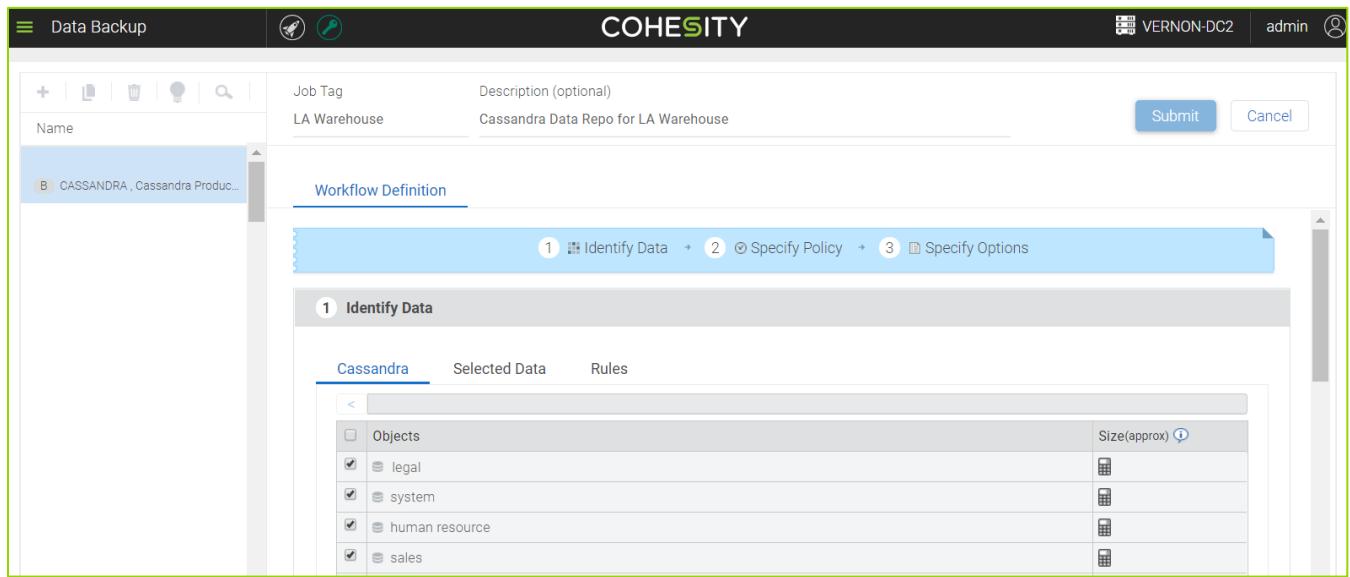
```
system
system_auth
system_distributed
system_schema
system_traces
dse_leases
dse_perf
dse_security
dse_system
```

solr_admin

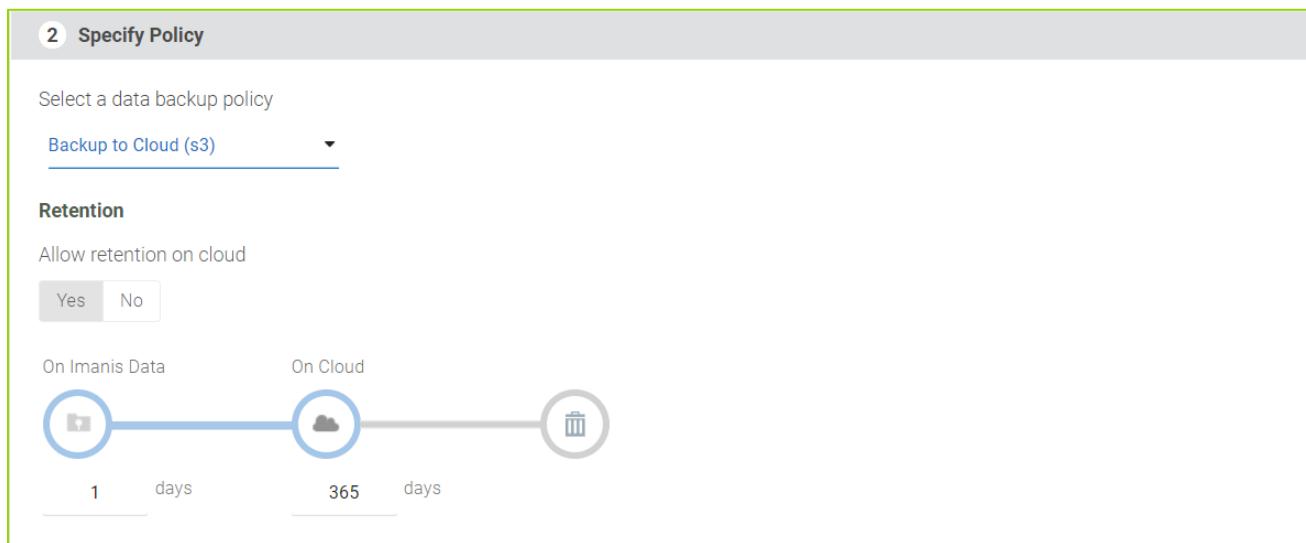
Restore of system keyspaces such as OpsCenter is only supported for the same cluster from which it was backed up. To enable backup for any of the preceding keyspaces, remove the particular keyspace from the list of system keyspaces defined in the following configuration file: \$INSTALL_DIR/conf/cassandra-conf.xml

To start a backup workflow for Cassandra, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Backup**.
2. On the **Data Backup** page, click the  button or the  icon. The **New Data Backup Workflow** dialog box appears.
3. In the **New Data Backup Workflow** dialog box, select a **Cassandra** data repository from the **Data Repository** drop-down menu. Imanis Data software auto-populates the **Application** field with **Cassandra** as it recognizes the type of data repository that is selected.

4. Click **OK**. All the keyspaces and tables available of the selected Cassandra data repository in the Cassandra data repository are displayed on the page.
5. In the **Data Backup** page, do the following:
 - a. Type a new job tag in the **Job Tag** field.
 - b. Type a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
6. In the **Identify Data** section, do the following:
 - In the **Cassandra** tab, identify the keyspaces and tables from which you want to backup the data by selecting the corresponding check boxes. You can also search for data objects by typing a regex in the search box. For example, if you type s*, Imanis Data software will display all objects starting with the alphabet 's'.
 - In the **Selected Data** tab, verify your selection or click the  icons to remove unwanted items.
 - In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section Appendix A: Rules for Data Inclusions and Exclusions.



7. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention. The following screenshot displays with cloud retention option.



8. In the **Specify Options** section, under **Data Centers**, select a data center or data centers from where you want to backup the data.

3 Specify Options

Data Centers

Select All
 DC1
 DC2

NOTE: Imanis Data software supports backup from a specific data center only if the keyspace spans across data centers, that is, if the keyspace is created using replication strategy as 'NetworkTopologyStrategy'. However, if a specific data center for a keyspace created using non-NetworkTopologyStrategy is selected, then data from all the data centers is automatically backed up. For example, you have data centers 'D1', 'D2', and 'D3'. You select 'D1' for a keyspace created using SimpleStrategy, then data on 'D2' and 'D3' is also backed up.

9. In the **Specify Options** section under **Cloud options**, do one of the following:

- Select an S3 cloud repository from the **Data Repository** drop-down menu and select a bucket (S3) from the **Buckets** drop-down menu to retain data in the cloud:

Cloud options ⓘ

Data Repository Buckets

S3 Cloud Storage ▾ hadoop_production_backup ▾

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure) from the **Containers** drop-down menu to retain data in the cloud:

Cloud options ⓘ

Data Repository Containers

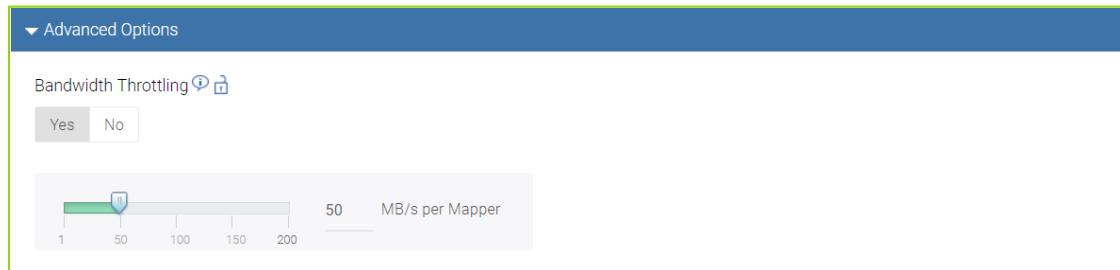
Azure Cloud Storage ▾ adls_stage_backup ▾

As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field.

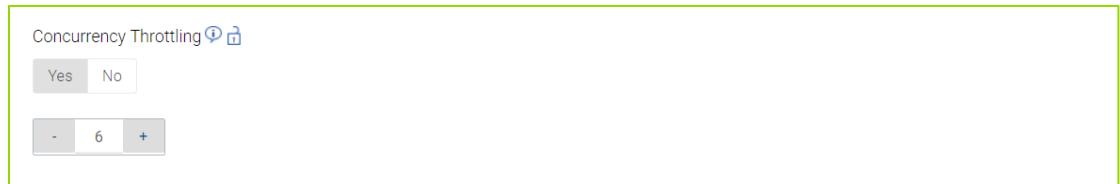
IMPORTANT: Data Centers and Cloud options are available ONLY if you have activated data centers in the Cassandra data repository or if you have selected a backup policy in which you have enabled the cloud retention feature respectively.

10. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

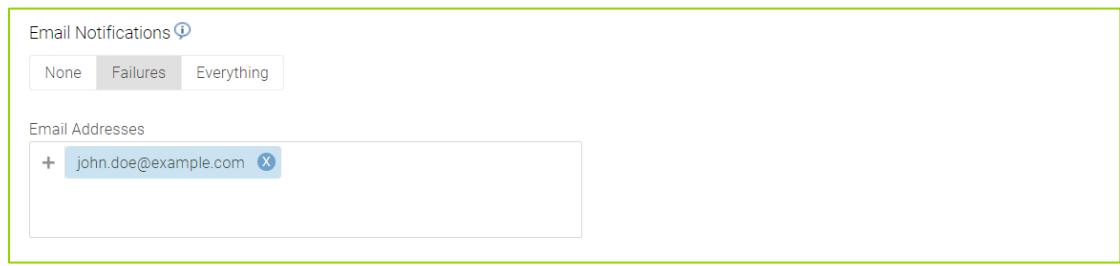
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



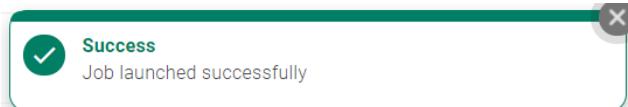
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the Email Addresses field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

11. Click **Submit**. Imanis Data software will then start executing your workflow. A confirmation message will be displayed on the page indicating the job is successfully launched. You can monitor the progression of the job on the Dashboard while its running.



IMPORTANT: Ensure that SSH access is allowed between all the Cassandra nodes and Imanis Data cluster. In case, Puppet/Chef is installed on either Imanis Data cluster or Cassandra nodes, make sure that the Puppet/Chef rules (setup by the user) do not prevent any changes made by the user to allow SSH access from Imanis Data cluster to Cassandra nodes.

WARNING: For Cassandra users, it is highly recommended to ensure that "gc_grace_seconds" is set higher than the backup frequency. In case the gc_grace_seconds is set low, delete mutations will be missed if any rows are deleted and compaction happens on primary cluster between two backup runs.

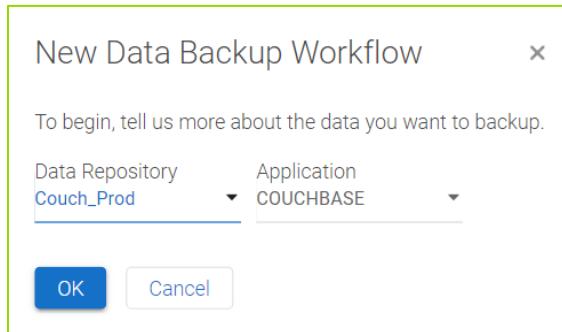
IMPORTANT: To do backup and recovery of DSE graph keyspace, it is advised to skip the 'keyspace_pvt' while executing the backup workflow. For example, if the name of the graph is warehouse, only 'warehouse' and 'warehouse_system' must be backed up. However before you execute the recovery workflow, first manually create the graph metadata using the Gremlin Console or DSE Studio and set the Overwrite option to 'Yes'. Ensure that the restored keyspaces are configured to have the appropriate names as per the created graph. For example, on the destination cluster if the graph has been renamed to 'storehouse' (previously known as 'warehouse'), then the keyspaces related to 'warehouse' must be appropriately renamed to 'storehouse' and 'storehouse_system' by using the Object Rename feature.

7.1.4 Backing Up Data for Couchbase

Imanis Data software enables you to back up data at the bucket level.

To start a backup workflow for Couchbase, do the following:

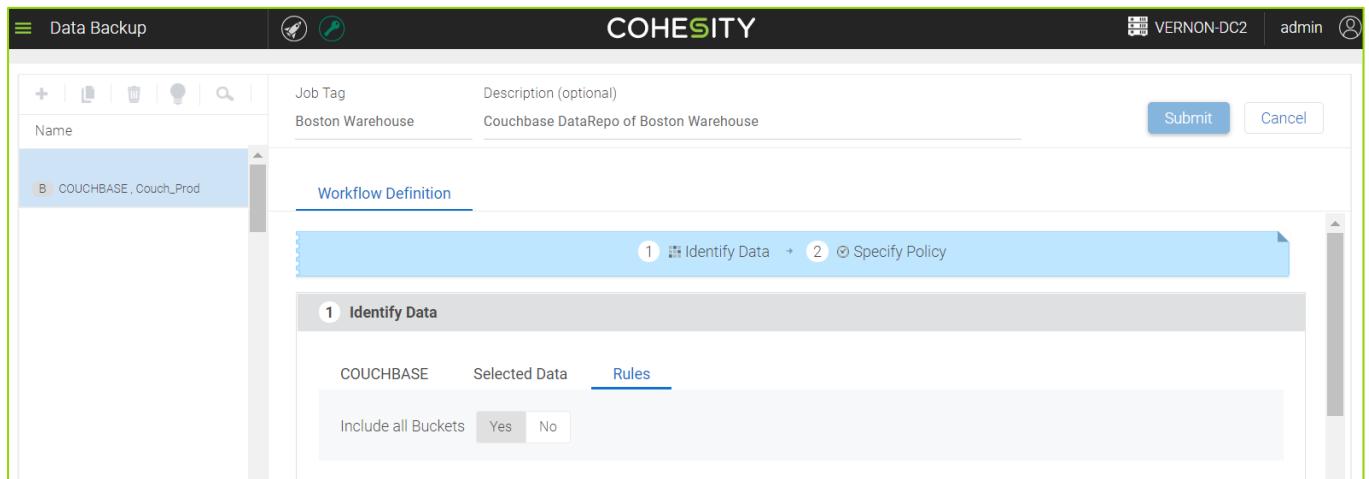
1. Click the **Main Menu** > **Data Management** > **Data Backup**.
2. On the **Data Backup** page, click the icon. The New Data Backup Workflow dialog appears.
In the **New Data Backup Workflow** dialog, select a **Couchbase** source data repository from the Data Repository drop-down menu and then click **OK**. All the buckets available in the Couchbase data repository that you selected earlier are displayed.



3. In the Data Backup page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
4. In the **Identify Data** section, do the following:
 - a. In the **Couchbase** tab, identify the buckets from which you want to backup data by selecting the corresponding check boxes.

IMPORTANT: Imanis Data software supports backup and recovery of Couchbase type buckets, however; you are not permitted to select and backup or recover ephemeral and memcached buckets.

- b. In the **Selected Data** tab, verify your selection or click the **X** icons to remove unwanted items.
- c. In the **Rules** tab, click **Yes** to include all buckets in the backup job. This will also include any buckets that get added in the future.



5. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention. The following screenshot displays with cloud retention option.

6. In the **Specify Options** section under **Cloud options**, do one of the following:

- Select an S3 cloud repository from the **Data Repository** drop-down menu and select a bucket (S3) from the **Buckets** drop-down menu to retain data in the cloud:

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure) from the **Containers** drop-down menu to retain data in the cloud:

The screenshot shows the 'Cloud options' pane. On the left, there's a 'Data Repository' section with 'Azure Cloud Storage' selected. On the right, there's a 'Containers' section with 'adls_stage_backup' selected. Both sections have dropdown menus below them.

As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field.

7. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

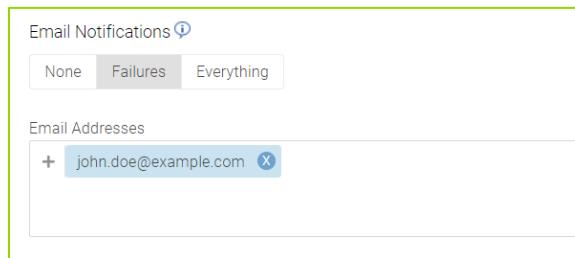
- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:

The screenshot shows the 'Advanced Options' pane expanded to the 'Bandwidth Throttling' section. It has two radio buttons: 'Yes' (selected) and 'No'. Below is a slider with a green bar set at 50, labeled 'MB/s per Mapper'.

- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:

The screenshot shows the 'Advanced Options' pane expanded to the 'Concurrency Throttling' section. It has two radio buttons: 'Yes' (selected) and 'No'. Below is a numeric input field with a value of 6, flanked by minus (-) and plus (+) buttons.

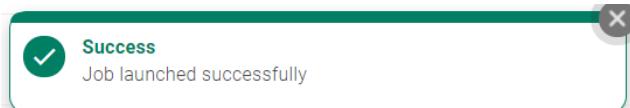
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

8. Click **Submit**. Imanis Data software will then start executing your workflow. A confirmation message will be displayed on the page indicating the job is successfully launched. You can monitor the progression of the job on the Dashboard while its running.



IMPORTANT: Imanis Data software does not support backup and recovery of a Couchbase cluster when one or more active master nodes in Couchbase cluster are unreachable. In case an active master node fails during the backup or recovery process, you can re-run the backup and recovery process once the rebalance or failover (auto or manual) is completed. Remove the failed node from the repository seed node list, re-verify the repository and then save it. To find out if an active node is unreachable, access the Couchbase Web Console. For more information, refer to the Couchbase documentation.

NOTE: The Couchbase Database Change Protocol (DCP) recommends that to derive optimum performance no more than six concurrent connections should be run from the same data repository at any given time. Therefore, by default, the Imanis Data agent invokes a maximum of six data mover processes in parallel. In case you need to execute more than one job concurrently on the same Couchbase cluster, then you must reduce the number of concurrent data movers. This change ensures that the total number connections do not exceed more than six connections.

To reduce the total number of concurrent connections, use the **Concurrency Throttling** option mentioned above.

NOTE: Imanis Data software does not support incremental backup of a bucket, if the user flushes the bucket after having previously taken a full backup. This is a limitation in the current version of the product. In such cases, the user will have to take a new full backup of the flushed bucket.

IMPORTANT: It is recommended that you either verify that all the backup and recovery jobs (including the recurring jobs) are completed before performing a rebalance or execute the backup or recovery jobs during an application's lowest traffic levels. For more information, refer to the information available in Couchbase documentation.

NOTE: Customers should set the backup frequency less than their metadata purge interval. Typically, it is recommended to keep the metadata purge intervals to 7 days just to be on safer side.

NOTE: If a bucket is deleted from the source cluster and a new bucket is created with the same name, then backup of that particular bucket is not supported.

For example, when a bucket is deleted from Couchbase source cluster, the bucket is still present in the source list of Imanis Data backup job. If a new bucket with the same name is created, the bucket is backed up automatically. As a user, you must edit the job to remove the bucket from the source list or not create a bucket with the same name as the deleted bucket.

7.1.5 Backing Up Data for MongoDB

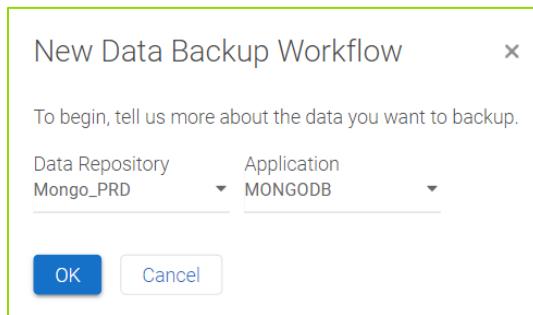
Imanis Data software enables you to back up data at the database and collection level.

IMPORTANT: MongoDB connector uses the MongoDB oplog to perform incremental backups. Backup policy should ensure that the backups are frequent, so that the oplog on any shard does not rollover between two consecutive backup runs.

For example: A workflow protects 2 collections on a 3 way sharded MongoDB cluster. Daily change rate (inserts, updates and deletes) on these two collections is 10 G each. So total change rate is 20Gb / day. Oplog size on each shard is 20Gb. The oplog will rollover in 24 hrs in this example. We recommend that the customer sets up a backup policy such that it is run every 6 hours.

To start a backup workflow for MongoDB, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Backup**.
2. On the **Data Backup** page, click the  icon. The New Data Workflow dialog box appears.
3. In the **New Data Backup Workflow** dialog, select a **MongoDB** source data repository from the **Data Repository** drop-down menu, and then click **OK**. All the databases and collections available in the MongoDB data repository that you selected earlier are displayed. A new page where you can set up a job is displayed.



4. In the **Data Backup Workflow** page, type a new job tag in the **Job Tag** field. The job tag name can include alphanumeric characters, numbers and/or special characters and then type a job tag description in the **Description** field to help others.
5. In the **Identify Data** section, do the following:
 - a. In the **MONGODB** tab, identify the databases and collections from which you want to backup data by selecting the corresponding check boxes.
 - b. In the **Selected Data** tab, verify your selection or click the **X** icons to remove unwanted items.
 - c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section **Appendix A: Rules for Data Inclusions and Exclusions**.

The screenshot shows the 'Data Backup' interface for MongoDB. The 'Workflow Definition' section is open, specifically the 'Identify Data' step. A table lists collections with checkboxes:

Objects	Size
<input checked="" type="checkbox"/> movielens	
<input checked="" type="checkbox"/> system	
<input checked="" type="checkbox"/> human resource	
<input checked="" type="checkbox"/> sales	
<input checked="" type="checkbox"/> payroll	

6. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention. The following screenshot displays with cloud retention option.

2 Specify Policy

Select a data backup policy

Backup to Cloud (s3)

Retention

Allow retention on cloud

Yes No

On Imanis Data On Cloud Delete

1 365

7. In the **Specify Options** section under **Cloud options**, do one of the following:

- Select an S3 cloud repository from the **Data Repository** drop-down menu and select a bucket (S3) from the **Buckets** drop-down menu to retain data in the cloud:

Cloud options

Data Repository	Buckets
S3 Cloud Storage	▼ hadoop_production_backup

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure) from the **Containers** drop-down menu to retain data in the cloud:

Cloud options

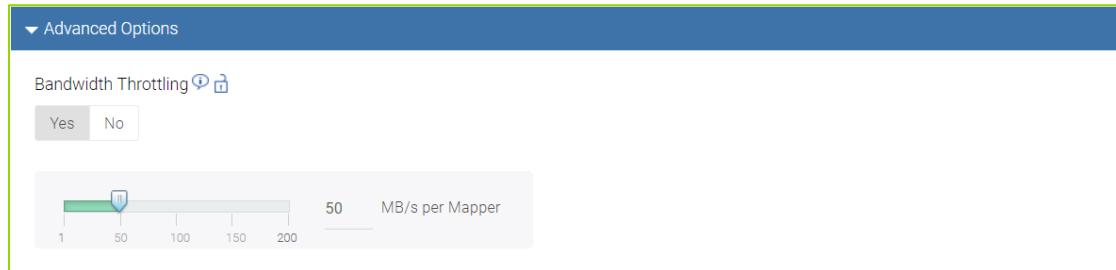
Data Repository	Containers
Azure Cloud Storage	▼ adls_stage_backup

As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field.

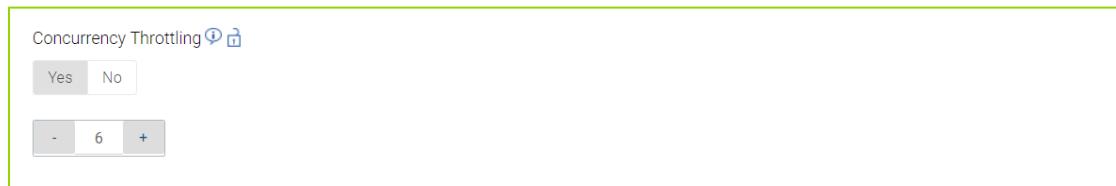
8. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each

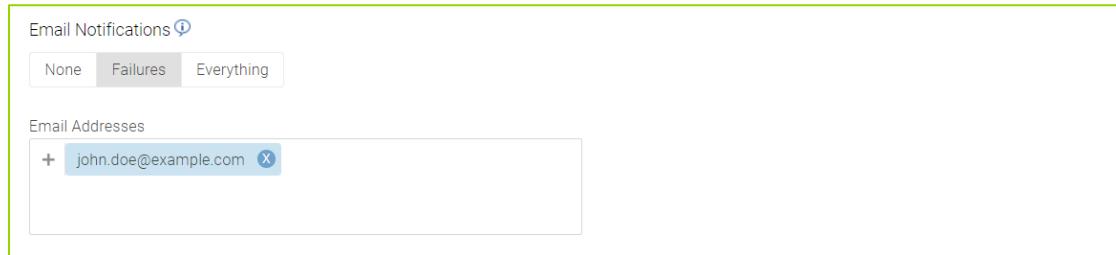
individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **in the MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



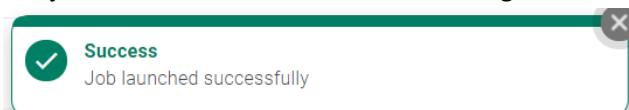
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. Imanis Data software will then start executing your workflow. A confirmation message will be displayed on the page indicating the job is successfully launched. You can monitor the progression of the job on the Dashboard while its running.



IMPORTANT: In case of full backup, the Sharded Cluster Balancer is automatically disabled. Disabling the balancer for the duration of the backup procedure is a recommended practice. For more information, refer to the MongoDB documentation on [disabling balancer during backups](#).

When the backup completes successfully or with errors, the balancer is automatically reactivated in case it was disabled by the MongoDB connector. However, in a rare scenario, if the MongoDB fails or if a running backup workflow is killed via the Imanis Data GUI, the balancer does not get reactivated. In this case, it is recommended that users manually enable the balancer on MongoDB cluster. Refer to the MongoDB documentation on Balancer [here](#).

8 Data Lifecycle Management

This section describes the features of the Data Lifecycle Management (DLM) menu of Imanis Data software.

8.1 Overview

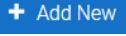
Imanis Data software enables you to manage the entire lifecycle of your data, through a user-defined policy, from capturing data in the primary data repositories to recovering data to a destination cluster and from archiving the data to Amazon Glacier to automatically deleting the cold data. Currently, you can manage the lifecycle of data for HDFS and Hive.

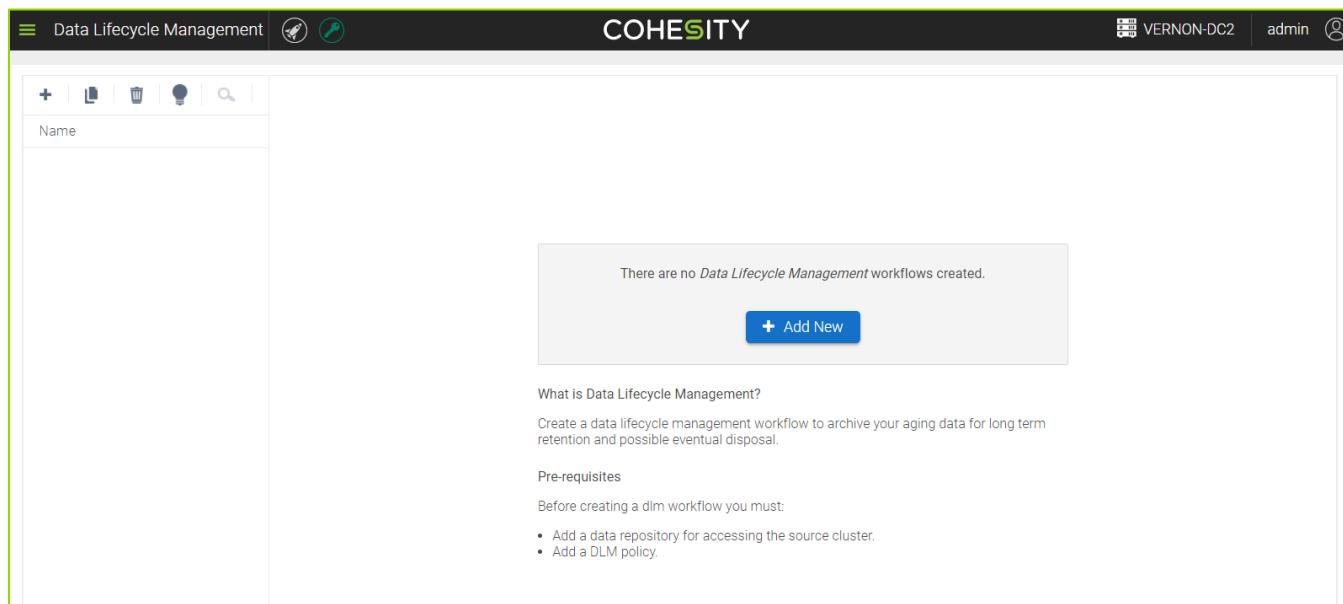
IMPORTANT: Before you get started, you must create a data repository and a data lifecycle policy through the Administrator menu.

8.1.1 Data Lifecycle Management for HDFS & Hive

Currently, you can manage the lifecycle of data for HDFS and Hive applications only.

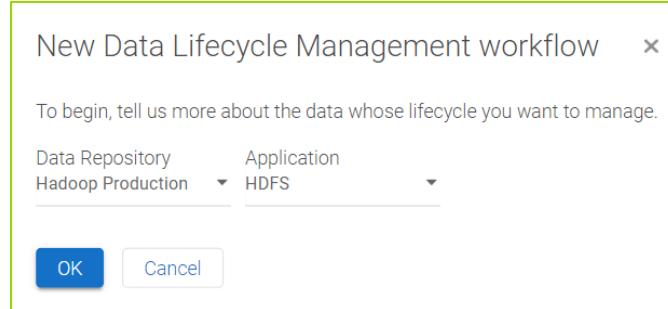
To manage the lifecycle of data on HDFS & Hive, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Lifecycle Management**.
2. Click the  button or the  icon. The **New Data Lifecycle Management Workflow** dialog box appears.



The screenshot shows the Cohesity Data Lifecycle Management interface. The top navigation bar includes 'Data Lifecycle Management', a search bar, and user information ('admin'). The main content area displays a table with columns for Name, Status, and Actions. A message box states 'There are no Data Lifecycle Management workflows created.' with a blue 'Add New' button. Below this, a 'What is Data Lifecycle Management?' section provides a brief description and a 'Pre-requisites' section with a bulleted list: 'Add a data repository for accessing the source cluster.' and 'Add a DLM policy.'

3. In the **New Data Lifecycle Management Workflow** dialog box, do the following:
- Select a data repository (from where you want to backup the data) from the **Data Repository** drop-down menu.
 - Select **HDFS** and **Hive** from the **Application** drop-down menu.
 - Click **OK**. All the files, directories, tables, databases, or partitions, available in the particular data repository that you selected earlier appear here.



4. In the **Data Lifecycle Management** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Identify Data** section, do the following:
- In the **HDFS** tab, identify the files, tables, directories, or databases from which you want to backup data by selecting the corresponding check boxes. You can search objects by typing a regex in the search box.
 - In the **Selected Data** tab, verify your selection or click the **X** icons to remove unwanted items.
 - In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section **Appendix A: Rules for Data Inclusions and Exclusions**.

Modified Time	Owner	Size
2019-09-06 01:30:35 ...	talena	2.8 GB
2019-09-06 01:30:35 ...	talena	29.7 GB
2019-09-06 01:30:35 ...	talena	937.6 MB
2019-09-06 01:30:35 ...	talena	2 GB
2019-09-06 01:30:35 ...	talena	5.7 GB

IMPORTANT: The Imanis Data does not support having blank spaces in the inclusion or exclusion regex.

NOTE: The regular expressions (regex) in the Exclusions and Inclusions field have to be separated by a new line. Regular expressions work as follows:

1. The inclusion filter generates a list of data objects on the selected data repository that match the regular expression. This list is in addition to the data objects that you have already selected in the Identify Data section.
2. The exclusion filter is applied to the selected objects (during the Identify Data step) and the objects that match the inclusion filter and exclude data objects that match the regex.

For example, a user has selected database “database_abc” in the **Identify Data** section:

1. In the Inclusion field, if the user specifies an inclusion filter as db_*, then the user has selected the database named “database_abc” and also all the databases from the data repository starting with name prefix ‘db_’.
2. In the Exclusion field, if the user specifies an exclusion filter as *table_tmp*, then the user has selected tables containing string table_tmp, which will be excluded from the list of data objects that have been selected using the Identify Data section and inclusion filter.

6. In the **Specify Policy** option, select a DLM policy. Only the Data Lifecycle Management (DLM) policies that you created earlier appear here. For example, the following graphic displays a DLM policy that stores data in the cloud.

2 Specify Policy

Select a data lifecycle management policy

Archive On Demand

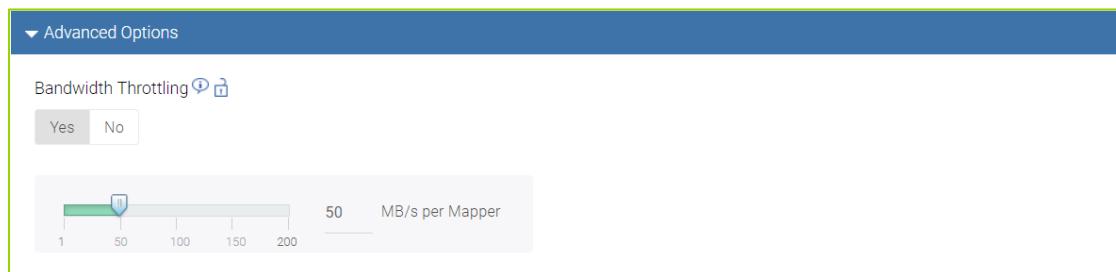
Lifecycle Management

Full Replica Reduced Replica On Imanis Data On Cloud

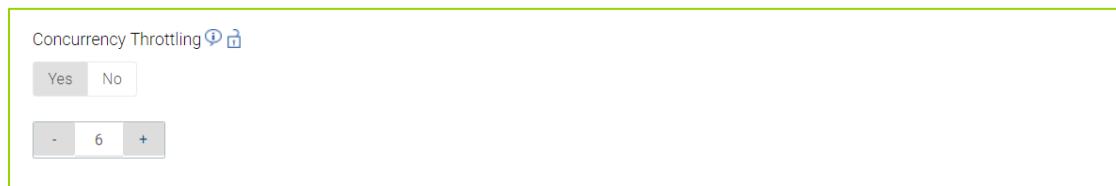
```
graph LR; A((Full Replica)) -- "90 days" --> B((Reduced Replica)); B -- "Skip" --> C((On Imanis Data)); C -- "Forever" --> D((On Cloud)); D -- "Skip" --> E((Delete))
```

Note: Priority defines the resource allocation that a job receives when the Imanis Data cluster is heavily loaded. The default priority for each job type is already set in the Imanis Data GUI. For example, the default priority for a recovery job is set to “High,” whereas the priority is set to “Medium” for Backup, DLM, and Data Mirroring jobs. Usually, the storage reduction jobs (which run in the background) would be set to low priority.

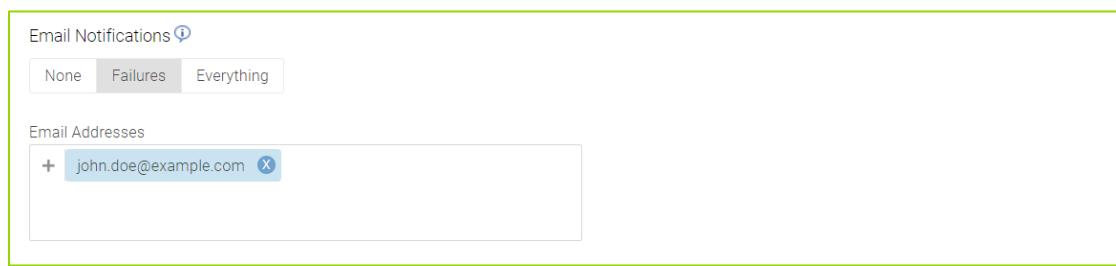
7. Select a data repository from the **Data Repository** drop-down menu. This option is available only if the cloud storage option (Amazon Glacier) is activated in the policy (that you selected in the preceding step).
8. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:
 - In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



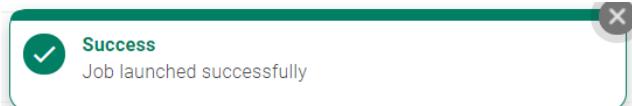
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

9. Click **Submit**. Imanis Data software will then start executing your workflow. A confirmation message will be displayed on the page indicating the job is successfully launched. You can monitor the progression of the job on the Dashboard while its running.



9 Data Recovery

This section describes the features of the Data Recovery menu of Imanis Data software.

9.1 Overview

The process of data recovery involves copying data that has been previously backed-up using Imanis Data software to the original, alternate location, or disaster recovery (DR) site.

9.2 Getting Started with Data Recovery

The Data Recovery workflow enables administrators to recover data from the Imanis Data cluster to the original or to an alternate primary data repository. When recovering backup data, time-specific restore points are available to recover from previous points in time.

NOTE: Prior to starting the recovery workflow, ensure that both 'read' and 'write' permissions are granted to the user that is used for application discovery.

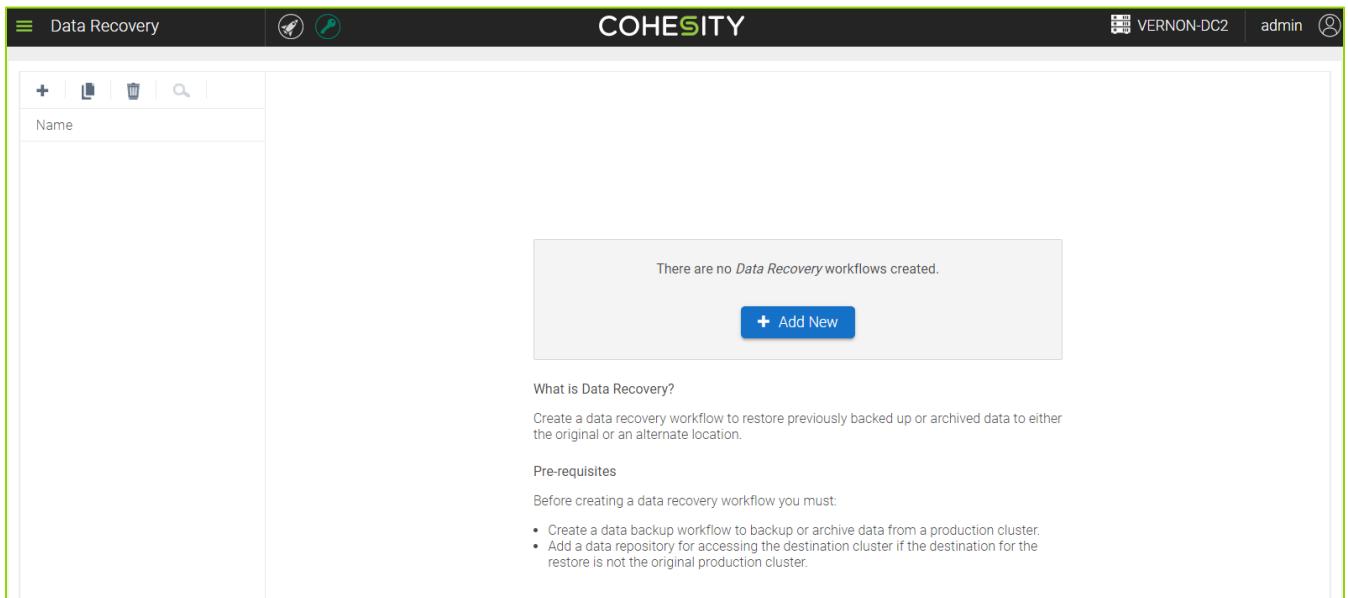
9.2.1 Recovering Data for HDFS

Imanis Data software supports HDFS recovery at the jobtag, file, and directory level.

To recover data for HDFS, do the following, do the following:

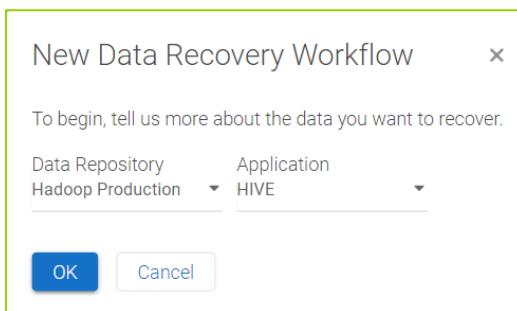
1. Click the **Main Menu**  > **Monitoring and Recovery** > **Data Recovery**.

The following page appears only if you are creating data recovery workflows for the first time:



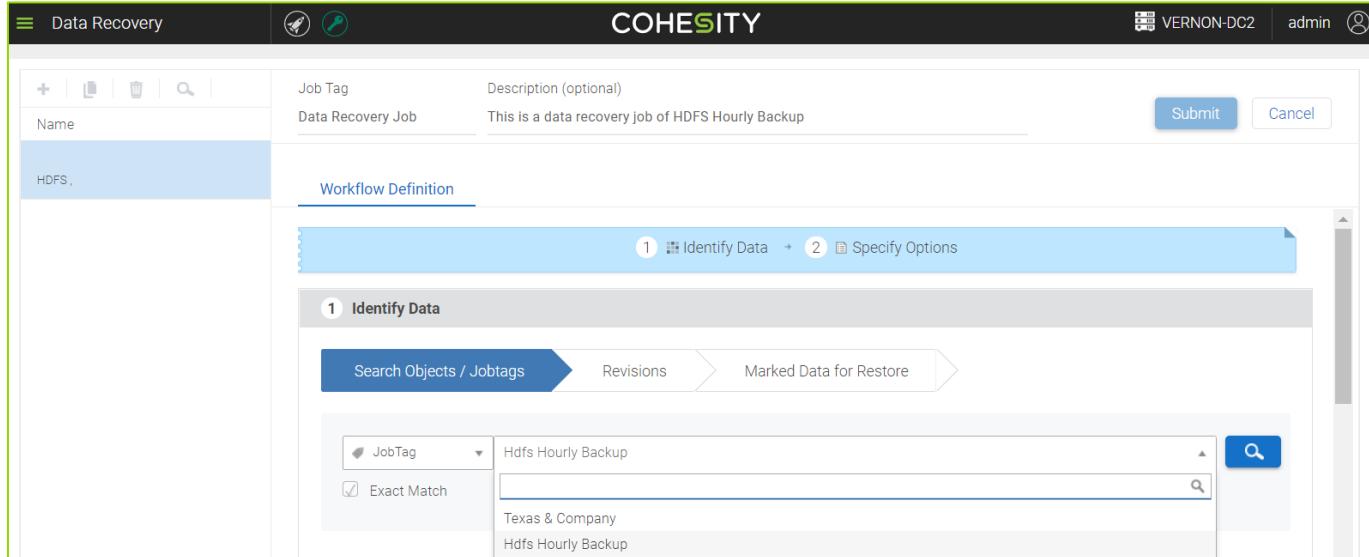
The screenshot shows the 'Data Recovery' page in the Cohesity interface. The top navigation bar includes the 'Data Recovery' tab, a search bar, and user information ('admin'). The main content area displays a message: 'There are no *Data Recovery* workflows created.' Below this is a blue 'Add New' button. To the right, there's a section titled 'What is Data Recovery?' with a brief description and a list of pre-requisites, which includes creating a backup workflow and adding a data repository.

2. On the Data Recovery page, click the  **+ Add New** button or the  icon to create data recovery workflows. The **New Data Recovery Workflow** dialog box appears.



3. In the **New Data Recovery Workflow** dialog box, select a HDFS data repository from the **Data Repository** drop-down menu and then click **OK**.
4. In the **Data Recovery** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.

5. In the **Search** tab, under the **Search Objects/Jogtags** label, click the search box, select a **JobTag** displayed by Imanis Data, and then click the  icon. Similarly, you can select **File** and **Directory** by typing the full or partial file name.

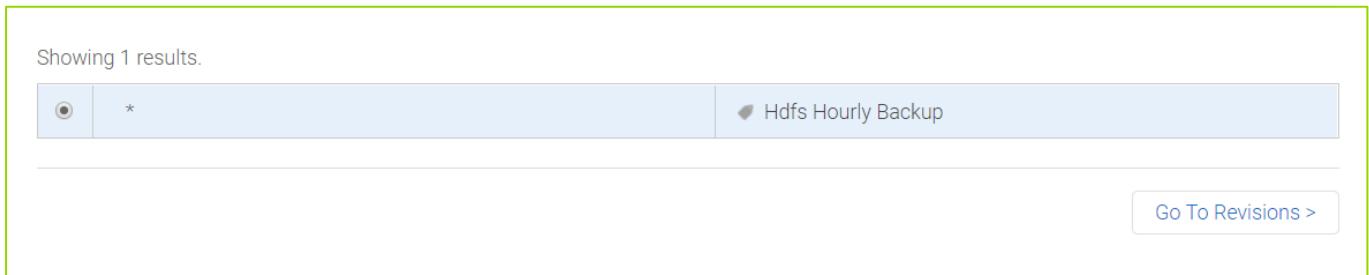


The screenshot shows the Cohesity Data Recovery interface. On the left, there's a sidebar with icons for Data Recovery, Protection, and Monitoring. The main area has tabs for 'Data Recovery' and 'Protection'. The 'Data Recovery' tab is active. A sub-menu for 'Data Recovery' is open, showing 'Job Tag' and 'Description (optional)' fields. The 'Description' field contains 'This is a data recovery job of HDFS Hourly Backup'. Below this is a 'Workflow Definition' section with two steps: '1 Identify Data' and '2 Specify Options'. Under 'Identify Data', there's a 'Search Objects / Jobtags' button, followed by 'Revisions' and 'Marked Data for Restore' buttons. A dropdown menu for 'Job Tag' is set to 'Hdfs Hourly Backup' and the 'Exact Match' checkbox is checked. The search results list 'Hdfs Hourly Backup' and 'Texas & Company Hdfs Hourly Backup'.

NOTE: A search that is based on partial term is enabled for File, Directory, Databases, Tables, and Partitions only. However, Imanis Data software does not support search terms containing the following special characters like file\$%#@#fs2.log.

NOTE: When using a partial or inexact term search, the message “No valid revisions available for this object at this time” may be displayed when selecting the object for recovery”. This behavior can be observed if the catalog is being updated by an actively running job. If you encounter this error, please rerun the search after job finishes completely.

6. Click the radio button of the **JobTag** search result and then click the **Go To Revisions** button.



The screenshot shows a search results page. It says 'Showing 1 results.' There is a single result listed: 'Hdfs Hourly Backup'. To the left of the result is a radio button and a wildcard character (*). To the right is a 'Go To Revisions >' button.

7. In the **Revisions** label, do one of the following:

- By default, the latest copy is selected which is indicated by the icon. You can click the **Next** button to restore the selected data object
- Click the data object icon to select a copy of data for a specific day and time. You can click the **Next** button to restore the selected data object

1 Identify Data

Search Objects / Jobtags Revisions Marked Data for Restore

Hdfs Hourly Backup *

Revisions
Select a revision to restore

< Re-select Next >

NOTE: Navigate all the data object revision by clicking the icons. You can also click the icon to jump to a specific revision in time by selecting a date and time or click the icon to jump to the currently selected revision.

8. In the **Marked Data for Restore** label, do one of the following:
- Verify the current revision selection
 - Click the **X** icon to remove the current selection and then click the **Change Revision** button to select a new Objects/Jobtags revision to restore

1 Identify Data

Search Objects / Jobtags > Revisions > Marked Data for Restore

Hdfs Hourly Backup (9 Sep 2019, 06:04 AM)

Selected Objects	
*	X

< Change Revision

Reset

9. In the **Specify Options** section, select an **Original Location** or **Alternate Location**.

ORIGINAL LOCATION:

1. Click the **Original Location** button.

2 Specify Options

Recover to

Original Location Alternate Location

2. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

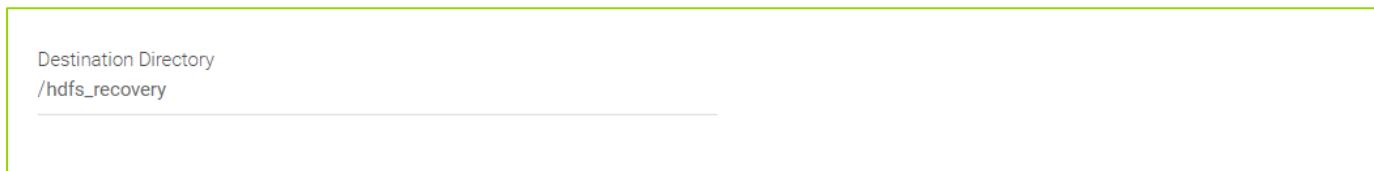
- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if data objects are already present on the destination the recovery job would fail

Additional Options

Overwrite Behavior ⓘ

Replace Keep

3. Type a name for the directory in the **Destination Directory** field. For example, /hdfs_recovery/new. Imanis Data software creates a new directory with this name and recovers data objects in this directory.

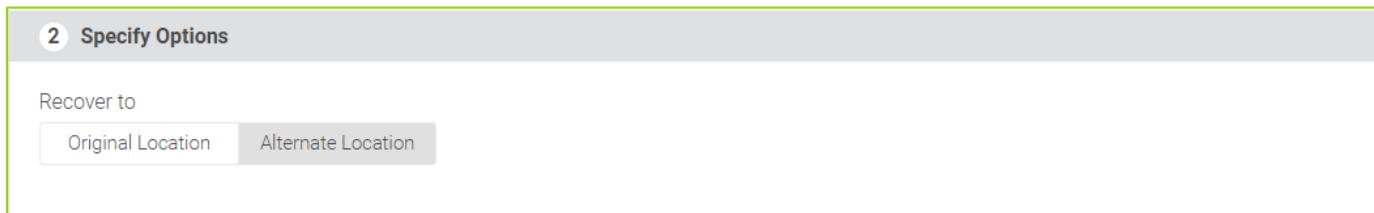


Destination Directory
/hdfs_recovery

NOTE: Recovery uses the entire available bandwidth if the throttling parameter is not specified.

ALTERNATE LOCATION:

1. Click the **Alternate Location** button.

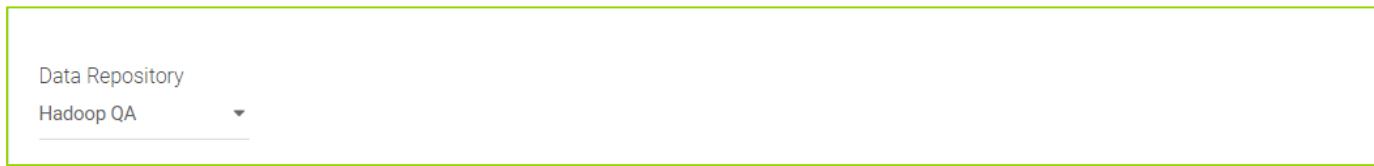


2 Specify Options

Recover to

Original Location **Alternate Location**

2. Select a data repository, from the **Data Repository** drop-down menu, where you want to recover the data.

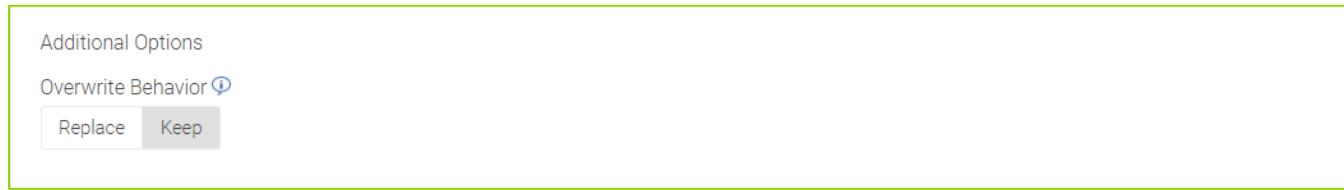


Data Repository

Hadoop QA

3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if data objects are already present on the destination the recovery job would fail



Additional Options

Overwrite Behavior ⓘ

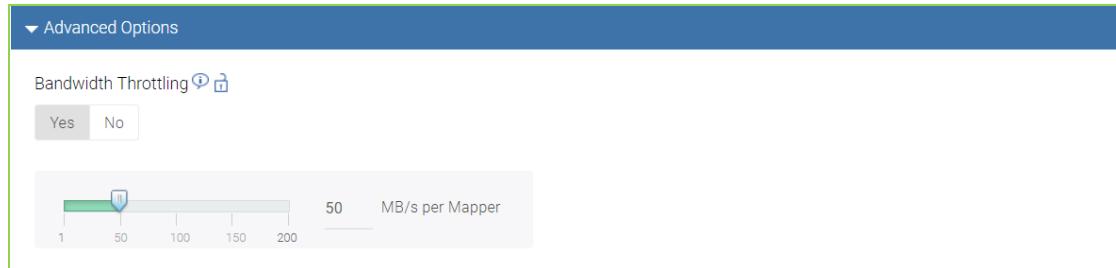
Replace **Keep**

4. Type a name for the directory in the **Destination Directory** field. For example, /hdfs_recovery/new. Imanis Data software creates a new directory in alternate location with this name and recovers data objects in this directory.

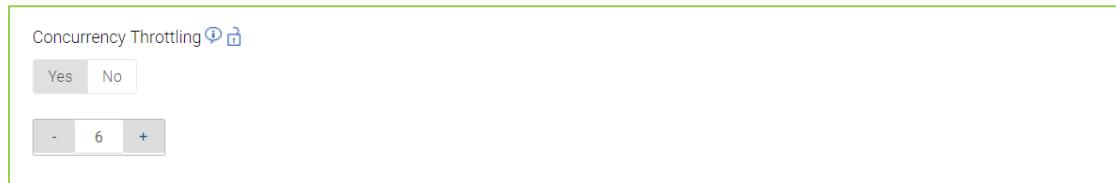
Destination Directory
/hdfs_recovery

11. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

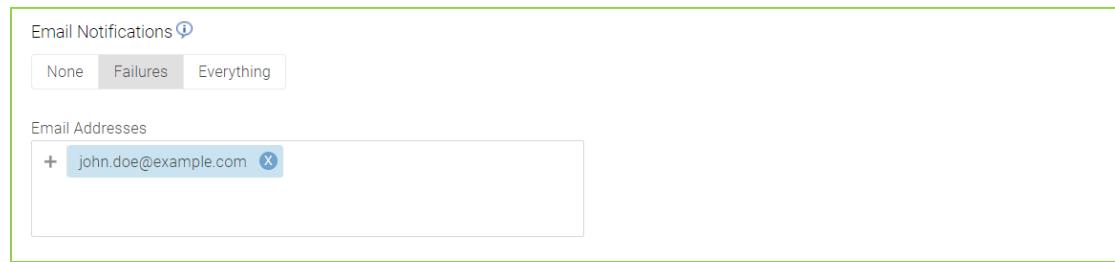
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field: In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



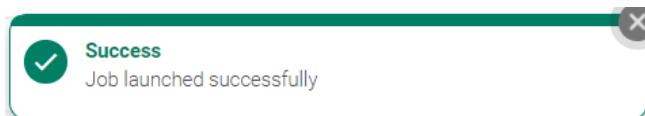
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

12. Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.

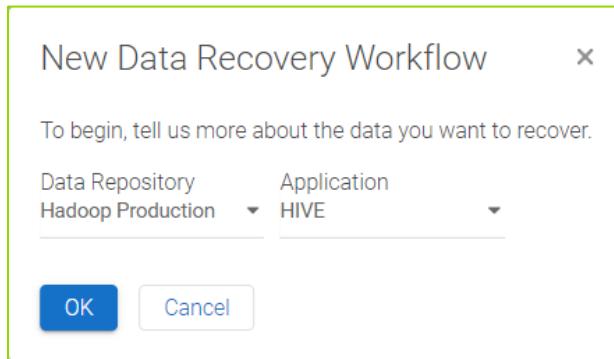


9.2.2 Recovering Data for Hive

Imanis Data software supports Hive recovery at the jobtag, database, table, and partition level.

To recover data for Hive, do the following:

1. Click the **Main Menu**  > **Monitoring and Recovery** > **Data Recovery**.
2. On the **Data Recovery** page, click the  or the  to create data recovery workflows. The **New Data Recovery Workflow** dialog box appears.



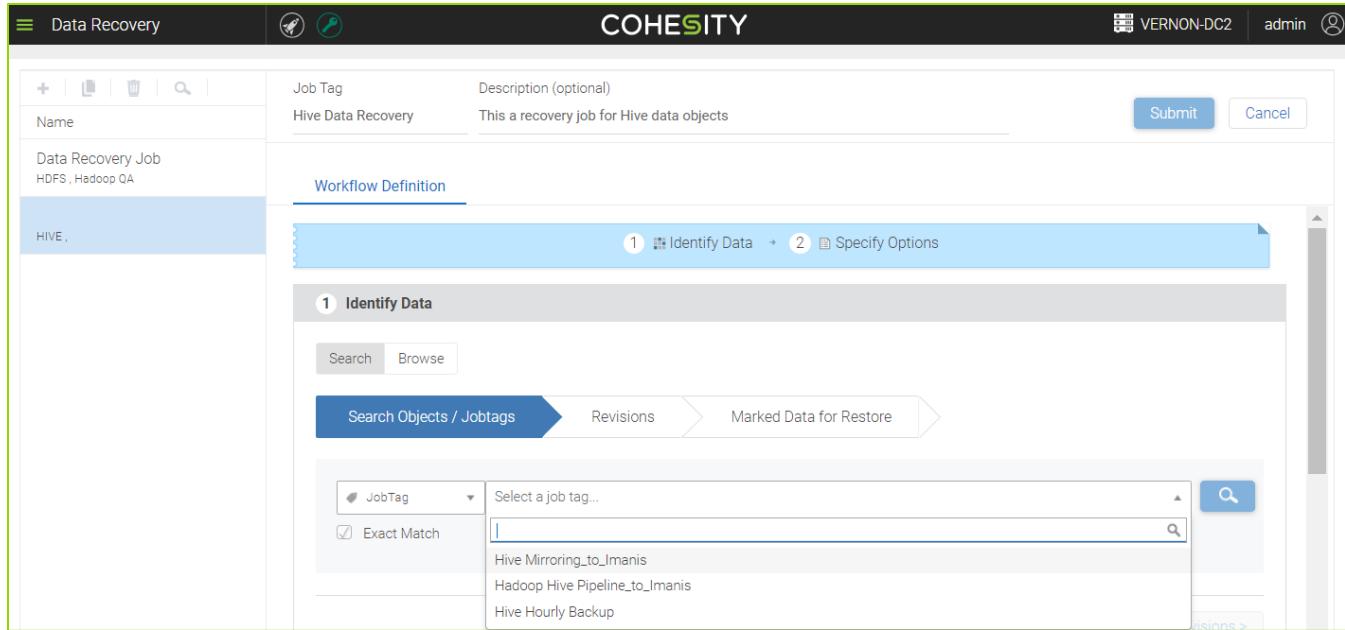
3. In the **New Data Recovery Workflow** dialog box, select a **Hive** data repository (from where the data to be recovered originated) from the **Data Repository** drop-down menu and then click **OK**.
4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.

5. In the **Identify Data** section, **Search** and **Browse** tabs are displayed. You can use the Search and Browse button as per your requirement:

SEARCH	BROWSE
Use when you know the data object that you wish to recover	Use when you want to view data objects and select specific data objects within a JobTag revision
Search specific Jobtag, Database, Table, or Partition	Browse the data objects catalog and select or deselect multiple data objects at a single time

Search Browse:

6. In the **Search Objects/Jobtags** label, click the search box, select a **JobTag** displayed by Imanis Data, and then click the  icon. JobTag search result is displayed.



Similarly, you can select **Database**, **Table**, and **Partition** from the drop-down list by typing the full or partial file name.

7. Click the radio button of the JobTag search result and then click the **Go To Revisions** button.

The screenshot shows the 'Identify Data' tab selected. At the top, there are 'Search' and 'Browse' buttons. Below them is a navigation bar with three steps: 'Search Objects / Jobtags' (highlighted in blue), 'Revisions', and 'Marked Data for Restore'. The main search area has a dropdown set to 'JobTag' and a text input field containing 'Hive Mirroring_to_Imanis'. A checkbox for 'Exact Match' is checked. Below the search bar, it says 'Showing 1 results.' followed by a table row with a radio button icon, an asterisk icon, and the object name 'Hive Mirroring_to_Imanis'. At the bottom right of the search area is a 'Go To Revisions >' button.

8. In the **Revisions** tab, do one of the following:

- By default, the latest copy is selected which is indicated by the icon. You can click the **Mark For Restore** button to restore the selected data object
- Click the data object icon to select a copy of data for a specific day and time. You can click the **Mark For Restore** button to restore the selected data object

1 Identify Data

Search Browse

Search Objects / Jobtags Revisions Marked Data for Restore

Hive Mirroring_to_Imanis *

Revisions
Select a revision to restore

< Re-select Next >

NOTE: Navigate all the data object revision by clicking the icons. You can also click the icon to jump to a specific revision in time by selecting a date and time or click the icon to jump to the currently selected revision.

9. In the **Marked Data for Restore** label, do one of the following:

- Verify the selected data object revision and move on to the next step
- Click the **Change Revision** button to reselect a data object revision to restore

The screenshot shows the 'Identify Data' process at step 1. It includes tabs for 'Search' and 'Browse'. A navigation bar at the top has three steps: 'Search Objects / Jobtags', 'Revisions', and 'Marked Data for Restore' (which is highlighted in blue). Below the navigation bar, a list item 'Hive Mirroring_to_Imanis 9 Sep 2019, 07:05 AM' is shown. Underneath it, there's a table titled 'Selected Objects' with one entry '*'. A red 'X' icon is in the top right corner of the table, and a 'Reset' button is in the bottom right corner. At the bottom left, there's a link '[< Change Revision](#)'.

You can continue with specifying the recovery options by referring to Step #17.

Browse Tab

- In the **Browse Jobtags** label, select a JobTag from the **JobTag** list, and then click the **Go To Revisions** button.

The screenshot shows the 'Identify Data' process at step 1. It includes tabs for 'Search' and 'Browse'. A navigation bar at the top has four steps: 'Browse Jobtags' (highlighted in blue), 'Revisions', 'Browse Objects', and 'Marked Data for Restore'. Below the navigation bar, there's a search bar labeled 'Enter job tag here to filter...' with a magnifying glass icon. Underneath it, a section titled 'JobTags List' shows a table with three rows:

<input checked="" type="radio"/>	Hive Mirroring_to_Imanis
<input type="radio"/>	Hadoop Hive Pipeline_to_Imanis
<input type="radio"/>	Hive Hourly Backup

At the bottom right of the interface, there's a link '[Go To Revisions >](#)'.

8. In the **Browse** tab, under the **Revisions** label, select a revision of the **JobTag** revision that you want to restore and then click the **Browse Objects** button.

The screenshot shows the 'Identify Data' interface with the 'Revisions' tab selected. A timeline displays several revision points for a JobTag named 'Hive Mirroring_to_Imanis'. The revision at 06:05 am on Sep 09, 2019, is highlighted with a red circle. Navigation arrows and a search/browsing bar are visible above the timeline.

Revision Time	Revision ID
Invalid date	
07:05 am	Sep 09, 2019
06:05 am	Sep 09, 2019
05:05 am	Sep 09, 2019
04:05 am	Sep 09, 2019
03:04 am	Sep 09, 2019
02:05 am	Sep 09, 2019
01:04 am	Sep 09, 2019

9. In the **Browse** tab, under the **Browse Objects** label, do one of the following:
- Select databases or tables that you want to restore and then click the **Next** button
 - Click the **Change revision** button to go back to the **Revisions** tabs and select a new revision of the JobTag

1 Identify Data

Search Browse

Browse Jobtags Revisions **Browse Objects** Marked Data for Restore

Hive Mirroring_to_Imanis -

Select the desired objects and click on "Mark For Restore" button to restore the objects.

	/
<input checked="" type="checkbox"/>	Objects
<input checked="" type="checkbox"/>	sales
<input checked="" type="checkbox"/>	payroll
<input checked="" type="checkbox"/>	legal

< Change revision Next >

10. In the **Browse** tab, under the **Marked Data for Restore** label, do one of the following:

- Click the **Reset** button or click the **X** icon to remove any data object from the list
- Click the **Browse Objects** button to go back and go through all the objects again

1 Identify Data

Search Browse

Browse Jobtags Revisions **Browse Objects** **Marked Data for Restore**

Hive Mirroring_to_Imanis - 4 AM

Selected Objects	
payroll	X
sales	X

Reset

< Browse Objects

11. In the **Specify Options** section, select one of **Original Location** or **Alternate Location**:

ORIGINAL LOCATION:

1. Click the **Original Location** button.

2 Specify Options

Recover to

Original Location Alternate Location

Additional Options

Overwrite Behavior ⓘ

Replace Keep

Suffix ⓘ

09092019

2. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if data objects are already present on the destination the recovery job would fail

Additional Options

Overwrite Behavior ⓘ

Replace Keep

3. Type a number and/or character in the **Suffix** field to add a suffix to the data objects being recovered from the Imanis Data cluster. For example, ImanisData11012014

Suffix ⓘ

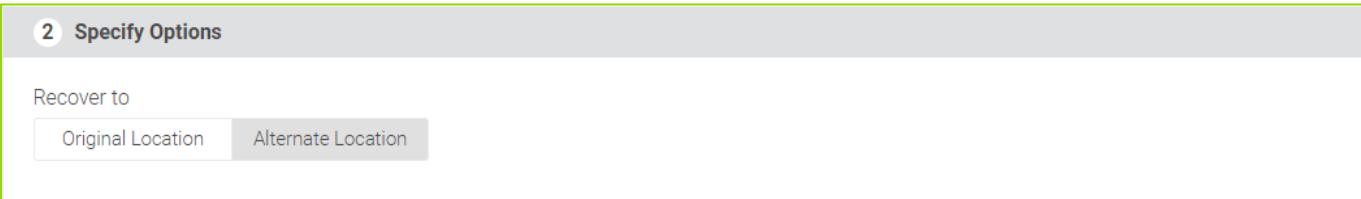
09092019

NOTE: Imanis Data software does not support recovery of Hive Views if the recovery is executed with an alternate name or alternate database. If a backed-up database contains Hive View, and if it is recovered with an alternate name, then Hive View will not be available.

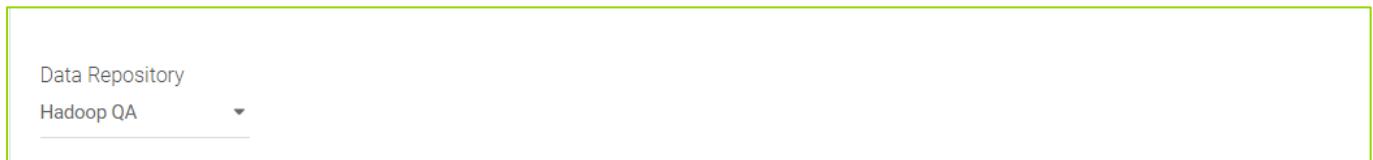
NOTE: Imanis Data software supports the use of alphanumeric characters in the suffix field, however; the use of uppercase is not recommended.

ALTERNATE LOCATION:

1. Click the **Alternate Location** button.

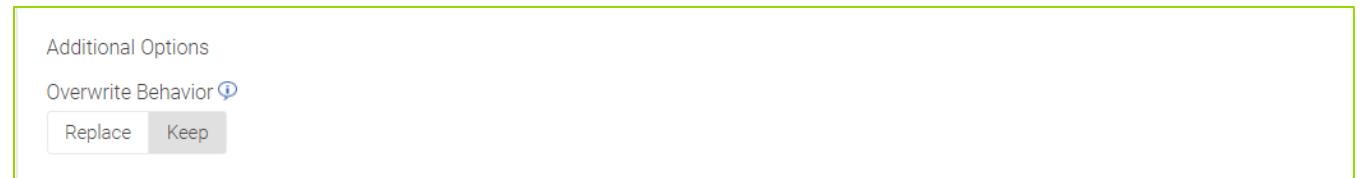


2. Select the destination data repository from the **Data Repository** drop-down menu where you want to recover the data.

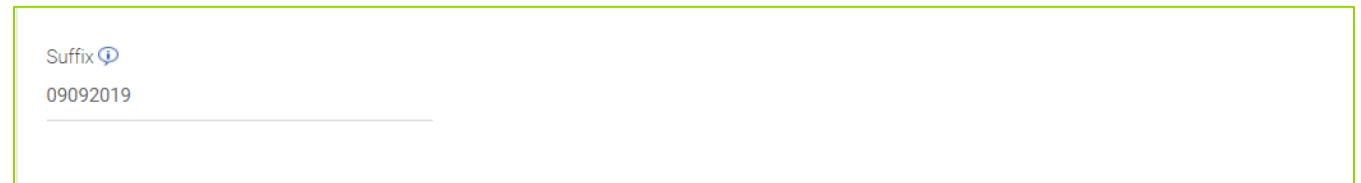


3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with existing data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if data objects are already present on the destination the recovery job would fail



4. Type a number and/or character in the **Suffix** field to add a suffix to the data objects being recovered from the Imanis Data cluster. For example, ImanisData11012014

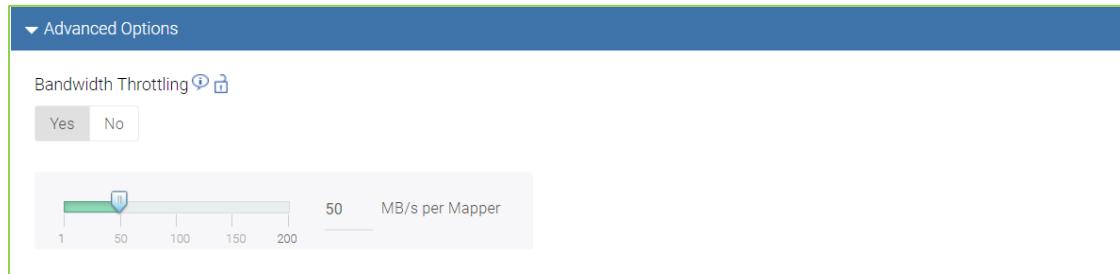


NOTE: Imanis Data software supports the use of alphanumeric characters in the suffix field; however, the use of uppercase is not supported.

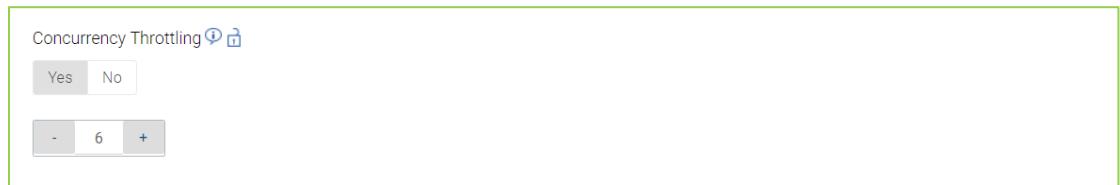
IMPORTANT: If you are restoring partition with suffix, then that suffix gets applied to partition value. Example, if a partition named database1.table1.dt=2014-11-11/hr=01 is being restored with suffix '_tmp', then restored partition name at destination will be database1.table1.dt=2014-11-11/hr=01_tmp.

12. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

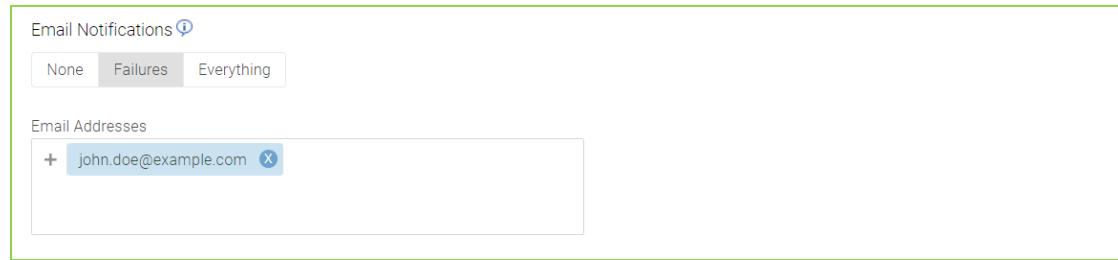
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

13. Click **Submit**.

9.2.3 Recovering Data for HBase

Imanis Data software supports HBase recovery at the **Jobtag**, **Namespace**, and **Table level**.

To recover data for HBase, do the following:

1. Click the **Main Menu** > **Monitoring and Recovery** > **Data Recovery**.
2. On the **Data Recovery** page, click the or the . The **New Data Recovery Workflow** dialog box appears.

There are no *Data Recovery* workflows created.

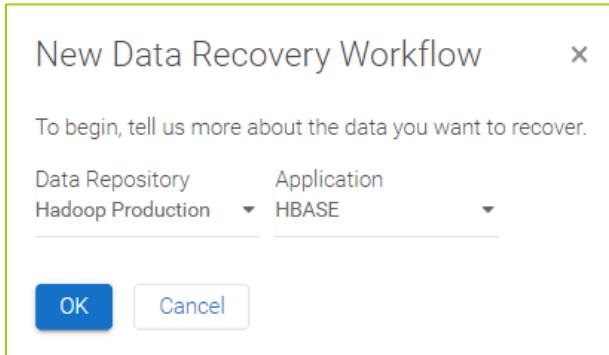
+ Add New

What is Data Recovery?
Create a data recovery workflow to restore previously backed up or archived data to either the original or an alternate location.

Pre-requisites
Before creating a data recovery workflow you must:

- Create a data backup workflow to backup or archive data from a production cluster.
- Add a data repository for accessing the destination cluster if the destination for the restore is not the original production cluster.

3. In the **New Data Recovery Workflow** dialog box, select a **Hadoop** source data repository from the **Data Repository** drop-down menu, select HBase from the **Application** field, and then click **OK**. The data recovery page for HBase appears.

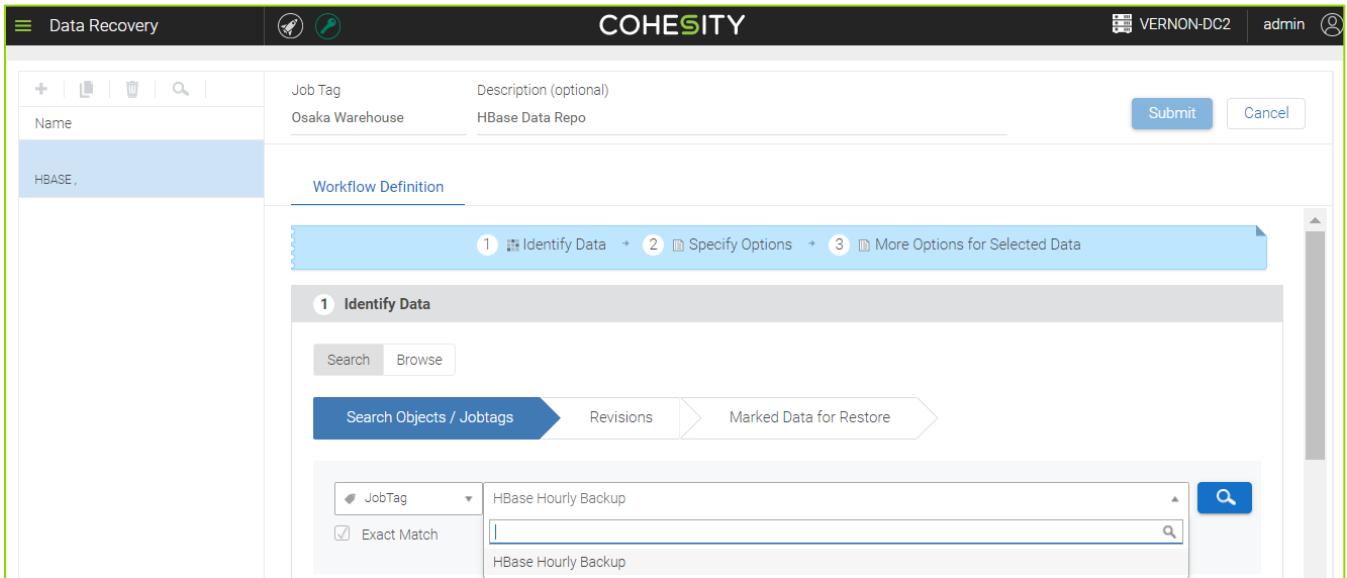


4. In the **Data Recovery** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Identify Data** section, **Search** and **Browse** buttons. You can use the Search and Browse button as per your requirement:

SEARCH	BROWSE
Use when you know the data object that you wish to recover	Use when you want to view data objects and select specific data objects within a JobTag revision
Search specific Jobtag, Namespace, Table	Browse the data objects catalog and select or deselect multiple data objects at a single time

Search Tab

6. In the **Search** tab, under the **Search Objects/Jobtags** label, click the search box, select a **JobTag** displayed by Imanis Data, and then click the  icon. JobTag search result is displayed.

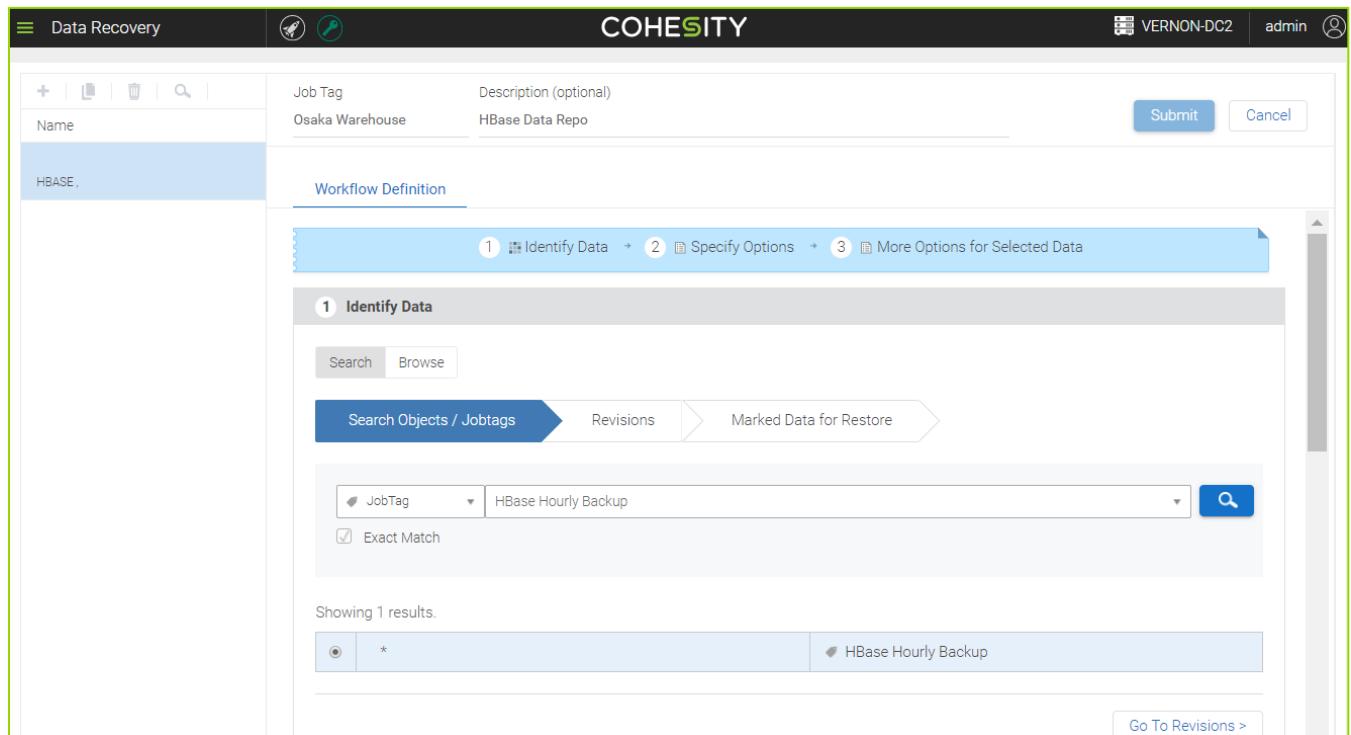


The screenshot shows the Cohesity Data Recovery interface. In the top navigation bar, the 'Data Recovery' tab is selected. On the left, there's a sidebar with options like '+', '|', 'trash', and a search icon. The main area has a title 'COHESITY' and a user 'admin'. Below the title, there's a form with 'Job Tag' set to 'Osaka Warehouse' and 'Description (optional)' set to 'HBase Data Repo'. There are 'Submit' and 'Cancel' buttons. The central part of the screen is titled 'Workflow Definition' and shows a step-by-step process: 1. Identify Data, 2. Specify Options, 3. More Options for Selected Data. Under 'Identify Data', there are tabs for 'Search' and 'Browse', and a sub-section for 'Search Objects / Jobtags'. It shows a dropdown menu with 'JobTag' selected and 'Exact Match' checked. The search input field contains 'HBase Hourly Backup', and the search button is highlighted with a blue border.

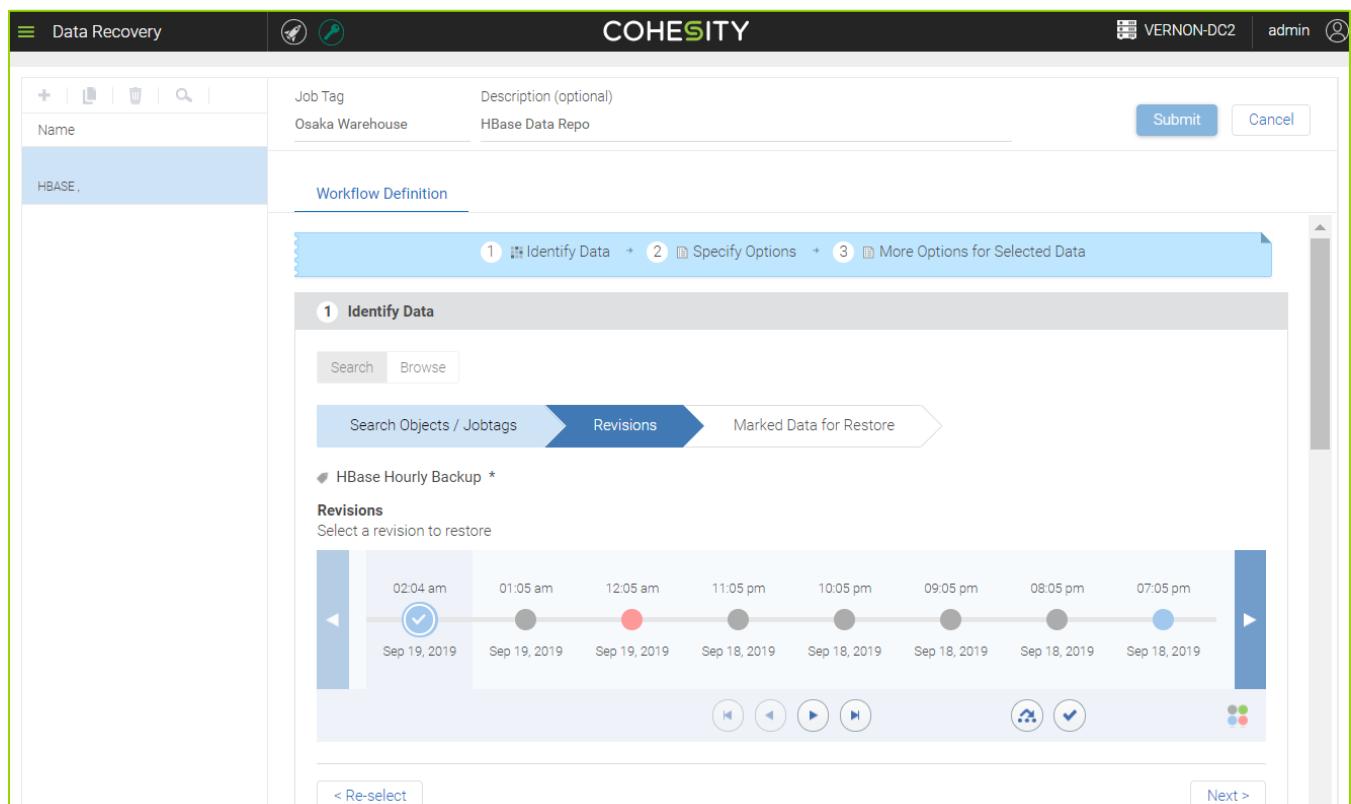
Similarly, you can search for a **Namespace** and **Table** by selecting the appropriate option from the drop-down list and by typing the full or partial file name. In order to search for a table, you have to specify the full path including the namespace name and table name

- For Exact match search, use : (colon) as the separator
- For Partial match search, use . (dot) as the separator

7. Click the radio button of the JobTag search result and then click the **Go To Revisions** button.



- In the **Search** tab, under the **Revisions** label, select a revision of the Objects/JobTag revision that you want to restore and then click the **Mark for Restore** button.



9. In the **Revisions** tab, do one of the following:

- By default, the latest copy is selected which is indicated by the icon. You can click the **Mark For Restore** button to restore the selected data object
- Click the data object icon to select a copy of data for a specific day and time. You can click the **Mark For Restore** button to restore the selected data object

NOTE: Navigate all the data object revision by clicking the icons. You can also click the icon to jump to a specific revision in time by selecting a date and time or click the icon to jump to the currently selected revision.

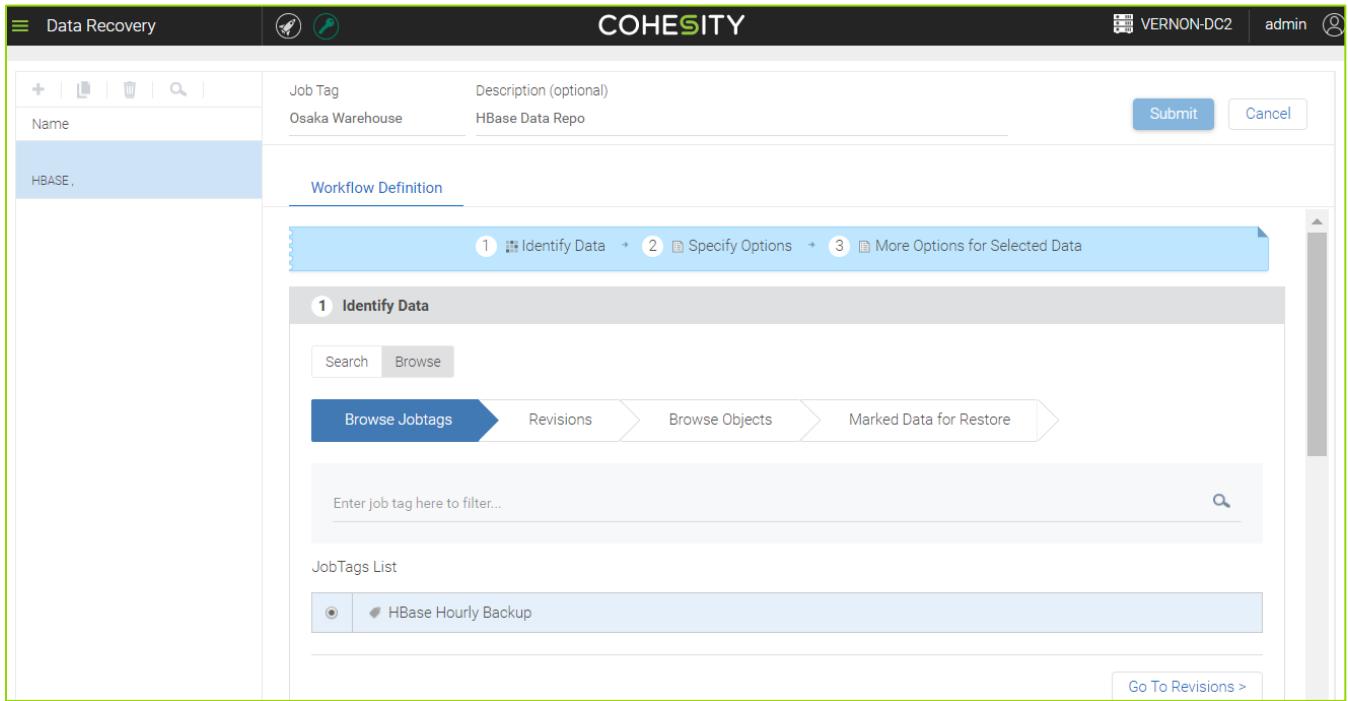
10. In the **Marked Data for Restore** label, do one of the following:

- Verify your selection
- Click the icon to remove the current selection and click the Change Revision button to reselect a new revision

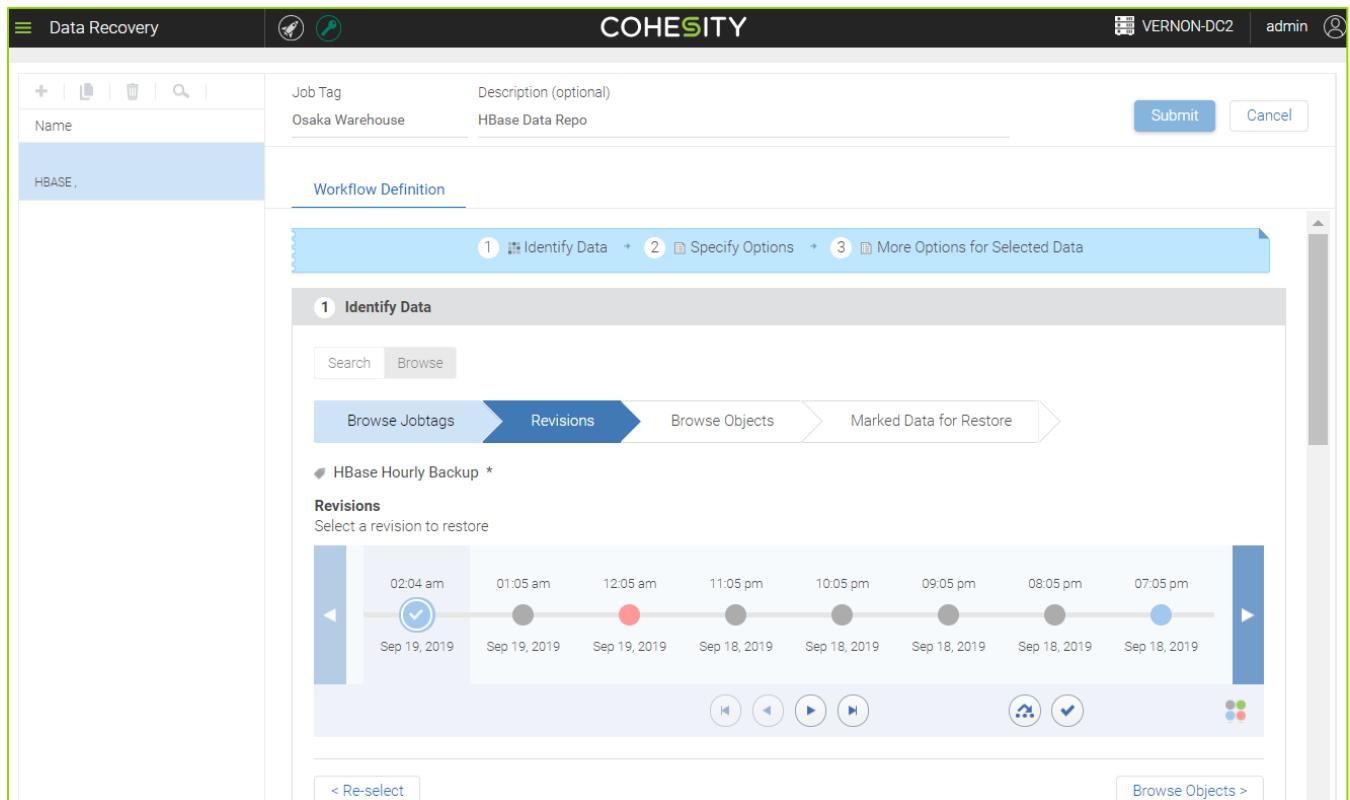
The screenshot shows the Cohesity Data Recovery interface. At the top, there's a navigation bar with 'Data Recovery' on the left, followed by several icons (refresh, search, etc.), the 'COHESITY' logo in the center, and user information ('VERNON-DC2' and 'admin') on the right. Below the navigation bar is a search bar with fields for 'Name' and 'Job Tag', and a 'Description (optional)' field containing 'Osaka Warehouse' and 'HBase Data Repo'. There are 'Submit' and 'Cancel' buttons. The main area is titled 'Workflow Definition' and contains three steps: 1. Identify Data, 2. Specify Options, and 3. More Options for Selected Data. Step 1 is currently active, showing a 'Search' tab selected. Below this, a breadcrumb navigation shows 'Search Objects / Jobtags' → 'Revisions' → 'Marked Data for Restore'. A specific backup entry is listed: 'HBase Hourly Backup' from '19 Sep 2019, 02:04 AM'. Underneath, a table titled 'Selected Objects' shows a single entry with a red 'X' icon to its right. At the bottom of the screen, there's a button labeled '< Change Revision'.

Browse Tab

11. In the **Browse Jobtags** label, select a **JobTag** from the **JobTag list**, and then click the **Go To Revisions** button.

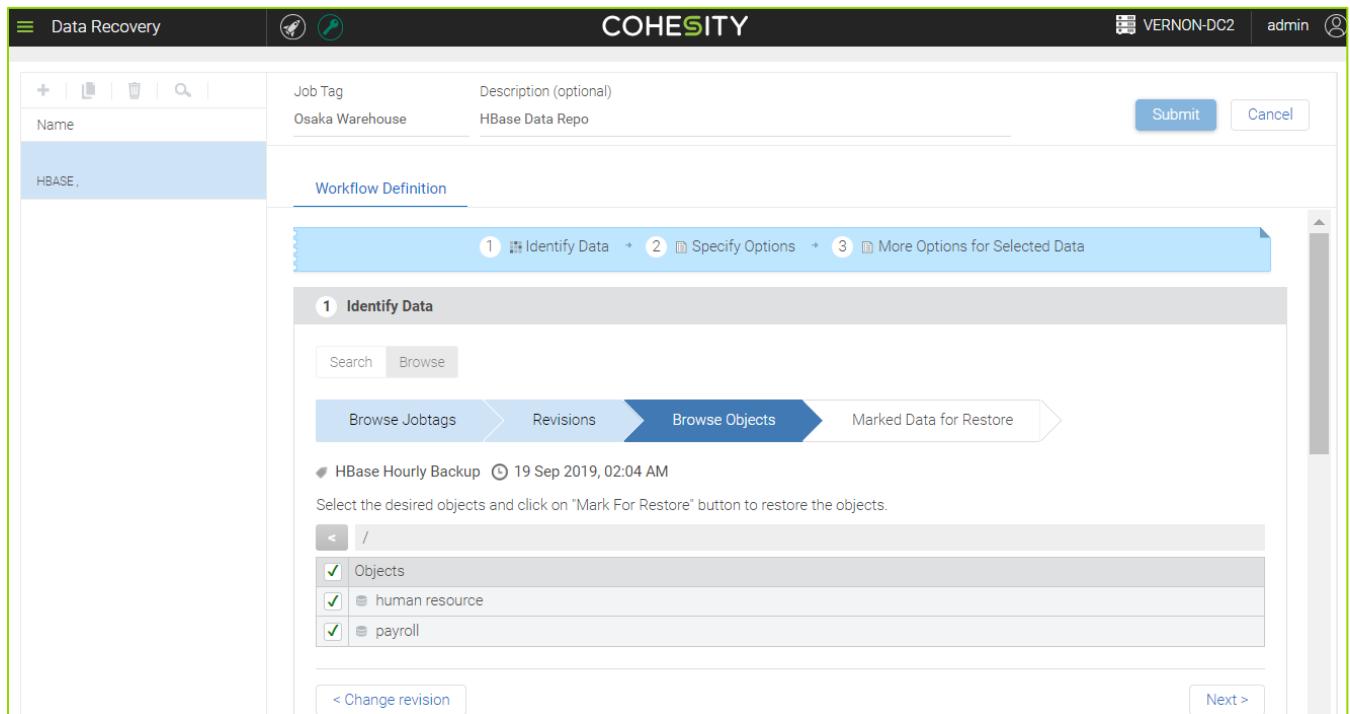


12. In the **Browse** tab, under the **Revisions** label, select a revision of the **JobTag** revision that you want to restore and then click the **Browse Objects** button.



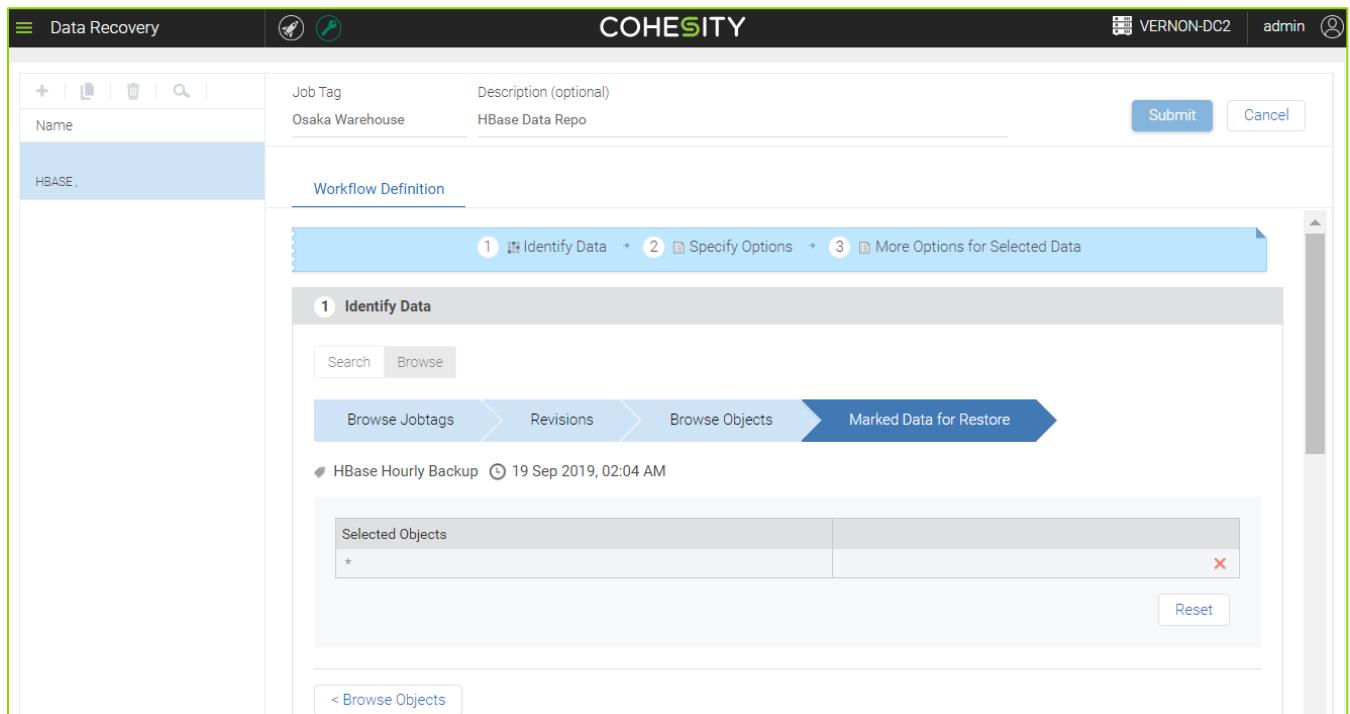
13. In the **Browse Objects** label, do one of the following:

- Select namespaces or tables that you want to restore and then click the **Next** button
- Click the **Change revision** button to go back to the Revisions tabs and select a new revision of the JobTag



14. In the **Marked Data for Restore label, do one of the following:**

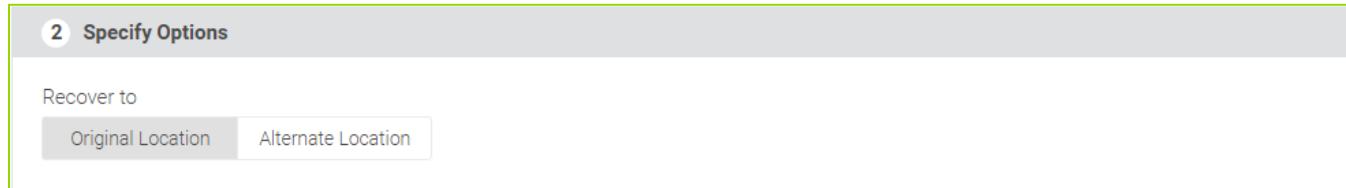
- Click the **X** icon to remove any data object from the list
- Click the **Browse Objects** button to go back and select the desired objects to restore



15. In the **Specify Options** section, select one of the following: **Original Location** or **Alternate Location**.

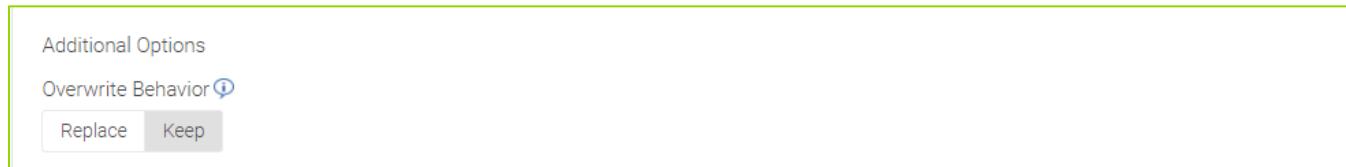
ORIGINAL LOCATION:

1. Click the **Original Location** button.

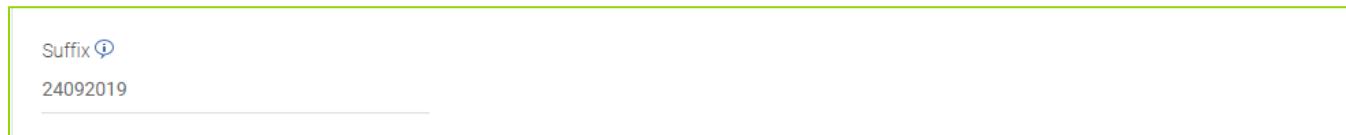


2. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with jobs existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied

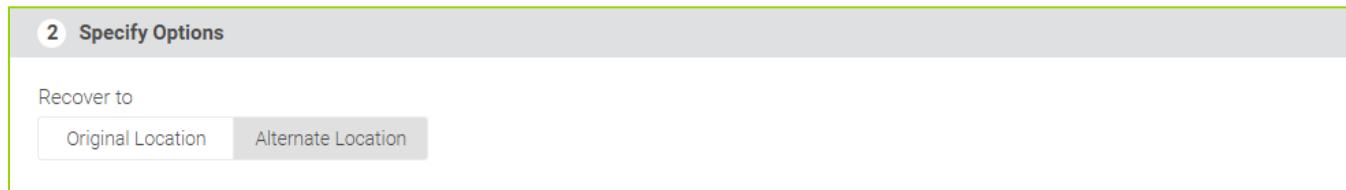


3. Type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster in the **Suffix** field. For example, _10102017.



ALTERNATE LOCATION:

1. Click the **Alternate Location** button.



2. Select a HBase data repository, from the **Data Repository** drop-down menu, where you want to recover the data.

Data Repository

Hadoop DR

3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click Replace to replace existing data with new data thus erasing any previously existing data
- Click Keep to retain existing data (if any). However, if there is no existing data then the new data will be copied

Additional Options

Overwrite Behavior 

Replace

Keep

4. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

Suffix 

24092019

12. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:

 Advanced Options

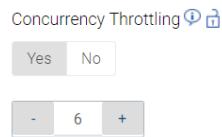
Bandwidth Throttling  

Yes

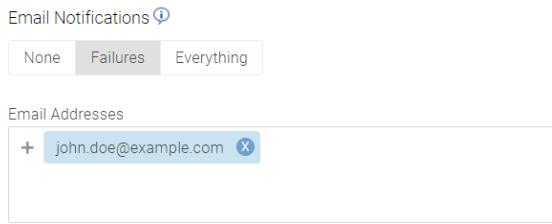
No



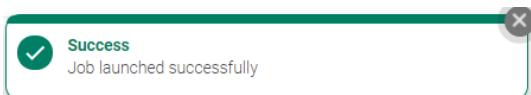
- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



IMPORTANT: Data recovery from one version to another version is not supported. For example, data recovery from 5.8.2 to 5.12.2 is not supported. This event is also applicable to data mirroring workflows where the source and destination versions are different. In case you upgrade to a new version of HBase, it is recommended to recreate all the workflows that are associated with the data repository.

NOTE: Imanis Data software supports the use of alphanumeric in the suffix field, however; the use of uppercase is not recommended.

9.2.4 Recovering Data for Cassandra

Imanis Data software supports recovery at the jobtag, keyspace, and the table level for Apache and DataStax Enterprise (DSE) Cassandra. Imanis Data software can do granular restores from a full backup. For example, if you had taken a full keyspace backup last night, then you can easily restore a specific table from it.

You can do data recovery to an alternate Cassandra cluster, which does not necessarily have to be the same size as the production Cassandra cluster. For example, if you have a six node Cassandra cluster, Imanis Data software enables you to restore to an alternative four node Cassandra cluster.

Imanis Data software automatically selects an optimal data recovery path if you perform recovery to the original location. This optimal recovery path will work only if the ssh user used at the time of Cassandra application discovery is 'root' or 'cassandra'. For more information, refer to the section on Cassandra application discovery.

You can also change the name and specific properties of the restored objects. As of now, only the following properties are supported: for Keyspace: replication, and for Table: compression and compaction.

IMPORTANT: If you have made any changes to the `cassandra.yaml` configuration file after adding a Cassandra data repository in Imanis Data software, then you must auto-discover the Cassandra data repository again. For example, if authentication or authorization type is changed in the `cassandra.yaml` file, then auto-discovering the Cassandra data repository is mandatory.

IMPORTANT: For DSE 5.0 and later releases, Imanis Data software does not restore permissions of tables or databases if the required set of roles are not present on destination restore cluster. Thus, the user must create the required set of roles on destination cluster before initiating the recovery workflow. The user must also ensure to create transitive role dependencies if any exists. For example, on the Source cluster, you have roles 'supervisor', 'admin', 'staff'.

The role 'admin' has "Select" privileges on database 'company'.

The role 'staff' has "Insert" and "Select" privileges on database 'company'.

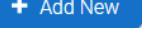
While the role 'supervisor' has transitive dependency wherein it inherits all the privileges of both the roles 'admin' and 'staff' with the following command:

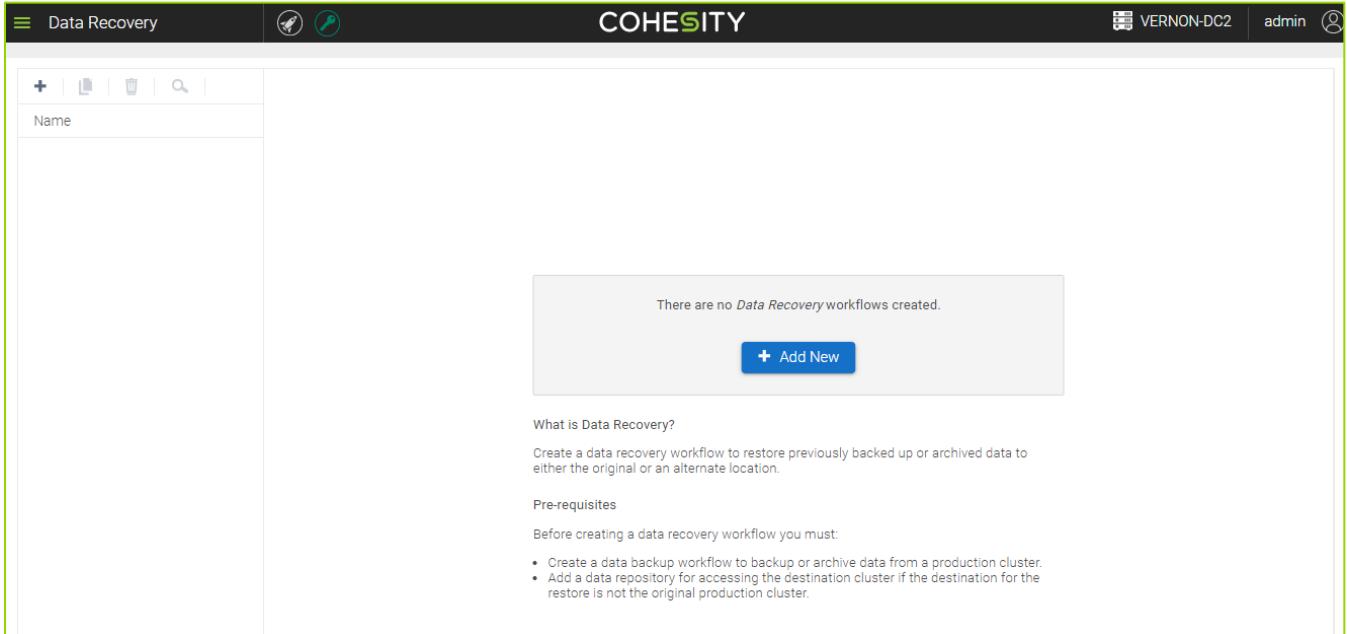
```
Grant admin to supervisor;
Grant staff to supervisor;
```

Thus, the user must ensure that such transitive dependent roles (like the role 'supervisor') are created on destination cluster before initiating recovery workflow.

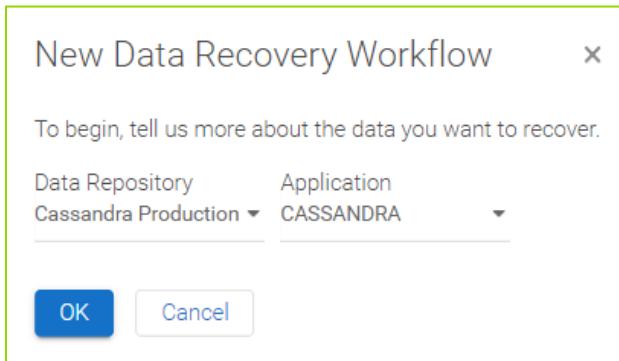
IMPORTANT: If Authorization is enabled on destination cluster, ensure that users present in Cassandra primary cluster are also present on destination cluster.

To start data recovery workflow, do the following:

1. Click the **Main Menu**  > **Monitoring and Recovering** > **Data Recovery**.
2. On the Data Recovery page, click the  or the  icon. The **New Data Recovery Workflow** dialog box appears.



3. In the **New Data Recovery Workflow** dialog box, select a **Cassandra** data repository from the **Data Repository** drop-down menu, and then click **OK**.



4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field.
5. In the **Identify Data** section, **Search** and **Browse** tabs are displayed. You can use the Search and Browse button as per your requirement:

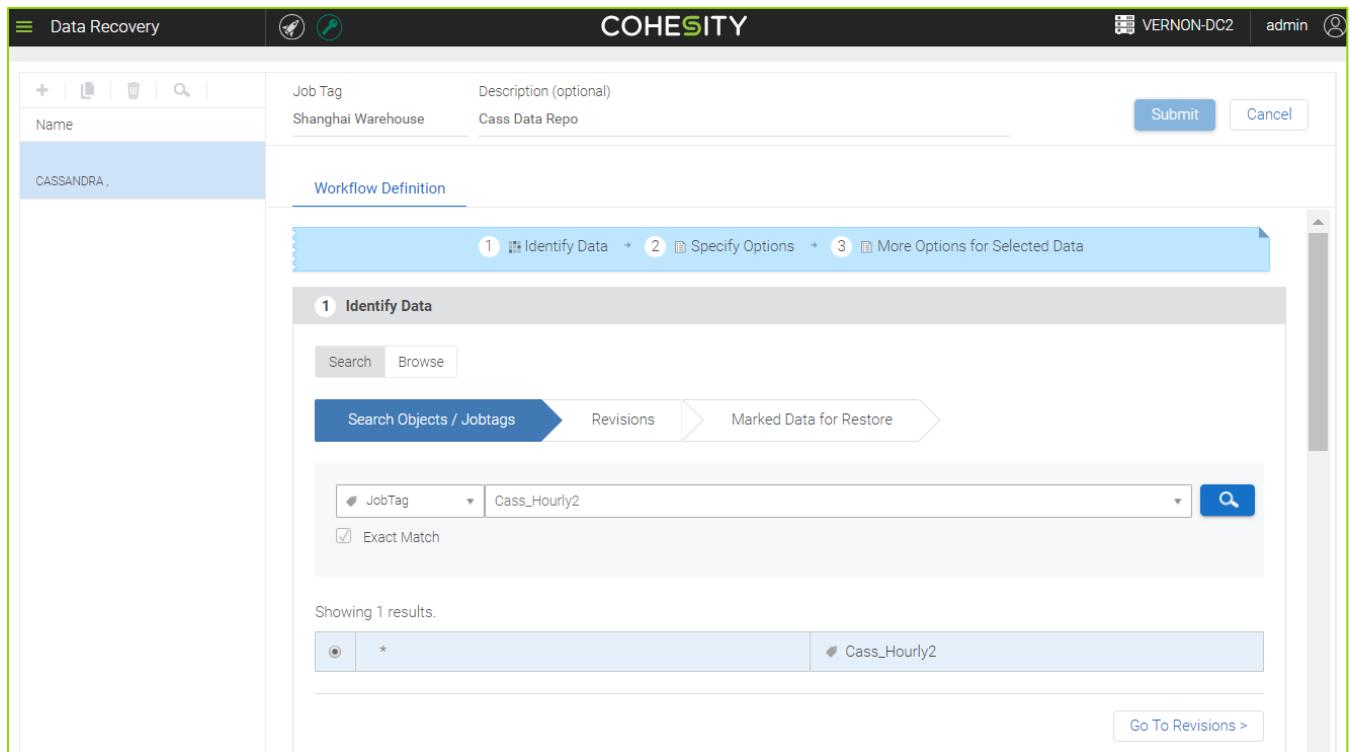
SEARCH	BROWSE
Use when you know the data object that you wish to recover	Use when you want to view data objects and select specific data objects within a JobTag revision
Search specific Jobtag, Keyspace, and Table	Browse the data objects catalog and select or deselect multiple data objects at a single time

Search Tab

6. In the **Search Objects/Jobtags** tab, click the search box and select a job tag displayed by Imanis Data, and then click the search icon. Imanis Data will display the jobtag data object as a search result.

The screenshot shows the Cohesity Data Recovery interface. The top navigation bar includes 'Data Recovery', a search icon, the 'COHESITY' logo, and user information ('VERNON-DC2' and 'admin'). The main area is titled 'Workflow Definition' and shows a three-step process: 'Identify Data', 'Specify Options', and 'More Options for Selected Data'. The 'Identify Data' step is active, with tabs for 'Search' and 'Browse'. Below this is a navigation bar with arrows pointing to 'Search Objects / Jobtags', 'Revisions', and 'Marked Data for Restore'. A dropdown menu shows 'JobTag' selected and 'Exact Match' checked. A search input field contains the placeholder 'Select a job tag...'. A list of job tags is displayed, including 'Cass_Hourly2', 'NY Warehouse', 'Test1', 'Cassandra Pipeline_to_Imanis', and 'CASS_Backup_Now'.

7. Select the radio button of the jobtag data object and click the **Go to Revisions** button.



- Select **Keyspace**, type the keyspace name, and then click the **Go To Revisions** button
- Select **Table**, type the table name, and then click the **Go To Revisions** button

NOTE: A search that is based on an exact or a partial term is enabled for Tables and Keyspaces only.

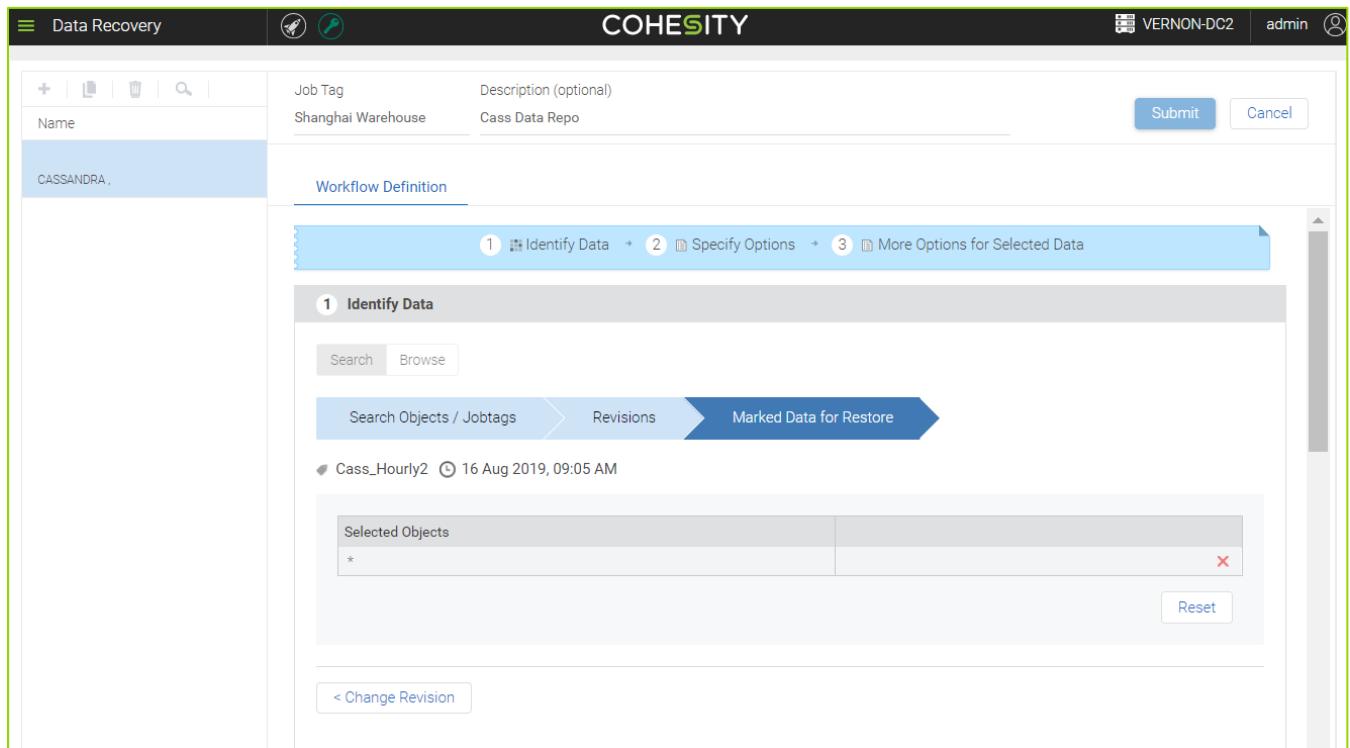
8. In the **Revisions** tab, click the **Browse Revisions**, do one of the following:

- By default, the latest copy is selected which is indicated by the icon. You can then click the **Next** button below to restore the selected data object
- Click the data object icon to select a copy of data for a specific day and time. You can then click the **Next** button to restore the selected data object

NOTE: Navigate all the data object revision by clicking the icons. You can also click the icon to jump to a specific revision in time by selecting a date and time or click the icon to jump to the currently selected revision.

9. In the **Marked Data for Restore** tab, do one of the following:

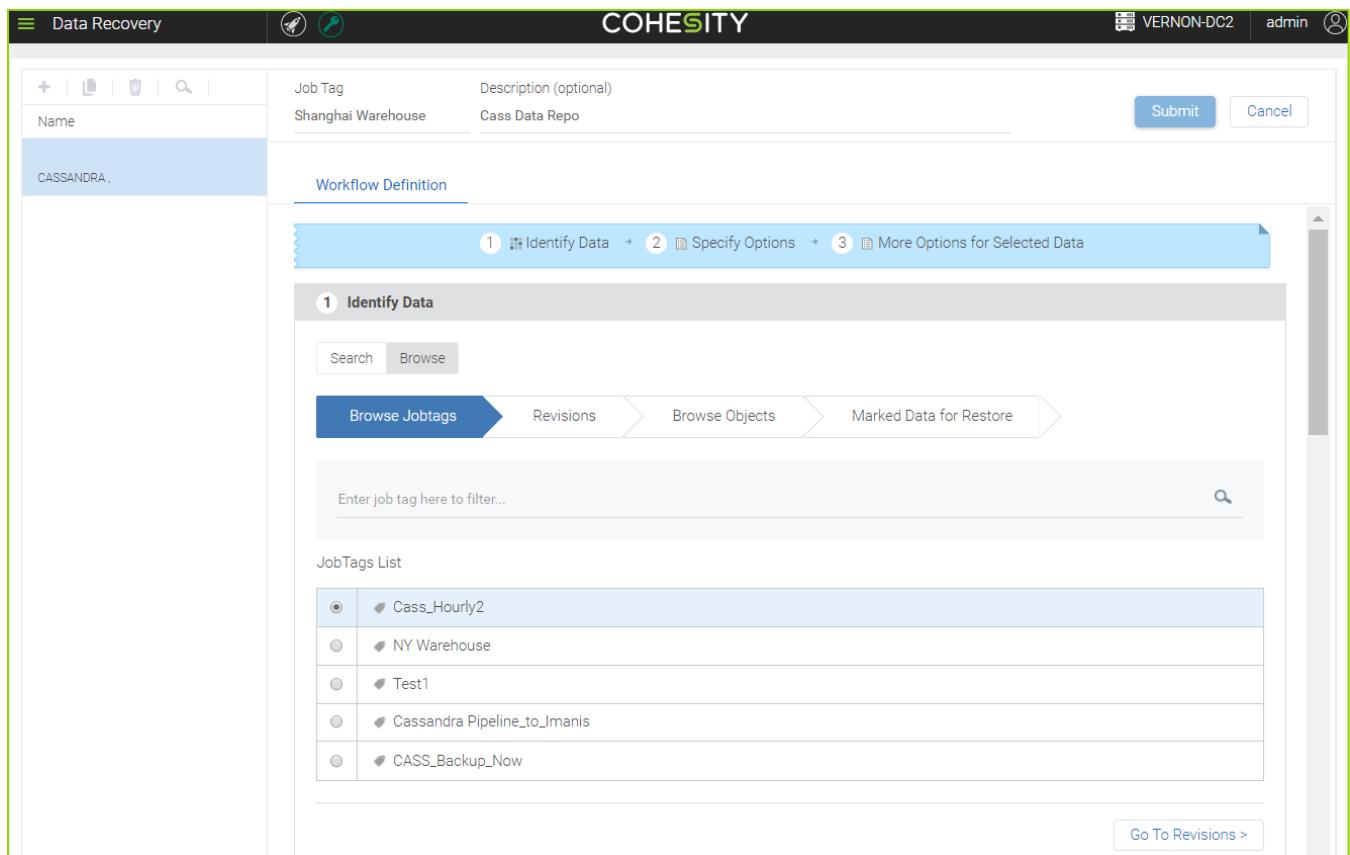
- Click the **Reset** button to go back to the Search Objects/Jobtags tab
- Click the **Change Revision** button to back to Revisions tab to reselect a data object revision to restore`



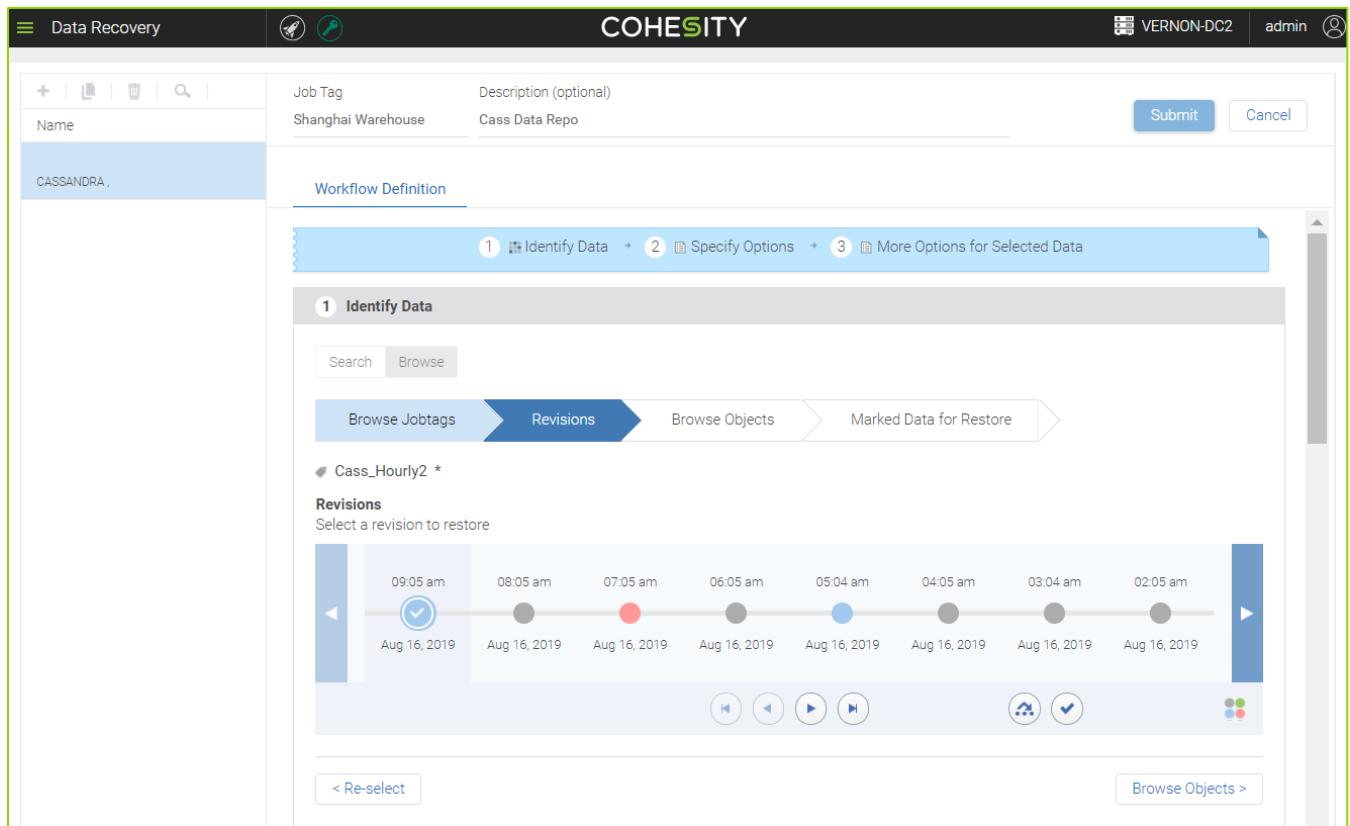
You can jump directly onto Step # 11 to continue Cassandra data recovery for JobTag.

Browse Tab

7. In the **Browse Jobtags** tab, select a **JobTag** from the **JobTag list**, and then click the **Go To Revisions** button.

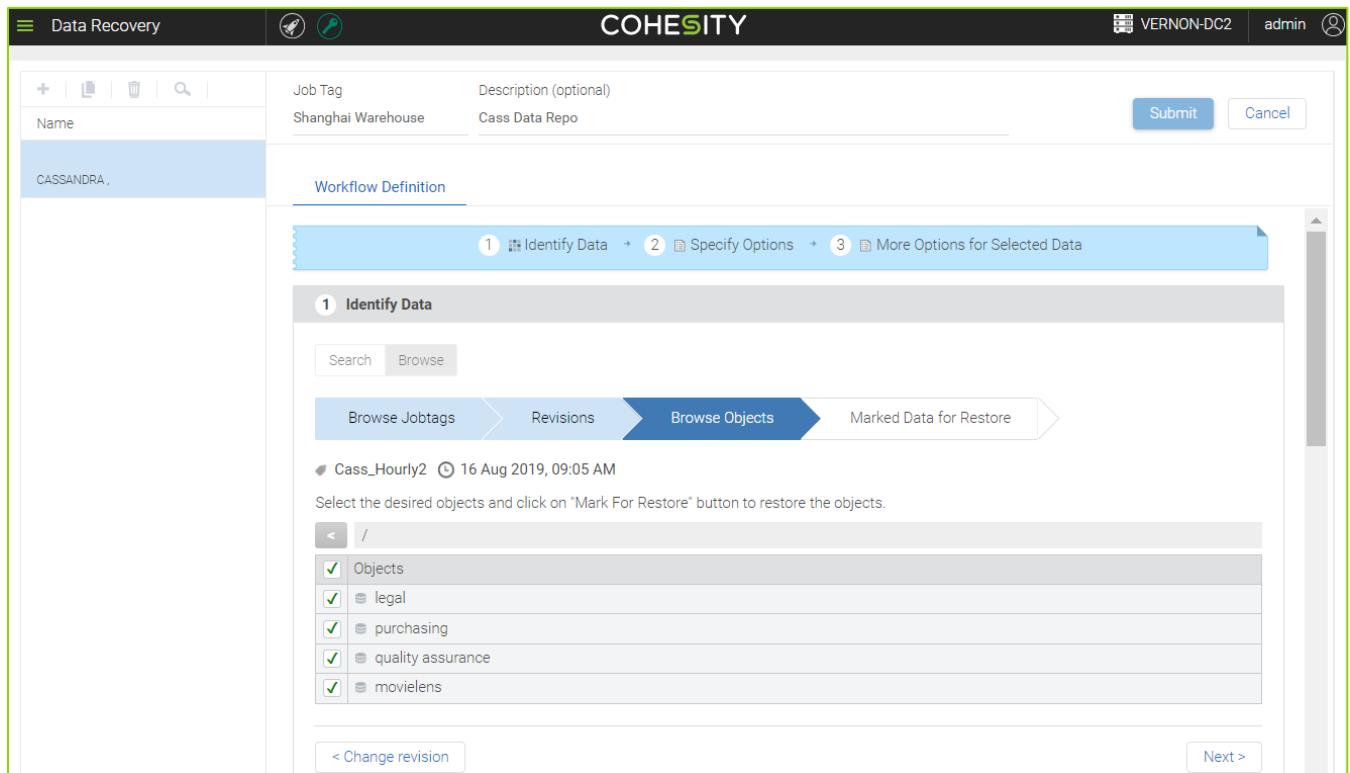


8. In the **Revisions** label, select a revision of the **JobTag** revision that you want to restore and then click the **Browse Objects** button.



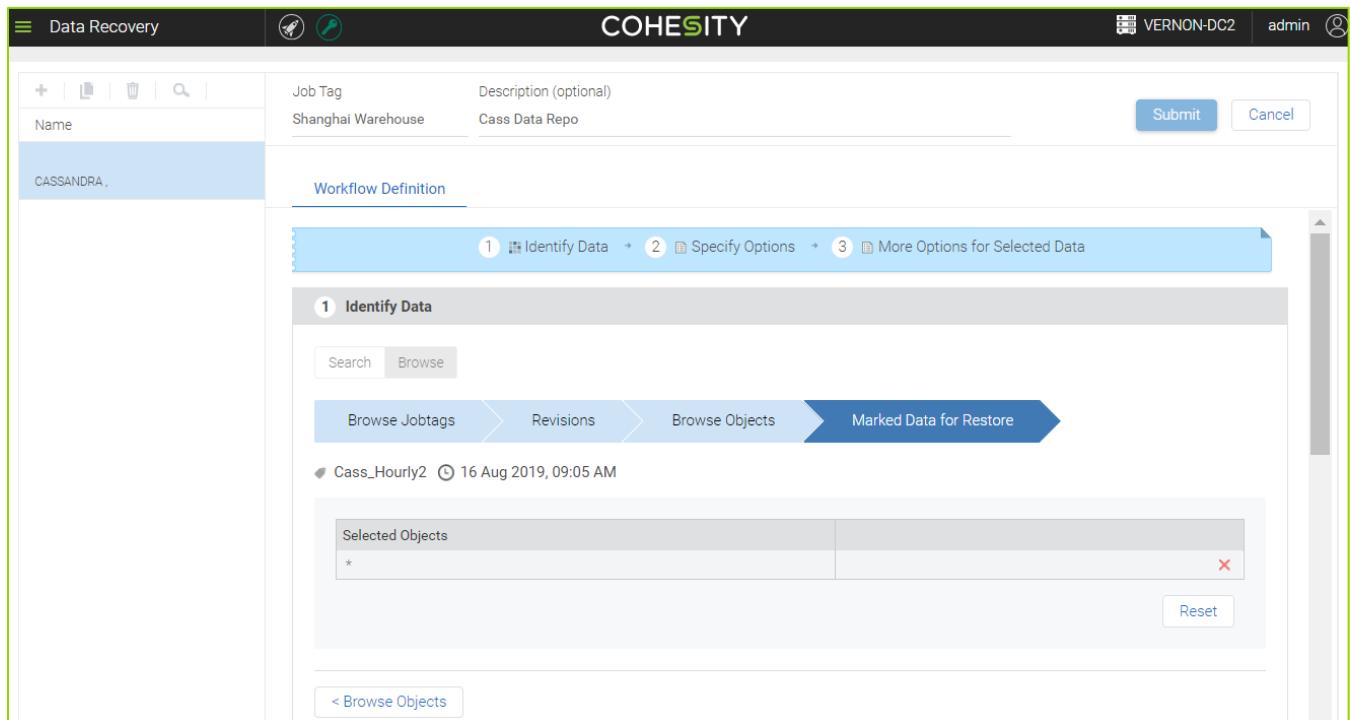
9. In the **Browse Objects** label, do one of the following:

- Select keyspace or tables that you want to restore and then click the **Next** button
- Click the **Change revision** button to go back to the Revisions tabs and select a new revision of the JobTag



10. In the **Marked Data for Restore** label, do one of the following:

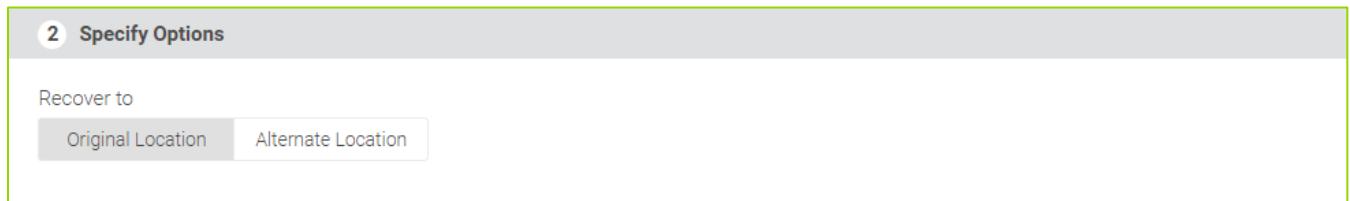
- Click the Reset button to go back to the Browse Jobtags tab
- Click the Browse Objects button to go back to the Revisions tab to reselect a data object revision to restore



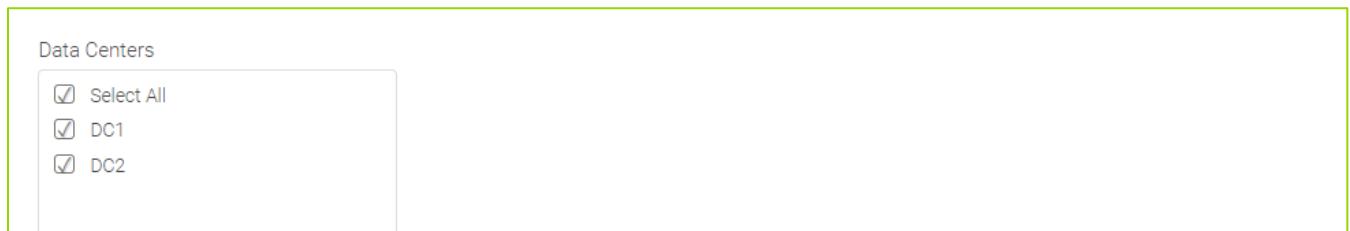
11. In the **Specify Options** section, under the **Recover To** area, select **Original Location** or **Alternate Location**.

ORIGINAL LOCATION:

1. Click the **Original Location** button.



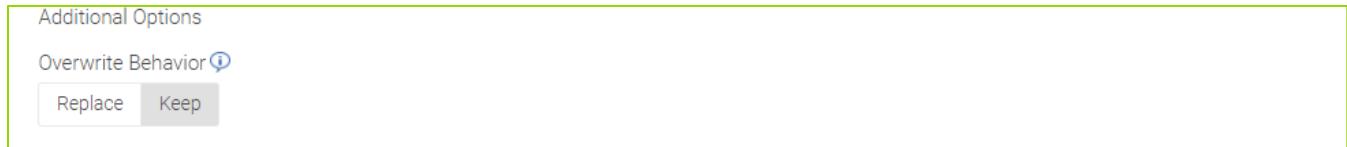
2. In the **Data Centers** area, select a data center or data centers where you want to recover the data.



NOTE: Imanis Data software supports data recovery to a specific data center only if the keyspace spans across data centers, that is, if the keyspace is created using replication strategy as 'NetworkTopologyStrategy'. However, if a specific data center for a keyspace created using non-NetworkTopologyStrategy is selected, then data is recovered automatically from all the data centers. For example, you have data centers 'D1', 'D2', and 'D3'. You select 'D1' for a keyspace created using SimpleStrategy, then data on 'D2' and 'D3' is also recovered.

3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

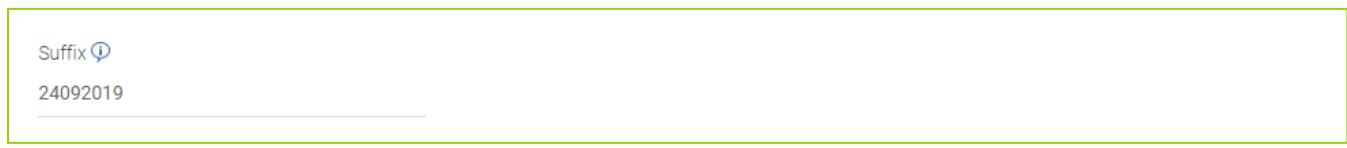
- Click **Replace** to replace existing data with existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if data objects are already present on the destination the recovery job would fail



Additional Options

Overwrite Behavior ⓘ

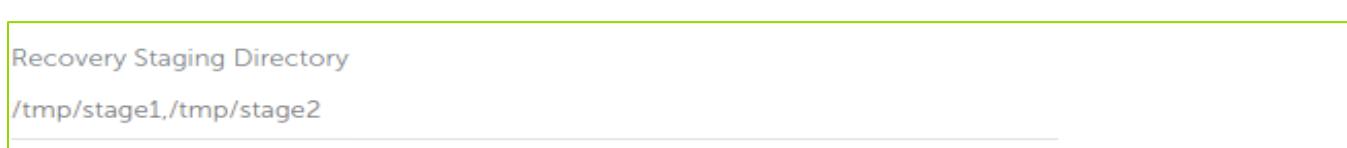
4. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _26102017.



Suffix ⓘ

24092019

5. For DSE 6.X, for **Recovery Staging Directory**, you can mention a temporary directory if you do not wish Imanis to use the Cassandra storage for staging. The Recovery Staging Directory field accepts a single directory or a comma separated list of directories. Ensure that the directories are present on all the nodes before executing the restore job.

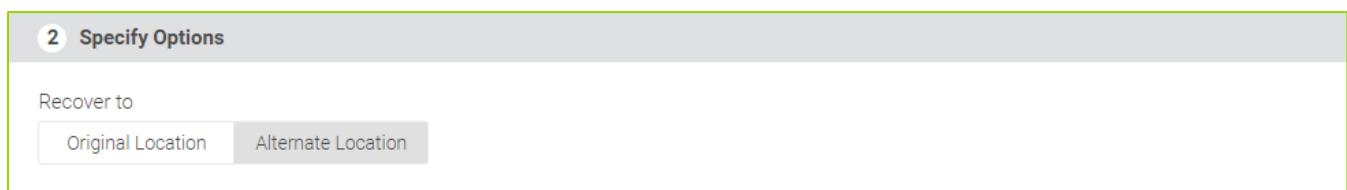


Recovery Staging Directory

/tmp/stage1,/tmp/stage2

ALTERNATE LOCATION:

1. Click the **Alternate Location** button.



2 Specify Options

Recover to

2. Select a Cassandra cluster name from the **Data Repository** drop-down menu.

Data Repository

Cassandra Dev

3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied

Additional Options

Overwrite Behavior 

Replace Keep

4. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

Suffix 

24092019

5. For DSE 6.X, for **Recovery Staging Directory**, you can mention a temporary directory if you do not wish Imanis to use the Cassandra storage for staging. The Recovery Staging Directory field accepts a single directory or a comma separated list of directories. Ensure that the directories are present on all the nodes before executing the restore job.

Recovery Staging Directory

/tmp/stage1,/tmp/stage2

NOTE: Imanis Data software supports the use of alphanumeric characters and uppercase in the suffix field.

12. In the **More Options for Selected Data** section, do the following:

3 More Options for Selected Data

Objects	Recover As	With Properties
sales.customers	sales.customers_example	+

- a. To rename restored objects, type the new name in the **Recover As** column.
- b. To change property of the restored object, click the  icon, and type the values in the Key and Value field. The key is auto-completed by the UI. The Value field must contain the complete value of the property. Only the following property changes are allowed: keyspace (only replication is allowed) and table (only compression and compaction is allowed).

For example, let's assume that a user wants to change replication for a source keyspace wherein the create query for the keyspace is as follows:

```
CREATE KEYSPACE tutorialspoint WITH replication = {'class':'SimpleStrategy', 'replication_factor': 3};
```

To change the replication factor to 1, the key value pair would be as follows:

KEY:replication

VALUE: {'class':'SimpleStrategy', 'replication_factor' : 1}

NOTE: The **More Options for Selected Data** field is supported for Keyspace and Table levels only.

NOTE: In Cassandra, incremental recovery partially fails if roles for which permissions to be set do not exist on the destination cluster. In this case, you have to manually restore all the permissions for roles on the destination cluster.

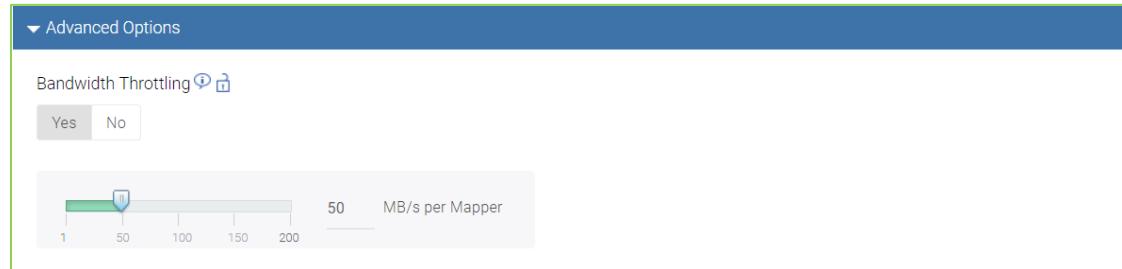
IMPORTANT: If the recovery process is executed to an '**Alternate Location**', Imanis Data software always uses 'SimpleStrategy' as the replication strategy unless it is overridden through the "**More Options for Selected Data**" section in the UI.

NOTE: In the current release, Imanis Data software does not support changing of compaction strategy to CFSSCompactionStrategy from any other strategy. For example, Imanis Data software does not support changing of compaction strategy from Leveled Compaction Strategy, Size Tiered Compaction Strategy or any other strategy to CFSSCompactionStrategy.

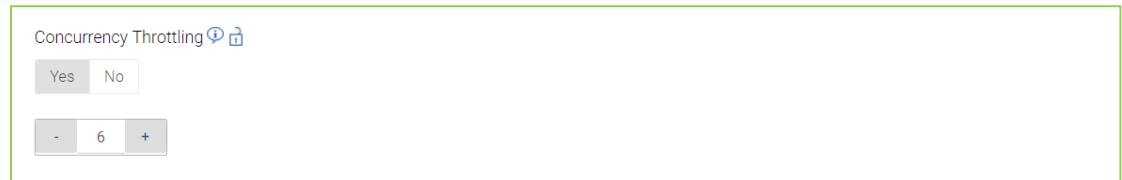
NOTE: While you are recovering Solr data to a Multi-DC setup, ensure that the replication strategy assigns at least one replica to the Solr DC.

13. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

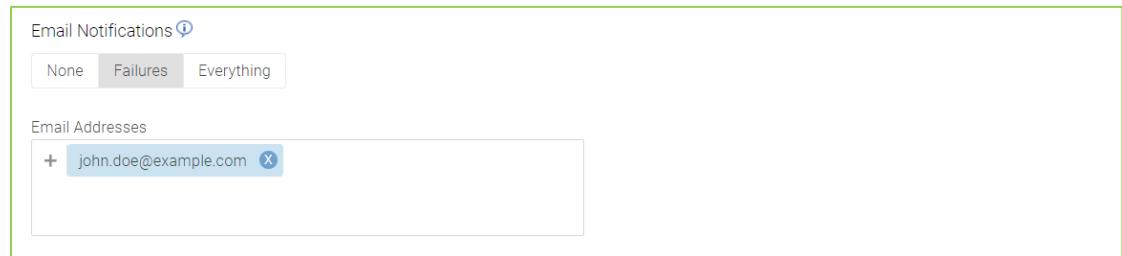
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



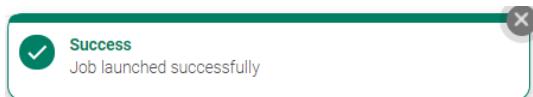
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



IMPORTANT: In the current release, recovery is not supported for tables containing ';' in column name.

IMPORTANT: Ensure that the system partitioner – a configurable parameter in the `cassandra.yaml` file that determines how data is distributed across the nodes in the cluster – at the source cluster matches with the system partitioner that is configured at the destination cluster. If you do not configure the system partitioner at the destination cluster, the data restore process will fail.

IMPORTANT: As part of the recovery process, the keyspace/table schemas are created (if required) on the destination cluster. However, this schema creation may fail if a data center(s) is down in a multi-dc environment. In such cases, the Cassandra data recovery process will fail.

IMPORTANT: In case of incremental restore, if the table has multiple primary keys, renaming of primary keys is not supported in Imanis Data software. Renaming is supported if the table has only one primary key.

IMPORTANT: As part of the data recovery process, the keyspace/table schemas are created (if required) on the destination cluster. However, this schema creation may fail if a data center(s) is down in a multi-dc environment. In such cases, the Cassandra data recovery process will fail.

IMPORTANT: To do backup and recovery of DSE graph keyspace, it is advised to skip the 'keyspace_pvt' while executing the backup workflow. For example, if the name of the graph is `warehouse`, only '`warehouse`' and '`warehouse_system`' must be backed up. However, before you execute the recovery workflow, first manually create the graph metadata using the Gremlin Console or DSE Studio and set the Overwrite option to 'Yes'. Ensure that the restored keyspaces are configured to have the appropriate names as per the created graph. For example, on the destination cluster if the graph has been renamed to '`storehouse`' (previously known as '`warehouse`'), then the keyspaces related to '`warehouse`' must be appropriately renamed to '`storehouse`' and '`storehouse_system`' by using the Object Rename feature.

IMPORTANT: Imanis Data software does not support recovering tables and keyspaces from a new version to an older version in Cassandra. For example, recovering tables and keyspaces in Cassandra from version 2.1.2 to version 2.0.12 is not supported.

9.2.4.1 JobTag PIT Restore

Prior to starting the Point-in-Time (PIT) recovery for Cassandra, you must ensure that you have created commit log restore directory on every Cassandra node with Read-Write-Execute (RWX) permissions for the Imanis Data recovery user. At the time of Cassandra data repository creation process, the Archive log directory is needed.

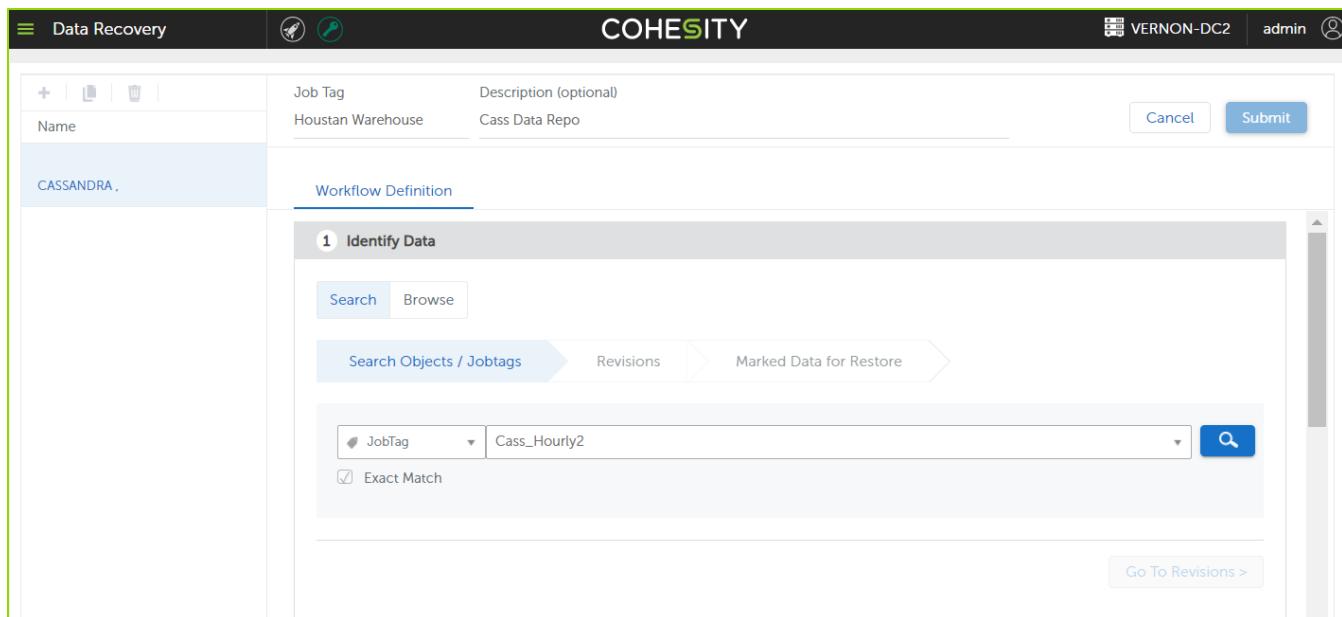
Similarly, you can select **Keyspace** or **Table** from the drop-down list for the Point-in-Time (PIT) recovery functionality. The steps are similar as mentioned in the following paragraph.

NOTE: Point-in-Time Recovery (PITR) feature is supported for restore to original location only. Also, the object renaming is not supported in PITR.

For more information, refer to the Cassandra Pre-requisites section. Once you have created the directories you can start the PIT recovery workflow.

To start Point-in-Time recovery workflow, do the following:

1. Click the **Main Menu**  > **Monitoring and Recovering** > **Data Recovery**.
2. On the **Data Recovery** page, click the  **+ Add New** button or the  icon. The **New Data Recovery Workflow** dialog box appears.
3. In the **New Data Recovery Workflow** dialog box, select a **Cassandra** data repository from the **Data Repository** drop-down menu, and then click **OK**.
4. Type a new job tag in the **JobTag** field and a job tag description in the **Description** field.
5. In the **Identify Data** section, click the **Search** button to access the PIT recovery menu.
6. In **Search Objects/Jobtags** label, click the search box, select a **JobTag** displayed by Imanis Data, and then click the  icon. JobTag search result is displayed.



Similarly, you can select **Keyspace** and **Table** from the drop-down list by typing the full or partial file name.

NOTE: A search that is based on an exact or a partial term is enabled for Tables and Keyspaces only.

7. Click the radio button of the JobTag search result and then click the **Go To Revisions** button.

The screenshot shows the 'Data Recovery' interface. In the top navigation bar, 'Data Recovery' is selected. The main area is titled 'COHESITY'. On the left, there's a sidebar with icons for Data Recovery, Protection, and Monitoring. The main content area has tabs for 'Workflow Definition' and 'Identify Data'. Under 'Identify Data', there are 'Search' and 'Browse' buttons. Below them is a search bar with 'JobTag' selected and 'Cass_Hourly2' typed in. There's also a checkbox for 'Exact Match'. The results section shows 'Showing 1 results.' with a list item 'Cass_Hourly2'. At the bottom right of the results area is a blue button labeled 'Go To Revisions >'. The top right of the interface shows the user 'admin' and the host 'VERNON-DC2'.

8. In the **Revisions** label, click the **Select PIT Revisions** button.

The screenshot shows the 'Revisions' step of the process. It features a search bar with 'Cass_Hourly2 *' and two buttons at the bottom: 'Browse Revisions' and 'Select PIT Revisions'. The 'Select PIT Revisions' button is highlighted with a yellow background, indicating it is the active or next step. The overall interface is clean and modern, with a light gray background and white text.

9. Set date and time in the **Select the data and time for PIT revision**. The PIT time that you set here must be Imanis Data time and not GMT format. Once set, click the **Go** button:

Select the date and time for PIT revision

2019-10-03

4:33 PM

Go

[< Re-select](#)

[Next >](#)

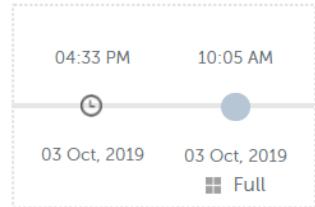
10. Verify the objects that are going to be recovered in the **Following objects will be restored** field:

Select the date and time for PIT revision

2019-10-03

4:33 PM

Go



Following objects will be restored

legal.reports,purchasing.procurements_2018,purchasing.procurements_2019
,quality
assurance.qa_reports,sales.customers,sales.orders,sales.transactions,sales.sal

11. Type the path where Cassandra commit log will be restored in the **Commitlog Restore Path** field and then click **Next**. Refer to the Cassandra Pre-requisites section for more information.

Commitlog Restore Path

/mnt/disk1/dse/commitlog

[< Re-select](#)

[Next >](#)

NOTE: Metadata operations cannot be replayed using **Commitlog** Restore. For instance, PITR restore does not recover any additional metadata changes with respect to the metadata state recorded in the selected snapshot.

12. In the **Marked Data for Restore** label, verify the selected objects and PIT date and time and go to the next step or click **Reset** button to start again.

1 Identify Data

Search Browse

Search Objects / Jobtags Revisions Marked Data for Restore

Cass_Hourly2 03 Oct 2019, 04:33 PM

Commitlog Restore Path : /mnt/disk1/dse/commitlog

Selected Objects

* Point-In-Time Copy

legal.reports,purchasing.procurements_2018,purchasing.procurements_2019,quality assurance.qa_reports,sales.customers,sales.orders,sales.transactions

Reset

< Change Revision

This screenshot shows the 'Identify Data' step of a restore process. At the top, there are 'Search' and 'Browse' buttons. Below them is a navigation bar with 'Search Objects / Jobtags', 'Revisions', and 'Marked Data for Restore'. A timestamp '03 Oct 2019, 04:33 PM' and a job name 'Cass_Hourly2' are displayed. The 'Commitlog Restore Path' is set to '/mnt/disk1/dse/commitlog'. The main area is titled 'Selected Objects' and contains a list with a single item: '* Point-In-Time Copy'. Underneath this item is a list of selected objects: 'legal.reports,purchasing.procurements_2018,purchasing.procurements_2019,quality assurance.qa_reports,sales.customers,sales.orders,sales.transactions'. There are scroll bars and a red 'X' button in the list area. A 'Reset' button is located at the bottom right of the list area.

13. In the **Specify Options** section, under the **Recover To** area, the **Original Location** option is already selected.

2 Specify Options

Recover to

Original Location Alternate Location

This screenshot shows the 'Specify Options' section. It features a 'Recover to' section with two buttons: 'Original Location' (which is highlighted in blue) and 'Alternate Location'. The background of the entire section is light gray.

14. In the **Data Centers** area, select a data center or data centers where you want to recover the data.

Data Centers

Select All
 DC1
 DC2

This screenshot shows the 'Data Centers' selection area. It contains a list of three items: 'Select All', 'DC1', and 'DC2', each preceded by a checkbox. The first checkbox is checked. The background of the entire section is light gray.

15. In the **Additional Options** area, the **Overwrite Behavior** option is automatically set to **Replace** and the **Suffix** option is disabled for PITR.

16. In the **More Options for Selected Data** section, do the following:

- a. Renaming the object is not supported in Cassandra PIT recovery
- b. To change property of the restored object, click the  icon, and type the values in the Key and Value field. The key is auto-completed by the UI. The Value field must contain the complete value of the property. Only the following property changes are allowed: keyspace (only replication is allowed) and table (only compression and compaction is allowed).

For example, let's assume that a user wants to change replication for a source keyspace wherein the create query for the keyspace is as follows:

```
CREATE KEYSPACE tutorialspoint WITH replication = {'class':'SimpleStrategy', 'replication_factor' : 3};
```

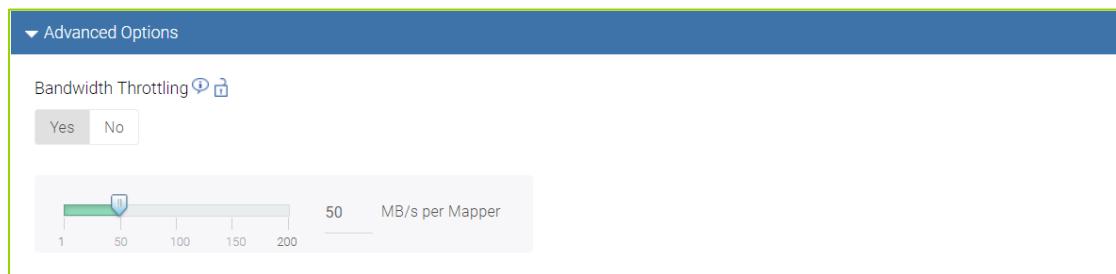
To change the replication factor to 1, the key value pair would be as follows:

KEY:replication

VALUE: {'class':'SimpleStrategy', 'replication_factor' : 1}

17. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:

Concurrency Throttling   Yes No

- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the Email Addresses field that will receive the job status notifications:

Email Notifications  None Failures Everything

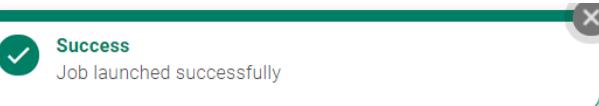
Email Addresses

john.doe@example.com 

IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. Imanis Data software will then start executing your workflow. A confirmation message will be displayed on the page indicating the job is successfully launched. You can monitor the progression of the job on the Dashboard while its running.



- Make sure that this job is completed successfully by verifying the status on Dashboard.
- Log on to Cassandra Primary cluster, access commitlog_archiving.properties file on each Cassandra node and make the following changes:

- Command to execute to make an archived commitlog live again. For example:

```
restore_command=/bin/cp -f %from %to
```

- Directory to scan the recovery files in. Set this path to Commitlog Restore Path Directory. For example:

```
restore_directories=/mnt/disk1/dse/commitlogresdir/
```

- Restore mutations created up to and including this timestamp in GMT. For example:

```
restore_point_in_time=2019:03:18 07:31:00
```

The date and time set must be equal to the Imanis Data PIT date and Time (in GMT) that was mentioned during PIT revision selection.

21. Access the `cassandra-env.sh` on every Cassandra node and add below property:

```
JVM_OPTS="$JVM_OPTS -Dcassandra.replayList=casst4.tab1"
```

You can specify list of comma separated tables on which PITR is to be performed. The names of tables should be in `keyspacename.tablename` format.

22. Restart Cassandra primary cluster.

23. Once the PITR recovery process is completed, remove the changes made so that future Cassandra restarts are not affected by these settings.

- From `commitlog_archiving.properties`, remove the following changes:

```
restore_command=/bin/cp -f %from %to
restore_directories=/mnt/disk1/dse/commitlogresdir/
restore_point_in_time=2019:03:18 07:31:00
```

- From `cassandra-env.sh`, remove the following changes:

```
JVM_OPTS="$JVM_OPTS -Dcassandra.replayList=casst4.tab1"
```

24. Delete the data present in the `commitlogresdir`.

9.2.5 Recovering Data for Couchbase

Imanis Data software supports Couchbase recovery at the Jobtag and Bucket level. In the latest release, you can also select the date and time for a specific Point in Time (PIT) revision copy of the data and recover it.

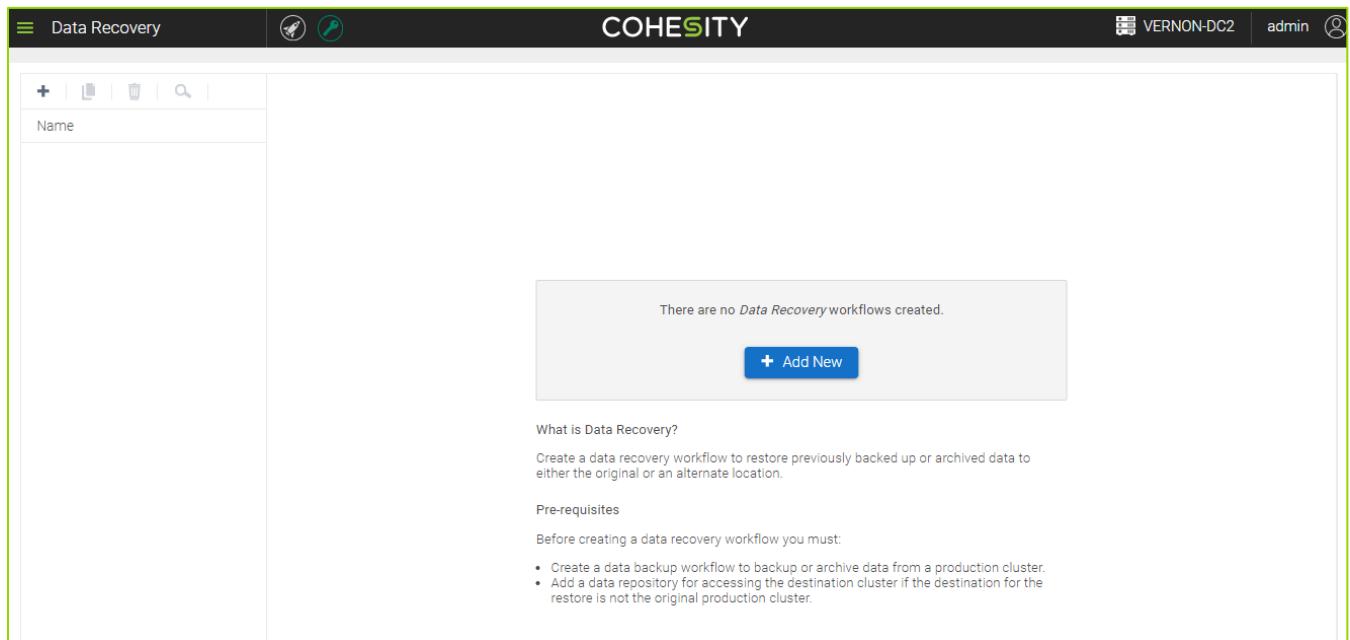
To know how to do Point in Time (PIT) restores for Jobtag and Bucket, click JobTag PITR and Bucket PITR. Also, refer to Appendix E to know limitation of Couchbase PIT Recovery.

9.2.5.1 JobTag Restore

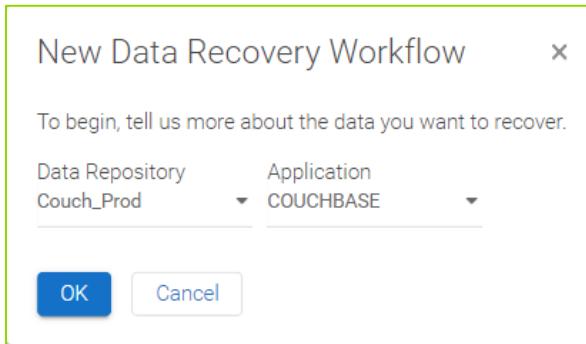
The following section discusses how to recover a data objects for a Couchbase JobTag.

To start a recovery workflow for Couchbase, do the following:

1. Click the **Main Menu**  > **Monitoring and Recovering** > **Data Recovery**.
2. On the Data Recovery page, click the  **Add New** button or the  icon. The **New Data Recovery Workflow** dialog appears.



- In the **New Data Recovery Workflow** dialog, select a **Couchbase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.

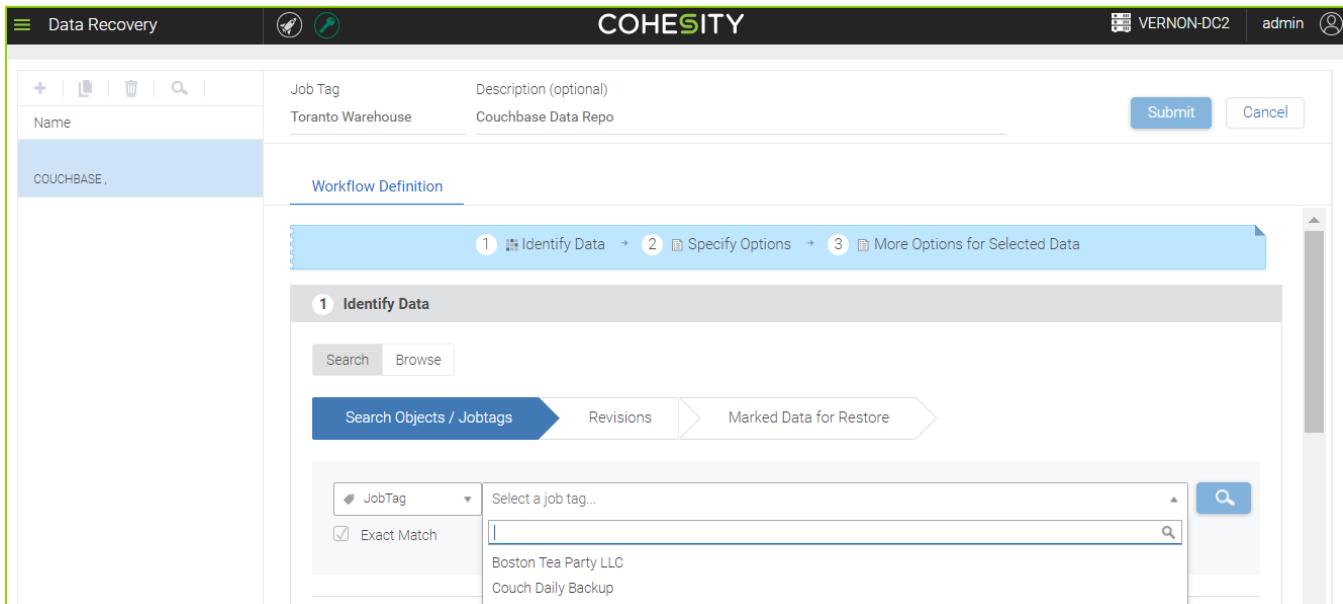


- Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field.
- In the **Identify Data** section, **Search** and **Browse** tabs are displayed. You can use the Search and Browse button as per your requirement:

SEARCH	BROWSE
Use when you know the data object that you wish to recover	Use when you want to view data objects and select specific data objects within a JobTag revision
Search specific Jobtag, and Bucket	Browse the data objects catalog and select or deselect multiple data objects at a single time

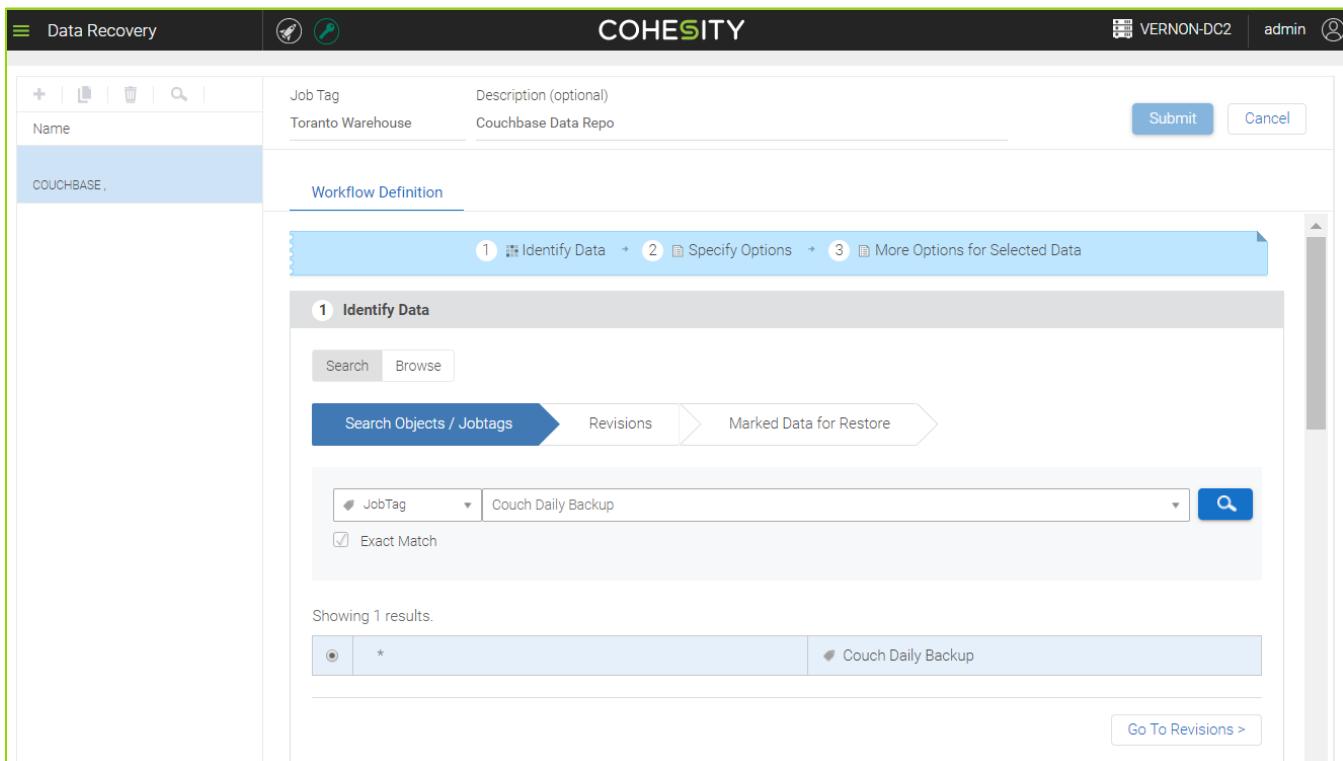
Search Tab

6. In the **Search Objects/Jobtags** tab, click the search box and select a job tag displayed by Imanis Data, and then click the  icon. Imanis Data will display the jobtag data object as a search result.



The screenshot shows the Cohesity Data Recovery interface. On the left, there's a sidebar with icons for Data Recovery, Job Tags, and a magnifying glass. The main header says "COHESITY". On the right, it shows "VERNON-DC2" and "admin". Below the header, there's a "Data Recovery" tab and a "Cancel" button. The main area has a "Name" input field containing "COUCHBASE," and a "Job Tag" input field with "Toronto Warehouse" and "Description (optional)" "Couchbase Data Repo". There are "Submit" and "Cancel" buttons. Under "Workflow Definition", there's a "1 Identify Data" step with "Search" and "Browse" buttons. A progress bar shows steps 1, 2, and 3. Below the progress bar, there's a search interface with a dropdown set to "JobTag" and "Exact Match" checked. The search term "Boston Tea Party LLC" is entered, and the results list shows "Boston Tea Party LLC" and "Couch Daily Backup".

7. Select the radio button of the jobtag data object and click the **Go to Revisions** button.



The screenshot shows the same interface as the previous one, but the search results have changed. The search term "Boston Tea Party LLC" is no longer present; instead, "Couch Daily Backup" is selected in the dropdown. The results list now only shows "Couch Daily Backup". At the bottom right, there's a "Go To Revisions >" button.

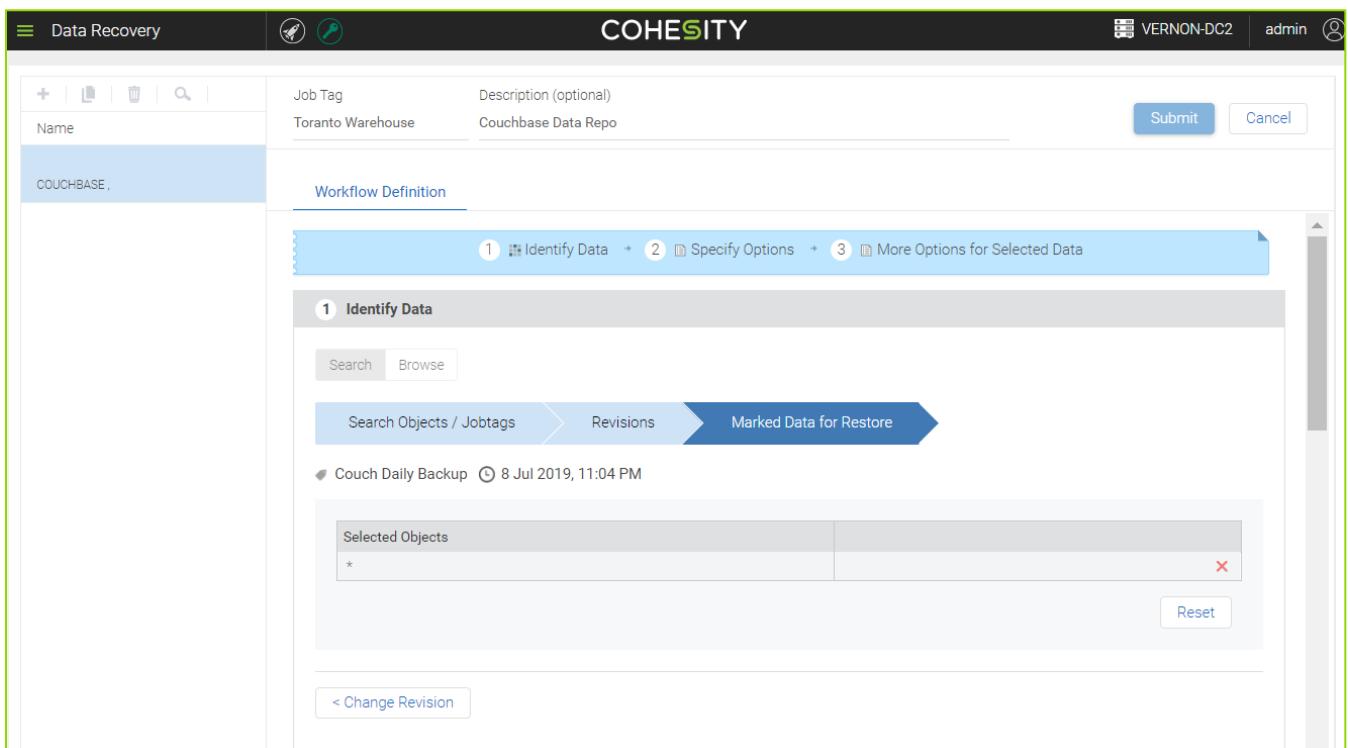
8. In the **Revisions** tab, click the **Browse Revisions**, do one of the following:

- By default, the latest copy is selected which is indicated by the  icon. You can then click the Next button below to restore the selected data object
- Click the data object  icon to select a copy of data for a specific day and time. You can then click the Next button to restore the selected data object

NOTE: Navigate all the data object revision by clicking the , , ,  icons. You can also click the  icon to jump to a specific revision in time by selecting a date and time or click the  icon to jump to the currently selected revision.

9. In the **Marked Data for Restore** tab, do one of the following:

- Click the **Reset** button to go back to the Search Objects/Jobtags tab
- Click the **Change Revision** button to back to Revisions tab to reselect a data object revision to restore



The screenshot shows the Cohesity Data Recovery interface. The top navigation bar includes 'Data Recovery', a search bar, and user information ('VERNON-DC2 | admin'). The main area is titled 'COHESITY'. On the left, a sidebar lists 'COUCHBASE...' under 'Workflow Definition'. The central workspace displays a workflow step titled '1 Identify Data'. It features tabs for 'Search' and 'Browse', and a breadcrumb navigation: 'Search Objects / Jobtags' → 'Revisions' → 'Marked Data for Restore'. Below this, a backup entry for 'Couch Daily Backup' from '8 Jul 2019, 11:04 PM' is shown. A table titled 'Selected Objects' contains a single row with an asterisk (*). A 'Reset' button is located at the bottom right of the step panel.

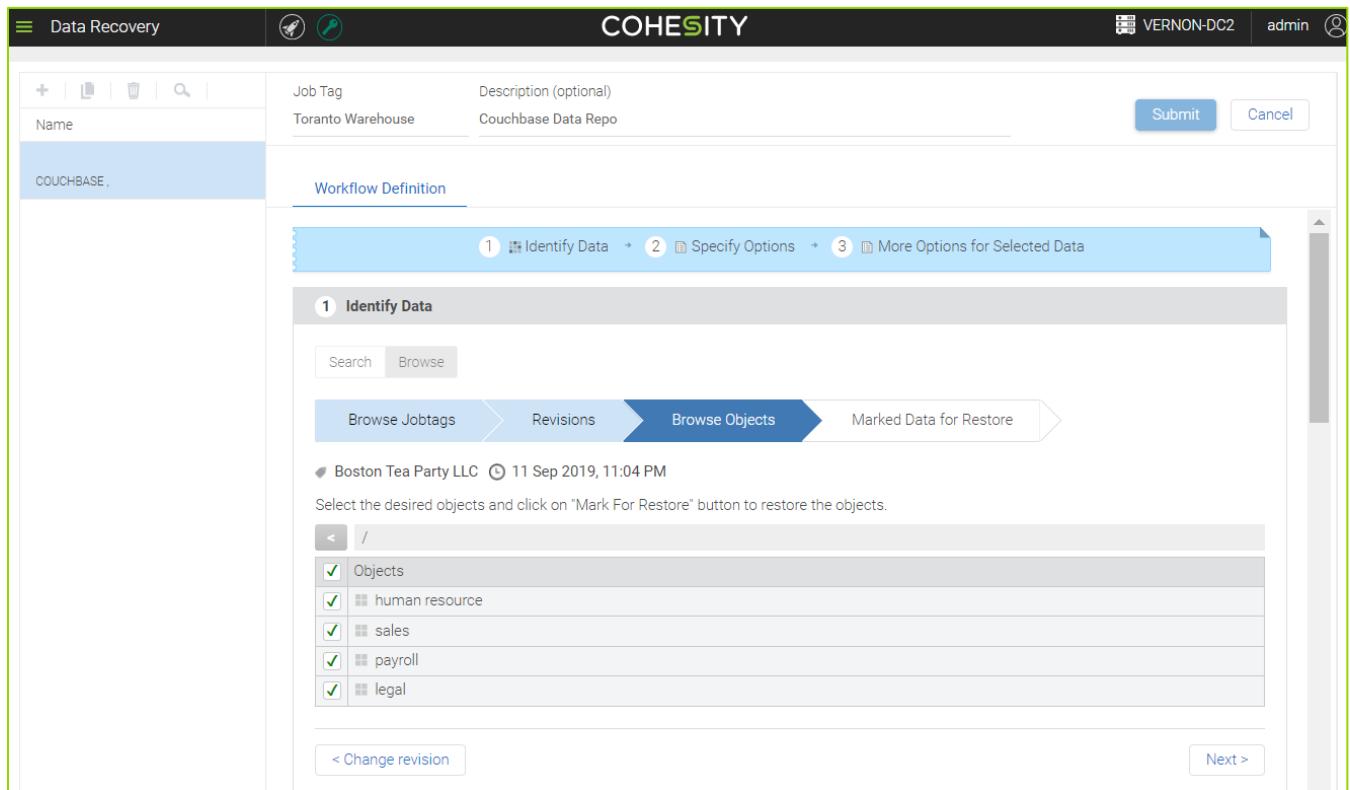
You can jump directly onto Step # 11 to continue Couchbase data recovery for JobTag.

Browse Tab

7. In the **Browse Jobtags** tab, select a **JobTag** from the JobTag list, and then click the **Go To Revisions** button.

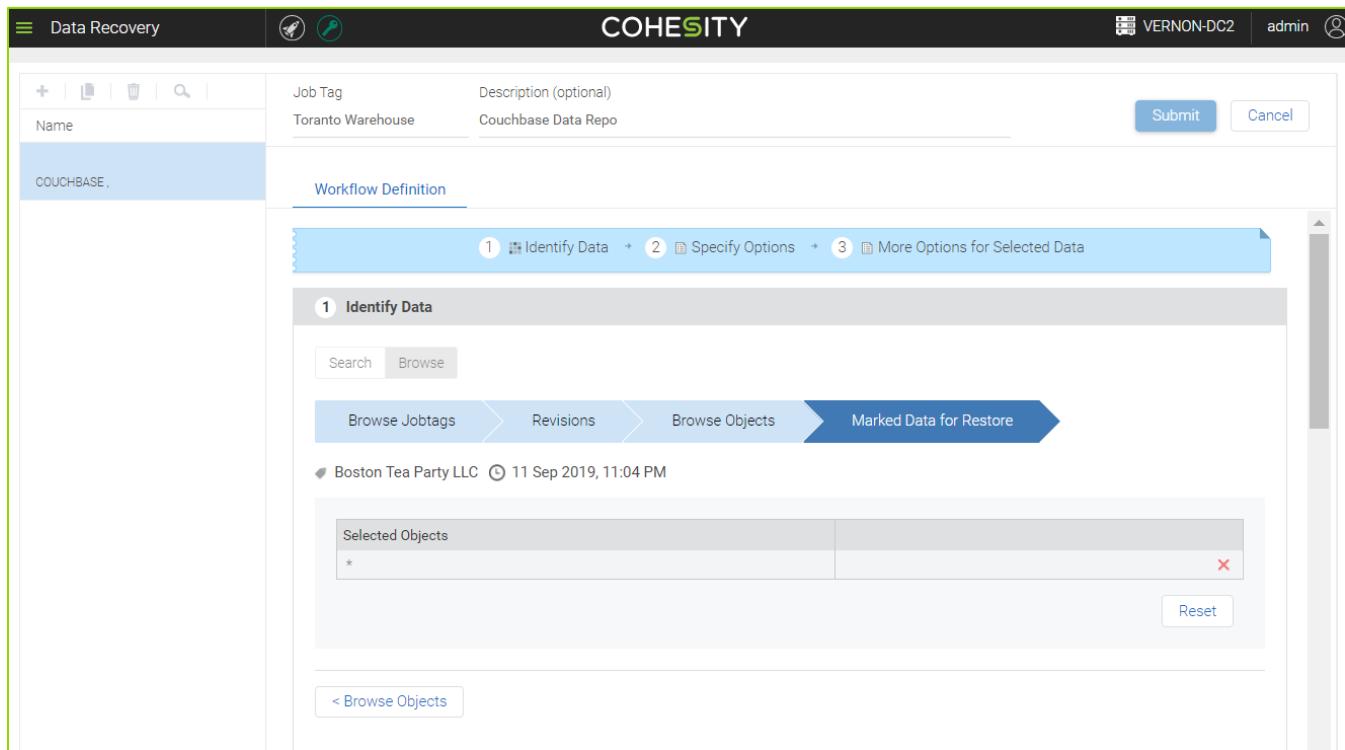
The screenshot shows the Cohesity Data Recovery interface with the 'Data Recovery' tab selected. In the center, there's a 'Workflow Definition' section with three steps: 1. Identify Data, 2. Specify Options, and 3. More Options for Selected Data. Step 1 is currently active. Below this, there's a 'Browse Jobtags' section with a search bar and a list of job tags: Boston Tea Party LLC and Couch Daily Backup. At the bottom right of the main area, there's a blue button labeled 'Go To Revisions >'. The top right corner of the interface shows the user 'admin' and the host 'VERNON-DC2'.

8. In the **Revisions** label, select a revision of the **JobTag** revision that you want to restore and then click the **Browse Objects** button.
9. In the **Browse Objects** label, do one of the following:
 - Select bucket you want to restore and then click the **Next** button
 - Click the **Change revision** button to go back to the **Revisions** tabs and select a new revision of the JobTag



10. In the **Marked Data for Restore** label, do one of the following:

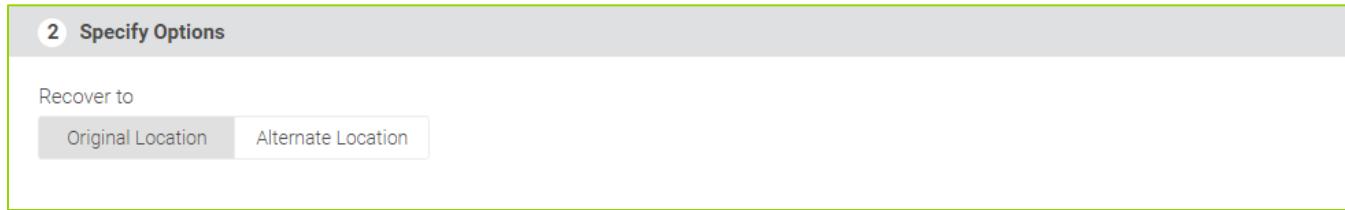
- Click the Reset button to go back to the Browse Jobtags tab
- Click the Browse Objects button to go back to the Revisions tab to reselect a data object revision to restore



11. In the **Specify Options** section, under the **Recover To** area, select **Original Location** or **Alternate Location**.

ORIGINAL LOCATION:

1. Click the **Original Location** button.



2. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:
 - Click **Replace** to replace existing data with existing data with new data thus erasing any previously existing data
 - Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
 - Click **Append** to add new data to an existing bucket

Additional Options

Overwrite Behavior 

3. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

Suffix 

29092019

ALTERNATE LOCATION:

1. Click the **Alternate** Location button.

Specify Options

Recover to

2. Select a Couchbase cluster name from the **Data Repository** drop-down menu.

Data Repository

Couch_Stage

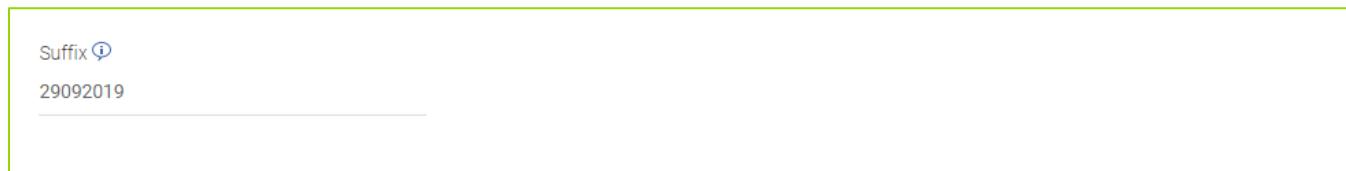
3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data:
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
- Click **Append** to add new data to an existing bucket

Additional Options

Overwrite Behavior 

4. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.



The screenshot shows a single-line input field labeled "Suffix" with a blue information icon. The value "29092019" is entered into the field. The entire input area is enclosed in a light gray border.

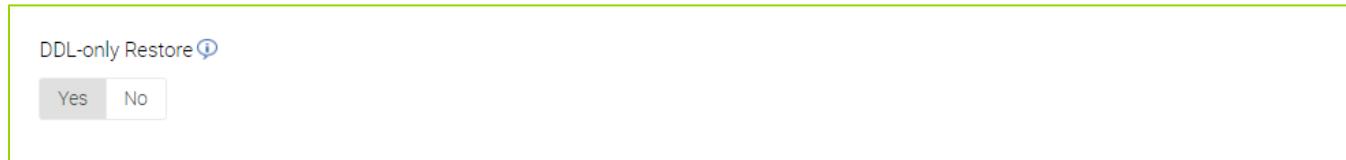
12. In the **More Options for Selected Data** section, edit the object name and rename it. The edited object takes precedence over the object name with suffix.
13. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Overwrite Users** option, click **Yes** to confirm overwriting of users with new users.



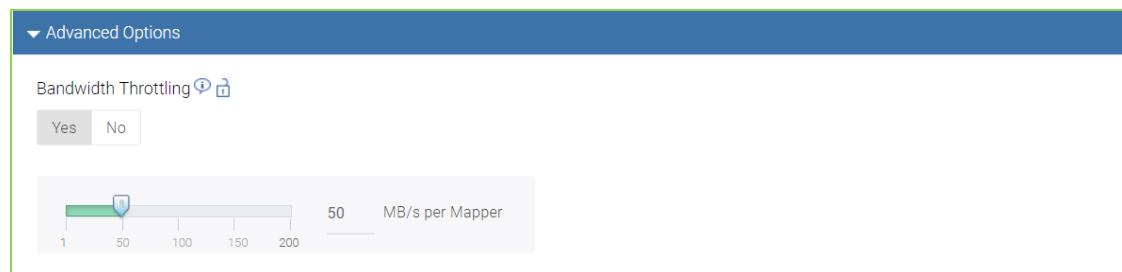
The screenshot shows a section titled "Advanced Options" with a blue header bar. Below it is a sub-section titled "Overwrite Users" with a blue information icon. There are two buttons: "Yes" (selected) and "No". The entire section is enclosed in a light gray border.

- In the **DDL-only Restore** option, click **Yes** to confirm to restore just the buckets. When you select this option, only the buckets will be restored and no the documents:



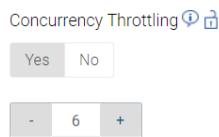
The screenshot shows a section titled "Advanced Options" with a blue header bar. Below it is a sub-section titled "DDL-only Restore" with a blue information icon. There are two buttons: "Yes" (selected) and "No". The entire section is enclosed in a light gray border.

- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:

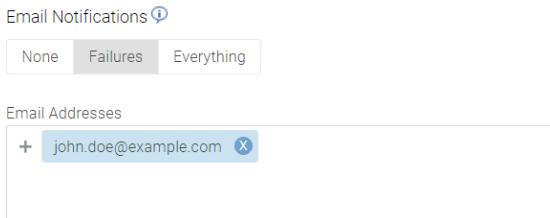


The screenshot shows a section titled "Advanced Options" with a blue header bar. Below it is a sub-section titled "Bandwidth Throttling" with a blue information icon. There are two buttons: "Yes" (selected) and "No". Below the buttons is a horizontal slider with a blue track and a green slider handle. The slider is positioned at the 50 mark. To the right of the slider, the text "50 MB/s per Mapper" is displayed. The entire section is enclosed in a light gray border.

- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



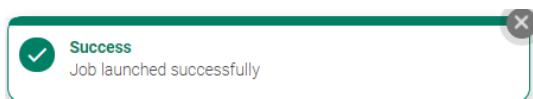
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.

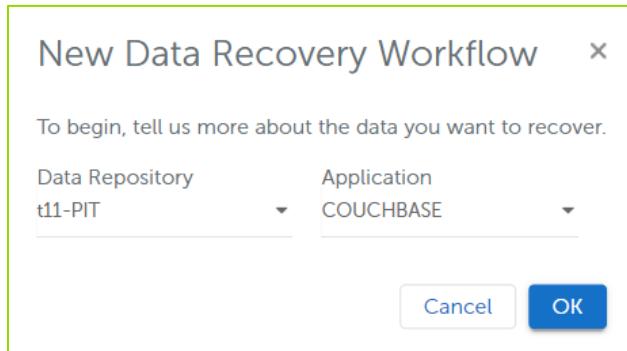


9.2.5.1.1 JobTag PIT Restore

The following section discusses how to Point-in-Time (PIT) recovery of data objects for a Couchbase JobTag.

To start a Point-in-Time (PIT) recovery workflow for Couchbase, do the following:

- Click the **Main Menu**  > **Monitoring and Recovering** > **Data Recovery**.
- On the **Data Recovery** page, click the  button or the  icon. The **New Data Recovery Workflow** dialog appears.
- In the **New Data Recovery Workflow** dialog, select a **Couchbase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.



4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field.
5. In the **Identify Data** section under the **Search** tab in the **Search Objects/Jobtags** tab, click the search box, select a JobTag displayed by Imanis Data, and then click the icon. Imanis Data will display the jobtag data object as a search result.

The screenshot shows the 'Data Recovery' interface. On the left, there's a sidebar with a tree view showing 'COUCHBASE' is selected. The main area has a 'Workflow Definition' header with a '1 Identify Data' step. Under 'Identify Data', there are tabs for 'Search' (which is selected) and 'Browse'. Below the tabs is a search bar with the placeholder 'Select a job tag...'. To the right of the search bar is a dropdown menu with 'JobTag' selected and 'Exact Match' checked. A search results list shows 'CouchPIT' and 'Couchbase_PITR'. At the top of the main area, there are fields for 'Job Tag' (set to 'Stockholm Warehouse') and 'Description (optional)' (set to 'Couchbase PITR'). On the far right, there are 'Cancel' and 'Submit' buttons.

6. Select the radio button of the jobtag data object and click the **Go to Revisions** button.

The screenshot shows the 'Workflow Definition' step in the Cohesity Data Recovery interface. The search bar is set to 'JobTag' and contains 'CouchPIT'. The results section shows one result: 'CouchPIT'. At the bottom right, there is a 'Go To Revisions >' button.

7. In the **Revisions** tab, click the **Select PIT Revisions**, do the following:

- Under the **Select the date and time for PIT revision** option, set a date and time to select a revision copy of PIT data of a specific day and time and then click **Go**.

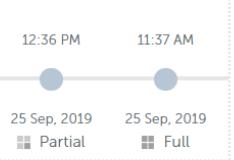
Select the date and time for PIT revision

2019-09-25 11:50 AM Go

- Select one of the PIT revisions copy of the data which is nearest to the date and time that you set in the previous step and then click the **Next** button to restore the selected data object.

Found following nearest PIT revision(s). Select desired revision.

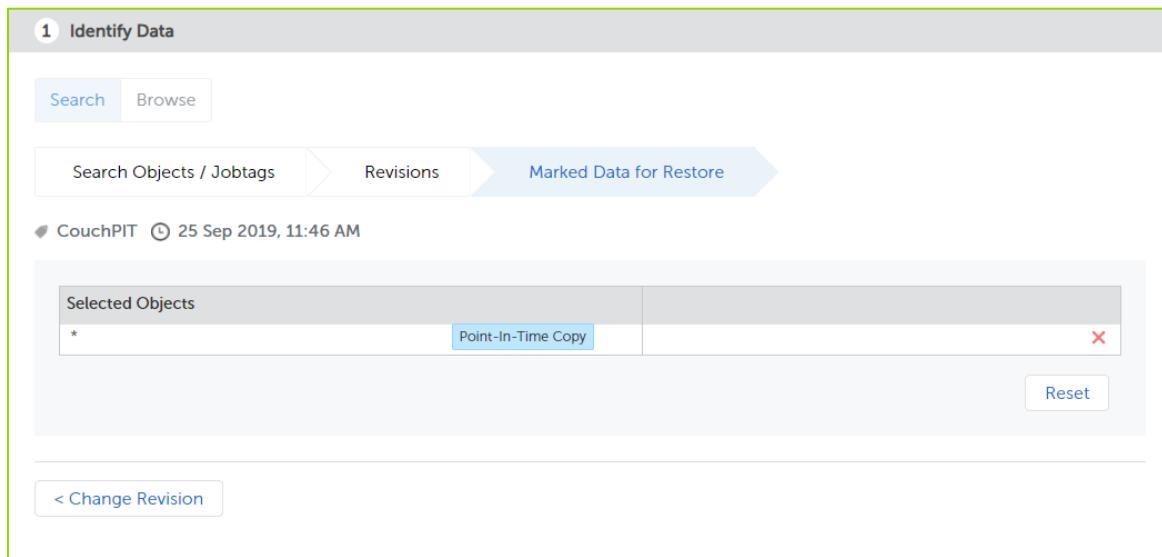
11:46 AM, 25 Sep 2019 12:01 PM, 25 Sep 2019



< Re-select

Next >

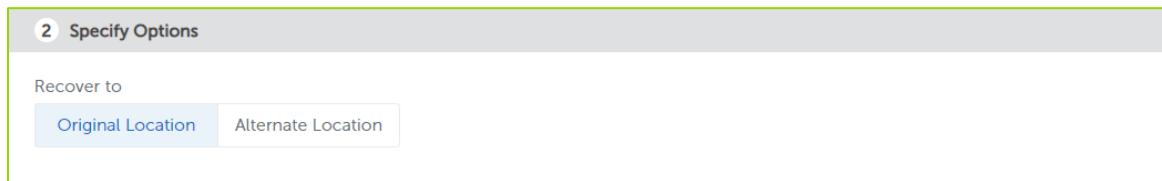
8. In the **Marked Data for Restore** tab, do one of the following:
- Click the **Reset** button to go back to the **Search Objects/Jobtags** tab
 - Click the **Change Revision** button to back to **Revisions** tab to reselect a data object revision to restore



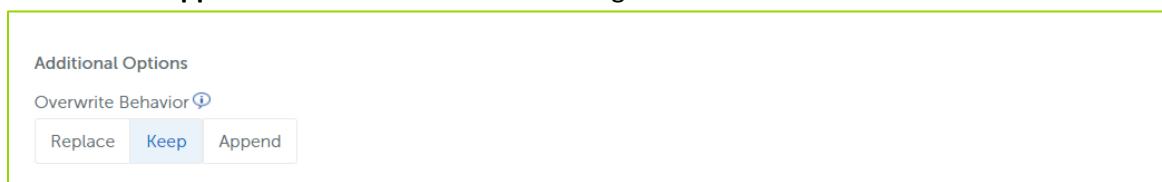
9. In the **Specify Options** section, under the **Recover To** area, select **Original Location** or **Alternate Location**.

ORIGINAL LOCATION:

- Click the **Original Location** button.



- In the **Additional Options** area, under **Overwrite Behavior** do one of the following:
- Click **Replace** to replace existing data with existing data with new data thus erasing any previously existing data
 - Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
 - Click **Append** to add new data to an existing bucket



3. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

A screenshot of a software interface showing a 'Suffix' input field. The field contains the text '_10102017'. Above the input field, there is a small blue information icon with a question mark. The entire input field is enclosed in a light gray border.

ALTERNATE LOCATION:

1. Click the **Alternate Location** button.

A screenshot of a software interface titled 'Specify Options'. Below the title, there is a section labeled 'Recover to' with two buttons: 'Original Location' and 'Alternate Location'. The 'Alternate Location' button is highlighted with a blue background and white text. The entire interface is enclosed in a light gray border.

2. Select a **Couchbase** cluster name from the **Data Repository** drop-down menu.

A screenshot of a software interface showing a 'Data Repository' dropdown menu. The menu is open and displays the option 'COUCH-Talena'. The entire dropdown menu is enclosed in a light gray border.

A screenshot of a software interface showing a 'Data Repository' dropdown menu. The menu is open and displays the option 'couchbase-talena10/'. The entire dropdown menu is enclosed in a light gray border.

3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied
- Click **Append** to add new data to an existing bucket

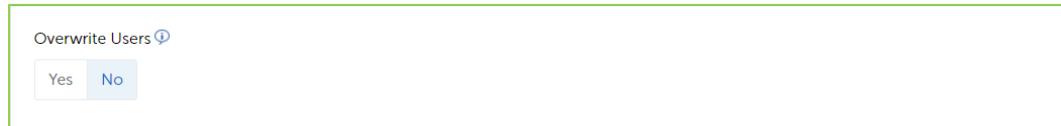
A screenshot of a software interface titled 'Additional Options'. Below the title, there is a section labeled 'Overwrite Behavior' with three buttons: 'Replace', 'Keep', and 'Append'. The 'Keep' button is highlighted with a blue background and white text. The entire interface is enclosed in a light gray border.

4. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

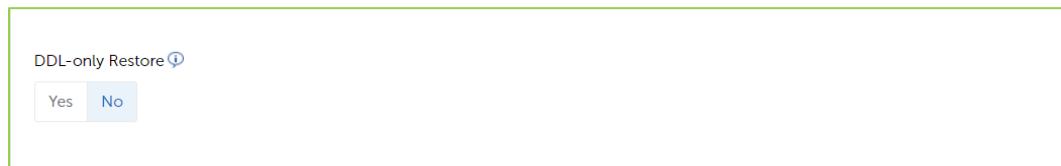
A screenshot of a software interface showing a 'Suffix' input field. The field contains the text '_10102017'. Above the input field, there is a small blue information icon with a question mark. The entire input field is enclosed in a light gray border.

5. In the **More Options for Selected Data** section, edit the object name and rename it. The edited object takes precedence over the object name with suffix.
6. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Overwrite Users** option, click **Yes** to overwrite the existing users.

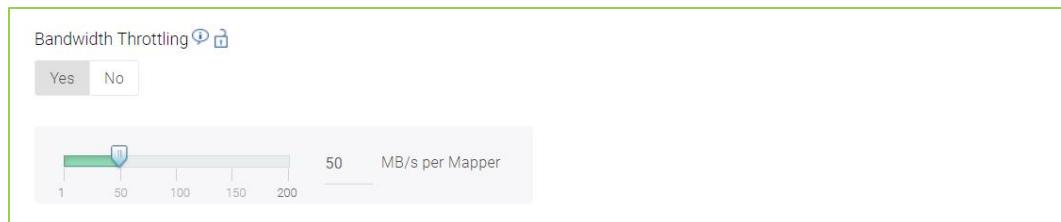


- In the **DDL-only Restore** option, click **Yes** to confirm to restore just the buckets. When you select this option, only the buckets will be restored and no the documents.

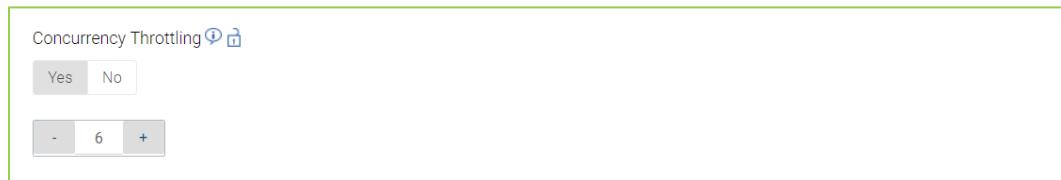


NOTE: DDL-only Restore does not support the recovery of index and views associated with the bucket.

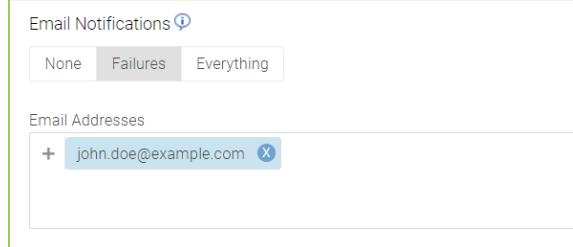
- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **in the MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



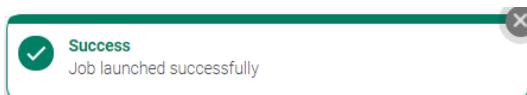
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.

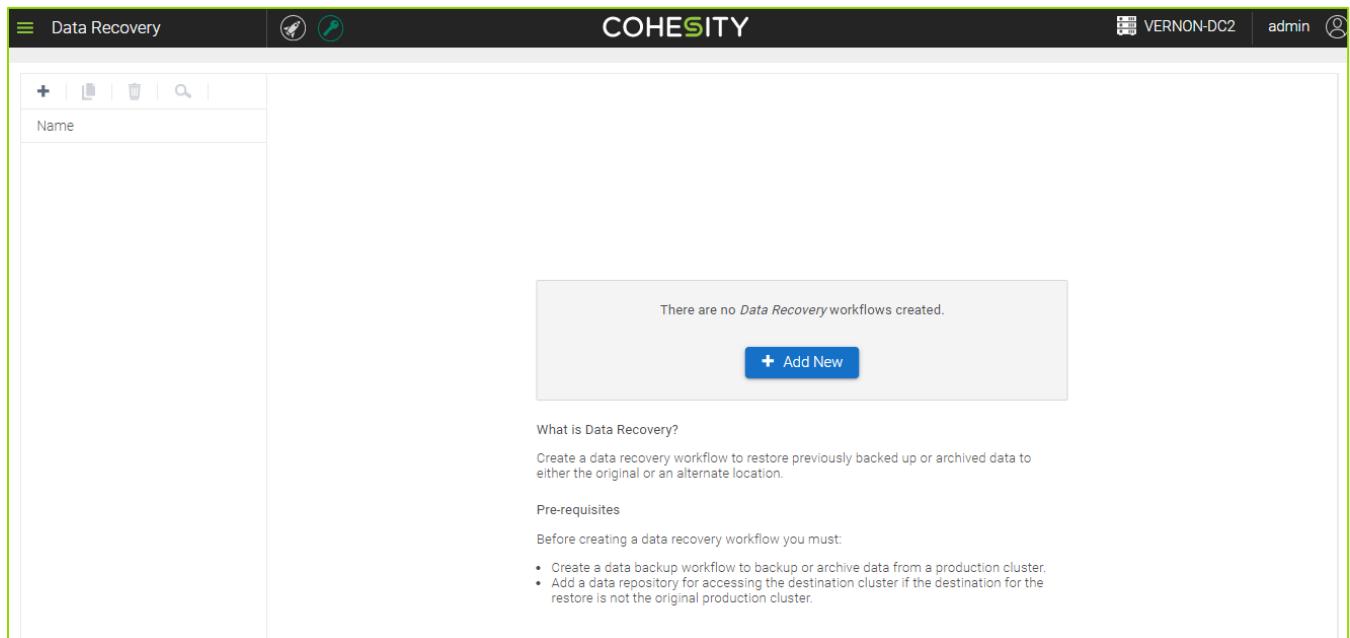


9.2.5.2 Bucket Restore

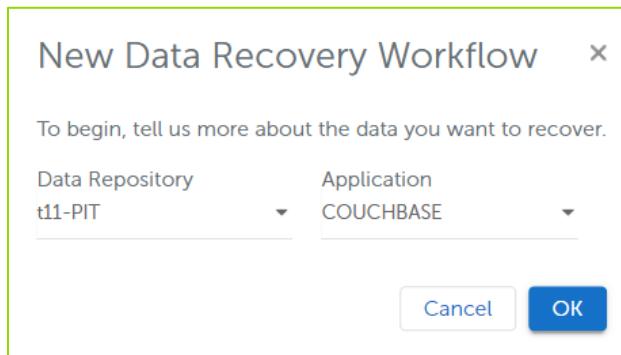
The following section discusses how to recover a data objects for a Couchbase Bucket.

To start a recovery workflow for Couchbase, do the following:

- Click the **Main Menu**  > **Monitoring and Recovering** > **Data Recovery**.
- On the **Data Recovery** page, click the  button or the  icon. The **New Data Recovery Workflow** dialog appears.



- In the **New Data Recovery Workflow** dialog, select a **Couchbase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.



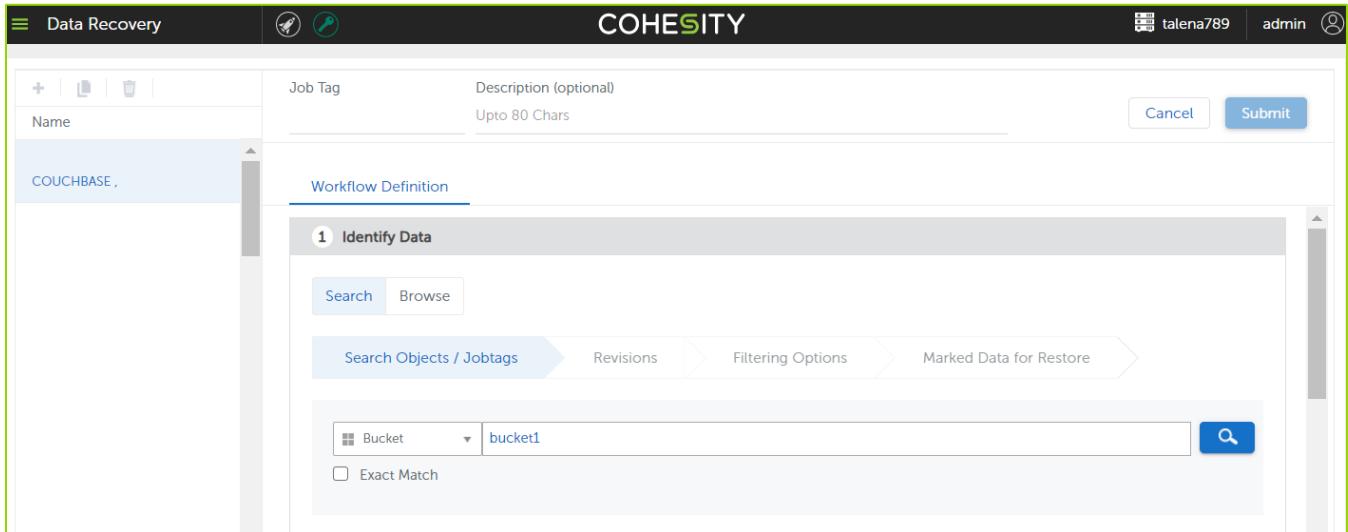
- Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field.
- In the **Identify Data** section, **Search** and **Browse** tabs are displayed. The table below illustrates how you can use both the Search and Browse tabs as per your requirement:

SEARCH	BROWSE
Use when you know the data object that you wish to recover	Use when you want to view data objects and select specific data objects within a JobTag revision
Search specific Jobtag and Bucket	Browse the data objects catalog and select or deselect multiple data objects at a single time

The following sections on Search and Browse show the step-by-step procedure to use these tabs.

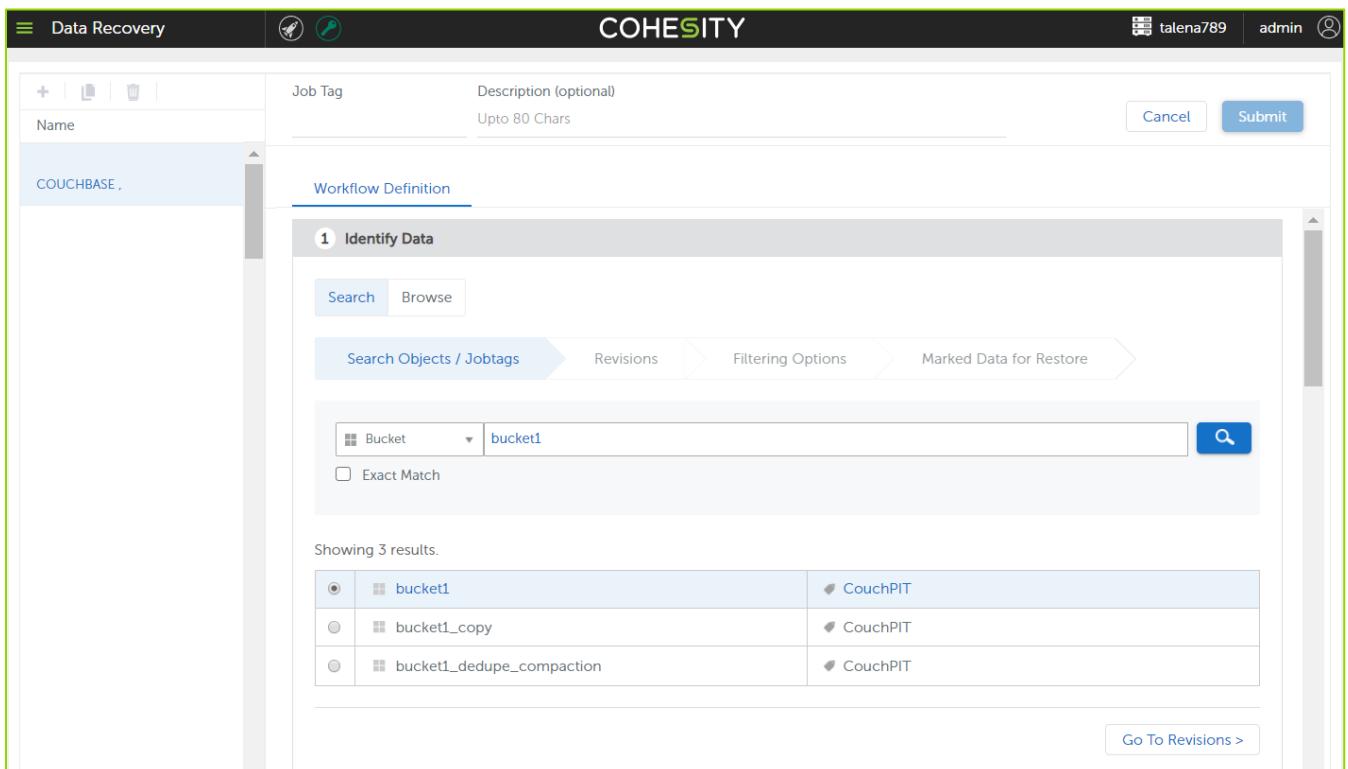
Search Tab

6. In the **Search Objects/Jobtags** tab, select **Bucket**, type the full bucket name, click the  icon.



This screenshot shows the 'Search Objects / Jobtags' tab of the Cohesity Data Recovery interface. The search bar at the bottom has 'Bucket' selected and contains the value 'bucket1'. The search button to the right of the input field is highlighted with a blue border, indicating it is the active or next step to be clicked.

7. Select the radio button of the bucket and then click the **Go to Revisions** button.



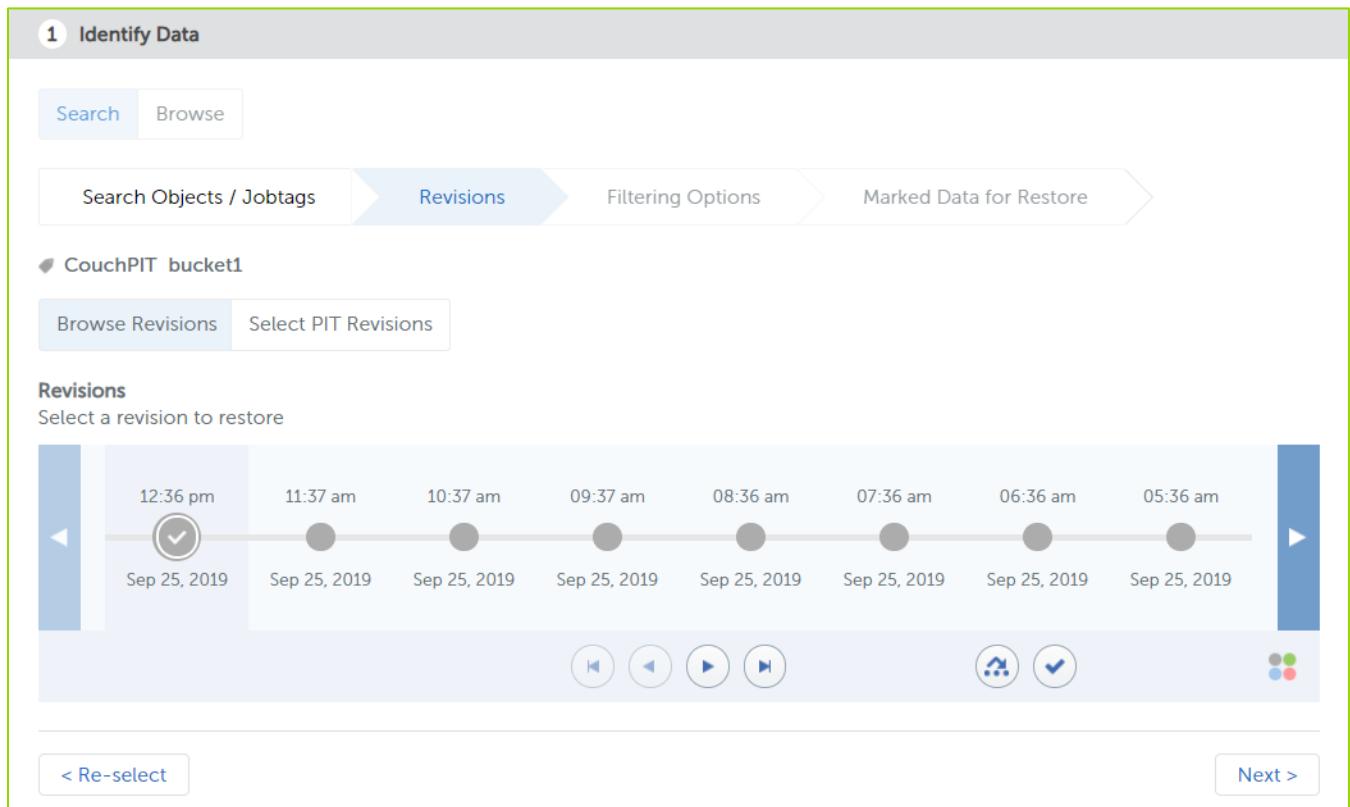
This screenshot shows the results of the search for 'bucket1'. The results table displays three entries:

Result	Description
bucket1	CouchPIT
bucket1_copy	CouchPIT
bucket1_dedupe_compaction	CouchPIT

A 'Go To Revisions >' button is visible at the bottom right of the results area.

8. In the **Revisions** tab, under **Browse Revisions**, do one of the following:

- By default, the latest copy is selected which is indicated by the  icon. You can then click the **Next** button below to restore the selected data object
- Click the data object  icon to select a copy of data for a specific day and time. You can then click the **Next** button to restore the selected data object



1 Identify Data

Search Browse

Search Objects / Jobtags Revisions Filtering Options Marked Data for Restore

CouchPIT bucket1

Browse Revisions Select PIT Revisions

Revisions
Select a revision to restore

12:36 pm 11:37 am 10:37 am 09:37 am 08:36 am 07:36 am 06:36 am 05:36 am

Sep 25, 2019 Sep 25, 2019

< Re-select Next >

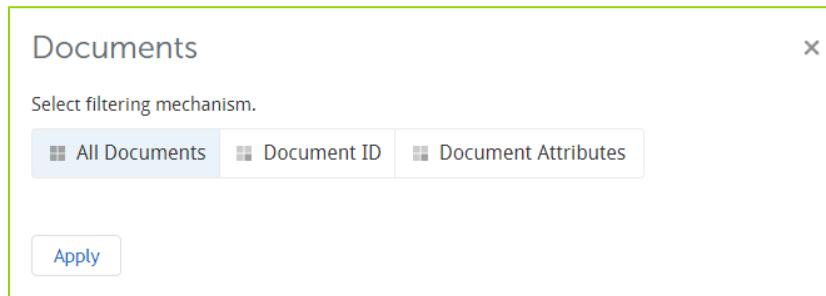
NOTE: Navigate all the data object revision by clicking the     icons. You can also click the  icon to jump to a specific revision in time by selecting a date and time or click the  icon to jump to the currently selected revision.

9. In the **Filtering Options** tab, click the edit icon. Documents dialog box will be displayed.

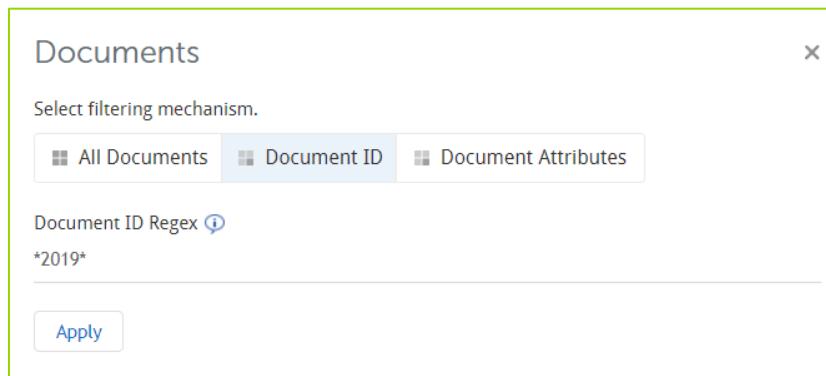
The screenshot shows the 'Identify Data' interface. At the top, there are 'Search' and 'Browse' buttons. Below them is a navigation bar with four tabs: 'Search Objects / Jobtags', 'Revisions', 'Filtering Options' (which is highlighted in blue), and 'Marked Data for Restore'. Underneath the tabs, a timestamp 'CouchPIT 25 Sep 2019, 12:36 PM' is displayed. A table titled 'Selected Objects' contains one item: 'bucket1'. To the right of this table is a 'Filtering Options' section with a 'All Documents' button and an edit icon. At the bottom left is a '< Back' button, and at the bottom right is a 'Mark For Restore >' button.

10. In the **Documents** dialog box, do one of the following:

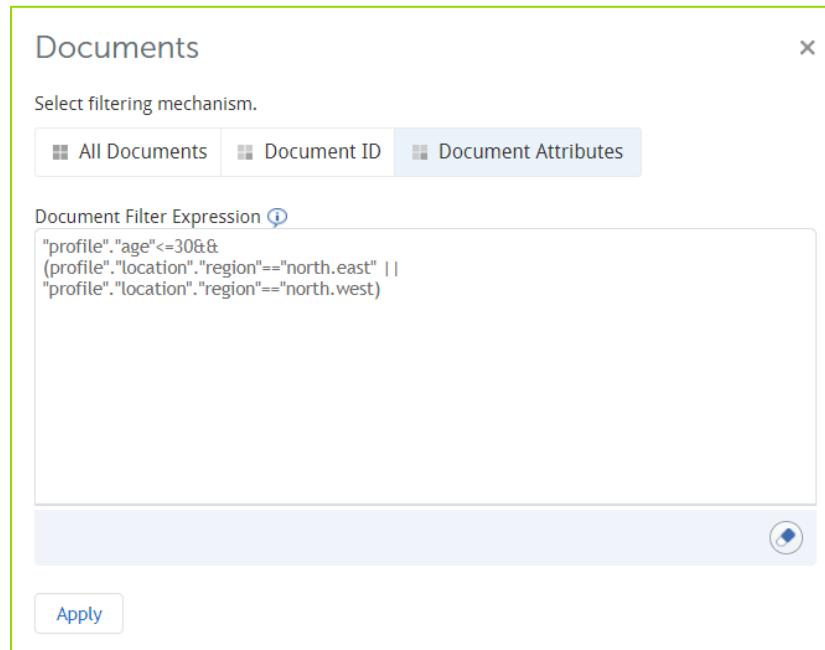
- By default, the **All Documents** button is selected. All you have to do is click the **Apply**:



- Click the **Document ID** button and then type a regular expression to match documents IDs (keys) to restore (only Java regular expression set is supported) and then click the **Apply** button:



- Click the **Document Attributes** button and then type a filter expression to match document attributes to restore and then click the **Apply** button:



NOTE: Imanis Data software supports filtering only on document keys and document content.

In full data recovery, both document key and document content filtering is supported:

This option is visible only if "bucket" is selected on full data recovery.

In Couchbase Recovery module, under **Documents**, under **Select Filtering mechanism**, click **Document ID** and type a regular express to match document IDs.

In incremental data recovery, only document content filtering is supported:

All terms must be separated by one space

Comparisons with "<=", "<", ">=", ">", "==", "!=" are supported. &&, || for and, or respectively.

Filtering on nested items like {"i1" : {"i2" : 43}} can be specified as "i1"."i2" > 21

Any condition where variables are not found in json is treated as false. Json docs are restored only if filter expression evaluates to true. Binary docs are always restored.

Parenthesis are also supported, however, whitespace are not permitted after opening brace and before closing brace. For example,

"ibu" > 0 && ("type" == "beer" || "abv" > 5).

11. In the **Marked Data for Restore** tab, do one of the following:

- Click the **Reset** button to go back to the **Search Objects/Jobtags** tab
- Click the **Search and Add More Objects** to back to the Search Objects/Jobtags tab to reselect a data object revision to restore

- Click the **Change Revision** button to go back to the **Revisions** tab

The screenshot shows the 'Identify Data' step of the Cohesity process. At the top, there are 'Search' and 'Browse' buttons. Below them is a navigation bar with four tabs: 'Search Objects / Jobtags', 'Revisions', 'Filtering Options', and 'Marked Data for Restore'. Underneath the tabs, a timestamp 'CouchPIT 25 Sep 2019, 12:36 PM' is displayed. The main area is titled 'Selected Objects' and contains a table with one row: 'bucket1'. To the right of the table are 'Filtering Options' (set to 'All Documents') and a delete 'X' button. At the bottom of this section are 'Reset' and 'Search and Add More Objects' buttons. A 'Change Revision' button is located at the very bottom left.

You can jump directly onto Step # 11 to continue Couchbase data recovery for Bucket.

Browse Tab

7. In the **Search Objects / Jobtags** tab, select a **JobTag** from the JobTag List, and then click the **Go To Revisions** button.

The screenshot shows the 'Identify Data' step with the 'Browse Jobtags' tab selected. At the top, there are 'Search' and 'Browse' buttons. Below them is a navigation bar with four tabs: 'Browse Jobtags', 'Revisions', 'Browse Objects', and 'Marked Data for Restore'. A search bar 'Enter job tag here to filter...' is present. The 'JobTags List' section shows two entries: 'CouchPIT' (selected) and 'Couchbase_PITR'. At the bottom right is a 'Go To Revisions >' button.

8. In the **Revisions** tab, select a revision of the JobTag revision that you want to restore and then click the **Browse Objects** button.

The screenshot shows the 'Identify Data' step of a process. At the top, there are 'Search' and 'Browse' buttons. Below them is a navigation bar with four tabs: 'Browse Jobtags', 'Revisions' (which is highlighted in blue), 'Browse Objects', and 'Marked Data for Restore'. Underneath the tabs, there is a section labeled 'CouchPIT *' with a radio button next to it. Below this is a 'Revisions' section with the instruction 'Select a revision to restore'. A horizontal timeline displays several revisions of a JobTag from Sep 25, 2019, at various times: 12:36 pm, 11:37 am, 10:37 am, 09:37 am, 08:36 am, 07:36 am, 06:36 am, and 05:36 am. The first revision (12:36 pm) has a checkmark icon inside a circle and is highlighted. Navigation arrows are located on the left and right ends of the timeline. At the bottom of the screen, there are two buttons: '< Re-select' on the left and 'Browse Objects >' on the right.

9. In the **Browse Objects** label, do one of the following:

- Select bucket you want to restore and then click the **Next** button
- Click the **Change revision** button to go back to the Revisions tabs and select a new revision of the JobTag

1 Identify Data

Search Browse

Browse Jobtags Revisions Browse Objects Marked Data for Restore

CouchPIT 25 Sep 2019, 12:36 PM

Select the desired objects and click on "Mark For Restore" button to restore the objects.

< /

<input checked="" type="checkbox"/> Objects
<input checked="" type="checkbox"/> b2
<input checked="" type="checkbox"/> bucket1_copy
<input checked="" type="checkbox"/> bucket1_dedupe_compaction
<input checked="" type="checkbox"/> bucket1
<input checked="" type="checkbox"/> 300million

< Change revision Next >

10. In the **Marked Data for Restore** label, do one of the following:

- Click the **Reset** button to go back to the **Browse Jobtags** tab
- Click the **Browse Objects** button to go back to the **Revisions** tab to reselect a data object revision to restore

1 Identify Data

Search Browse

Browse Jobtags Revisions Browse Objects Marked Data for Restore

CouchPIT 25 Sep 2019, 12:36 PM

Selected Objects	
*	X

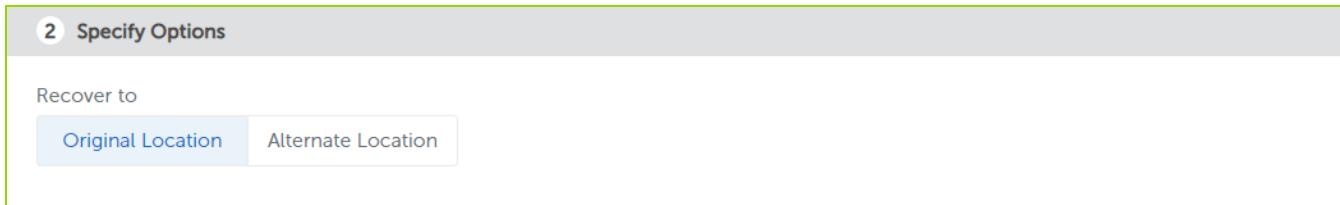
Reset

< Browse Objects

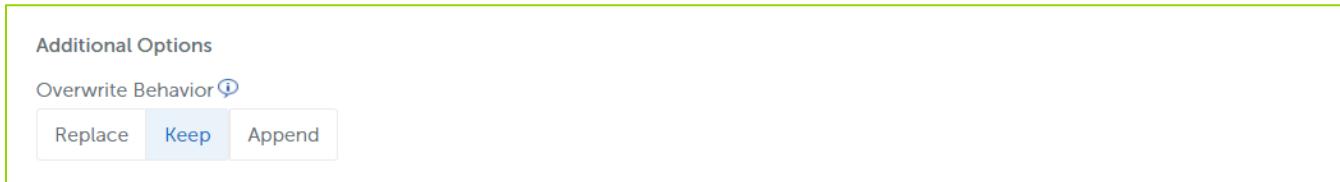
11. In the **Specify Options** section, under the **Recover To** area, select **Original Location or Alternate Location**.

ORIGINAL LOCATION:

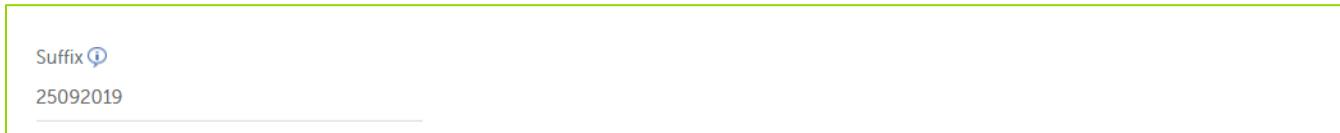
1. Click the **Original Location** button.



2. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:
 - Click **Replace** to replace existing data with existing data with new data thus erasing any previously existing data
 - Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
 - Click **Append** to add new data to an existing bucket

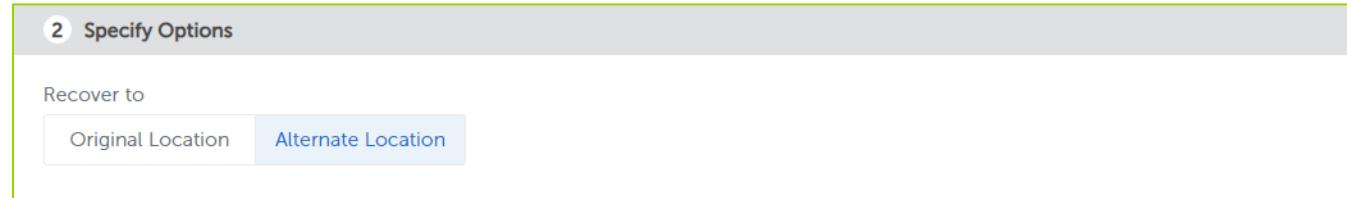


3. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

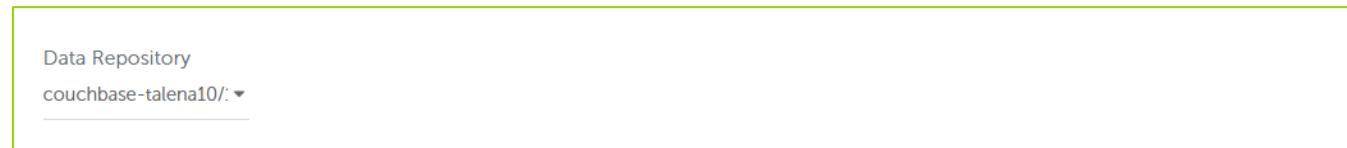


ALTERNATE LOCATION:

1. Click the **Alternate Location** button.

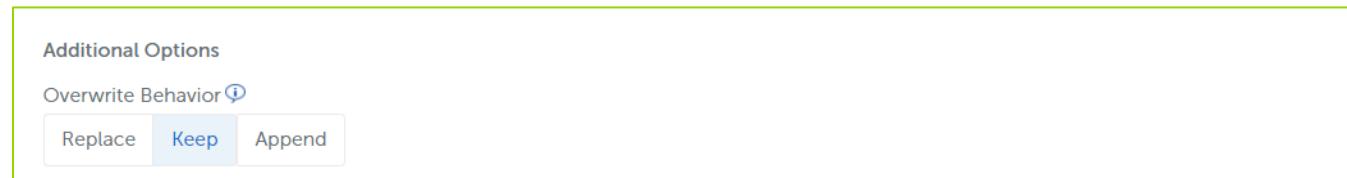


2. Select a Couchbase cluster name from the **Data Repository** drop-down menu.

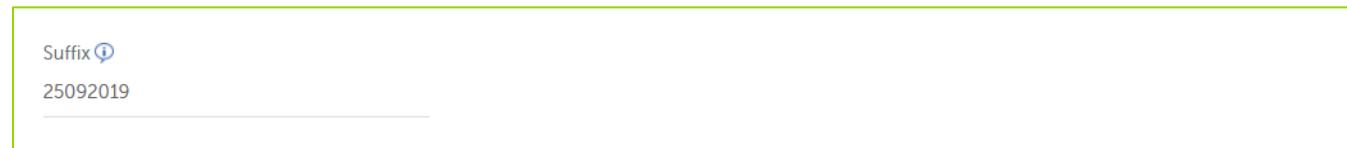


3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
- Click **Append** to add new data to an existing bucket



4. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.



12. In the **More Options for Selected Data** section, edit the object name and rename it. The edited object takes precedence over the object name with suffix.

13. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Overwrite Users** option, click **Yes** to overwrite the existing users. When you select, all the existing users are overwritten by new users.

Overwrite Users 

- In the **DDL-only Restore** option, click **Yes** to confirm to restore just the buckets. When you select this option, only the buckets will be restored and no the documents:

DDL-only Restore 

NOTE: DDL-only Restore does not support the recovery of index and views associated with the bucket.

- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:

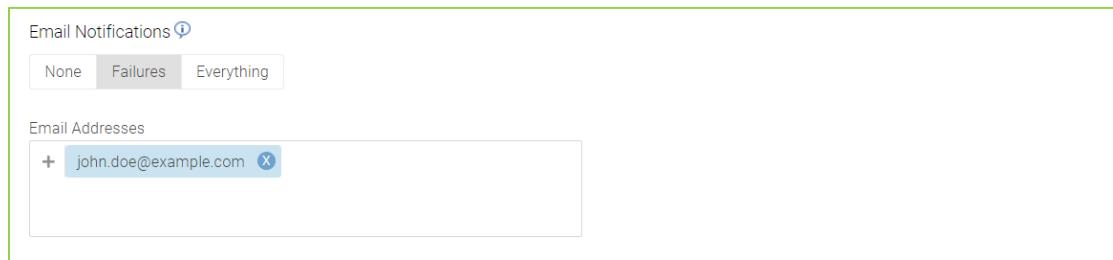
Bandwidth Throttling  



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:

Concurrency Throttling  

- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

13. Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



IMPORTANT: Imanis Data software does not support backup and recovery of a Couchbase cluster when one or more active master nodes in Couchbase cluster are unreachable. In case an active master node fails during the backup or recovery process, you can re-run the backup and recovery process once the rebalance or failover (auto or manual) is completed. Remove the failed node from the repository seed node list, re-verify the repository and then save it. To find out if an active node is unreachable, access the Couchbase Web Console. For more information, refer to the Couchbase documentation.

NOTE: The Couchbase Database Change Protocol (DCP) recommends that to derive optimum performance no more than six concurrent connections should be run from the same data repository at any given time. Therefore, by default, the Imanis Data agent invokes a maximum of six data mover processes in parallel. In case you need to execute more than one job concurrently on the same Couchbase cluster, then you must reduce the number of concurrent data movers. This change ensures that the total number connections do not exceed more than six connections.

NOTE: Currently, Imanis Data software restores the following properties on bucket create: Index replicas, Conflict resolution mode, Bucket Max Time-To-Live, Compression Mode, threadsNumber, evictionPolicy, enableFlush, quota, and replicas. If a bucket is already created and some properties are changed at source cluster, then those changes will not be restored onto the destination cluster. The properties which are not restored by Imanis Data software are recovered onto the destination cluster with their default value.

NOTE: Recovery of Ephemeral and Memcached buckets is not supported.

NOTE: The Couchbase Recovery workflow attempts to restore a bucket on the destination cluster with the same RAM as it is available on the source cluster. In such cases, the Couchbase sizing guidelines recommend that you must have the required RAM on the destination cluster before attempting to restore a bucket. For more information, refer to the information available in Couchbase documentation about [Sizing Guidelines](#).

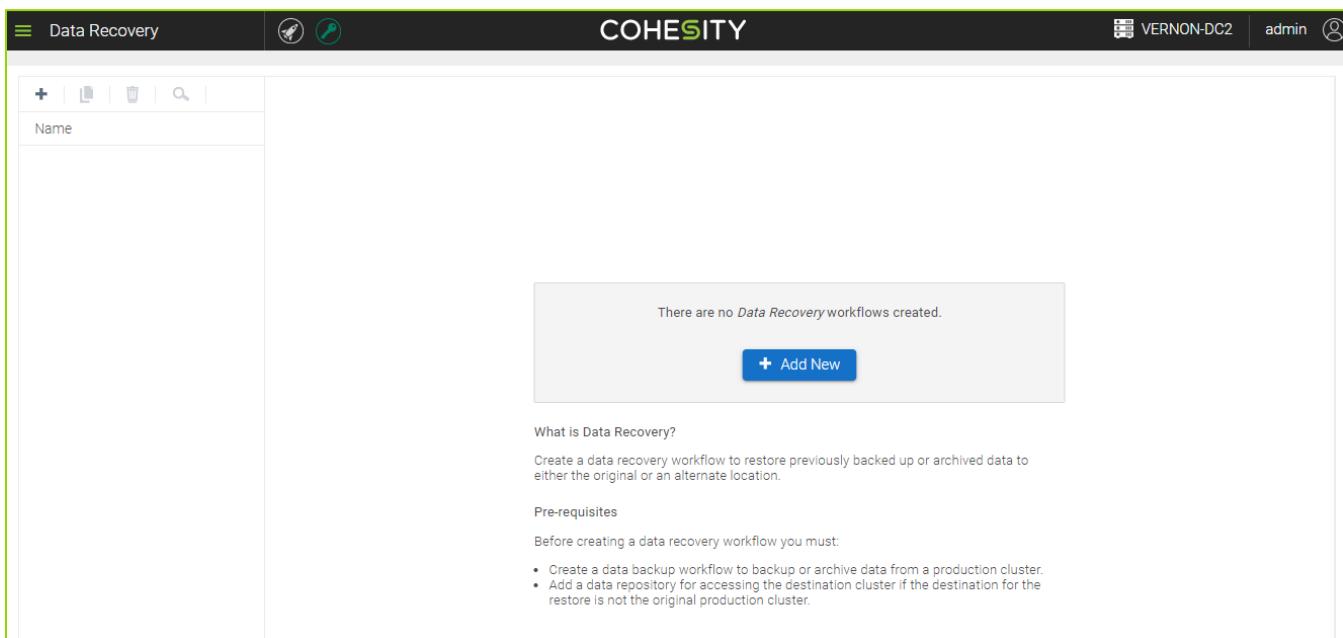
IMPORTANT: Couchbase best practice suggests that you perform a rebalance during an application's lowest traffic levels. Rebalance increases the overall load and resource utilization of a system hence certain environments may notice degradation. If during rebalance, backup and recovery jobs - including the recurring jobs – are being executed; it may increase the overall load of the Couchbase application. Therefore, it is recommended that you either verify that all the backup and recovery jobs (including the recurring jobs) are completed before performing a rebalance or execute the backup or recovery jobs during an application's lowest traffic levels. For more information, refer to the information available in [Couchbase documentation](#).

9.2.5.2.1 Bucket PIT Restore

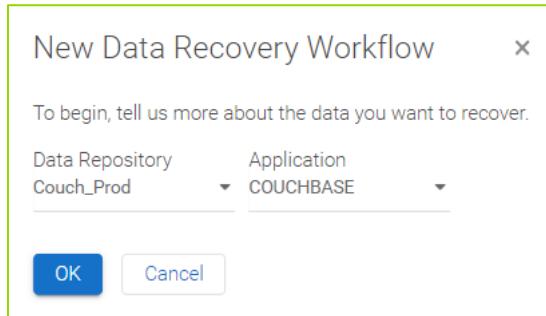
The following section discusses how to Point-in-Time (PIT) recovery of data objects for a Couchbase Bucket.

To start a recovery workflow for Couchbase, do the following:

1. Click the **Main Menu**  > **Monitoring and Recovering** > **Data Recovery**.
2. On the **Data Recovery** page, click the  button or the  icon. The **New Data Recovery Workflow** dialog appears.



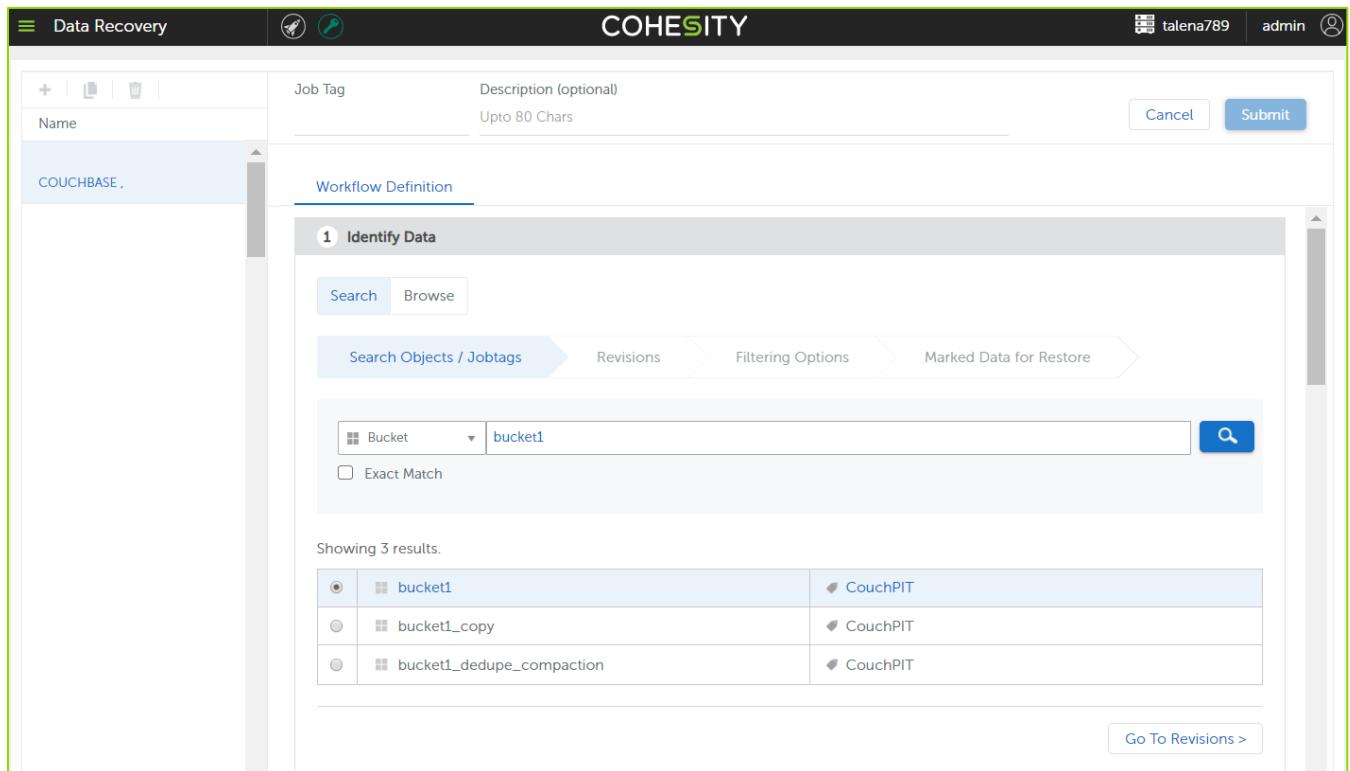
3. In the **New Data Recovery Workflow** dialog, select a **Couchbase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.



4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field.
5. In the **Identify Data** section, under the **Search** tab in the **Search Objects/Jobtags** tab, select **Bucket**, type the bucket name and then click the icon. Imanis Data will display the bucket data object as a search result.

The screenshot shows the Cohesity Imanis Data interface with a "Data Recovery" workflow definition. The "Identify Data" step is selected. Under the "Search Objects / Jobtags" tab, the "Bucket" dropdown is set to "bucket1" and the search icon is highlighted with a blue border. Other tabs include "Revisions", "Filtering Options", and "Marked Data for Restore".

6. Select the radio button of the bucket data object and click the **Go to Revisions** button.



7. In the **Revisions** tab, click the **Select PIT Revisions** tab, do the following:

- Under the **Select the date and time for PIT revision** option, set a date and time to select a revision copy of PIT data of a specific day and time and then click **Go**.

Select the date and time for PIT revision

2019-09-25 01:37 PM Go

- Select one of the PIT revisions copy of the data which is nearest to the date and time that you set in the previous step and then click the **Next** button to restore the selected data object.

Found following nearest PIT revision(s). Select desired revision.

11:31 AM, 25 Sep 2019 11:46 AM, 25 Sep 2019

12:36 PM 11:37 AM

25 Sep, 2019 25 Sep, 2019

Partial Full

< Re-select

Next >

8. The **Filtering Options** tab will be displayed, however, in the current release filtering options are NOT supported in PIT restore.

The screenshot shows the 'Identify Data' process step 1. The navigation bar includes 'Search' and 'Browse' buttons. Below the bar are tabs: 'Search Objects / Jobtags', 'Revisions', 'Filtering Options' (which is highlighted in blue), and 'Marked Data for Restore'. A timestamp 'CouchPIT 25 Sep 2019, 11:46 AM' is displayed. The main area shows a table with two columns: 'Selected Objects' and 'Filtering Options'. Under 'Selected Objects', there is a row for 'bucket1'. Under 'Filtering Options', there are two buttons: 'Point-In-Time Copy' and 'All Documents'. A red 'X' icon is located at the far right of the table. At the bottom left is a '[< Back](#)' link, and at the bottom right is a '[Mark For Restore >](#)' button.

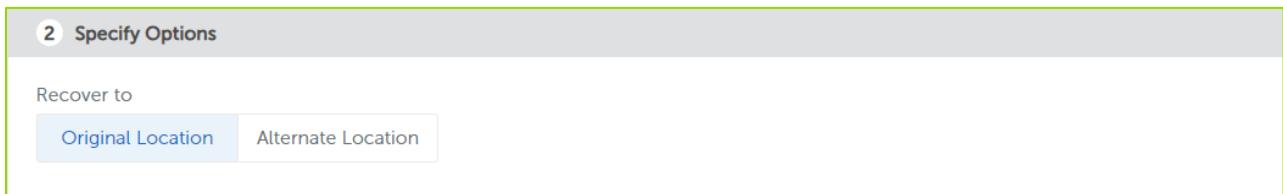
9. In the **Marked Data for Restore** tab, do one of the following:

- Click the **Reset** button to go back to the **Search Objects/Jobtags** tab
- Click the **Change Revision** button to back to **Revisions** tab to reselect a data object revision to restore

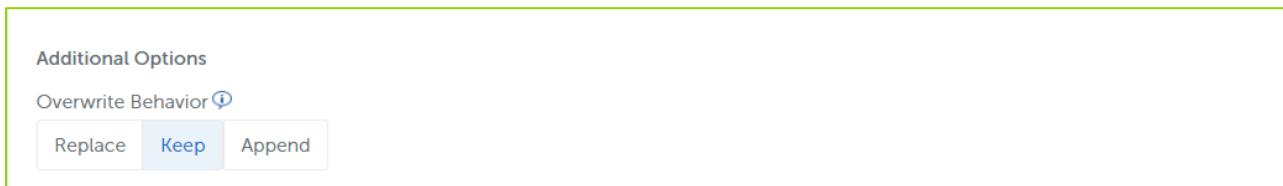
The screenshot shows the 'Identify Data' process step 1. The navigation bar includes 'Search' and 'Browse' buttons. Below the bar are tabs: 'Search Objects / Jobtags', 'Revisions', 'Filtering Options', and 'Marked Data for Restore' (which is highlighted in blue). A timestamp 'CouchPIT 25 Sep 2019, 11:46 AM' is displayed. The main area shows a table with two columns: 'Selected Objects' and 'Filtering Options'. Under 'Selected Objects', there is a row for 'bucket1'. Under 'Filtering Options', there are two buttons: 'Point-In-Time Copy' and 'All Documents'. A red 'X' icon is located at the far right of the table. At the bottom right are two buttons: 'Reset' and 'Search and Add More Objects'. At the bottom left is a '[< Change Revision](#)' link.

ORIGINAL LOCATION:

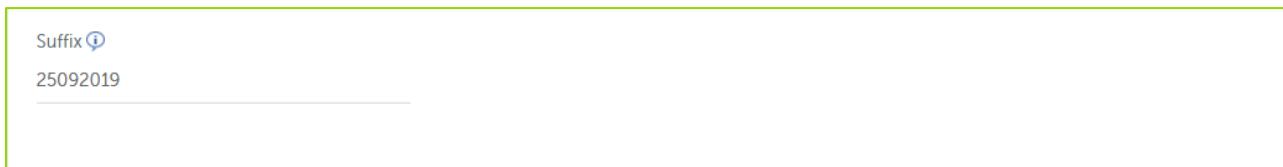
1. Click the **Original Location** button.



2. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:
 - Click **Replace** to replace existing data with new data thus erasing any previously existing data
 - Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
 - Click **Append** to add new data to an existing bucket

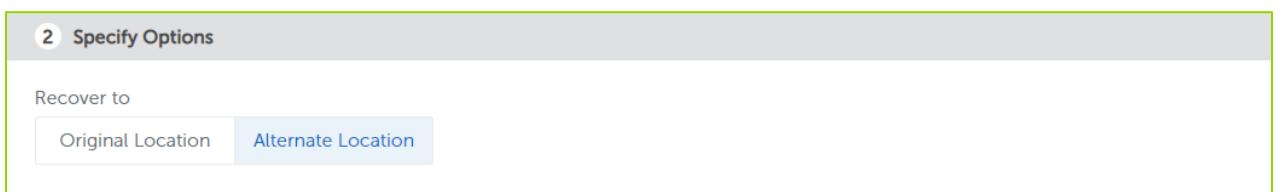


3. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

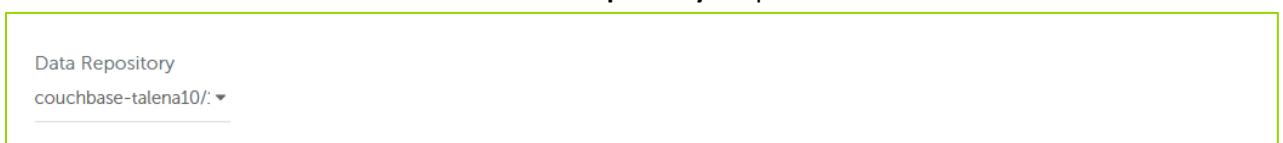


ALTERNATE LOCATION:

1. Click the **Alternate Location** button.

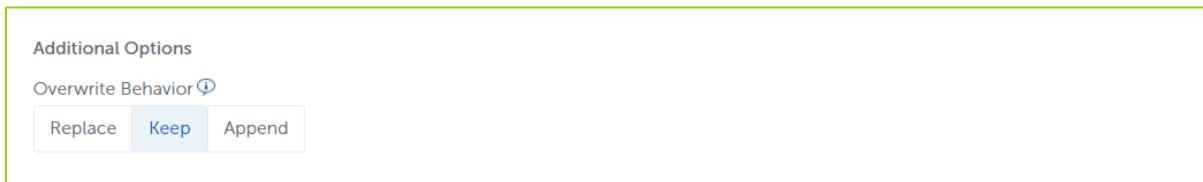


2. Select a Couchbase cluster name from the **Data Repository** drop-down menu.

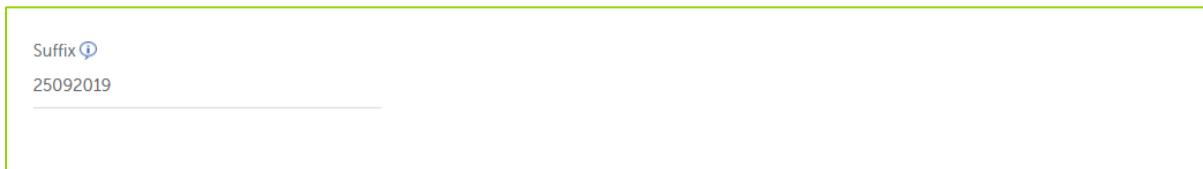


3. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
- Click **Append** to add new data to an existing bucket



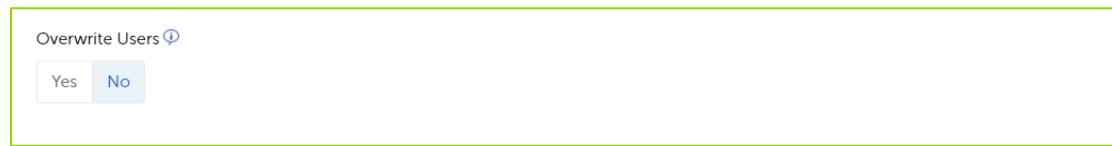
4. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.



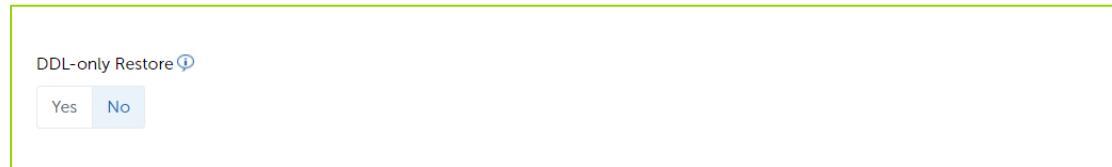
11. In the **More Options for Selected Data** section, edit the object name and rename it. The edited object takes precedence over the object name with suffix.

12. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Overwrite Users** option, click **Yes** to overwrite the existing users. When you select, all the existing users are overwritten by new users.



- In the **DDL-only Restore** option, click **Yes** to confirm to restore just the buckets. When you select this option, only the buckets will be restored and no the documents:

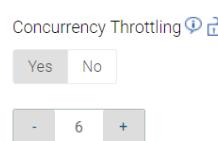


NOTE: DDL-only Restore does not support the recovery of index and views associated with the bucket.

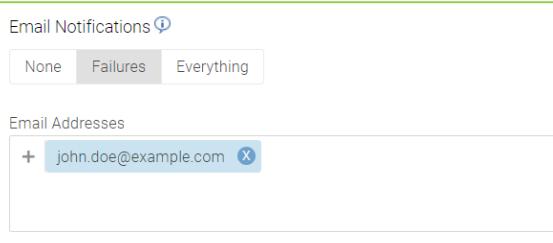
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **in the MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



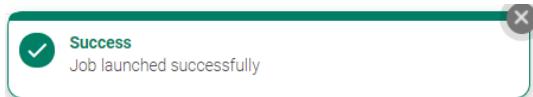
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.

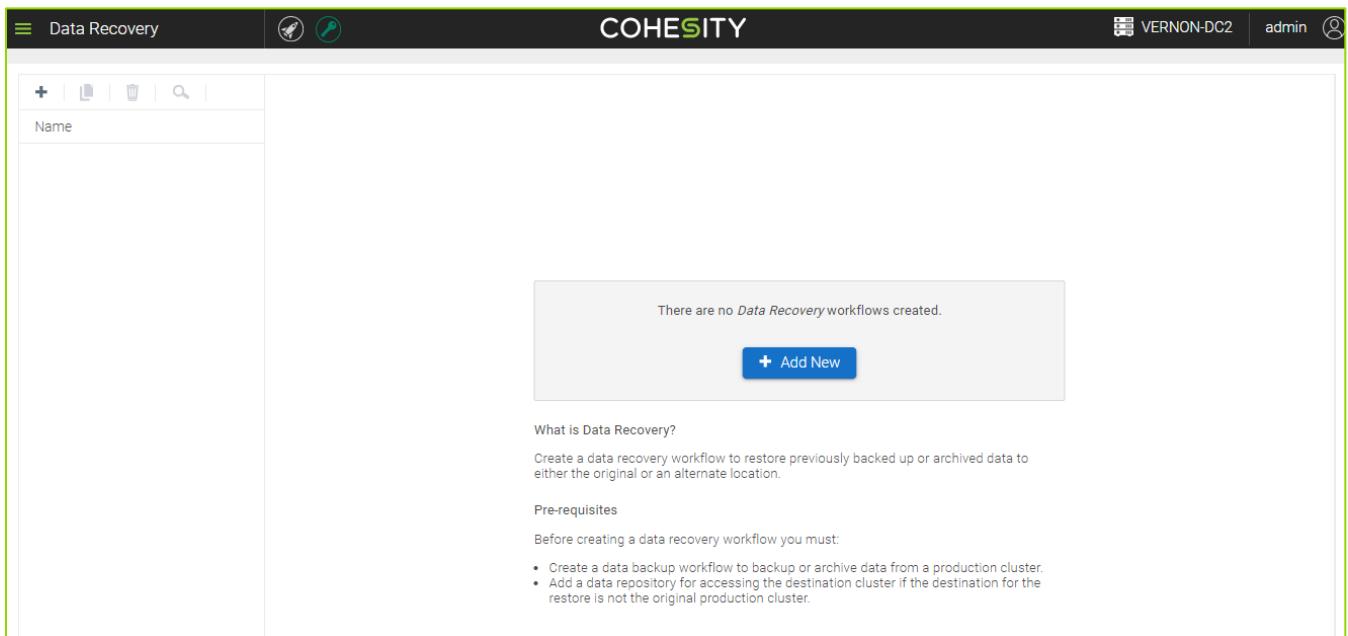


9.2.6 Recovering Data for MongoDB

Imanis Data software supports MongoDB recovery at the database and collections level.

To start a recovery workflow for MongoDB, do the following:

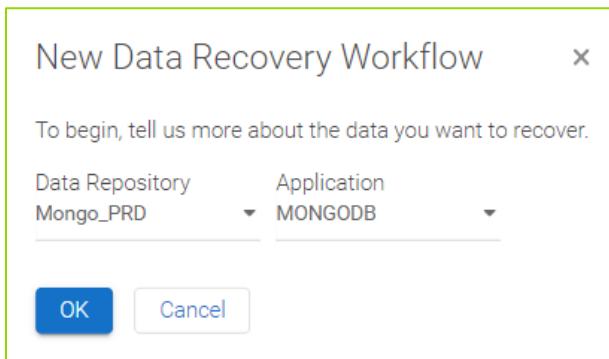
1. Click the **Main Menu**  > **Monitoring and Recovering** > **Data Recovery**.
2. On the **Data Recovery page**, click the  button or the  icon. The **New Data Recovery Workflow** dialog appears.



The screenshot shows the Cohesity Data Recovery interface. At the top, there is a navigation bar with the title "Data Recovery". Below the navigation bar, there is a toolbar with icons for creating, deleting, and searching. The main content area displays a message: "There are no Data Recovery workflows created." Below this message is a blue button labeled "+ Add New". Further down, there is a section titled "What is Data Recovery?" with a brief description: "Create a data recovery workflow to restore previously backed up or archived data to either the original or an alternate location." There is also a "Pre-requisites" section with a list of requirements:

- Create a data backup workflow to backup or archive data from a production cluster.
- Add a data repository for accessing the destination cluster if the destination for the restore is not the original production cluster.

3. In the **New Data Recovery Workflow** dialog, select a **MongoDB** source data repository from the **Data Repository** drop-down menu, and then click **OK**.



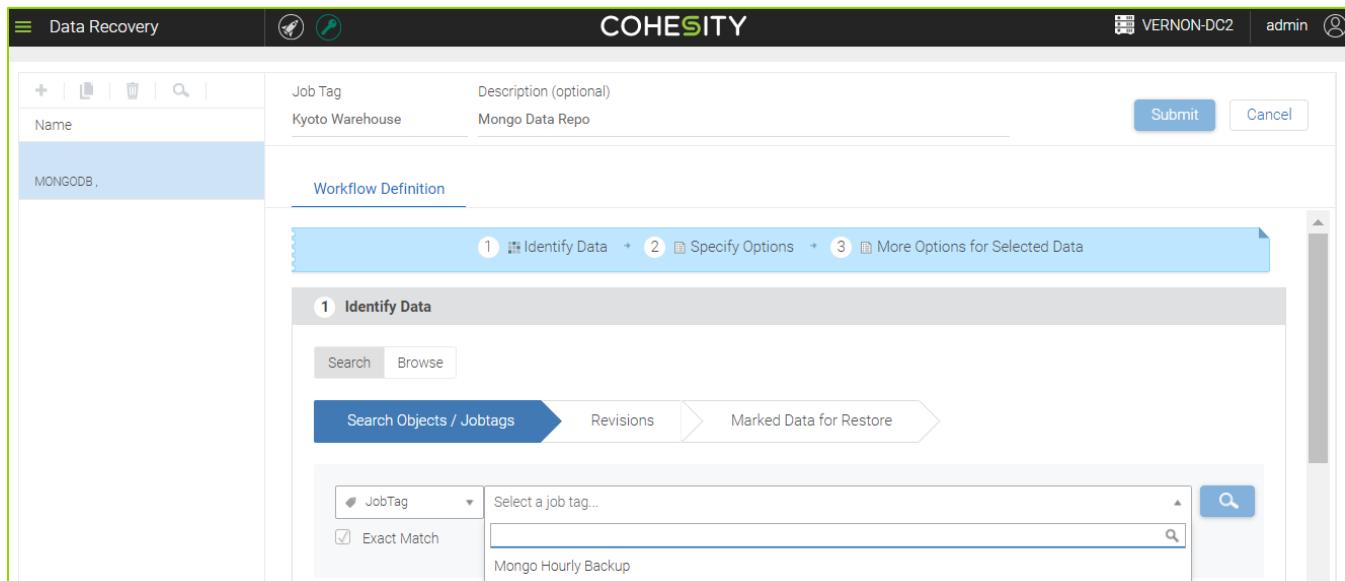
4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field.
5. In the **Identify Data** section, **Search** and **Browse** tabs are displayed. You can use the Search and Browse button as per your requirement:

SEARCH	BROWSE
Use when you know the data object that you wish to recover	Use when you want to view data objects and select specific data objects within a JobTag revision
Search specific Jobtag, Database, Table, or Partition	Browse the data objects catalog and select or deselect multiple data objects at a single time

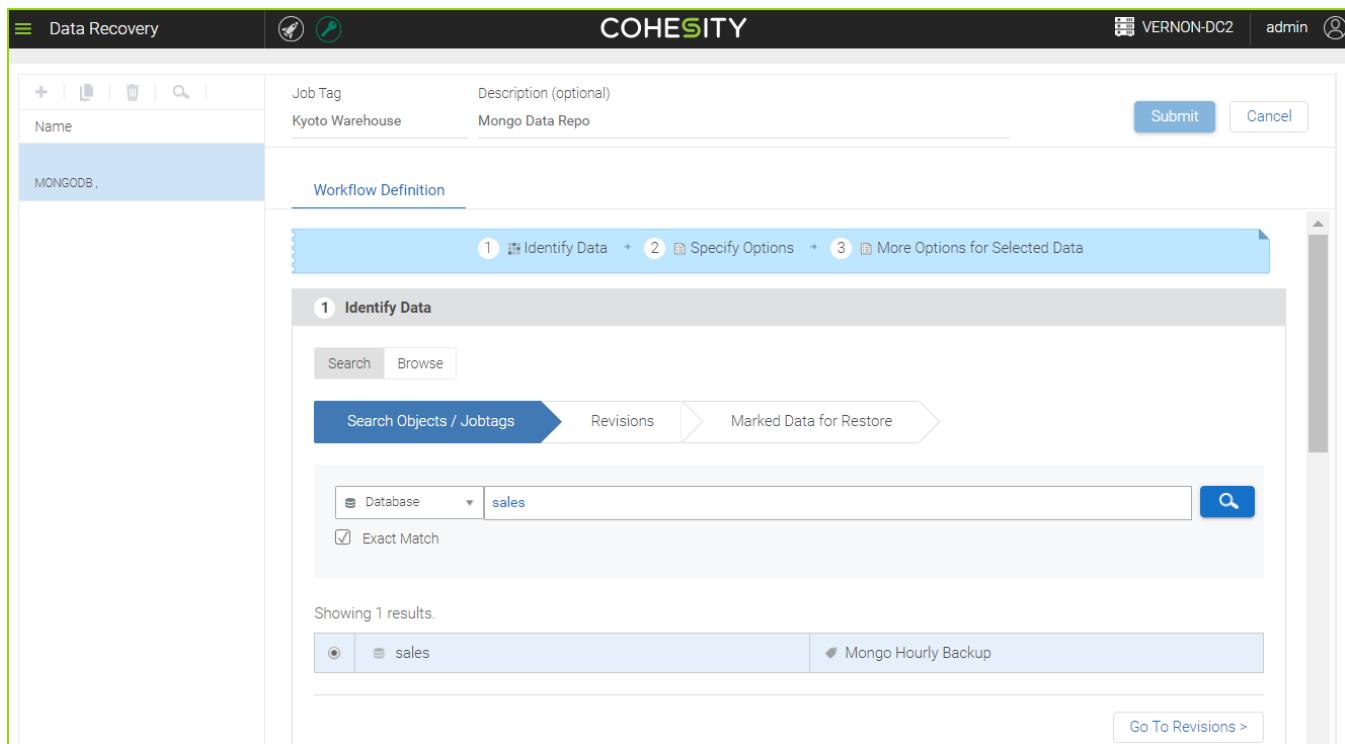
Search Tab

6. In the **Search Objects/Jobtags** tab, do one of the following:

- Select a **JobTag**, click the search box, select displayed by Imanis Data, and then click the icon



- Select **Database**, type the database name, and then click the  icon



- Select **Collection**, type the collection name, and then click the  icon

To know how to specify Collection specific properties ('shardKeyJson' and 'unique'), refer to this section: **More Options for Selected Data**.

What exactly are the Collection specific properties shardKeyJson & unique?

* {{shardKeyJson}}: The index specification document to use as the shard key. The shard key determines how MongoDB distributes the documents among the shards. This should be specified as a JSON.

Example: {{\{_id: "hashed"\}}}

* {{unique}}: When true, the unique option ensures that the underlying index enforces a unique constraint. Hashed shard keys do not support unique constraints. This should be specified as either {{true}} or {{false}}

For details please refer to MongoDB documentation at

<https://docs.mongodb.com/manual/reference/method/sh.shardCollection/>

7. In the **Revisions** tab, do one of the following:

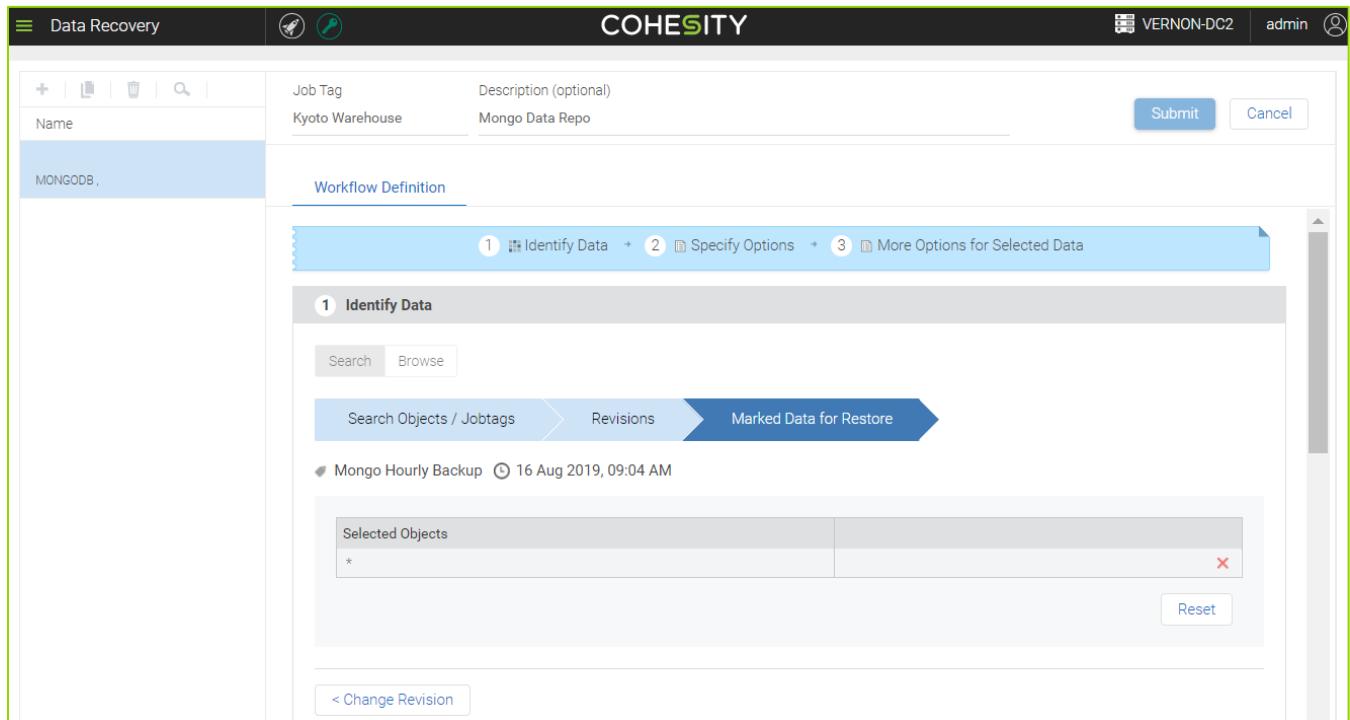
- By default, the latest copy is selected which is indicated by the  icon. You can then click the Next button to restore the selected data object
- Click the data object  icon to select a copy of data for a specific day and time. You can then click the Next button to restore the selected data object

The screenshot shows the Cohesity Data Recovery interface. At the top, there are tabs for 'Data Recovery' and 'Vault'. The main area is titled 'COHESITY' with a 'Job Tag' field containing 'Kyoto Warehouse' and a 'Description (optional)' field containing 'Mongo Data Repo'. Below this is a 'Workflow Definition' section with three steps: 1. Identify Data, 2. Specify Options, and 3. More Options for Selected Data. The 'Identify Data' step is active, showing a 'Search' tab selected. A timeline displays data object revisions from 09:04 am to 02:05 am on Aug 16, 2019. The revision at 09:04 am is highlighted with a blue circle and checked. Other revisions are marked with grey circles. Navigation icons for previous and next revisions are available. A 'Re-select' button is at the bottom left, and a 'Next >' button is at the bottom right.

NOTE: Navigate all the data object revision by clicking the icons. You can also click the icon to jump to a specific revision in time by selecting a date and time or click the icon to jump to the currently selected revision.

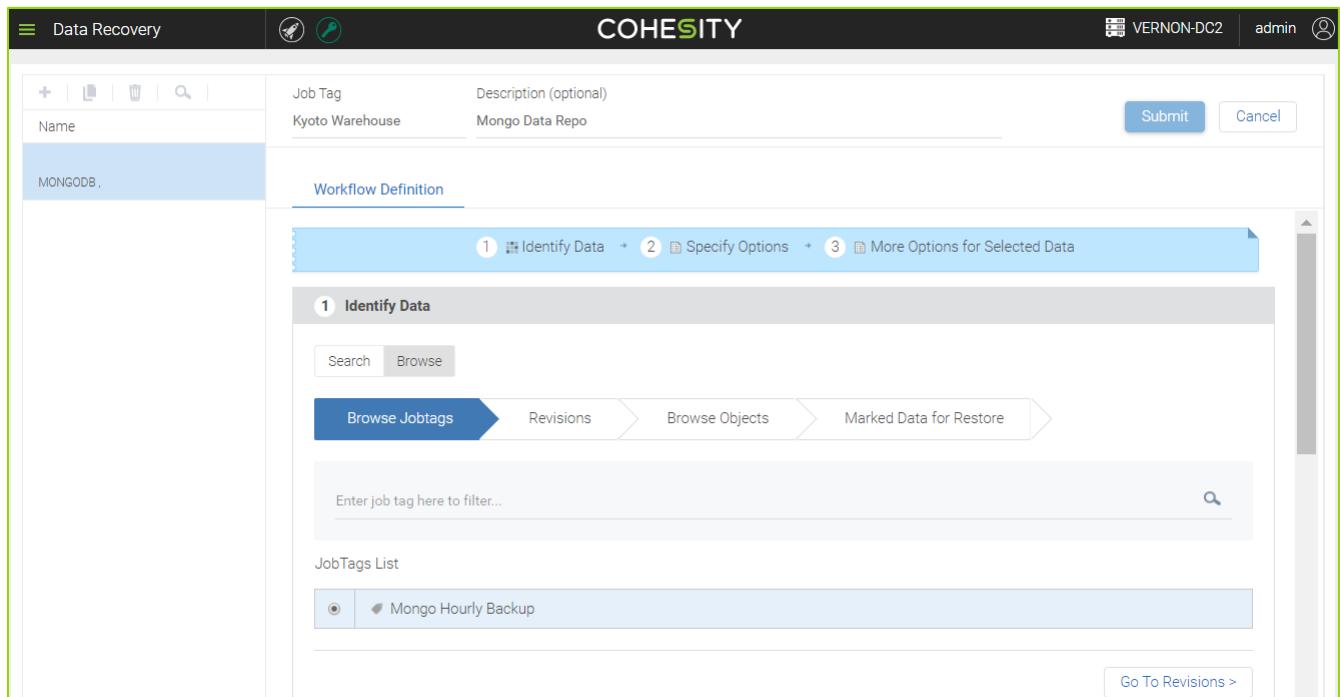
8. In the **Marked Data for Restore** tab, do one of the following:

- Verify your selection
- Click the **X** icon to remove the current selection and click the **Change Revision** button to reselect a new revision

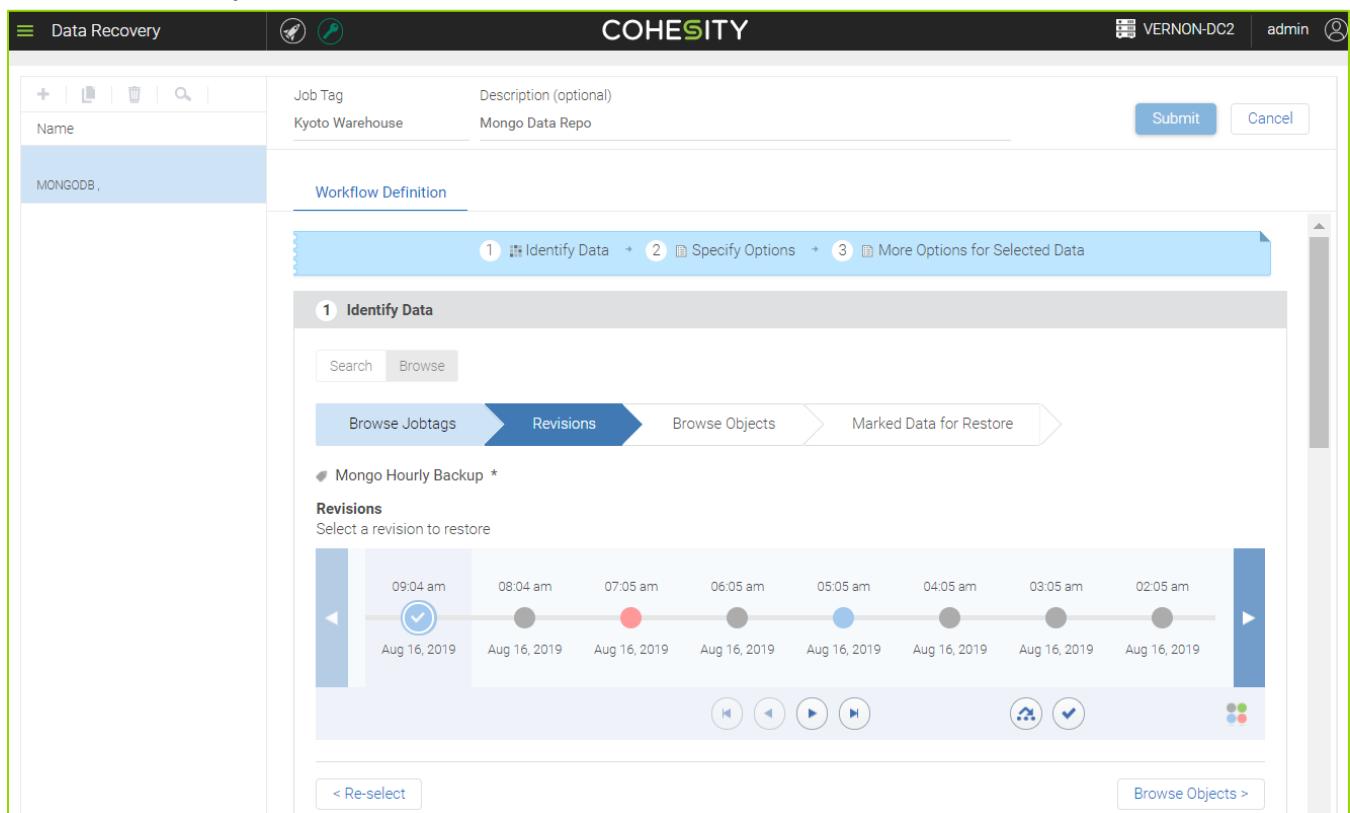


Browse Tab

7. In the **Browse Jobtags** label, select a **JobTag** from the JobTag List, and then click the **Go To Revisions** button.



- In the **Revisions** label, select a revision of the JobTag revision that you want to restore and then click the **Browse Objects** button.



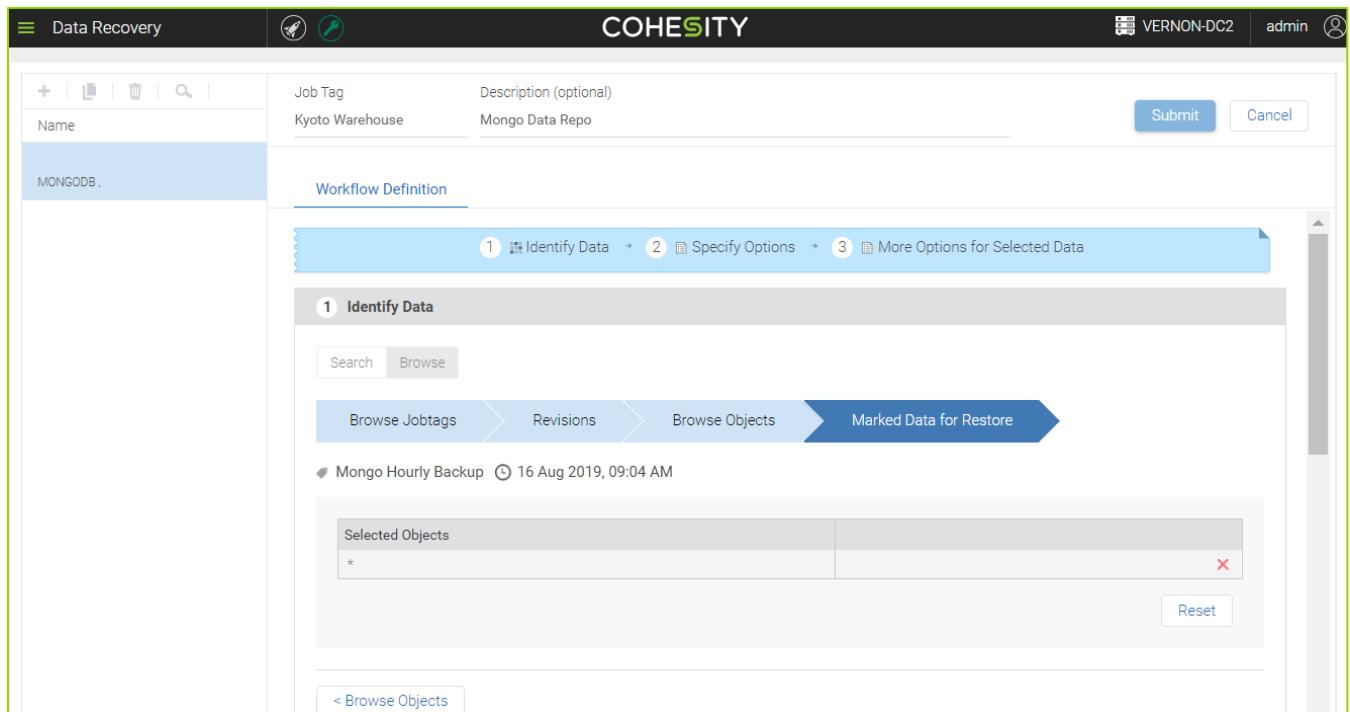
9. In the **Browse** tab, under the **Browse Objects** label, do one of the following:

- Select database or collection you want to restore and then click the **Next** button
- Click the **Change revision** button to go back to the Revisions tabs and select a new revision of the JobTag

The screenshot shows the Cohesity Data Recovery interface. At the top, there's a navigation bar with 'Data Recovery' on the left and a user icon on the right. The main area has a title 'COHESITY' and a sub-section 'Workflow Definition'. A progress bar at the top indicates three steps: '1 Identify Data', '2 Specify Options', and '3 More Options for Selected Data'. Step 1 is currently active. Below the progress bar, the 'Identify Data' section is titled '1 Identify Data'. It includes a 'Search' and 'Browse' button, and a breadcrumb navigation: 'Browse Jobtags' → 'Revisions' → 'Browse Objects' (which is highlighted in blue) → 'Marked Data for Restore'. Underneath, it says 'Mongo Hourly Backup 16 Aug 2019, 09:04 AM'. A note says 'Select the desired objects and click on "Mark For Restore" button to restore the objects.' A list of selected objects is shown in a table with checkboxes: 'Objects' (checked) and 'sales' (checked). At the bottom of the section are buttons for '< Change revision' and 'Next >'.

10. In the **Marked Data for Restore** label, do one of the following:

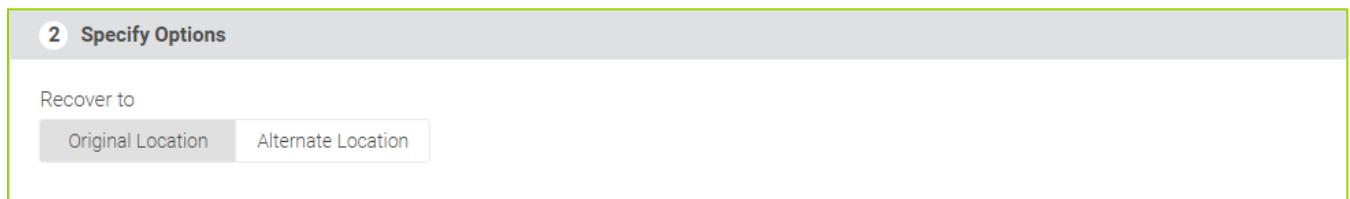
- Click the **X** icon to remove any data object from the list
- Click the **Browse Objects** button to go back and go through all the objects again



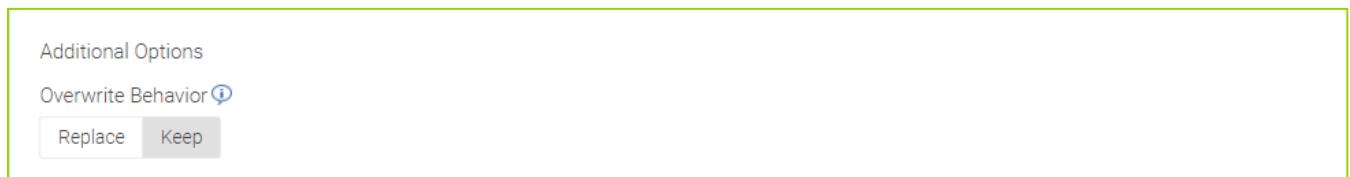
- In the **Specify Options** section, under the **Recover To** area, select **Original Location** or **Alternate Location**.

ORIGINAL LOCATION:

- Click the **Original Location** button.



- In the **Additional Options** area, under **Overwrite Behavior** do one of the following:
 - Click **Replace** to replace existing data with existing data with new data thus erasing any previously existing data
 - Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied



- Type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster in the **Suffix** field. For example, _10102017.

A screenshot of a software interface showing a single-line input field. The field is labeled "Suffix" with a blue information icon. The value "24092019" is typed into the field. The entire input area is enclosed in a light gray border.

ALTERNATE LOCATION:

- Click the **Alternate Location** button.

A screenshot of a software interface showing a step titled "2 Specify Options". Below it is a section labeled "Recover to" with two buttons: "Original Location" and "Alternate Location". The "Alternate Location" button is highlighted with a blue border, indicating it is selected. The entire interface is enclosed in a light gray border.

- Select a MongoDB cluster name from the **Data Repository** drop-down menu where you want to recover the data set.

A screenshot of a software interface showing a dropdown menu labeled "Data Repository". The option "Mongo_QA" is selected and highlighted with a blue border. The entire dropdown menu is enclosed in a light gray border.

- In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

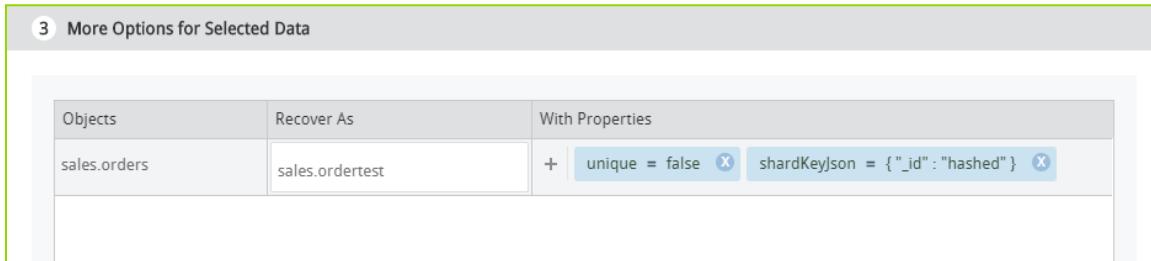
- Click **Replace** to replace existing data with new data with the existing data and thus erasing any previously existing data
- Click **Keep** to retain the new data along with the existing data thereby retaining both the new and old data sets

A screenshot of a software interface showing an "Additional Options" section. It includes a label "Overwrite Behavior" with a blue information icon, and two buttons: "Replace" and "Keep". The "Replace" button is highlighted with a blue border, indicating it is selected. The entire options area is enclosed in a light gray border.

- Type a number and/or character as a suffix to the MongoDB data objects being recovered from the Imanis Data cluster in the **Suffix** field. For example, _10102017.

A screenshot of a software interface showing a single-line input field. The field is labeled "Suffix" with a blue information icon. The value "24092019" is typed into the field. The entire input area is enclosed in a light gray border.

14. In the **More Options for Selected Data**, you can specify Collection specific properties ('shardKeyJson' and 'unique') in the GUI. See the following screenshot:

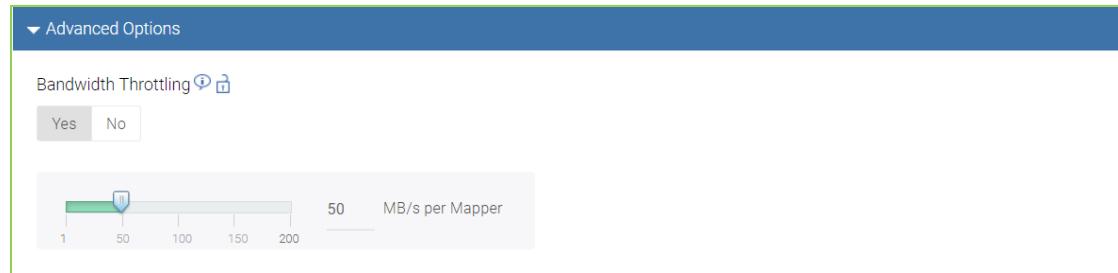


IMPORTANT: The following information regarding limitations of specifying collection specific properties ('shardKeyJson' and 'unique') must be noted:

1. Currently, these properties are not auto detected from Primary in the case of Mirroring workflow. However, these properties can be specified in the 'Rename with Properties' section while creating a Mirroring job. These values will then override any existing configuration on the relevant collections.
2. In case of Restore, the existing values for these properties, if any, are displayed, that is, the values for these properties when the collections were backed up are displayed. You can choose to add, update, delete these property values.
3. The Imanis Data GUI does not enforce any relationship between these two properties while collecting this information from the user. The user is expected to input a valid combination of the two properties. If any of the parameters left blank, corresponding value from the source collection would be used. The Imanis Data GUI validates that the user has entered a valid Json as the value for shardKeyJson. However, the schema for this Json is not validated. Similarly, GUI enforces that as the value for 'unique' property, the only two values that are permitted to be entered are true or false.

15. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

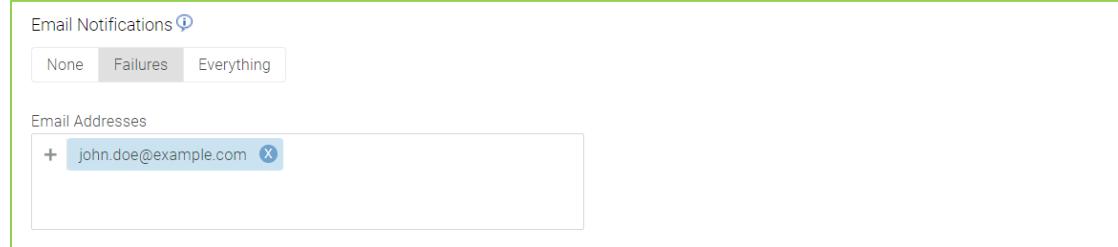
- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



9.3 Recovery Sandbox

You can create a recovery sandbox workflow to test the data recoverability of your restore points. This workflow lets you verify that data of restore points is recoverable by replaying essential components of an actual recovery workflow inside a sandbox environment. The entire execution happens as a dry run inside the Imanis cluster and no data is sent outside.

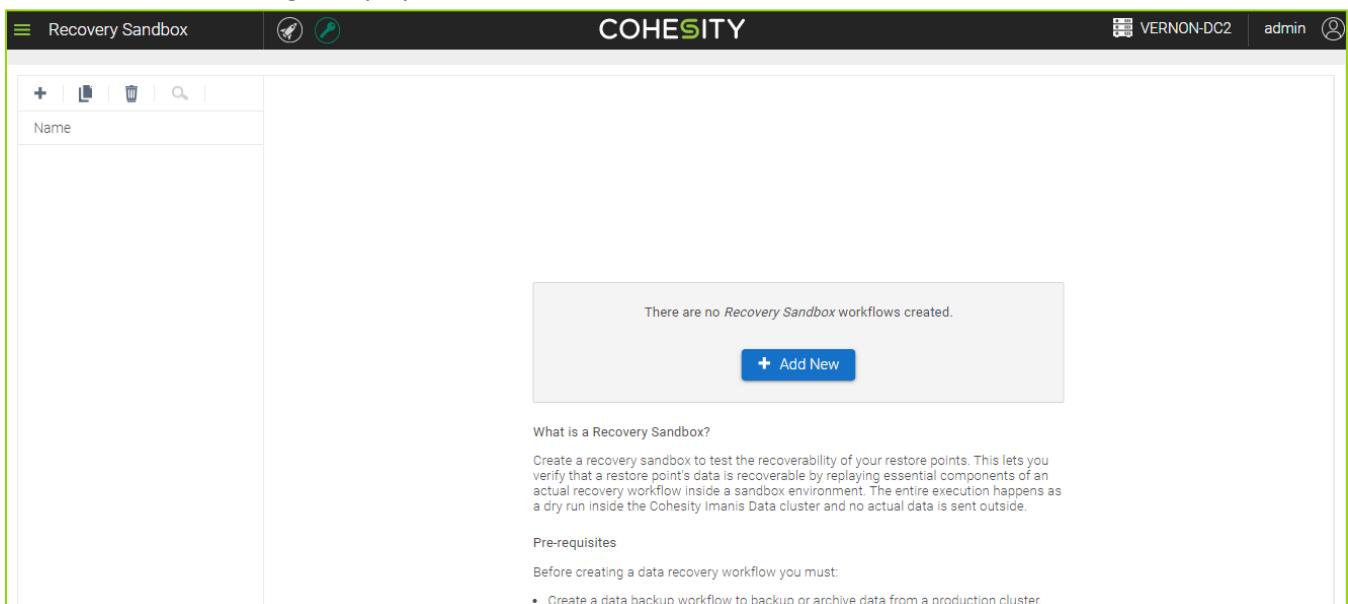
In the latest release, you can test the data recoverability with the Point in Time (PIT) feature inside a sandbox environment for Couchbase and Cassandra only.

9.3.1 Couchbase

The following section discusses the steps to use Sandbox recovery feature for Couchbase.

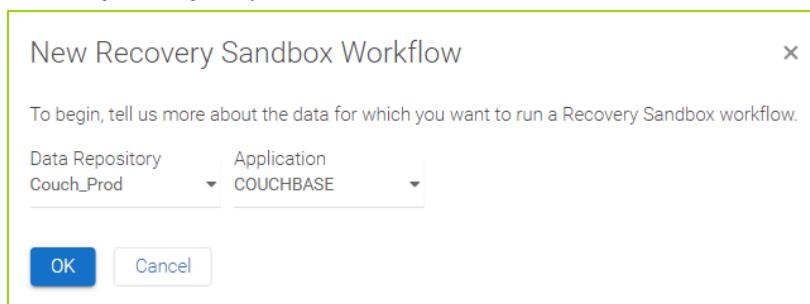
To run recovery sandbox, do the following:

1. Click the **Main Menu**  > **Monitoring and Recovering** > **Recovery Sandbox**
2. On the Data Recovery page, click the  **+ Add New** button or the  icon. The **New Recovery Sandbox Workflow** dialog is displayed.



The screenshot shows the Cohesity UI with the title bar "COHESITY". In the top navigation bar, there are icons for Home, Monitoring, Recovering, and Recovery Sandbox. The "Recovery Sandbox" icon is highlighted. The main content area is titled "Recovery Sandbox". On the left, there is a sidebar with icons for Create, Edit, Delete, and Search. The main pane shows a table with one row, which has been collapsed. A message box states: "There are no Recovery Sandbox workflows created." with a "+ Add New" button. Below this, a "What is a Recovery Sandbox?" section explains its purpose and provides prerequisites: "Before creating a data recovery workflow you must:" followed by a bullet point: "Create a data backup workflow to backup or archive data from a production cluster."

3. In the **New Recovery Sandbox Workflow** dialog, select a **Couchbase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.



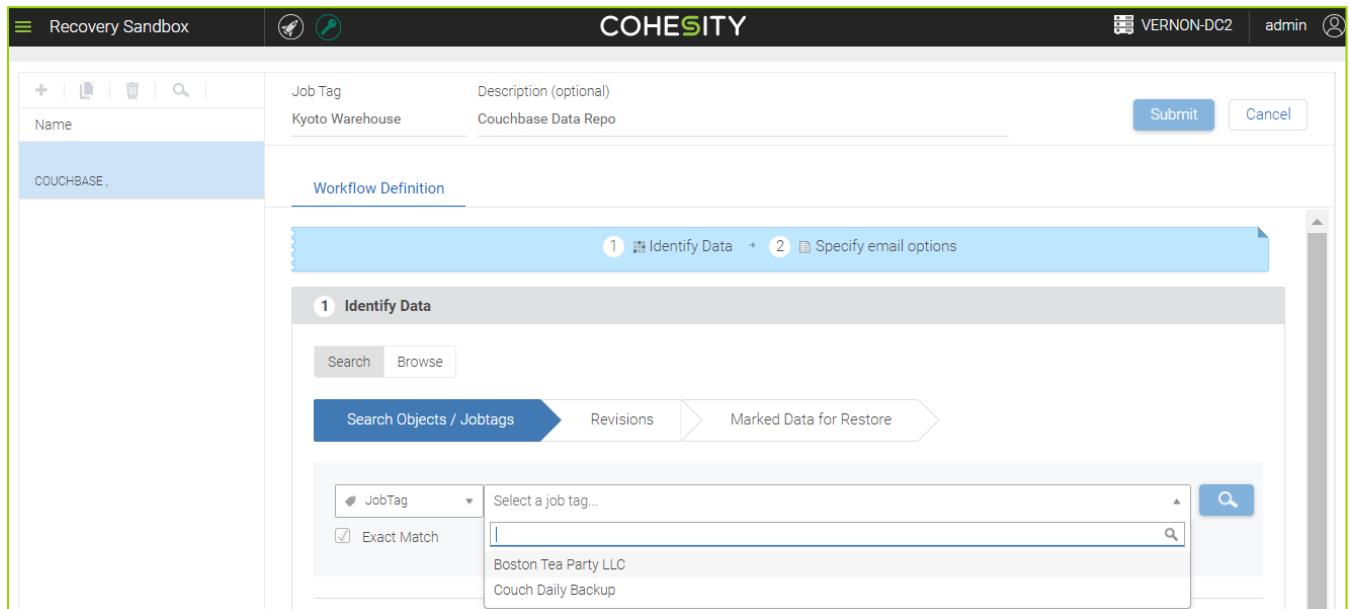
4. In the **Recovery Sandbox** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field.

5. In the **Identify Data** section, Search and Browse tabs are displayed. You can use the **Search** and **Browse** button as per your requirement:

SEARCH	BROWSE
Use when you know the data object that you wish to recover	Use when you want to view data objects and select specific data objects within a JobTag revision
Search specific Jobtag and Bucket	Browse the data objects catalog and select or deselect multiple data objects at a single time

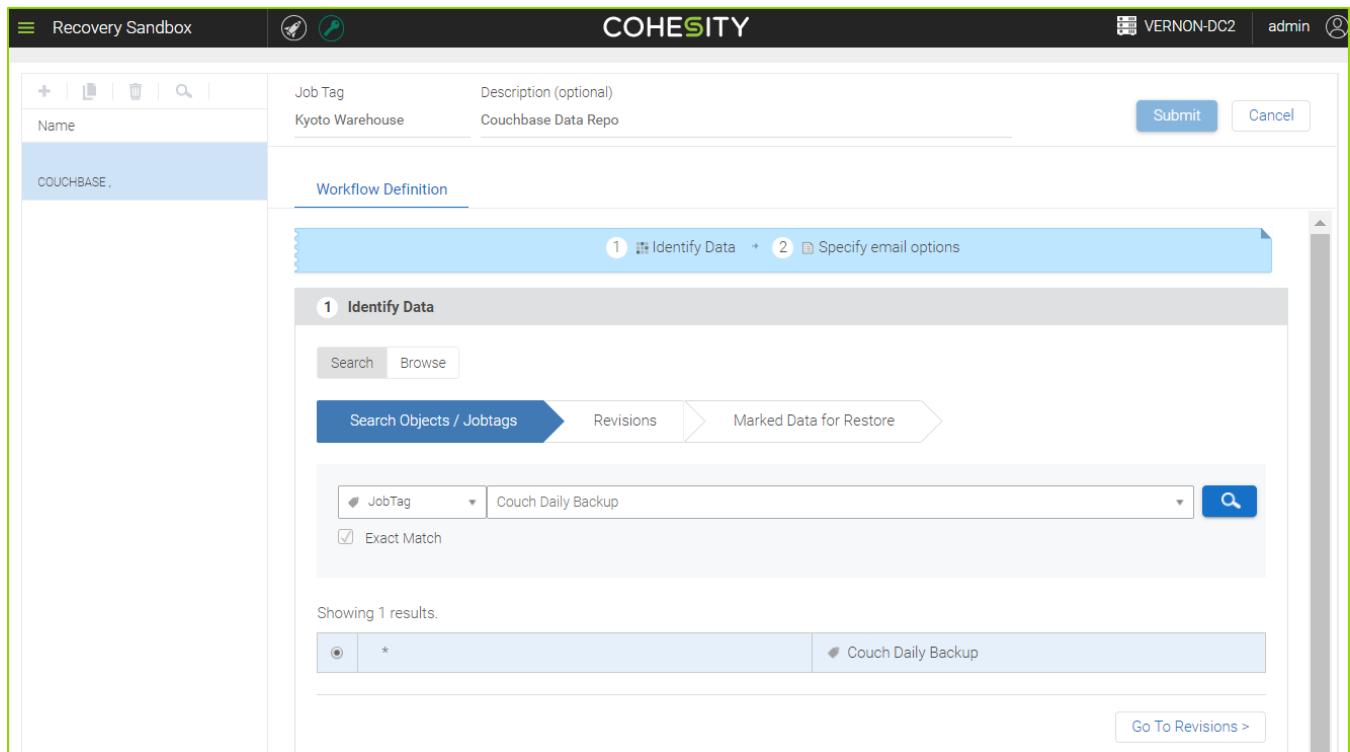
Search Tab

6. In the **Search Objects/Jobtags** tab, click the search box and select a job tag displayed by Imanis Data, and then click the  icon. Imanis Data will display the jobtag data object as a search result.



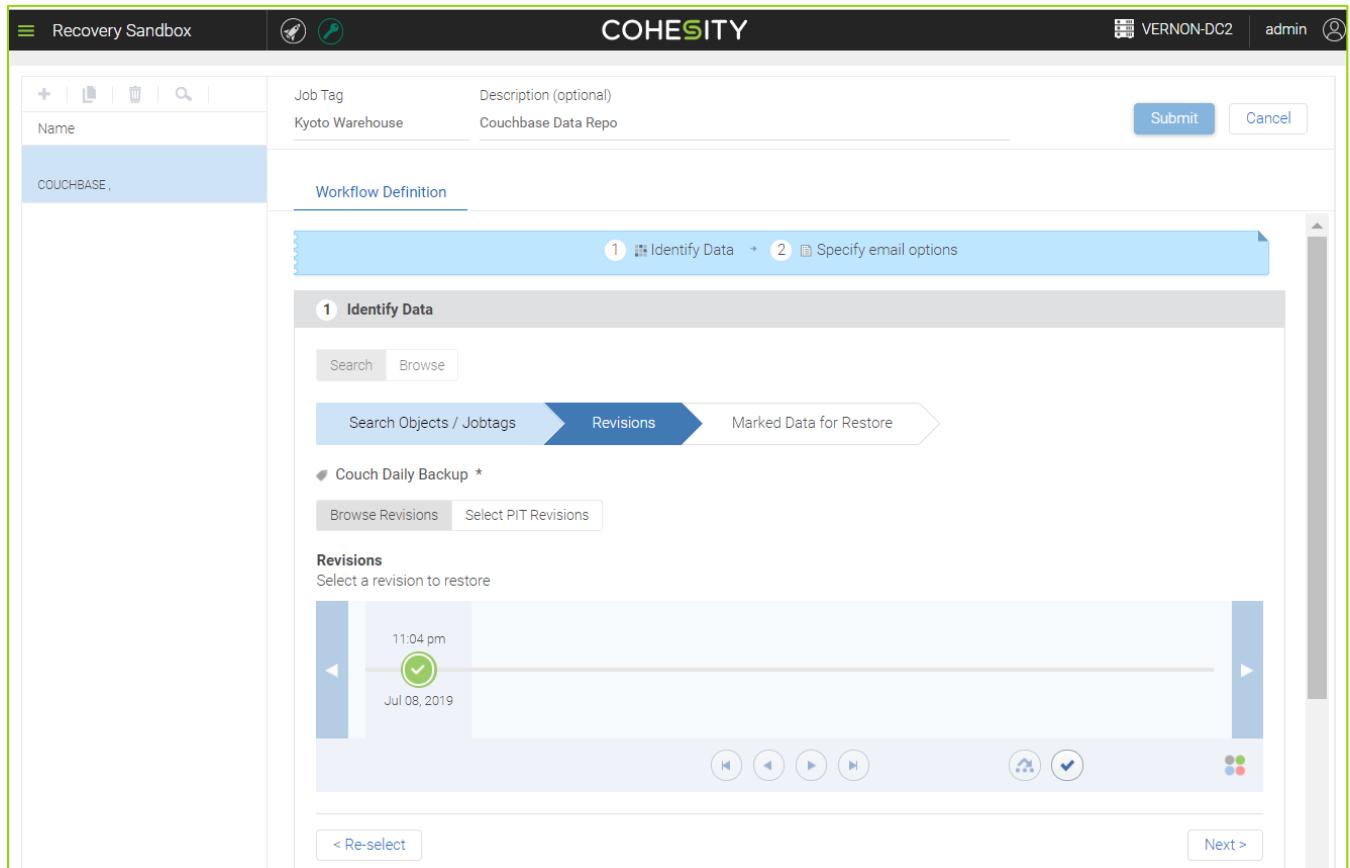
The screenshot shows the Cohesity Recovery Sandbox interface. At the top, there's a navigation bar with 'Recovery Sandbox', a search icon, and the Cohesity logo. On the right, it shows 'VERNON-DC2' and 'admin'. Below the navigation, there's a form for creating a new job tag, with fields for 'Name' (set to 'COUCHBASE'), 'Job Tag' (set to 'Kyoto Warehouse'), and 'Description (optional)' (set to 'Couchbase Data Repo'). There are 'Submit' and 'Cancel' buttons. Underneath this, a 'Workflow Definition' section is visible, showing a blue header bar with steps: 1. Identify Data → 2. Specify email options. The main area is titled '1 Identify Data' and contains a 'Search' tab selected. It shows a search bar with 'Select a job tag...' and a dropdown menu with 'JobTag' and 'Exact Match' checked. A search result list shows 'Boston Tea Party LLC' and 'Couch Daily Backup'. A search icon is located to the right of the search bar.

7. Select the radio button of the jobtag data object and click the **Go to Revisions** button.



8. In the **Revisions** tab, click the **Browse Revisions**, do one of the following:

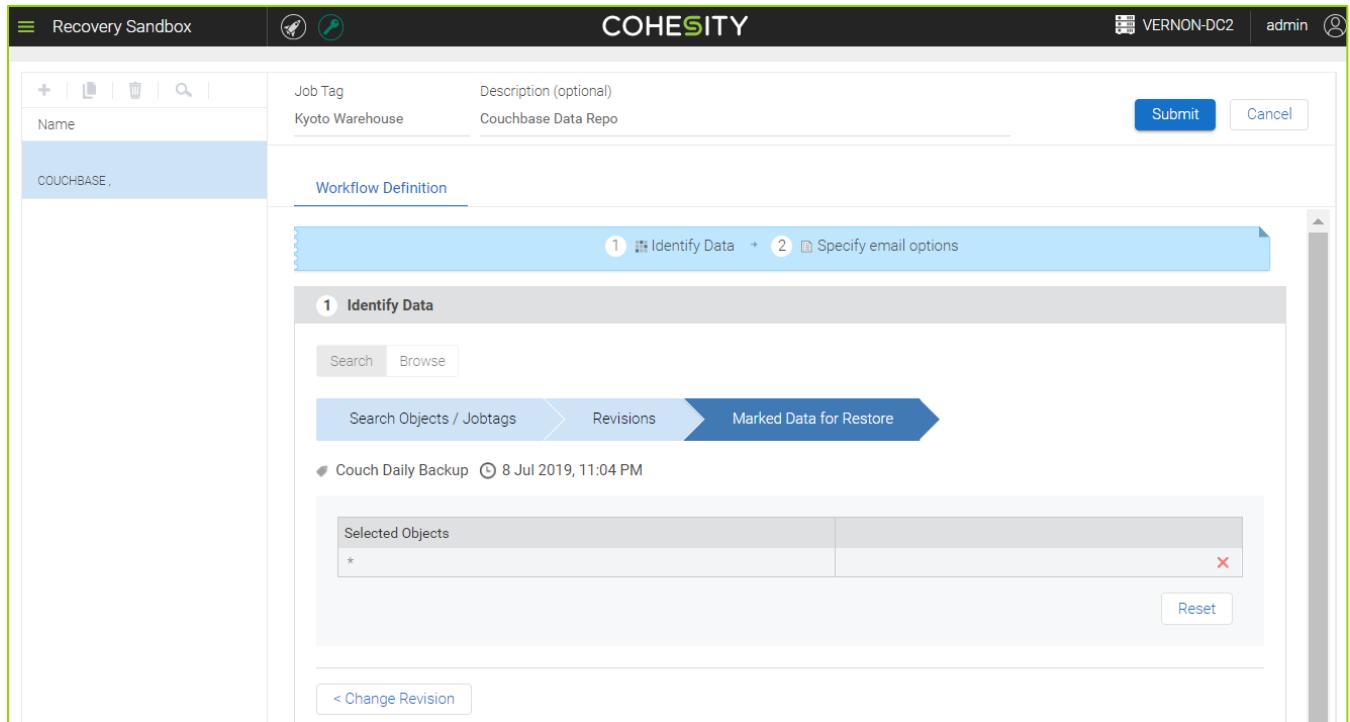
- By default, the latest copy is selected which is indicated by the icon. You can then click the **Next** button below to restore the selected data object
- Click the data object icon to select a copy of data for a specific day and time. You can then click the **Next** button to restore the selected data object



NOTE: Navigate all the data object revision by clicking the icons. You can also click the icon to jump to a specific revision in time by selecting a date and time or click the icon to jump to the currently selected revision.

9. In the **Marked Data for Restore** tab, do one of the following:

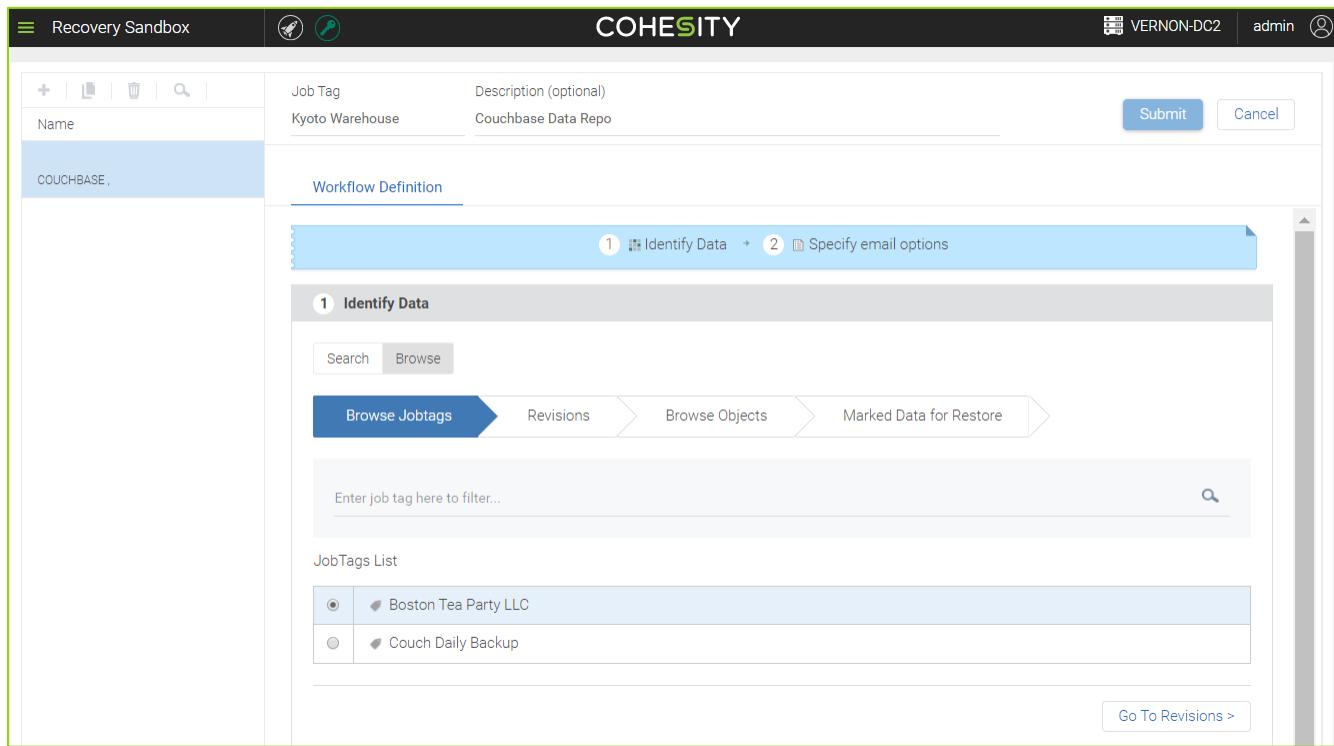
- Click the **Reset** button
- Click the **Change revision** button to go back to the Revisions tabs and select a new revision of the JobTag



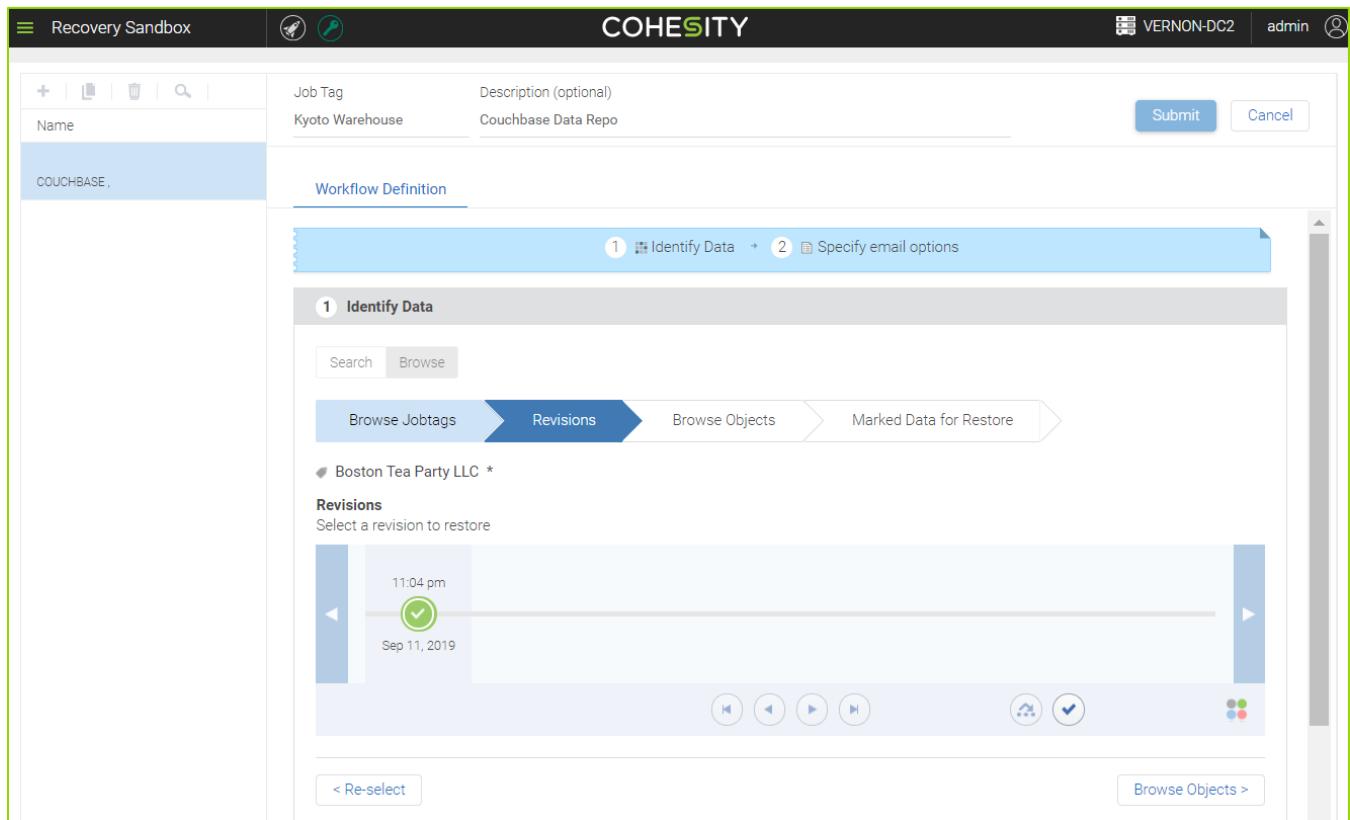
You can jump directly onto Step # 12 to continue Couchbase sandbox data recovery for JobTag.

Browse Tab

7. In the **Browse Jobtags** tab, select a **JobTag** from the **JobTag** list, and then click the **Go To Revisions** button.

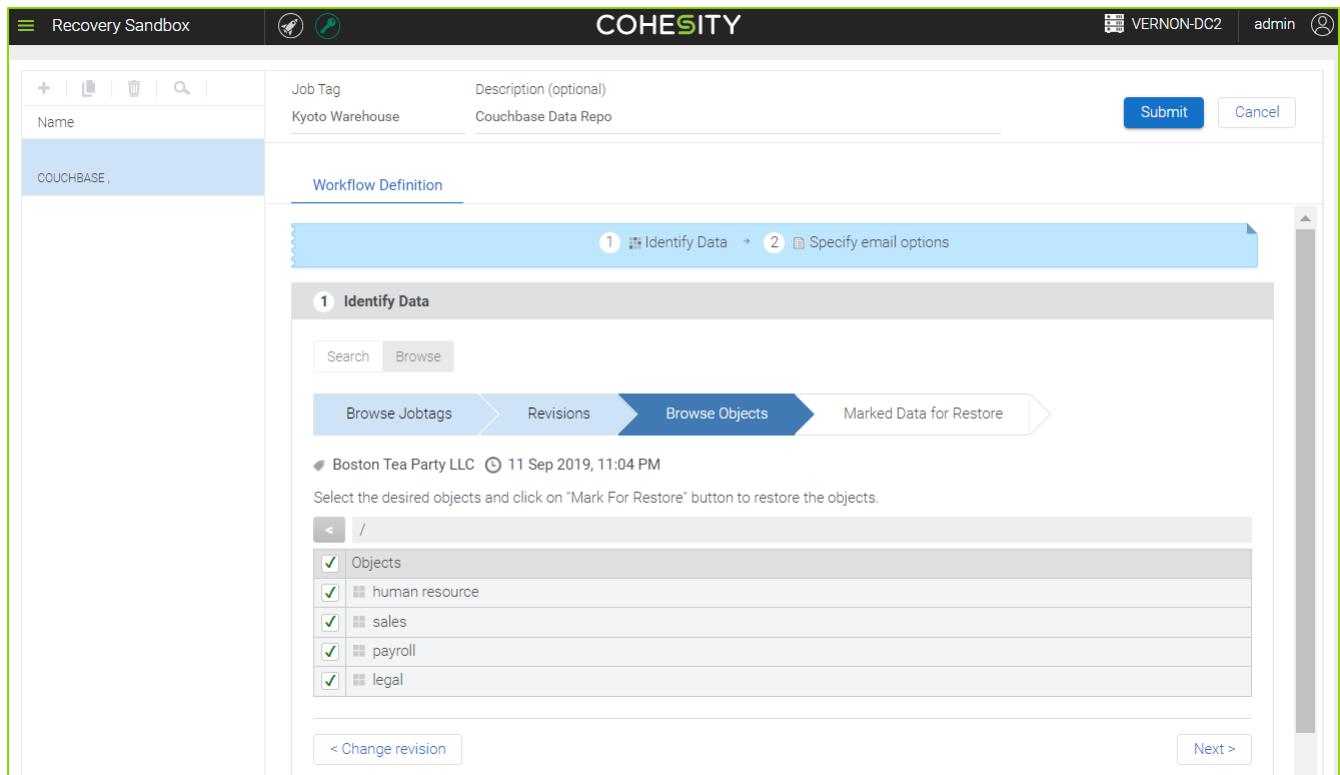


8. In the **Revisions** label, select a revision of the JobTag revision that you want to restore and then click the **Browse Objects** button.



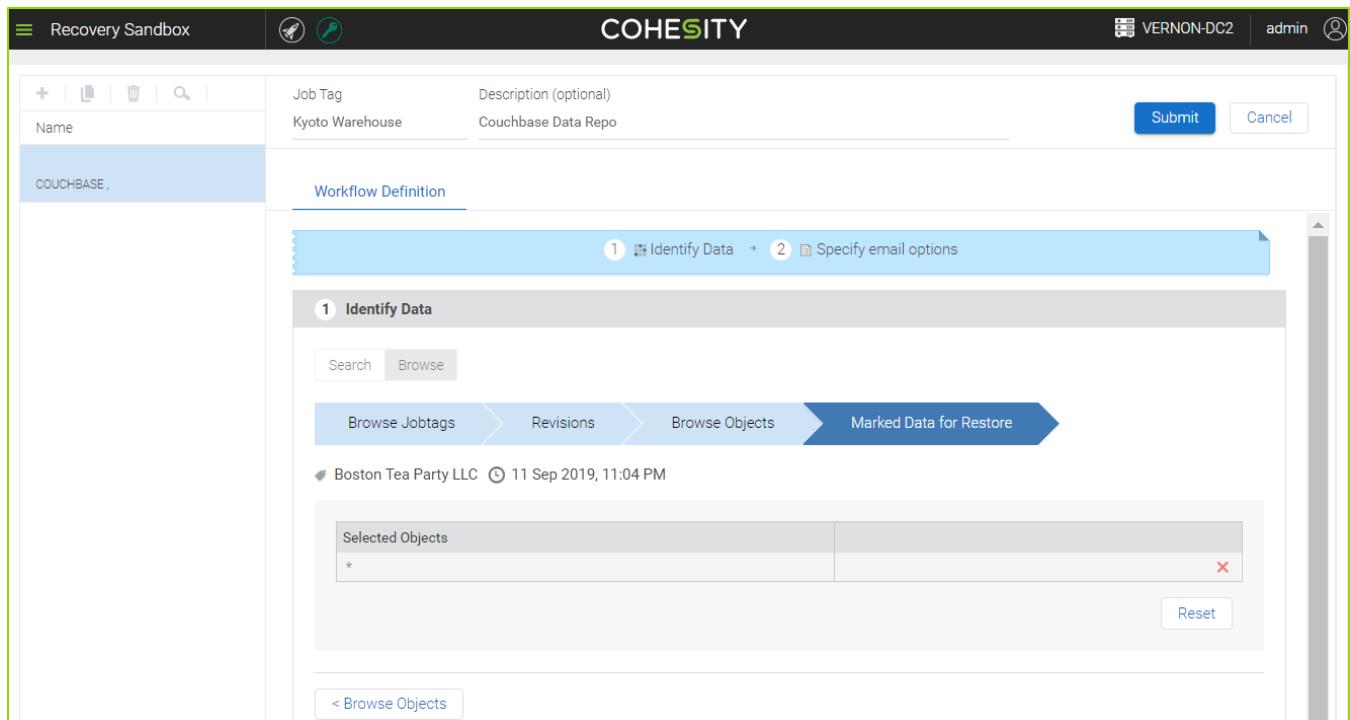
9. In the **Browse Objects** label, do one of the following:

- Select or clear the check boxes of the data objects that you want to restore and then click the **Next** button
- Click the **Change revision** button to go back to the Revisions tabs and select a new revision of the JobTag.



10. In the **Marked Data for Restore** label, do one of the following:

- Click the **Reset** button to go back to the **Browse Jobtags** tab
- Click the **Browse Objects** button to go back to the Revisions tab to reselect a data object revision to restore



11. In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:

Email Notifications ?

None Failures Everything

Email Addresses

+ john.doe@example.com X

12. Click **Submit**.

9.3.2 Cassandra

The following section discusses the steps to use Sandbox recovery feature for Cassandra. Point-in-Time (PIT) recovery is also supported in Sandbox recovery. The Sandbox recovery feature is supported for DSE 5.x (Cassandra 3.x) onwards only.

To run recovery sandbox, do the following:

1. Click the **Main Menu** > **Monitoring and Recovering** > **Data Recovery**.
2. On the **Data Recovery** page, click the **+ Add New** button or the icon. **New Data Recovery Workflow** dialog box appears.

The screenshot shows the Cohesity Imanis Data interface. At the top, there is a navigation bar with icons for search, refresh, and user profile, and the text "COHESITY" in the center. On the far right of the navigation bar, it says "VERNON-DC2" and "admin". Below the navigation bar, the main content area has a header "Recovery Sandbox". To the left of the main content area, there is a sidebar with a search bar and a table header "Name". The main content area contains a message box stating "There are no Recovery Sandbox workflows created." with a blue "Add New" button. Below this message box, there is a section titled "What is a Recovery Sandbox?" followed by a detailed description of what a recovery sandbox is and how it works. At the bottom of this section, there is a "Pre-requisites" heading and a list of requirements.

There are no Recovery Sandbox workflows created.

[+ Add New](#)

What is a Recovery Sandbox?

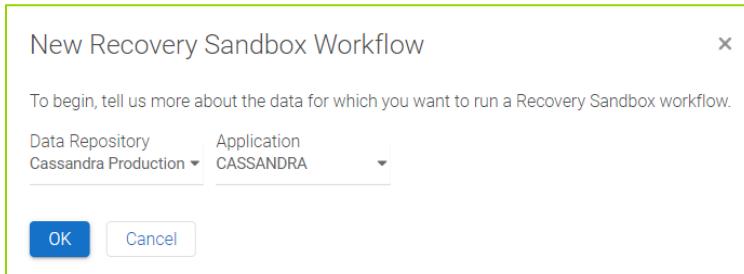
Create a recovery sandbox to test the recoverability of your restore points. This lets you verify that a restore point's data is recoverable by replaying essential components of an actual recovery workflow inside a sandbox environment. The entire execution happens as a dry run inside the Cohesity Imanis Data cluster and no actual data is sent outside.

Pre-requisites

Before creating a data recovery workflow you must:

- Create a data backup workflow to backup or archive data from a production cluster.

3. In the **New Data Recovery Workflow** dialog box, select a **Cassandra** data repository from the **Data Repository** drop-down menu, select Cassandra from the **Application** field, and then click **OK**.

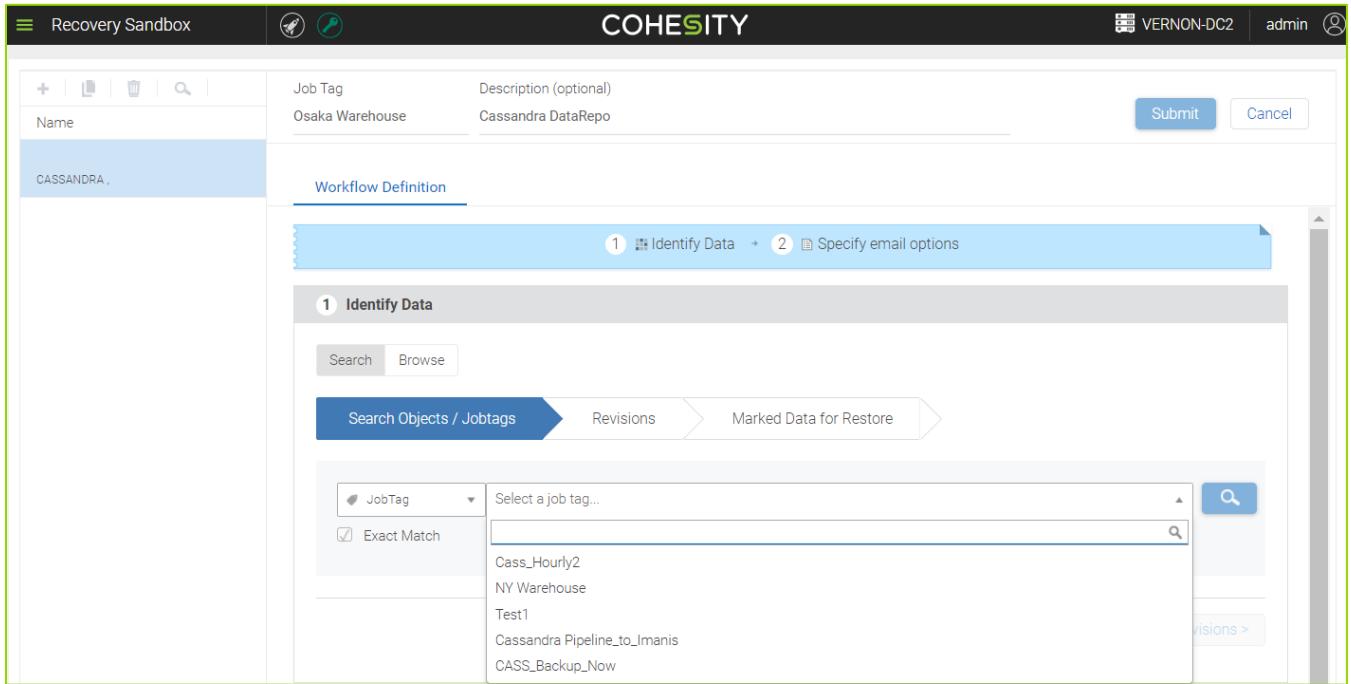


4. In the Recovery Sandbox page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field.
5. In the **Identify Data** section, Search and Browse tabs are displayed. You can use the Search and Browse button as per your requirement:

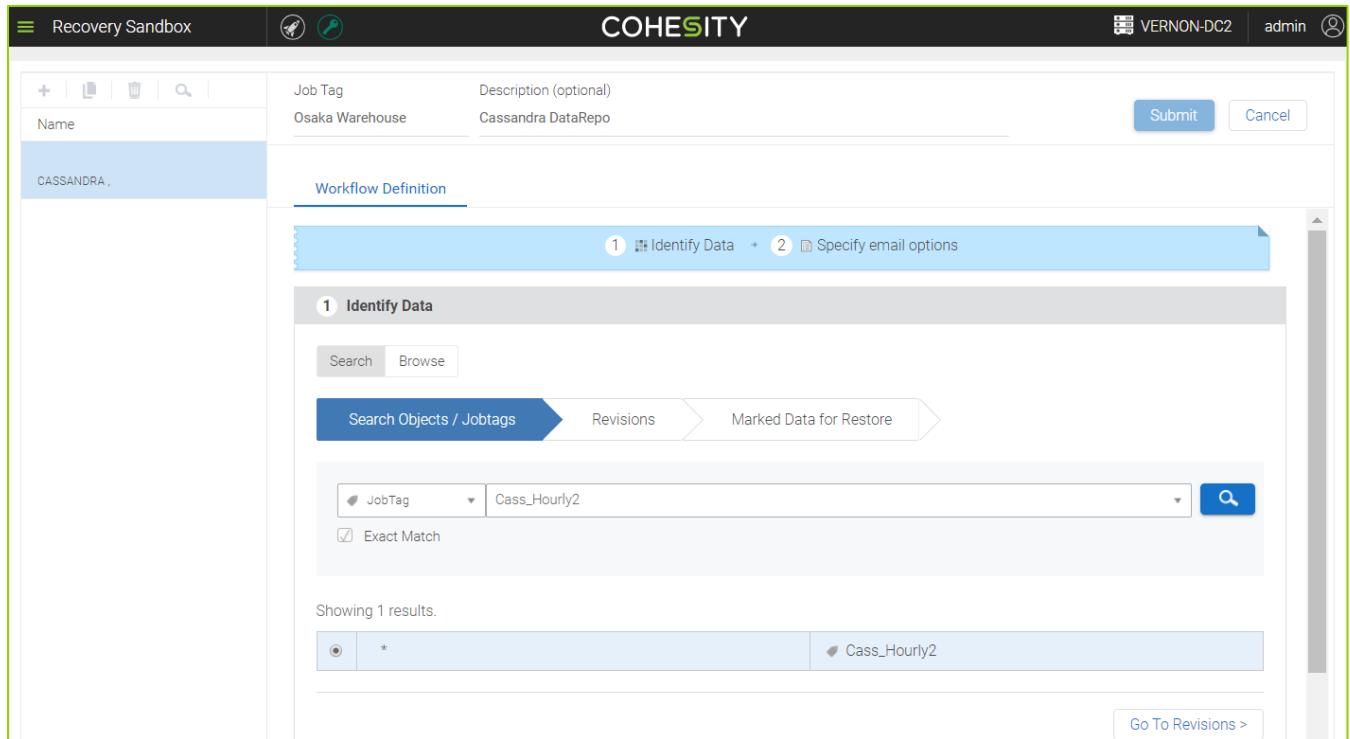
SEARCH	BROWSE
Use when you know the data object that you wish to recover.	Use when you want to view data objects and select specific data objects within a JobTag revision.
Search specific Jobtag, Keyspace, and Table.	Browse the data objects catalog and select or deselect multiple data objects at a single time.

Search Tab

6. In the **Search Objects/Jobtags** tab, click the search box and select a job tag displayed by Cohesity Imanis Data, and then click the search icon. Cohesity Imanis Data will display the jobtag data object as a search result.



7. Select the radio button of the jobtag data object and click the **Go to Revisions** button.

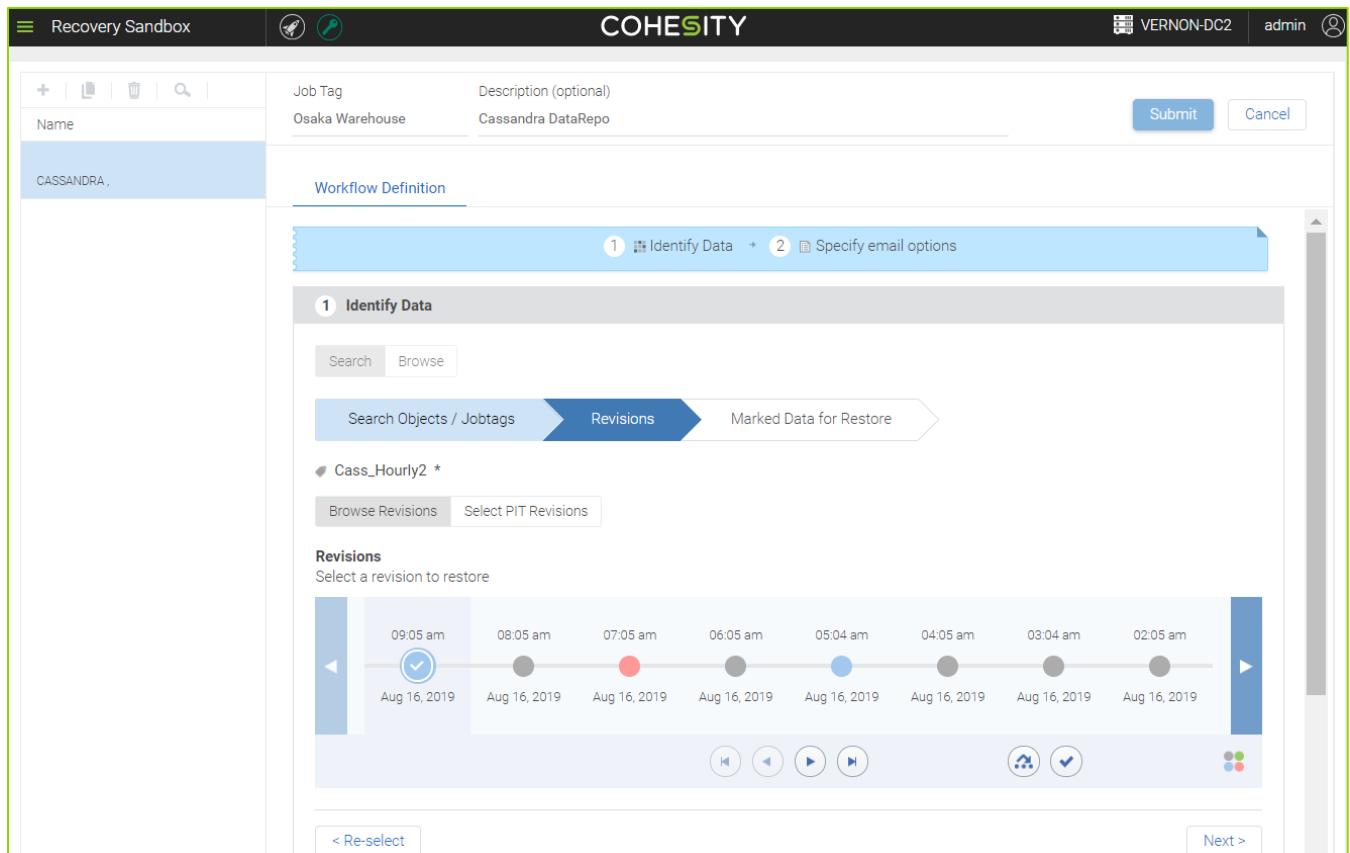


- Select **Keyspace**, type the keyspace name, and then click the Go To Revisions button
- Select **Table**, type the table name, and then click the Go To Revisions button

NOTE: A search that is based on an exact or a partial term is enabled for Tables and Keyspaces only.

8. In the **Revisions** tab, click the **Browse Revisions**, do one of the following:

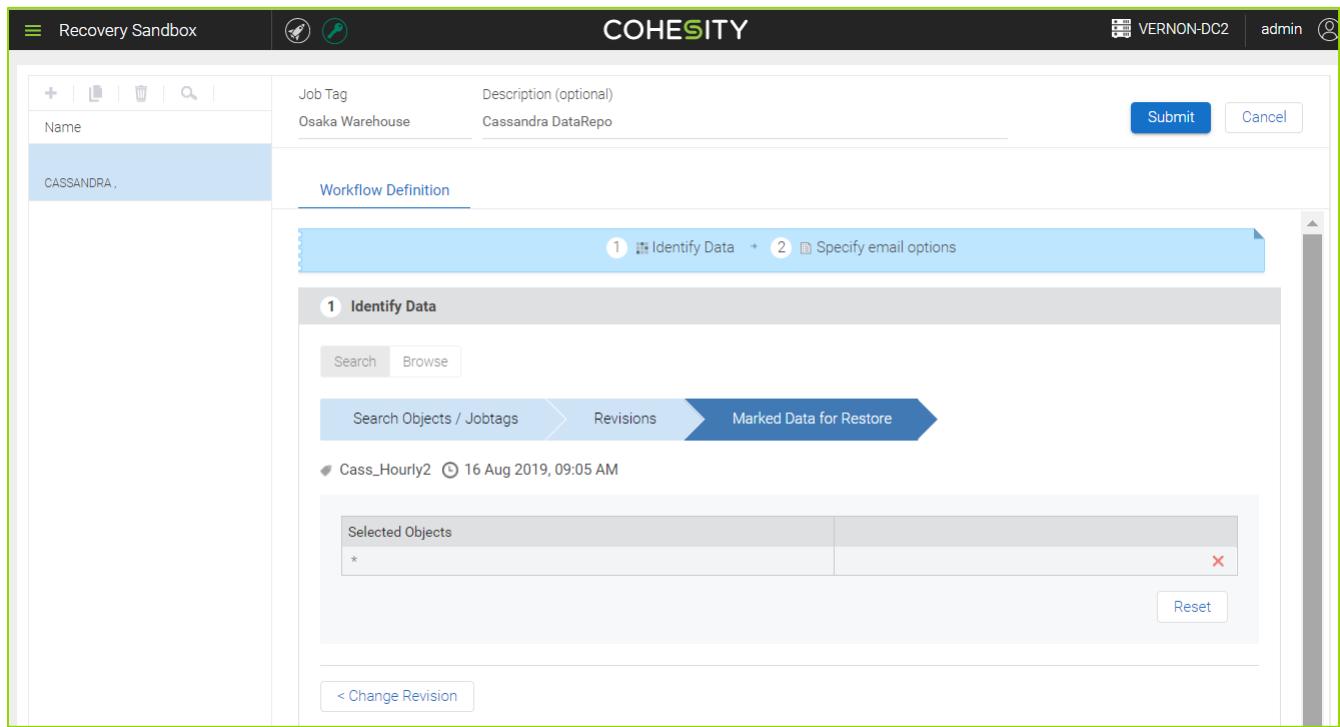
- By default, the latest copy is selected which is indicated by the icon. You can then click the **Next** button below to restore the selected data object
- Click the data object icon to select a copy of data for a specific day and time. You can then click the **Next** button to restore the selected data object



NOTE: Navigate all the data object revision by clicking the icons. You can also click the icon to jump to a specific revision in time by selecting a date and time or click the icon to jump to the currently selected revision.

9. In the **Marked Data for Restore** tab, do one of the following:

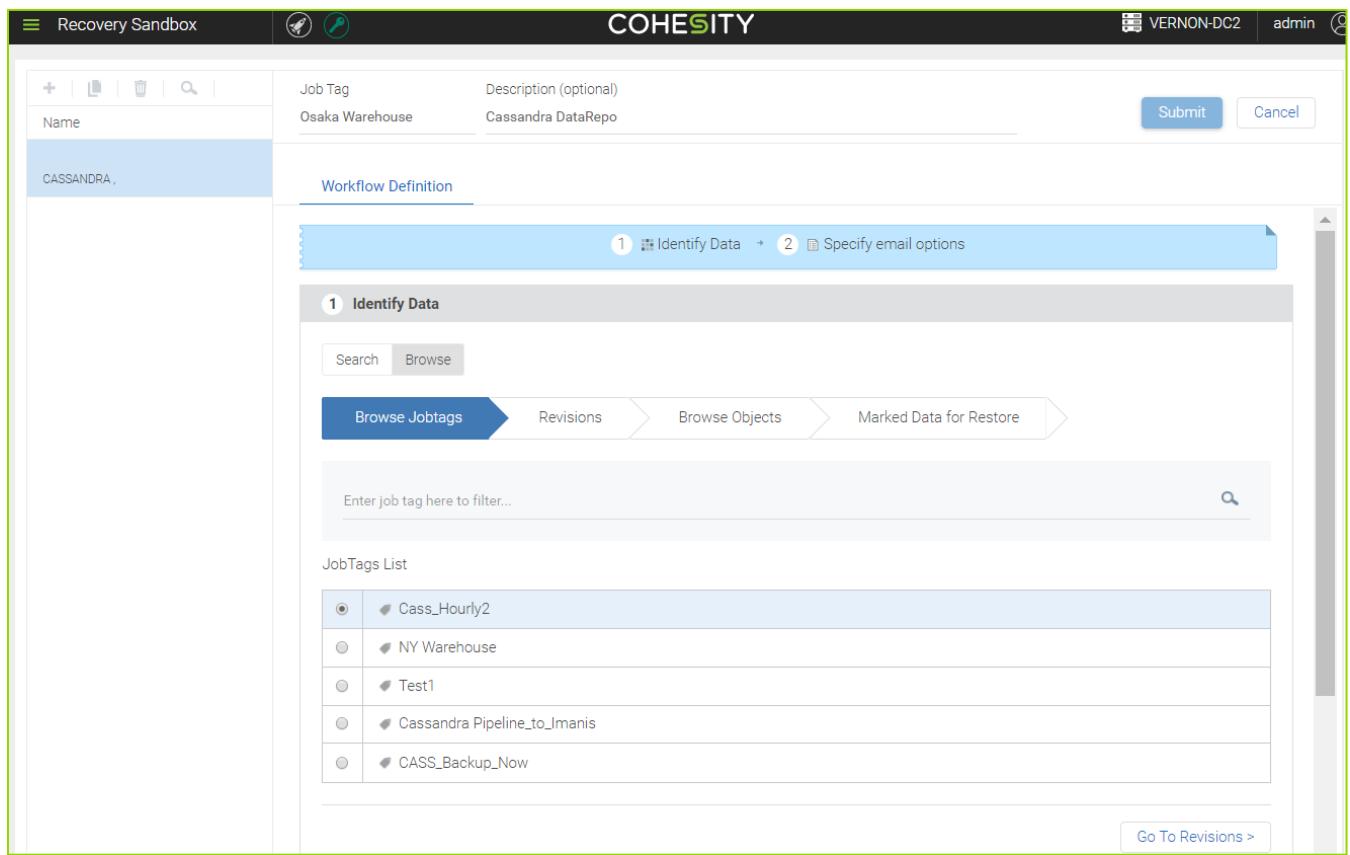
- Click the **Reset** button to go back to the **Search Objects/Jobtags** tab
- Click the **Change Revision** button to back to **Revisions** tab to reselect a data object revision to restore



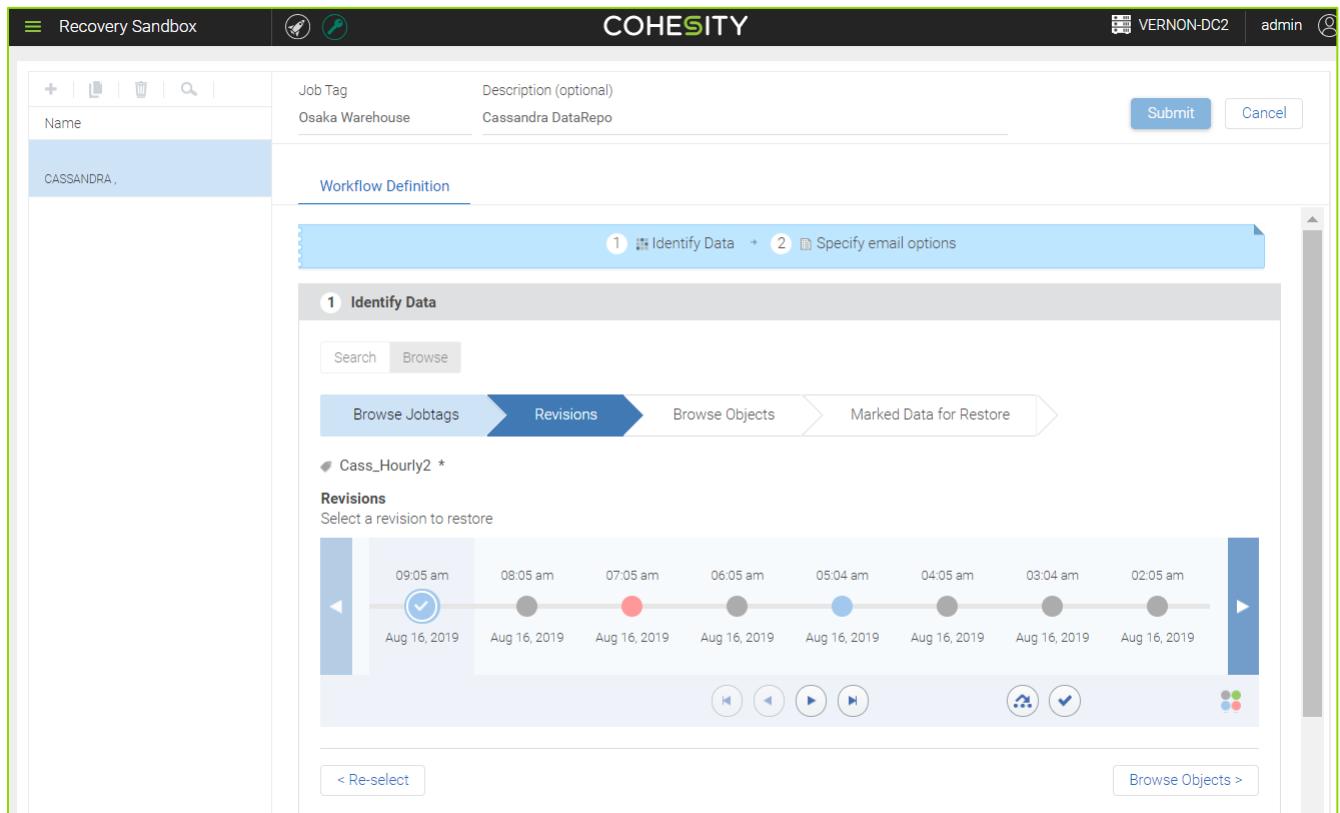
You can jump directly onto Step # 11 to continue Cassandra data recovery for JobTag.

Browse Tab

7. In the **Browse Jobtags** tab, select a **JobTag** from the **JobTag** list, and then click the **Go To Revisions** button.

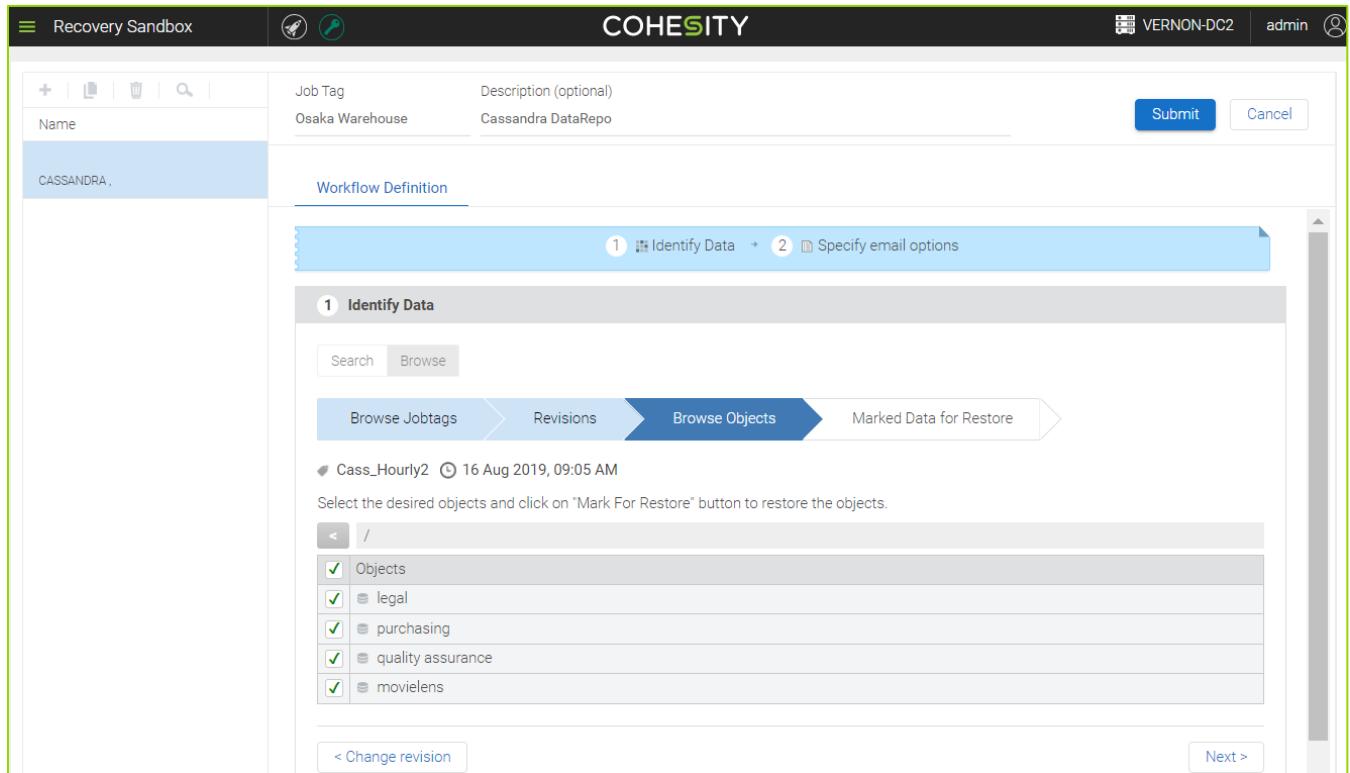


8. In the **Revisions** label, select a revision of the JobTag revision that you want to restore and then click the **Browse Objects** button.



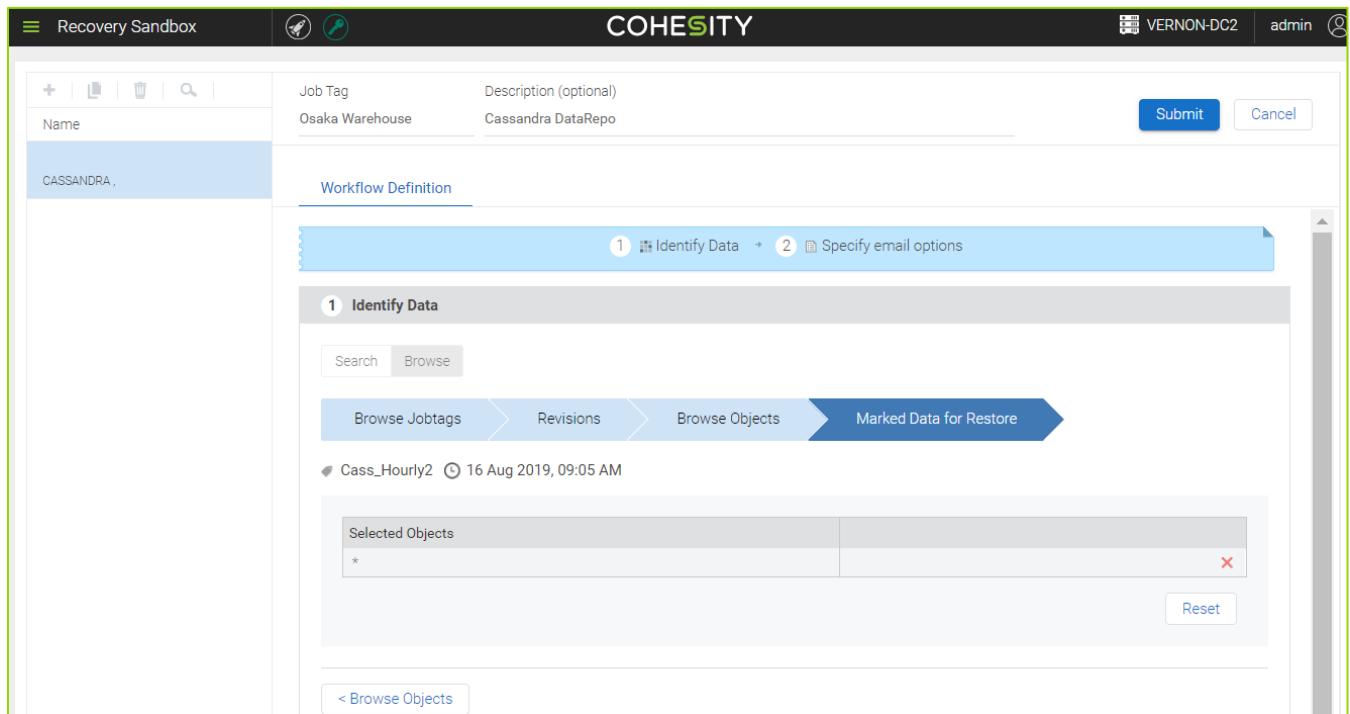
9. In the **Browse Objects** label, do one of the following:

- Select keyspaces or tables that you want to restore and then click the **Next** button
- Click the **Change revision** button to go back to the Revisions tabs and select a new revision of the JobTag



10. In the **Marked Data for Restore** label, do one of the following:

- Click the **Reset** button to go back to the Browse Jobtags tab
- Click the **Browse Objects** button to go back to the Revisions tab to reselect a data object revision to restore



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:

Email Notifications ⓘ

None Failures Everything

Email Addresses

+ john.doe@example.com X

- Click **Submit**.

9.3.3 Hadoop

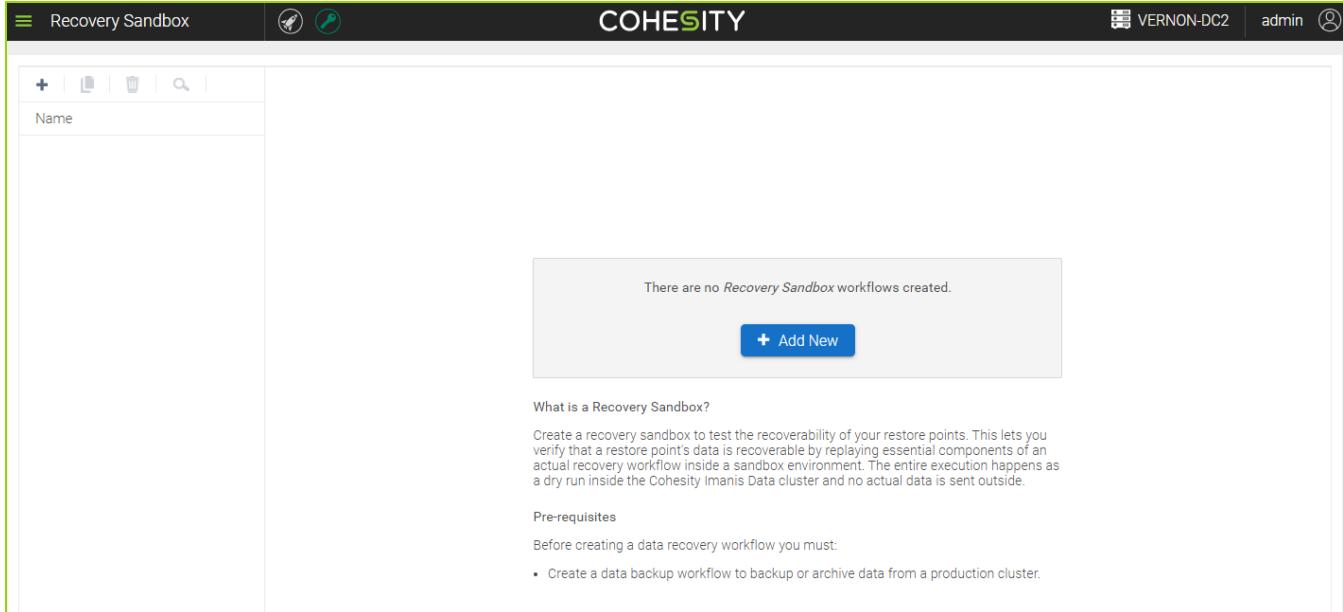
The following section discusses the steps to use Sandbox recovery feature for Hadoop.

You can create a recovery sandbox workflow to test the data recoverability of your restore points. This workflow lets you verify that data of restore points is recoverable by replaying essential components of an actual recovery workflow inside a sandbox environment. The entire execution happens as a dry run inside the Imanis cluster and no data is sent outside.

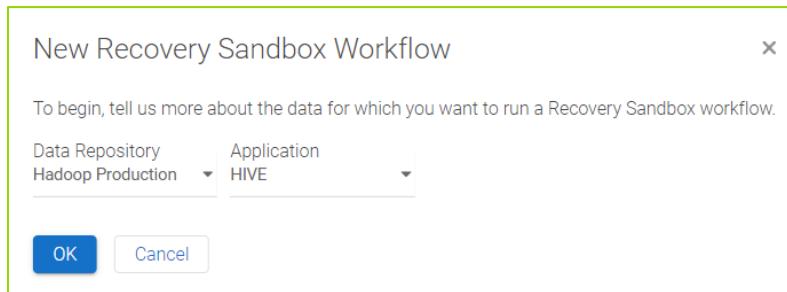
The follow example discusses Hive; however, you can test data recoverability of your restore points for HDFS and HBase too. The Browse tab is not supported for HDFS.

To run recovery sandbox, do the following:

1. Click the **Main Menu** > **Monitoring and Recovering** > **Recovery Sandbox**
2. On the **Data Recovery** page, click the **+ Add New** button or the **+** icon. The **New Recovery Sandbox Workflow** dialog is displayed.



3. In the **New Recovery Sandbox Workflow** dialog, select a **Hadoop** source data repository from the **Data Repository** drop-down menu, select one **HDFS/Hive/HBase** from the **Application** field, and then click **OK**.



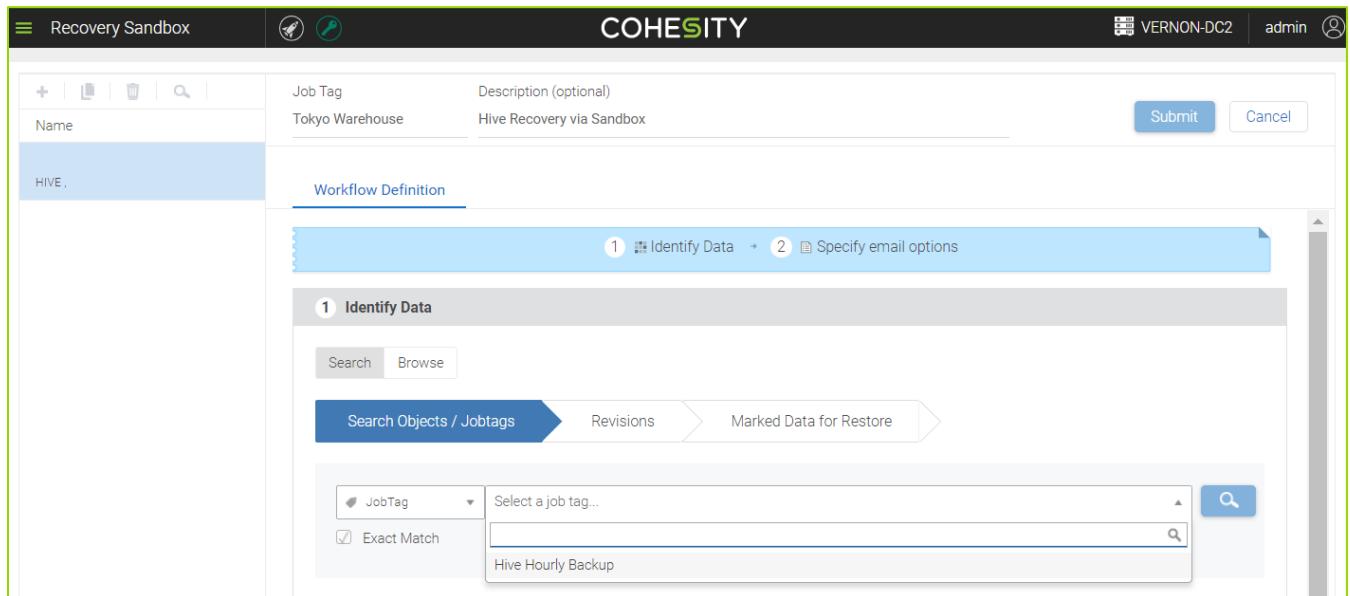
4. In the Recovery Sandbox page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.

5. In the **Identify Data** section, **Search** and **Browse** tabs are displayed. You can use the **Search** and **Browse** button as per your requirement:

SEARCH	BROWSE
Use when you know the data object that you wish to recover	Use when you want to view data objects and select specific data objects within a JobTag revision
Search specific Jobtag, Database, Table, or Partition	Browse the data objects catalog and select or deselect multiple data objects at a single time

Search Tab

6. In the **Search Objects/Jobtags** label, click the search box, select a **JobTag** displayed by Imanis Data, and then click the  icon. JobTag search result is displayed.



The screenshot shows the Cohesity Recovery Sandbox interface. At the top, there's a navigation bar with icons for Home, Recovery, and Support, followed by the COHESITY logo and user information (admin). Below the navigation is a search bar with fields for Name and Description (optional), and buttons for Submit and Cancel. A sidebar on the left lists categories like HIVE, SERVERS, and VOLUMES. The main area is titled 'Workflow Definition' and shows step 1: Identify Data. It includes tabs for Search and Browse, and a progress bar indicating step 1 (Identify Data) is active. Below this, there's a search interface with a dropdown for 'JobTag' (set to 'JobTag') and a search input field containing 'Hive Hourly Backup'. There are also checkboxes for 'Exact Match' and a search icon.

Similarly, you can select **Database**, **Table**, and **Partition** from the drop-down list by typing the full or partial file name.

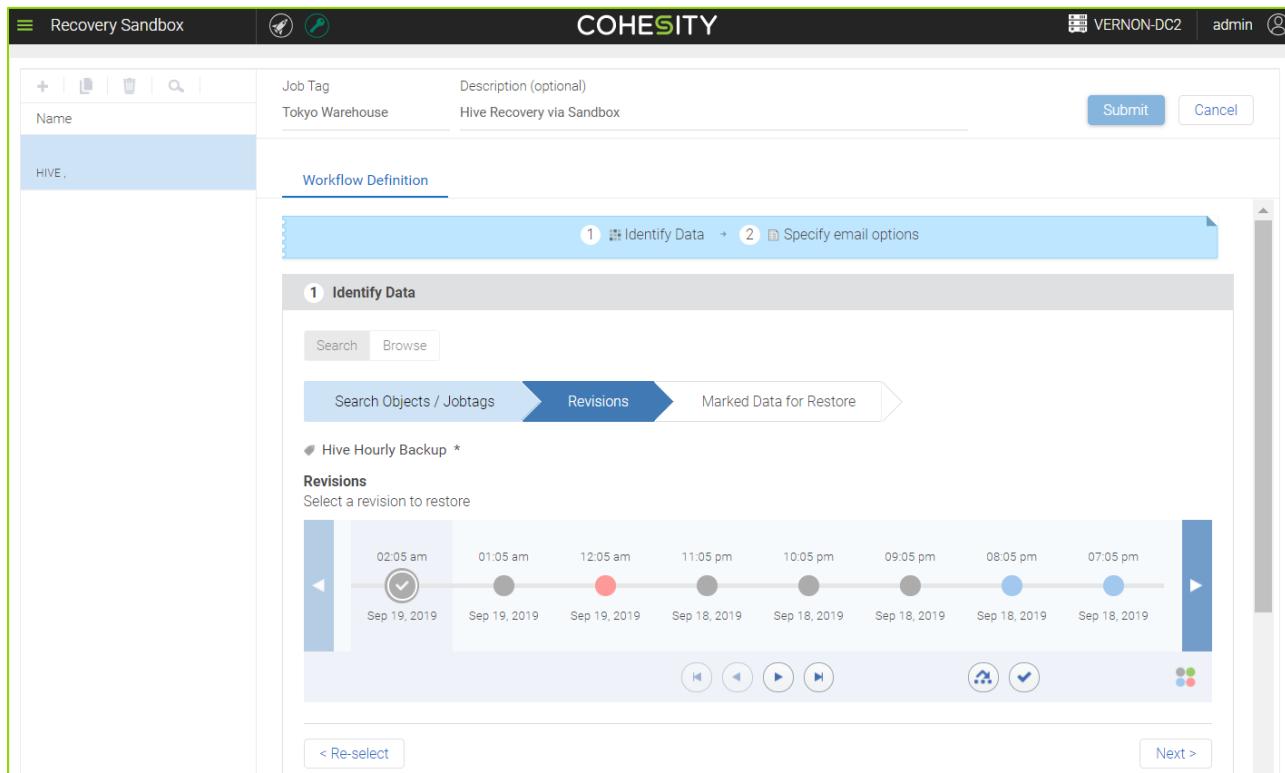
7. Click the radio button of the JobTag search result and then click the **Go To Revisions** button.

The screenshot shows the Cohesity Recovery Sandbox interface. At the top, there's a navigation bar with 'Recovery Sandbox', a search icon, and the Cohesity logo. On the right, it shows 'VERNON-DC2' and 'admin'. Below the navigation bar, there's a form for creating a new job tag. The 'Name' field is filled with 'Tokyo Warehouse', 'Job Tag' is 'Tokyo Warehouse', and 'Description (optional)' is 'Hive Recovery via Sandbox'. There are 'Submit' and 'Cancel' buttons. The main area is titled 'Workflow Definition' and shows a step-by-step process: '1 Identify Data' (selected), '2 Specify email options'. Under 'Identify Data', there are tabs for 'Search' and 'Browse', and a search bar with 'JobTag' selected and 'Hive Hourly Backup' entered. An 'Exact Match' checkbox is checked. Below the search bar, it says 'Showing 1 results.' and lists one result: 'Hive Hourly Backup'. At the bottom right of this section is a 'Go To Revisions >' button.

8. In the **Revisions** tab, do one of the following:

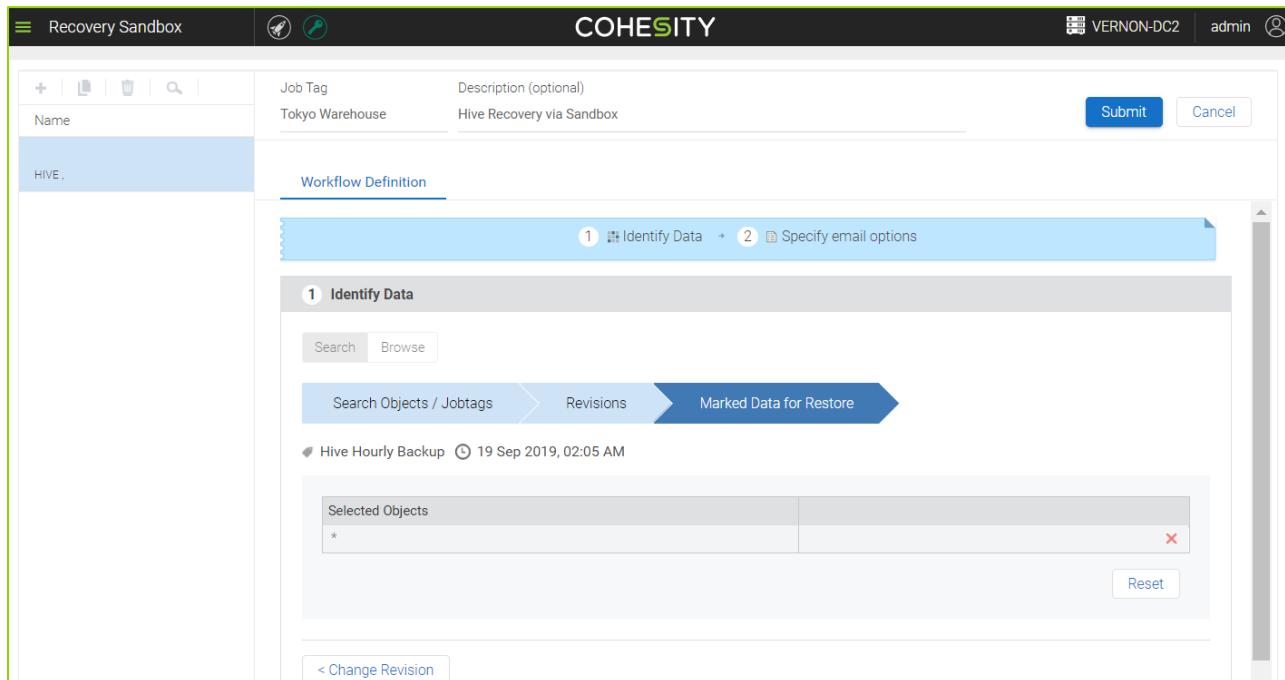
- By default, the latest copy is selected which is indicated by the icon. You can click the **Mark For Restore** button to restore the selected data object
- Click the data object icon to select a copy of data for a specific day and time. You can click the **Mark For Restore** button to restore the selected data object

NOTE: Navigate all the data object revision by clicking the icons. You can also click the icon to jump to a specific revision in time by selecting a date and time or click the icon to jump to the currently selected revision.



9. In the **Marked Data for Restore** label, do one of the following:

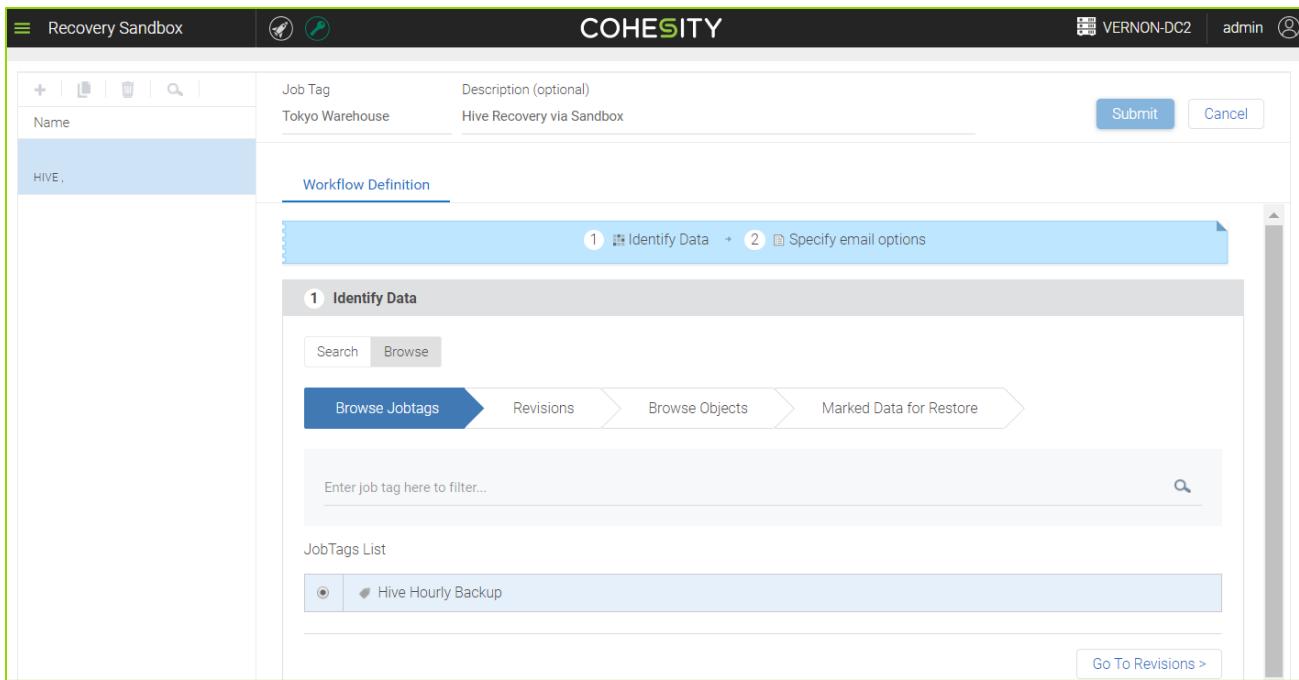
- Verify the selected data object revision and move on to the next step
- Click the **Change Revision** button to reselect a data object revision to restore



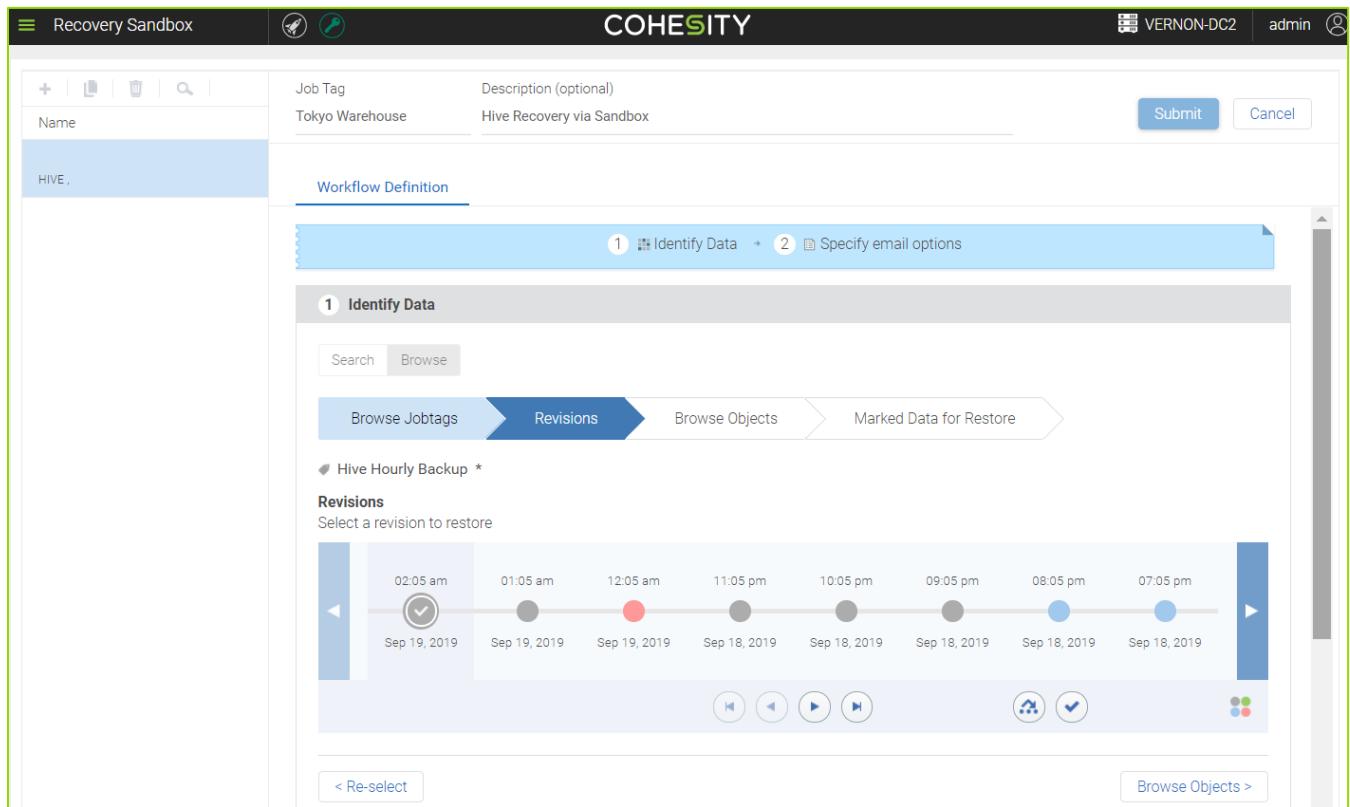
You can continue with specifying the recovery options by referring to Step #17.

Browse Tab

7. In the **Browse** tab, under the **Browse Jobtags** label, select a **JobTag** from the **JobTag list**, and then click the **Go To Revisions** button.



8. In the **Browse** tab, under the **Revisions** label, select a revision of the **JobTag** revision that you want to restore and then click the **Browse Objects** button.



9. In the **Browse** tab, under the **Browse Objects** label, do one of the following:

- Select or clear the check boxes of the data objects that you want to restore and then click the **Next** button
- Click the **Change revision** button to go back to the Revisions tabs and select a new revision of the JobTag

Recovery Sandbox

Job Tag: Tokyo Warehouse Description (optional): Hive Recovery via Sandbox

Workflow Definition: 1 Identify Data + 2 Specify email options

1 Identify Data

Search Browse

Browse Jobtags > Revisions > **Browse Objects** > Marked Data for Restore

Hive Hourly Backup 19 Sep 2019, 02:05 AM

Select the desired objects and click on "Mark For Restore" button to restore the objects.

/

Objects
 sales
 payroll
 legal

< Change revision Next >

10. In the **Browse** tab, under the **Marked Data for Restore** label, do one of the following:

- Click the **Reset** button or click the **X** icon to remove any data object from the list
- Click the **Browse Objects** button to go back and go through all the objects again

Recovery Sandbox

Job Tag: Tokyo Warehouse Description (optional): Hive Recovery via Sandbox

Workflow Definition: 1 Identify Data + 2 Specify email options

1 Identify Data

Search Browse

Browse Jobtags > Revisions > **Browse Objects** > **Marked Data for Restore**

Hive Hourly Backup 19 Sep 2019, 02:05 AM

Selected Objects

*

Reset

< Browse Objects

11. In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:

The screenshot shows a configuration panel for 'Email Notifications'. At the top, there are three radio button options: 'None' (unselected), 'Failures' (selected), and 'Everything' (unselected). Below this is a section titled 'Email Addresses' containing a text input field. Inside the field, there is a blue button with a white plus sign (+) followed by the email address 'john.doe@example.com' and a small blue 'X' button to its right.

12. Click **Submit**.

10 Data Mirroring

This section describes the features of the Data Mirroring menu of Imanis Data software.

10.1 Overview

Data Mirroring is the process of copying data from one data repository to a different repository of your choice. In Data Mirroring, the information copied to the destination location is an exact copy of the data selected from the source data repository, meaning the data will be copied and includes the same data schema as appeared in the source repository, often referred to as ‘app-aware’ capability.

10.1.1 Getting Started with Data Mirroring

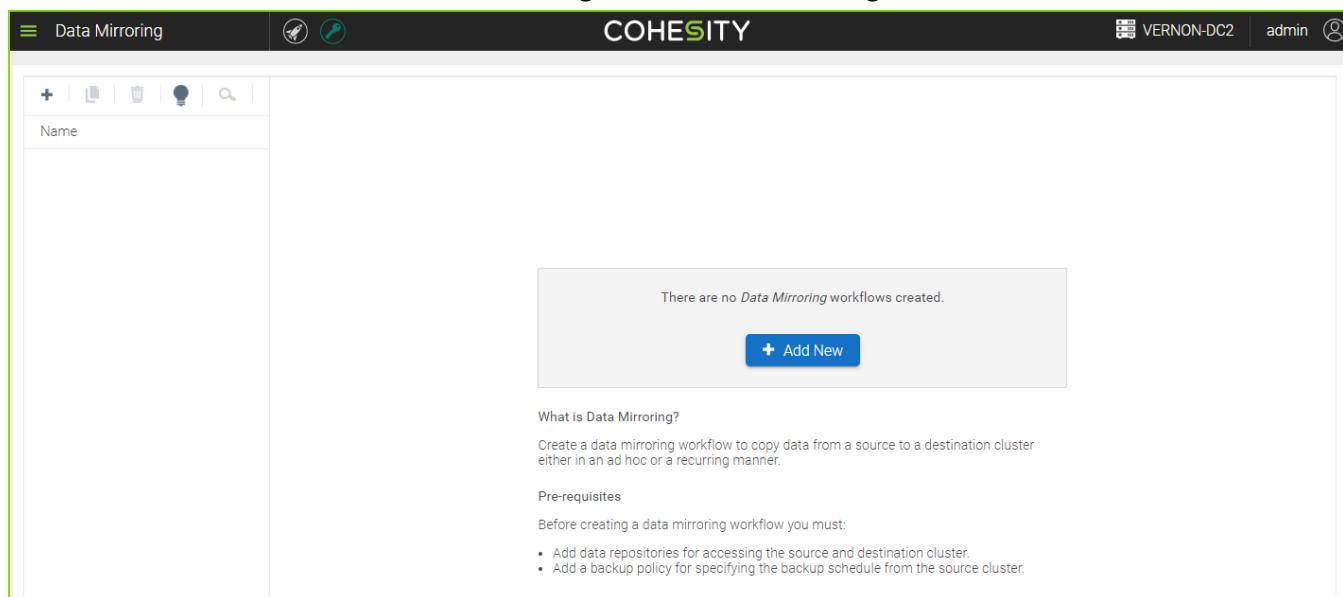
Data is captured from the primary data repository and backed up to another cluster referred to as the destination. In the Data Mirroring user interface, **Source** represents a data repository from which the data is copied, and **Destination** represents a data repository where the data will be eventually copied.

10.1.2 Data Mirroring for Hive

This section describes Hive data mirroring.

To start a data mirroring workflow:

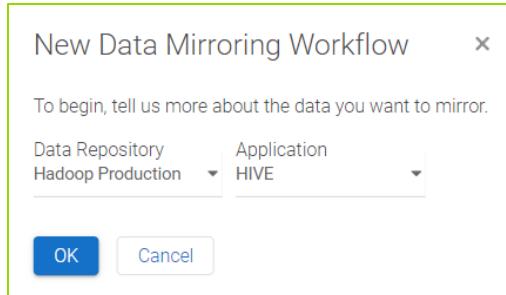
1. Click the Main Menu  > Data Management > Data Mirroring.



The screenshot shows the Cohesity Data Management interface with the following details:

- Header:** COHESITY, VERNON-DC2, admin, and a user icon.
- Left Sidebar:** A navigation bar with icons for Data Mirroring, Protection, Recovery, and Monitoring.
- Main Content Area:**
 - A message box states: "There are no *Data Mirroring* workflows created."
 - A blue "Add New" button is located below the message box.
 - A "What is Data Mirroring?" section defines it as "Create a data mirroring workflow to copy data from a source to a destination cluster either in an ad hoc or a recurring manner."
 - An "Pre-requisites" section lists requirements:
 - Before creating a data mirroring workflow you must:
 - Add data repositories for accessing the source and destination cluster.
 - Add a backup policy for specifying the backup schedule from the source cluster.

2. On the **Data Mirroring** page, click the **+ Add New** button  icon. The **New Data Mirroring Workflow** dialog box appears.
3. In the **New Data Mirroring Workflow** dialog box, select a **Hive** data repository from the **Data Repository** drop-down menu, and then click **OK**.



4. In the **Data Mirroring** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Identify Data** area, do the following:
 - In the **Hive** tab, identify the databases, tables, partitions that you want to backup by selecting the corresponding check boxes.
 - In the **Selected Data** tab, verify your selection or click the  icon to remove unwanted items.
 - In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section **Appendix A: Rules for Data Inclusions and Exclusions**.

Objects	Owner	Table Type	Size
movielens	--	--	
system	--	--	

6. In the **Specify Policy** section, under **Keep a copy on Imanis**, click **Yes**.

The **Yes** option activates the mirroring workflow thereby keeping a copy of your data on Imanis Data cluster. The **No** option activates the Direct Replication feature.

4 Specify Policy

Keep a copy on Cohesity Imanis Data cluster Yes No

Select a data backup policy

Backup to Cloud (s3)

Retention

Allow retention on cloud

Yes No

7. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without the cloud retention option.

8. In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

3 Specify Options

Cloud options [i](#)

Data Repository Buckets

S3 Cloud Storage hadoop_dr_archive

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).

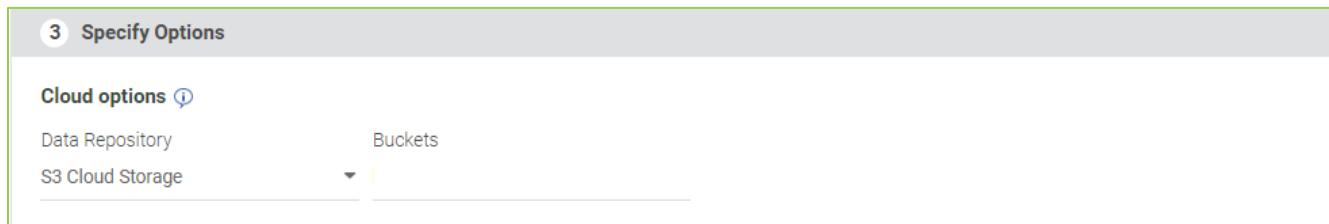
3 Specify Options

Cloud options [i](#)

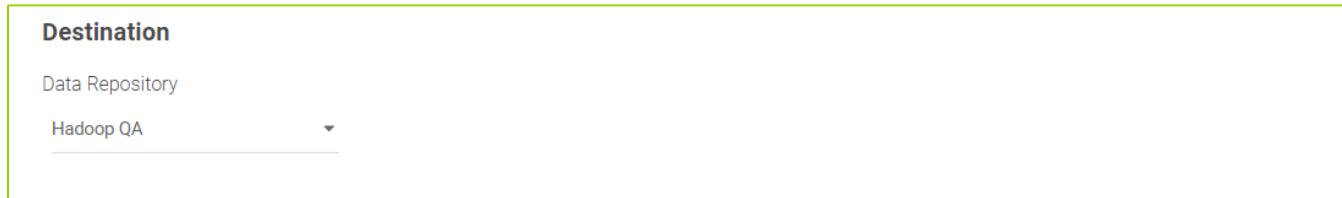
Data Repository Containers

Azure Cloud Storage hadoop_dr_backup

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

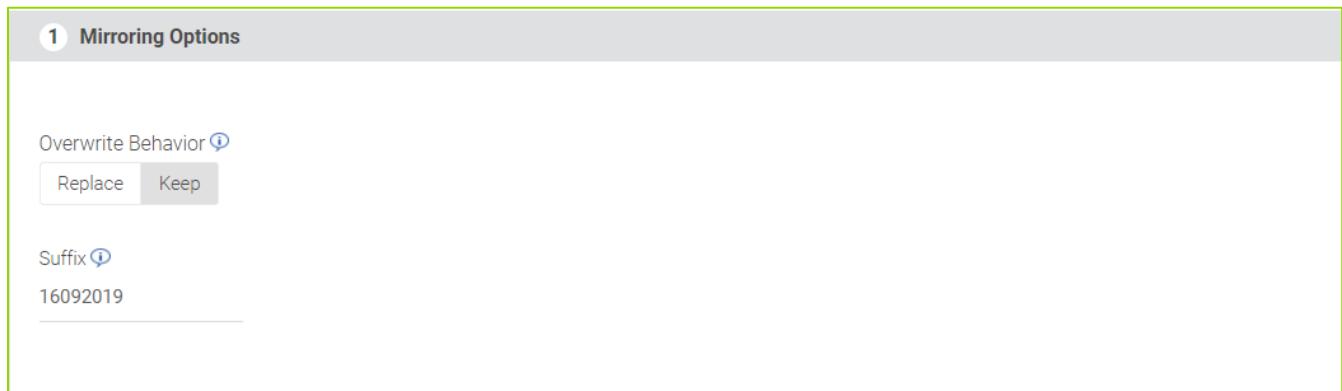


9. In the **Destination** section, from the Data Repository drop-down menu, select a data repository where the data will be moved to, that is, the destination cluster.



10. In the **Mirroring Options** area, under **Overwrite Behavior**, do one of the following:

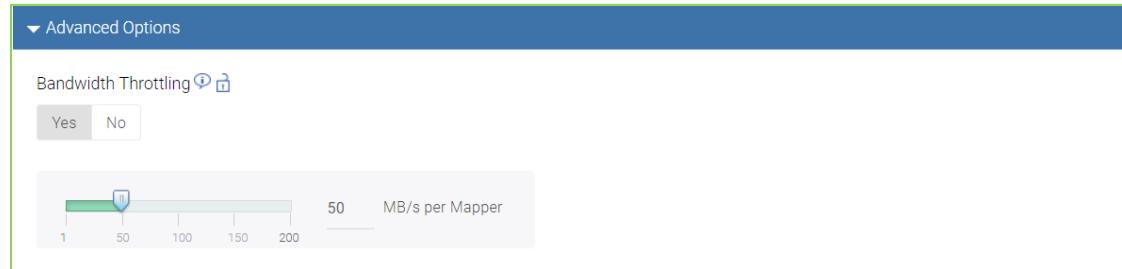
- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
- In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.



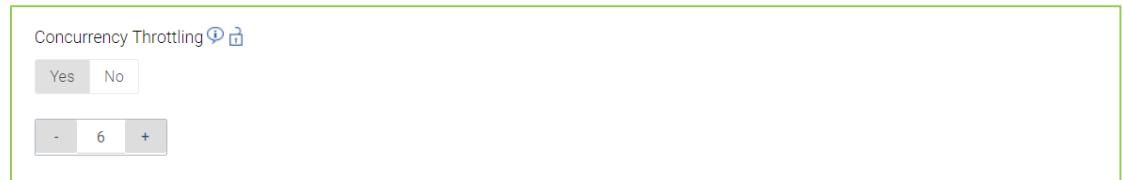
NOTE: Imanis Data software supports the use of alphanumeric characters in the suffix field; however, the use of uppercase is not recommended.

11. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



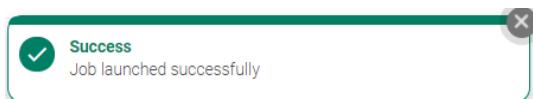
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.

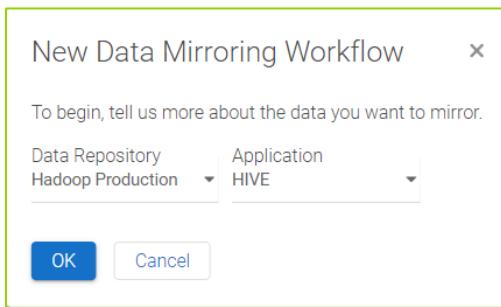


10.1.2.1 Data Masking & Sampling for Hive

This section describes the steps of using the sampling-masking feature in Hive.

To start data mirroring workflow for Hive, do the following:

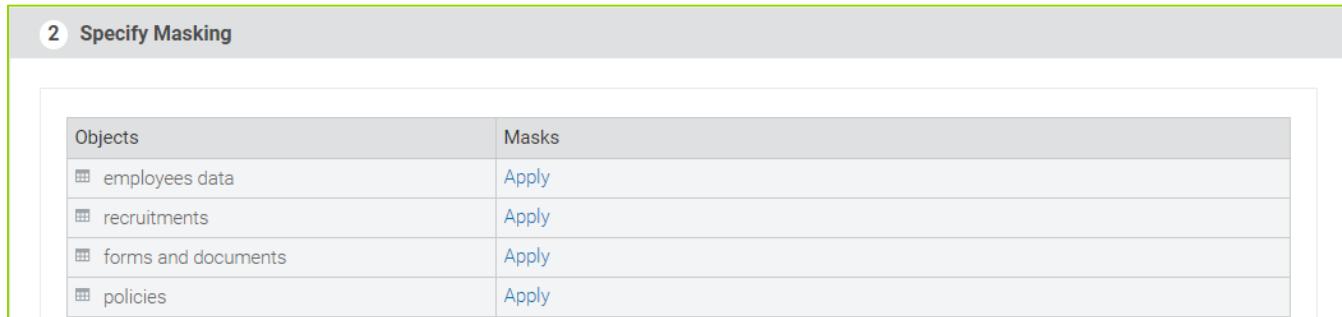
1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the Data Mirroring page, click the  **+ Add New** button or the  icon. The **New Data Mirroring Workflow** dialog box is displayed.
3. In the **New Data Mirroring Workflow** dialog box, select a **Hive** data repository from the **Data Repository** drop-down menu, and then click **OK**.



4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Identify Data** area, do the following:
 - a. In the **Hive** tab, identify the database that includes private and sensitive data, within tables, that you want to conceal (through masking) and/or down sample (through sampling) by selecting the corresponding check boxes. For example, click the **WareHouse_NYC** as shown in the above screenshot and then select **Design**.

	Owner	Table Type	Size
employees data	talena		
recruitments	talena		
forms and documents	talena		
policies	talena		

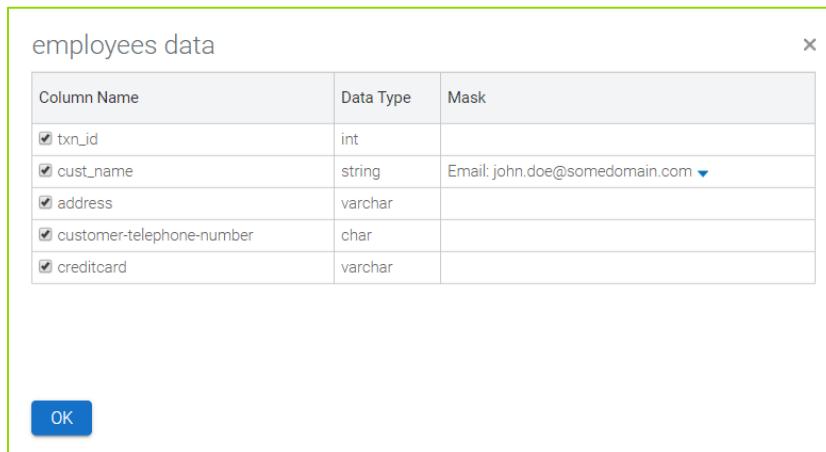
- b. In the **Selected Data** tab, verify your selection or click the  icon to remove unwanted items.
 - c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.
6. In the **Specify Masking** area, click the **Apply** link to apply masking for the tables selected in the preceding step. For example, click **Apply** for the **employees_data** table as displayed in the following screenshot:



Objects	Masks
employees data	Apply
recruitments	Apply
forms and documents	Apply
policies	Apply

7. Select the **Column Name** and **Mask**, and then click **OK**.

For example, for 'name' select mask as Full Name: John. E, Doe and for 'creditcard' select mask as Credit Card Number: 1234 5678 1234 5678. See the following screenshot:



Column Name	Data Type	Mask
<input checked="" type="checkbox"/> txn_id	int	
<input checked="" type="checkbox"/> cust_name	string	Email: john.doe@somedomain.com ▾
<input checked="" type="checkbox"/> address	varchar	
<input checked="" type="checkbox"/> customer-telephone-number	char	
<input checked="" type="checkbox"/> creditcard	varchar	

OK

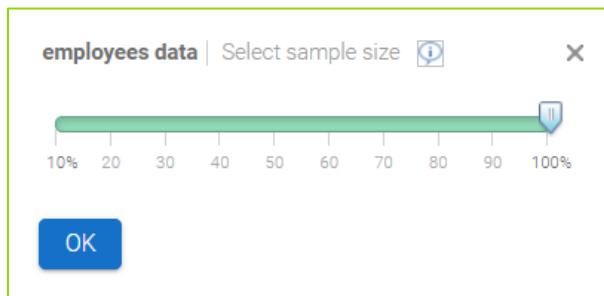
In the preceding screenshot, all the column names are pre-selected by default. You can manually clear the check boxes of column names that you do not want to mask. Optionally to remove masking that you may have applied, click the ‘Remove All’ link under the **Masks** column name (see the following screenshot):

2 Specify Masking	
Objects	Masks
employees data	Edit Remove All
recruitments	Apply
forms and documents	Apply
policies	Apply

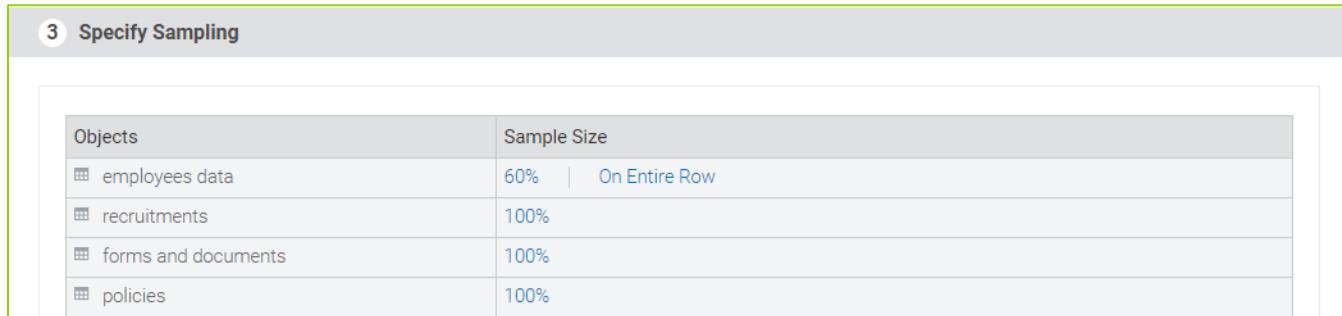
8. In the **Specify Sampling** area, click the link that denotes value **100%**.

3 Specify Sampling	
Objects	Sample Size
employees data	100%
recruitments	100%
forms and documents	100%
policies	100%

The following window appears. By default, the sample size is always set at 100%.



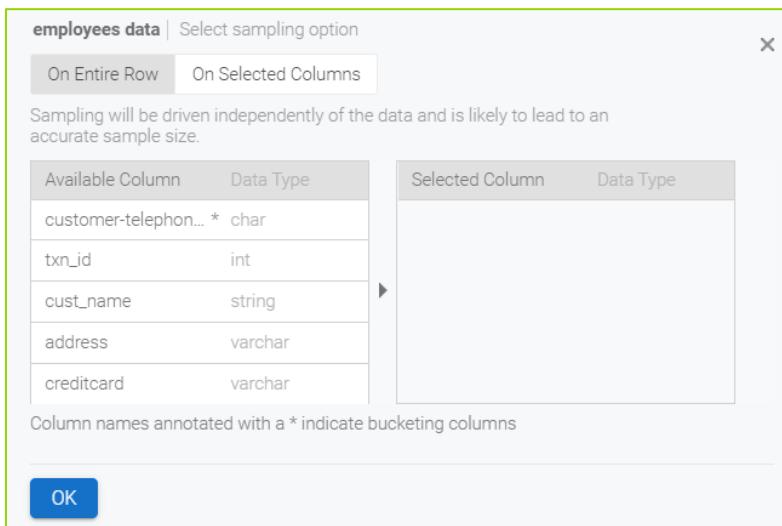
9. Move the slider to set a data sample size, and then click **OK**. Once you click OK, the table available in the **Specify Sampling** area displays a new link **On Entire Row**. See the following screenshot:



The screenshot shows a table titled "Specify Sampling" with the following data:

Objects	Sample Size
employees data	60% On Entire Row
recruitments	100%
forms and documents	100%
policies	100%

10. Click the **On Entire Row** link displayed in the above screenshot. The following page appears:
Cohesity Imanis Data software offers **On Entire Row** and **On Selected Columns** as sampling options.



The '**On Entire Row**' sampling option denotes that sampling is applied on the entire row instead of an individual column. While the '**On Selected Columns**' sampling option is based on the value of the column and denotes the set of columns on which the table is hash-partitioned or clustered on.

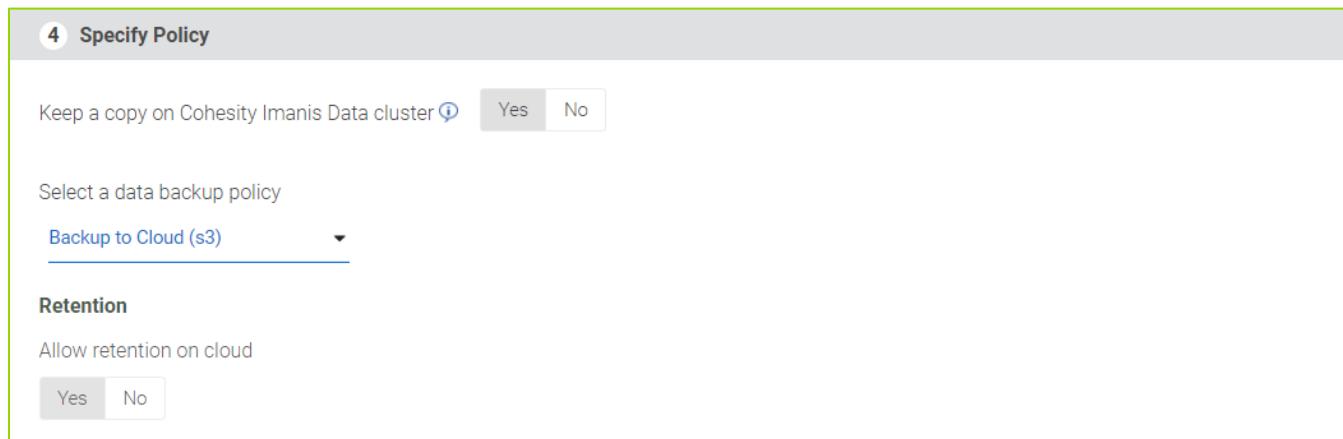
NOTE: In the current release, some data may be repeated in case of Sampling when you select the On Entire Row option.

11. Do one of the following:
- Select the **On Entire Row** option, and then click **OK**.

OR

- a. Select the **Selected Columns** option

- b. Identify the columns to select and click the  icon to transfer the selected column from the **Available Columns** section to the **Selected Columns** section. You can select maximum two columns only.
 - c. Click **OK**.
12. In the **Specify Policy** section, under **Keep a copy on Imanis**, click **Yes**.
The Yes option activates the mirroring workflow thereby keeping a copy of your data on Imanis Data cluster. The No option activates the Direct Replication feature.



4 Specify Policy

Keep a copy on Cohesity Imanis Data cluster  Yes No

Select a data backup policy

Backup to Cloud (s3)

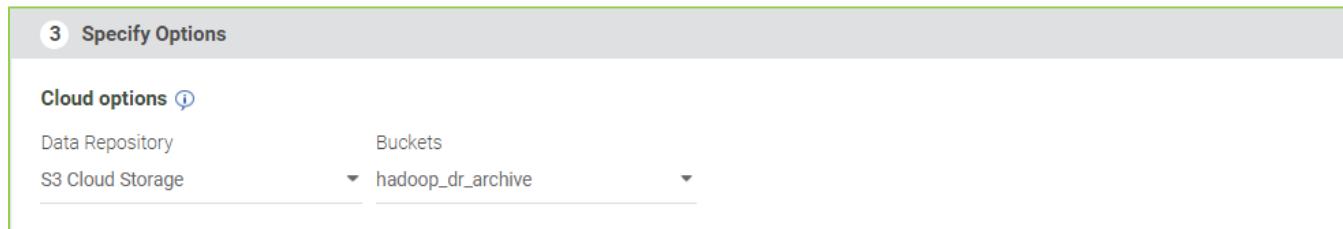
Retention

Allow retention on cloud

Yes No

13. In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

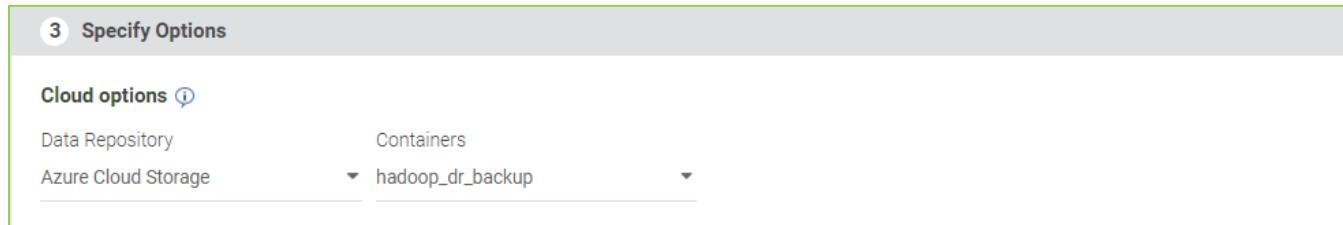


3 Specify Options

Cloud options 

Data Repository	Buckets
S3 Cloud Storage	 hadoop_dr_archive 

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).



3 Specify Options

Cloud options 

Data Repository	Containers
Azure Cloud Storage	 hadoop_dr_backup 

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

The screenshot shows a step titled "3 Specify Options". Under "Cloud options", there is a dropdown menu labeled "Data Repository" with "S3 Cloud Storage" selected. To the right of the dropdown is a "Buckets" input field.

14. In the **Destination 1** section, from the **Data Repository** drop-down menu, select a data repository where the data will be moved to, that is, the destination cluster.

The screenshot shows a section titled "Destination". A dropdown menu labeled "Data Repository" has "Hadoop DR" selected.

15. In the **Mirroring Options** area, under **Overwrite Behavior**, do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied

The screenshot shows a section titled "1 Mirroring Options". Under "Overwrite Behavior", there are two buttons: "Replace" (highlighted) and "Keep".

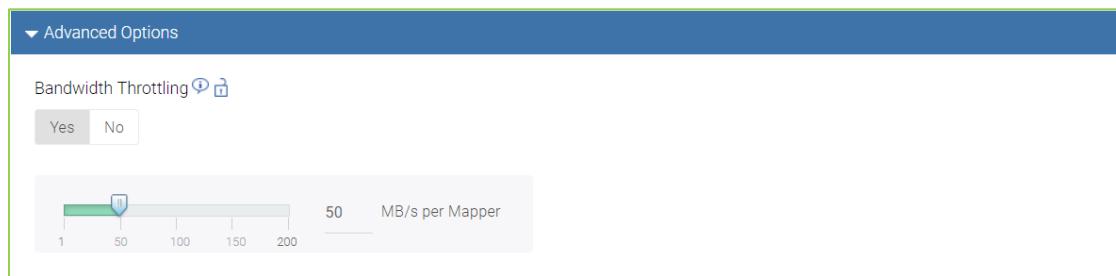
16. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

The screenshot shows a field titled "Suffix" with the value "19092019" entered.

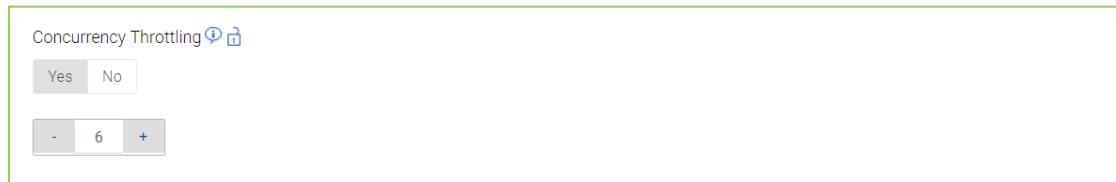
NOTE: Imanis Data software supports the use of alphanumeric characters in the suffix field; however, the use of uppercase is not recommended.

17. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

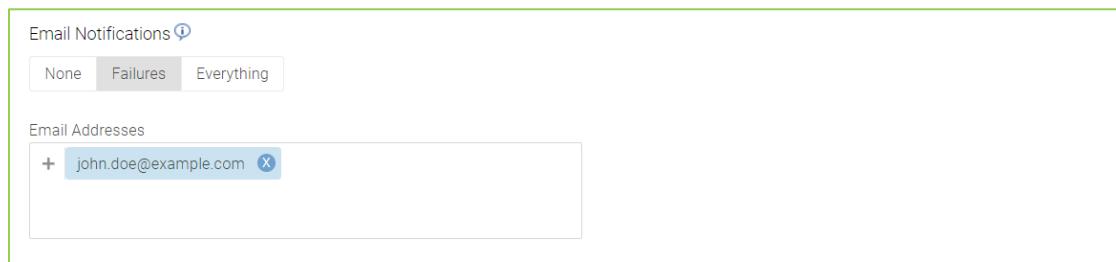
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

18. Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



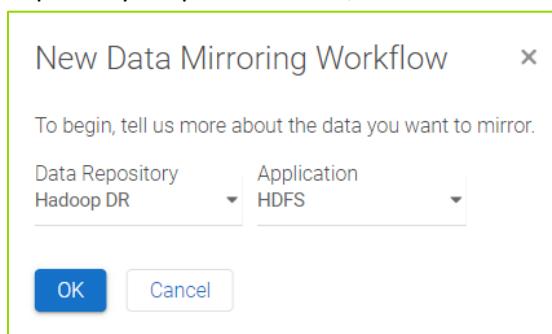
NOTE: By default, Imanis Data software restores a compressed file in source cluster with the same compression codec to the destination cluster. For example, if the compression codec of a file on source cluster is gzip, then the file will be restored to the destination cluster with the same compression codec.

10.1.3 Data Mirroring for HDFS

Imanis Data software supports data mirroring for HDFS data sets.

To start a data mirroring workflow, do the following:

1. Click the **Main Menu** > **Data Management** > **Data Mirroring**.
2. On the **Data Mirroring** page, click the or the . The **New Data Mirroring Workflow** dialog box appears.
3. In the **New Data Mirroring Workflow** dialog box, select a **HDFS** data repository from the Data Repository drop-down menu, and then click **OK**.



4. In the **Data Mirroring** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Source** section, in the **Identify Data** area, do the following:
 - a. The **HDFS** tab, Identify the files and directories that you want to backup by selecting the corresponding check boxes. Use regular expressions (regex) for primary repository browsing.
 - b. In the **Selected Data** tab, verify your selection or click the to remove unwanted items.

- c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

IMPORTANT: Do not use blank spaces in the inclusion or exclusion regex.

The screenshot shows the 'Data Mirroring' interface. In the 'Workflow Definition' section, there is a diagram showing three stages: 'Hadoop DR' (gray), 'Imanis Data' (blue), and a final stage (green). Below the diagram, a progress bar indicates '1 Identify Data' and '2 Specify Policy'. The 'Identify Data' section shows a table of HDFS objects:

	Objects	Modified Time	Owner	Size
<input type="checkbox"/>	movielens	2019-09-16 09:10:33 ...	talena	2.8 GB
<input checked="" type="checkbox"/>	system	2019-09-16 09:10:33 ...	talena	29.7 GB
<input checked="" type="checkbox"/>	human resource	2019-09-16 09:10:33 ...	talena	937.6 MB
<input checked="" type="checkbox"/>	sales	2019-09-16 09:10:33 ...	talena	2 GB

6. In the **Specify Policy** section, under **Keep a copy on Imanis**, click **Yes**. The Yes option activates the mirroring workflow thereby keeping a copy of your data on Imanis Data cluster before copying it to destination cluster. The **No** option activates the Direct Replication feature.

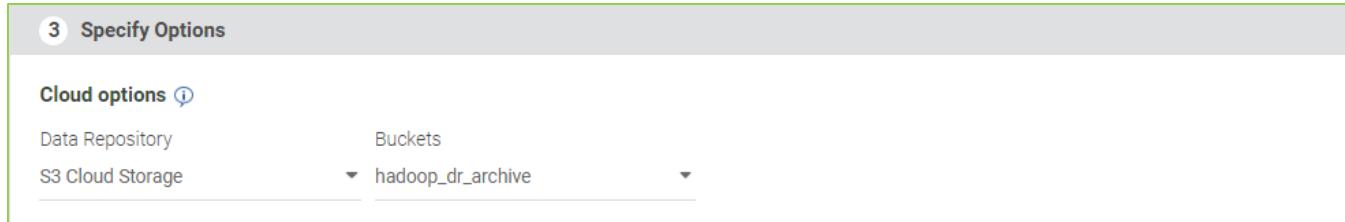
The screenshot shows the 'Specify Policy' section. A toggle switch labeled 'Keep a copy on Cohesity Imanis Data cluster' is set to 'Yes'. There is also a 'No' option available.

7. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

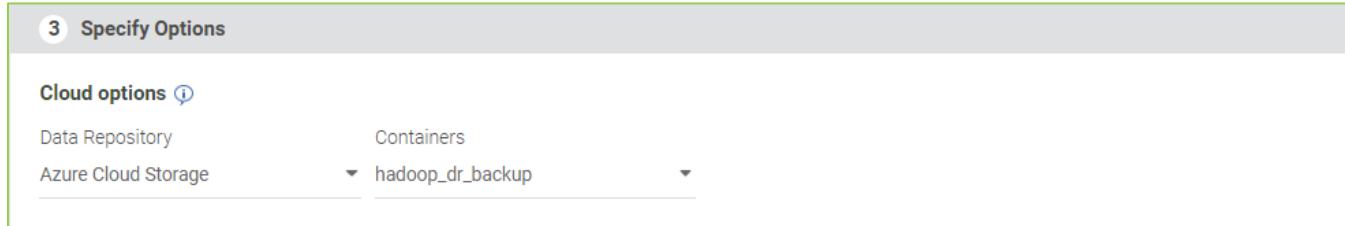
The screenshot shows a dropdown menu for selecting a data backup policy. The option 'Backup to Cloud (s3)' is highlighted.

8. In the **Specify Options** section, under **Cloud Options**, do one of the following:

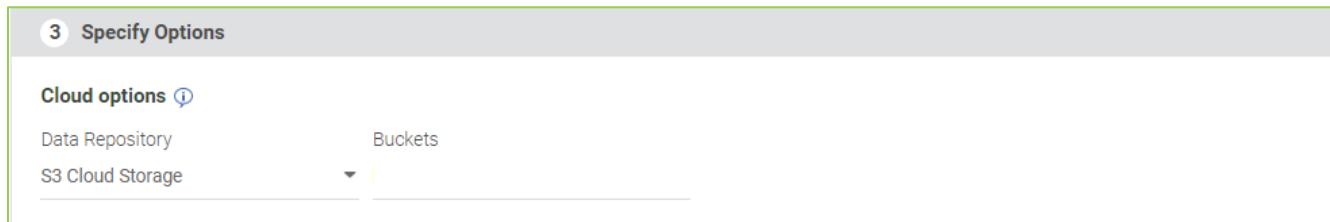
- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:



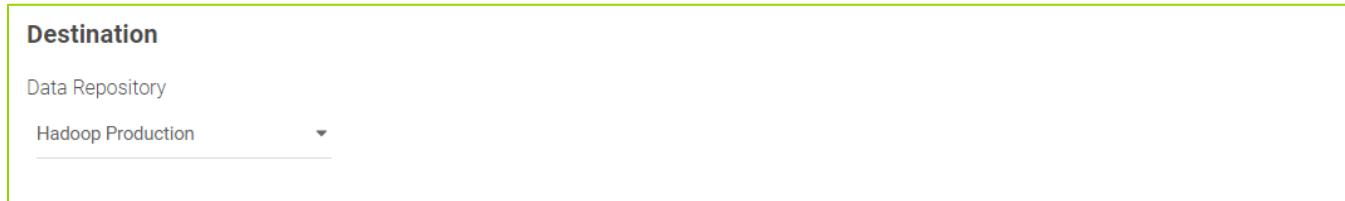
- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud.



- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:



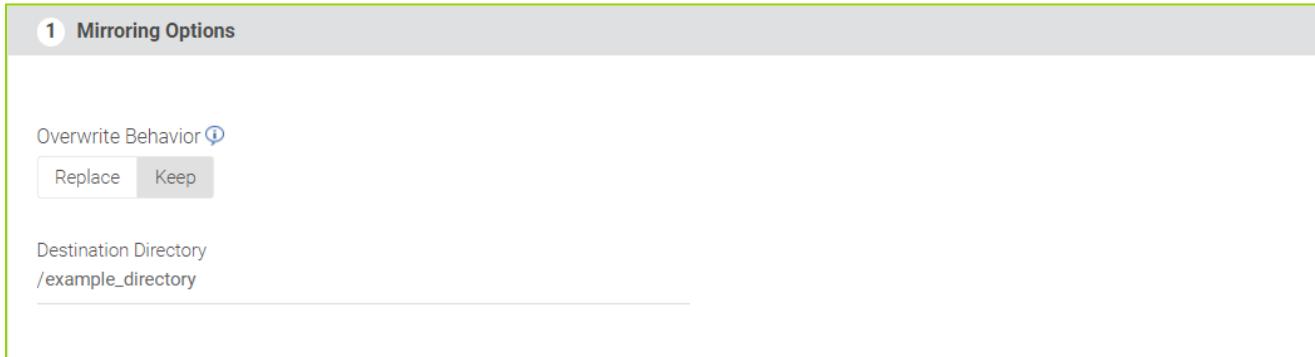
9. In the **Destination** section, select the destination data repository from the drop-down menu where you want to move the backed-up data.



10. In the **Mirroring Options**, do the following:

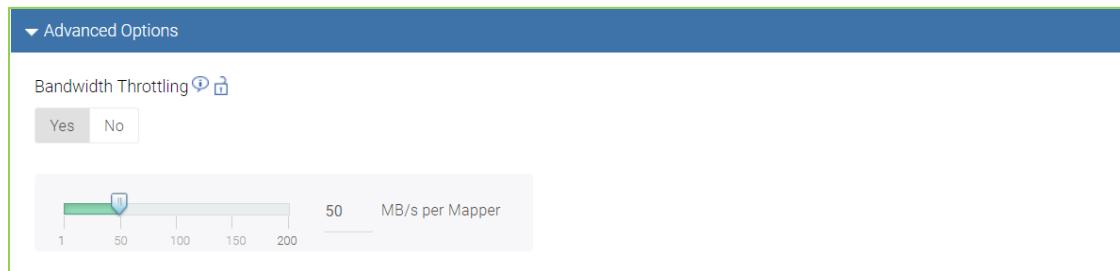
- a. Click **Replace** in the **Overwrite Behavior** option to overwrite old data with new data. By default, **Keep** is selected.

- b. Type a name for the directory in the **Destination Directory** field. For example, /example_directory. Imanis Data software creates a new directory with this name and data will be restored in this directory.

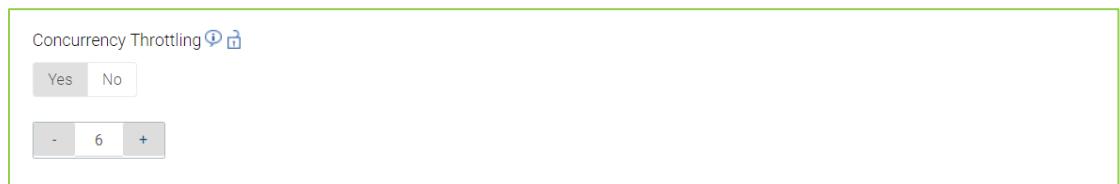


11. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** of both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

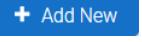
```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

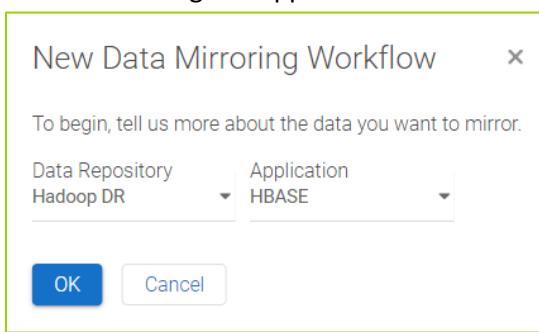
12. Click **Submit**.

10.1.4 Data Mirroring for HBase

Imanis Data software supports data mirroring for HBase data sets.

To start a data mirroring workflow, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the **Data Mirroring** page, click the  button or the  icon. The **New Data Mirroring Workflow** dialog box appears.



3. In the **New Data Mirroring Workflow** dialog box, select a **HBase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.

4. In the **Data Mirroring** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.

5. In the **Source** section, in the **Identify Data** area, do the following:
 - a. In the **HBase** tab, identify the namespaces and tables that you want to back up by selecting the corresponding check boxes. Use regular expressions (regex) for primary repository browsing.
 - b. In the **Selected Data** tab, verify your selection or click the **X** icon to remove unwanted items.
 - c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

The screenshot shows the 'Data Mirroring' configuration screen. At the top, there's a header with 'Data Mirroring', a save icon, and a cancel icon. Below the header, the 'Job Tag' is set to 'Berlin Warehouse' and the 'Description' is 'HBase Data Repo'. On the right, there are 'Submit' and 'Cancel' buttons. The main area is titled 'Workflow Definition' and shows a flow from 'Hadoop DR' to 'Imanis Data'. Under 'Source', there are two tabs: 'Identify Data' (selected) and 'Specify Policy'. In 'Identify Data', the 'HBASE' tab is active, showing a list of namespaces: 'Objects' (checkbox), 'movielens' (checkbox checked), and 'system' (checkbox checked). There are also 'Selected Data' and 'Rules' tabs.

6. In the **Specify Policy** section, under **Keep a copy on Imanis**, click **Yes**.
The **Yes** option activates the mirroring workflow thereby keeping a copy of your data on Imanis Data cluster before copying it to destination cluster. The **No** option activates the Direct Replication feature.

The screenshot shows the 'Specify Policy' section. It has a header '2 Specify Policy'. Below it is a question 'Keep a copy on Cohesity Imanis Data cluster' with a help icon. Two buttons are shown: 'Yes' (highlighted in blue) and 'No'.

7. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

The screenshot shows a dropdown menu for 'Select a data backup policy'. The options listed are 'Select a data backup policy' (disabled) and 'Backup to Cloud (s3)' (selected).

8. In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a configuration wizard. Under 'Cloud options', the 'Data Repository' dropdown is set to 'S3 Cloud Storage' and the 'Buckets' dropdown is set to 'hadoop_dr_archive'.

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud.

The screenshot shows the 'Specify Options' step of a configuration wizard. Under 'Cloud options', the 'Data Repository' dropdown is set to 'Azure Cloud Storage' and the 'Containers' dropdown is set to 'hadoop_dr_backup'.

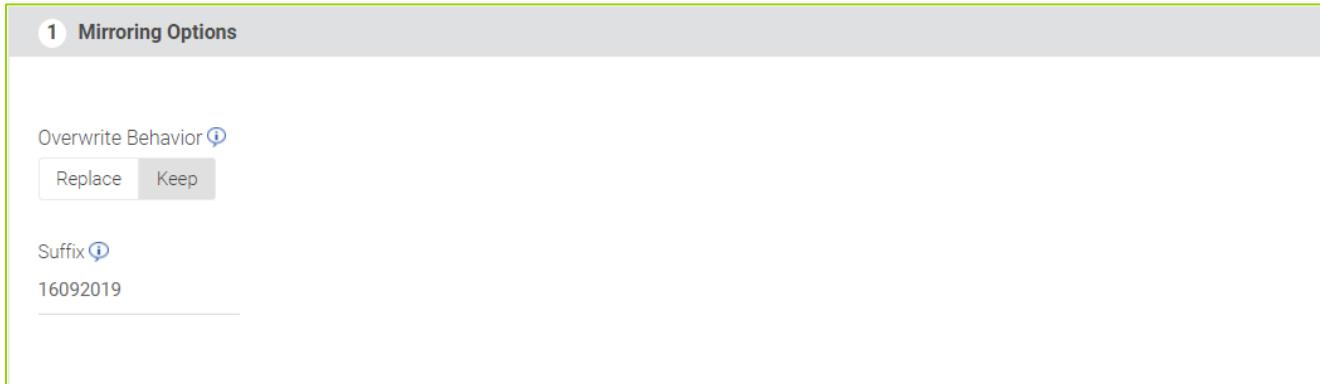
- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a configuration wizard. Under 'Cloud options', the 'Data Repository' dropdown is set to 'S3 Cloud Storage'. The 'Buckets' dropdown is empty, indicating a custom bucket name will be typed in.

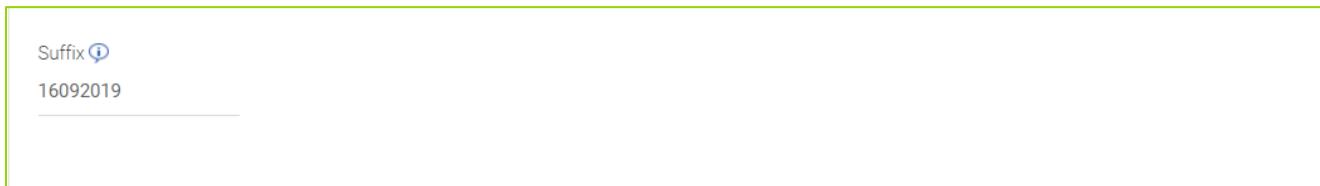
9. In the **Destination** section, select the destination data repository from the drop-down menu where you want to move the backed-up data.

The screenshot shows the 'Destination' step of a configuration wizard. The 'Data Repository' dropdown is set to 'Hadoop QA'.

10. In the **Mirroring Options**, click **Replace** to overwrite any previously existing data. By default, **Keep** is selected.



11. In the **Suffix** field, type



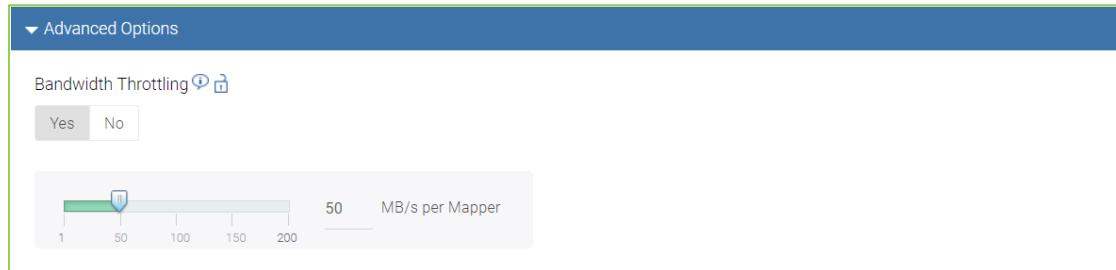
12. In the **More Options for Selected Data**,

2 More Options for Selected Data	
Objects	Recover As
movielens	movielens16092019
system	system16092019
human resource	human resource16092019
sales	sales16092019
payroll	payroll16092019

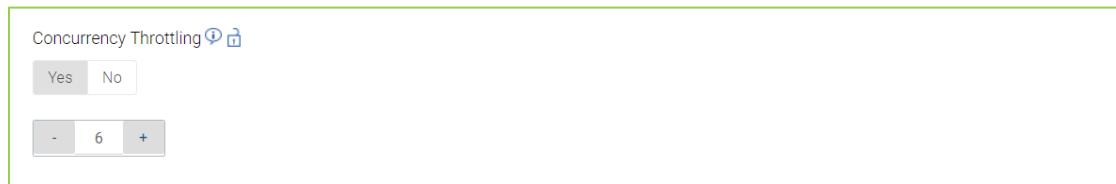
13. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each

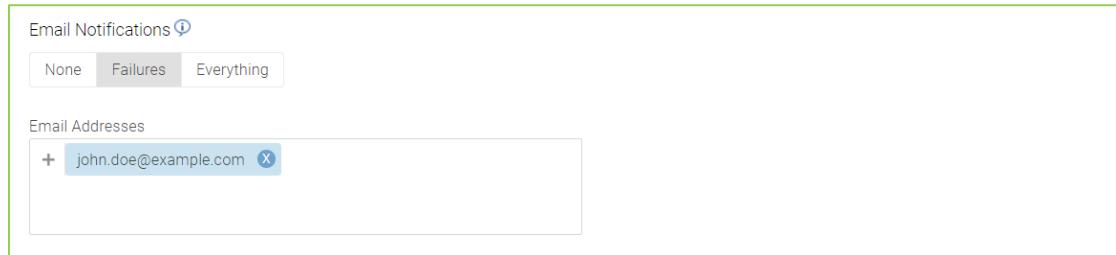
individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **in the MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

14. Click **Submit**.

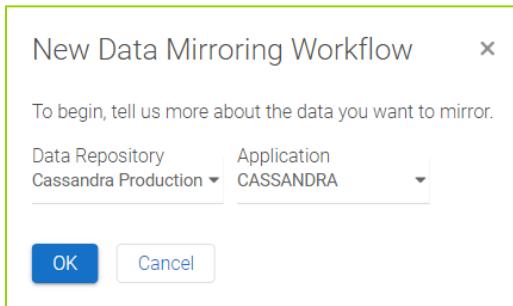
IMPORTANT: Data recovery from one version to another version is not supported. For example, data recovery from 5.8.2 to 5.12.2 is not supported. This event is also applicable to data mirroring workflows where the source and destination versions are different. In case you upgrade to a new version of HBase, it is recommended to recreate all the workflows that are associated with the data repository.

10.1.5 Data Mirroring for Cassandra

Imanis Data software enables you to create data mirroring workflows for Cassandra and Solr-enabled DataStax Enterprise (DSE) Cassandra. During the data mirroring process, Imanis Data software automatically identifies Solr-enabled Cassandra in both backup and recovery process.

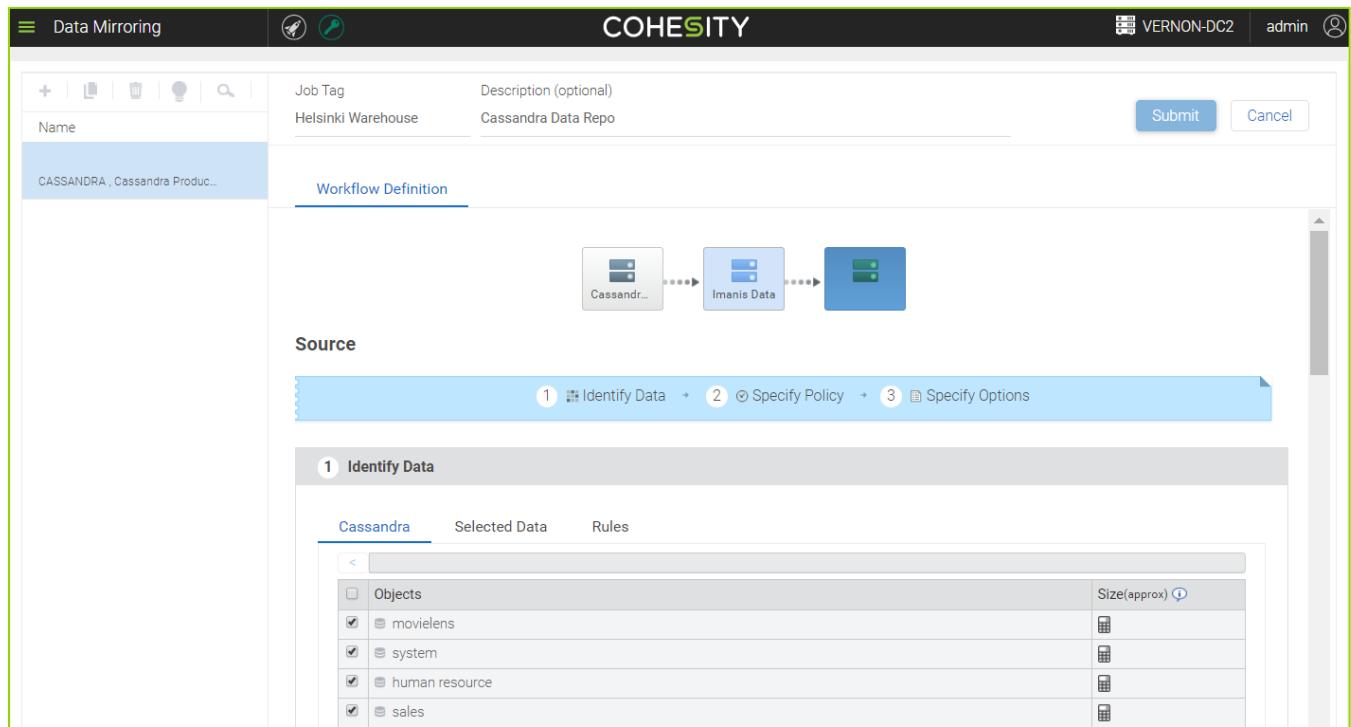
To start a data mirroring workflow, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the **Data Mirroring** page, click the  **+ Add New** button or the  icon. The **New Data Mirroring Workflow** dialog box appears.



3. In the **New Data Mirroring Workflow** dialog box, select a **Cassandra** data repository from the **Data Repository** drop-down menu, and then click **OK**.
4. In the **Data Mirroring** page, do the following:
 - Type a new job tag in the **Job Tag** field.
 - Type a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Source** area, under **Identify Data**, do the following:
 - In the **Cassandra** tab, identify the keyspaces and tables that you want to back up by selecting the corresponding check boxes. Use regular expressions (regex) for primary repository browsing.
 - In the **Selected Data** tab, verify your selection or click the  icon to remove unwanted items.
 - In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section **Appendix A: Rules for Data Inclusions and Exclusions**.

IMPORTANT: Avoid the use of blank spaces in the inclusion or exclusion regex.



6. Ignore the **Specify Masking** and **Specify Sampling** area as it is separately explained in this section: Data Masking & Sampling for Cassandra.
7. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

2 Specify Policy

Select a data backup policy

Backup to Cloud (s3)

Retention

Allow retention on cloud

Yes No

8. In the **Specify Options** section, select a data center or data centers from where you want to backup the data. This **Data Centers** option is displayed in Multi-DC setup only.

3 Specify Options

Data Centers

Select All
 DC1
 DC2

9. In the **Specify Options** section, under **Cloud options** area, do one of the following. This option is displayed if you selected a backup policy with the cloud retention option.

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

Cloud options ⓘ

Data Repository Buckets

S3 Cloud Storage ▾ cassandra_dev_backup ▾

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud.

Cloud options ⓘ

Data Repository Containers

Azure Cloud Storage ▾ cassandra_dev_backup ▾

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

3 Specify Options

Cloud options ⓘ

Data Repository Buckets

S3 Cloud Storage ▾ |

10. In the **Destination** section, select the destination data repository from the drop-down menu where you want to move the backed-up data.

Destination

Data Repository

Cassandra Production



11. In the **Mirroring Options**, do the following:

- In the Data Centers option, select one or multiple data centers, in the Data Centers option, where you want to recover the data. This option is displayed if you select the Source as destination data repository only.

1 Mirroring Options

Data Centers

- Select All
- DC1
- DC2

12. In the **Overwrite Behavior** section do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied

13. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

14. For DSE 6.X, for **Recovery Staging Directory**, you can mention a temporary directory if you do not wish Imanis to use the Cassandra storage for staging. The Recovery Staging Directory field accepts a single directory or a comma separated list of directories. Ensure that the directories are present on all the nodes before executing the restore job.

Recovery Staging Directory`/tmp/stage1,/tmp/stage2`

15. In the **More Options for Selected Data** section, do the following:

- To rename restored objects, type the new name in the **Recover As** column.
- To change property of the restored object, click the  icon, and type the values in the Key and Value field. Usually, the key is auto-completed by the UI. The Value field must contain the

complete value of the property. Only following property changes are allowed: keyspace (replication) and table (compression and compaction).

For example, let's assume that a user wants to change replication for a source keyspace wherein the create query for the keyspace is as follows:

```
CREATE KEYSPACE tutorialspoint WITH replication = {'class':'SimpleStrategy', 'replication_factor': 3};
```

To change the replication factor to 1, the key value pair would be as follows:

KEY: replication

VALUE: {'class':'SimpleStrategy', 'replication_factor' : 1}

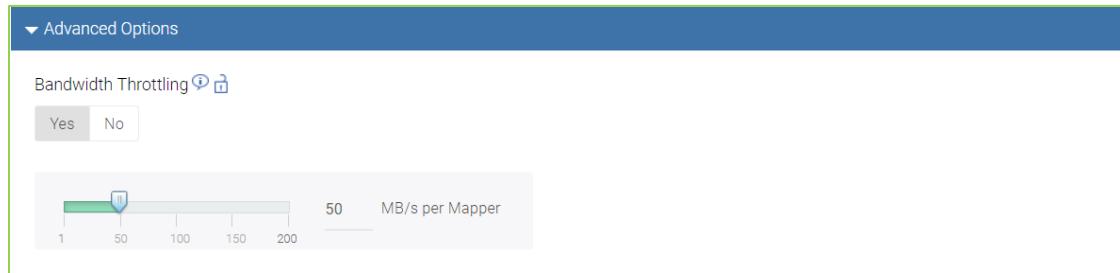
Objects	Recover As	With Properties
movielens	movielens16092019	+ replication = {'class':'Simple Strategy', 'replication_factor': 1} X
system	system16092019	+ X

IMPORTANT: If the recovery process is executed to an '**Alternate Location**', Imanis Data software always uses 'SimpleStrategy' as the replication strategy unless it is overridden through the "**More Options for Selected Data**" section in the UI.

NOTE: In the current release, Imanis Data software does not support changing of compaction strategy to CFSCompactionStrategy from any other strategy. For example, Imanis Data software does not support changing of compaction strategy from Leveled Compaction Strategy, Size Tiered Compaction Strategy or any other strategy to CFSCompactionStrategy.

16. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

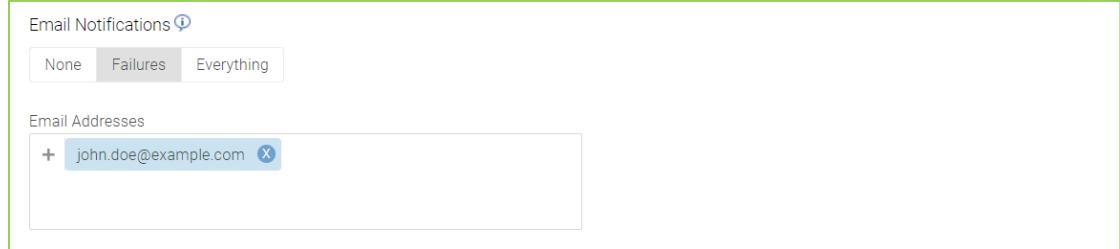
- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

17. Click **Submit**.

IMPORTANT: In the current release, alteration of schema at the source during incremental restore where the destination cluster is a different Cassandra version than the source cluster is not supported.

IMPORTANT: Ensure that the system partitioner – a configurable parameter in the `cassandra.yaml` file that determines how data is distributed across the nodes in the cluster – at the source cluster matches with the

system partitioner that is configured at the destination cluster. If you do not configure the system partitioner at the destination cluster, the data restore process will fail.

IMPORTANT: As part of the recovery process, the keyspace/table schemas are created (if required) on the destination cluster. However, this schema creation may fail if a data center(s) is down in a multi-dc environment. In such cases, the Cassandra data recovery process will fail.

NOTE: The content-aware storage reduction activity in Imanis Data software may result in a difference between backup and recovery data size stats.

IMPORTANT: For DSE 5.0 and later releases, Imanis Data software does not restore permissions of tables or databases if the required set of roles are not present on destination restore cluster. Thus, the user must create the required set of roles on destination cluster before initiating the recovery workflow. The user must also ensure to create transitive role dependencies if any exists. For example, on the Source cluster, you have roles 'supervisor', 'admin', 'staff'.

The role 'admin' has "Select" privileges on database 'company'.

The role 'staff' has "Insert" and "Select" privileges on database 'company'.

While the role 'supervisor' has transitive dependency wherein it inherits all the privileges of both the roles 'admin' and 'staff' with the following command:

```
Grant admin to supervisor;  
Grant staff to supervisor;
```

Thus, the user must ensure that such transitive dependent roles (like the role 'supervisor') are created on destination cluster before initiating recovery workflow.

IMPORTANT: If Authorization is enabled on destination cluster, ensure that users present in Cassandra primary cluster are also present on destination cluster.

NOTE: In the current release, multiple schema alteration queries executed on the single column on the source Cassandra cluster will not be propagated to the destination Cassandra cluster during a single invocation of data mirroring workflow.

For example, if the following schema alteration commands are used on a table:

1. ALTER TABLE newyork.transit_map DROP events [where events column is dropped]
2. ALTER TABLE newyork.transit_map ADD events text [where events column of type text is added again to the schema]

The preceding commands are performed on the Cassandra cluster and then the incremental data mirroring workflow is executed which moves data objects to the destination Cassandra cluster.

This results in an error with the incremental data mirroring workflow as schema alteration queries executed on the source Cassandra cluster will not be propagated to the destination Cassandra cluster during a single invocation of data mirroring workflow.

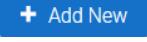
NOTE: Bulk Loader restore is supported for DSE 6.x. The process of restore involves copying of files to a temporary location on the primary cluster. The maximum temporary space can be up to the size of the data to be restored. The restore agent creates temporary directories in the Cassandra storage directories on the primary cluster for copying files. The temporary directories are created by using the restore job's UUID which are then deleted after the data restore completes.

10.1.6 Data Mirroring for Couchbase

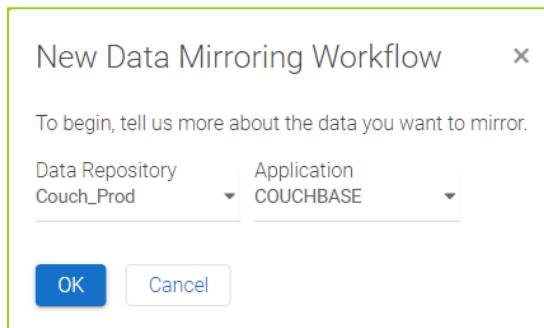
Imanis Data software supports data mirroring for Couchbase data sets at the jobtag and bucket level. Prior to starting the recovery process, you must first identify the bucket(s) that you want to recover at the source cluster and then manually create the bucket(s) with the same name(s) at the destination cluster.

For example, at the source cluster you identify test_bucket1, test_bucket2, and test_bucket3 that you want to recover. At the destination cluster, you must manually create test_bucket1, test_bucket2, and test_bucket3 before starting the recovery process.

To start a data mirroring workflow for Couchbase, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the **Data Mirroring** page, click the  **+ Add New** button or the  icon. The **New Data Mirroring** dialog appears.

3. In the **New Data Mirroring Workflow** dialog, select a source data repository from the **Data Repository** drop-down menu, select Couchbase from the **Application** drop-down menu and click **OK**.



4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Source** section, in the **Identify Data** area, do the following:
- In the **Couchbase** tab, identify the jobtag and bucket level data that you want to backup by selecting the corresponding check boxes.

The screenshot shows the 'Data Mirroring' interface. A workflow definition is being created between 'Couch_Prod' and 'Imanis Data'. The 'Workflow Definition' section shows a sequence of three nodes: 'Couch_Prod' (source), 'Imanis Data' (target), and another node (likely a destination or processing step). Below this, the 'Source' section is expanded, showing the 'Identify Data' step. Under the 'COUCHBASE' tab, a table lists objects: 'movielens' (selected), 'memcached' (unchecked), 'system' (unchecked), 'ephemeral' (unchecked), 'human resource' (checked), and 'sales' (checked). The 'Document Count' column for selected objects is empty.

Object	Document Count
movielens	-
memcached	-
system	-
ephemeral	-
human resource	
sales	

- b. In the **Selected Data** tab, verify your selection or click the **X** icon to remove unwanted items.

- c. In the **Rules** tab, click **Yes** to include all buckets in the backup job.

The screenshot shows the Cohesity Data Mirroring interface. At the top, there's a navigation bar with 'Data Mirroring' and other icons. The main area has a 'Workflow Definition' section with three nodes: 'Couchbase Pr...' (grey), 'Imanis Data' (blue), and another blue node. Below this is a 'Source' section with five numbered steps: 1. Identify Data, 2. Specify Filters, 3. Specify Sampling, 4. Specify Masking, and 5. Specify Policy. Step 1 is currently selected. Under 'Identify Data', there are tabs for 'COUCHBASE', 'Selected Data', and 'Rules'. The 'Rules' tab is active, showing a 'Include all Buckets' checkbox which is checked ('Yes').

6. In the **Specify Policy** section, under **Select a data backup** policy, select a backup policy with or without cloud retention.

This screenshot shows the 'Specify Policy' section. It starts with a heading 'Select a data backup policy' followed by a dropdown menu set to 'Backup to Cloud (s3)'. Below that is a 'Retention' section with the sub-instruction 'Allow retention on cloud'. Underneath is a 'Yes' button, which is highlighted, and a 'No' button.

7. In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

This screenshot shows the 'Specify Options' section. It has a 'Cloud options' section with a help icon. Below it are two dropdown menus: 'Data Repository' set to 'S3 Cloud Storage' and 'Buckets' set to 'hadoop_dr_archive'.

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).

The screenshot shows the 'Specify Options' step of a backup configuration. Under 'Cloud options', 'Data Repository' is selected, and 'Containers' is chosen from the dropdown. Below this, 'Azure Cloud Storage' is selected, and 'hadoop_dr_backup' is chosen from the dropdown.

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a backup configuration. Under 'Cloud options', 'Data Repository' is selected, and 'Buckets' is chosen from the dropdown. Below this, 'S3 Cloud Storage' is selected, and a dropdown menu is open, showing 'hadoop_dr_backup' as the current selection.

NOTE: In Couchbase, Tombstones are records of expired or deleted items that include item keys and metadata. Couchbase deletes tombstones permanently after metadata purge interval has elapsed. Due to this process, Couchbase Database Change Protocol (DCP) issues rollbacks when a client requests for mutations (incremental) after tombstones are removed, which may lead to a full backup. If the metadata purge interval and job interval is configured such that incremental backup always runs after purge, every backup may end up getting a rollback. Refer to Couchbase documentation here for information:
<https://developer.couchbase.com/documentation/server/3.x/admin/Concepts/concept-tombstone.html>

NOTE: Customers should set the backup frequency less than their metadata purge interval. Typically, it is recommended to keep the metadata purge intervals to 7 days just to be on safer side.

8. In the **Destination** section, select the destination data repository from the drop-down menu where you want to move the backed-up data.

The screenshot shows the 'Destination' section of a backup configuration. Under 'Data Repository', 'Couch_Stage' is selected from the dropdown.

9. In the **Mirroring Options**, in the **Additional Options** area, under **Overwrite Behavior** do one of the following:
- Click **Replace** to replace existing data with new data thus erasing any previously existing data
 - Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
 - Click **Append** to add new data to an existing bucket

The screenshot shows a user interface titled "1 Mirroring Options". Under the heading "Overwrite Behavior", three buttons are displayed: "Replace", "Keep" (which is highlighted in grey), and "Append".

10. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, 19092019.

The screenshot shows a user interface titled "2 More Options for Selected Data". Under the heading "Suffix", the value "19092019" is entered into a text input field.

11. In the **More Options for Selected Data**, edit the object name and rename it, if needed.

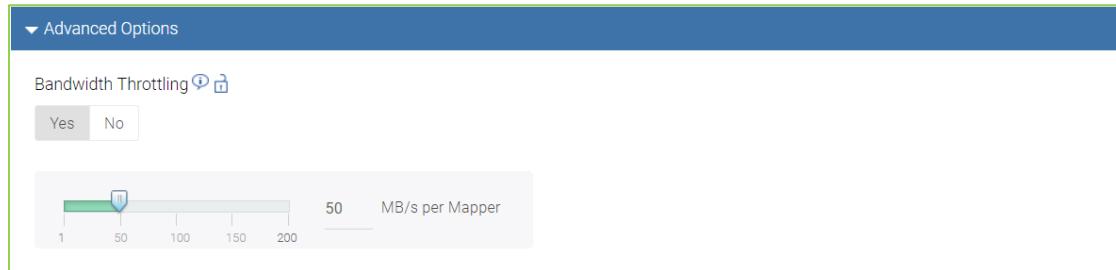
The screenshot shows a user interface titled "2 More Options for Selected Data". A table lists objects and their recovered names:

Objects	Recover As
human resource	human resource19092019
sales	sales19092019
payroll	payroll19092019
legal	legal19092019

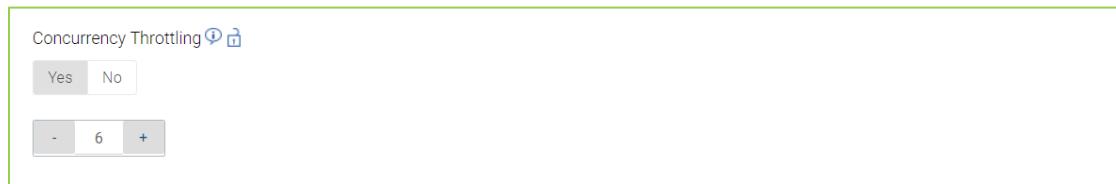
12. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each

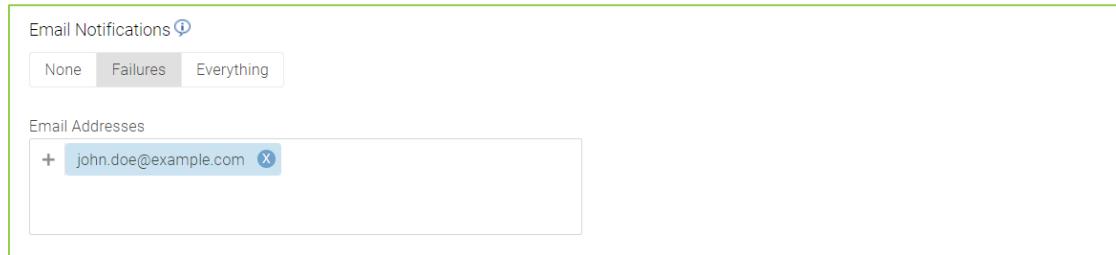
individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **in the MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



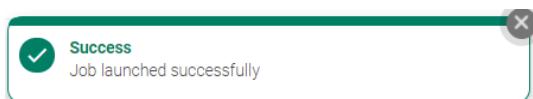
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



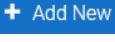
NOTE: Recovery of Ephemeral and Memcached buckets is not supported.

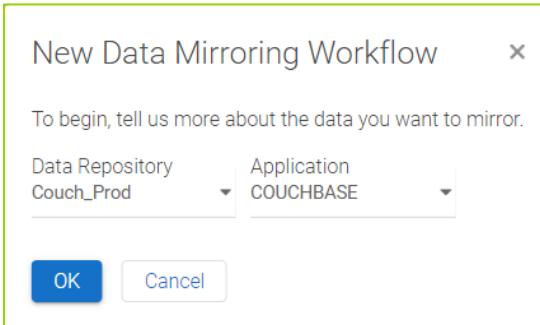
IMPORTANT: Imanis Data backs up the latest state of the data from Couchbase even if the data is not persisted.

10.1.6.1 Data Masking & Sampling for Couchbase

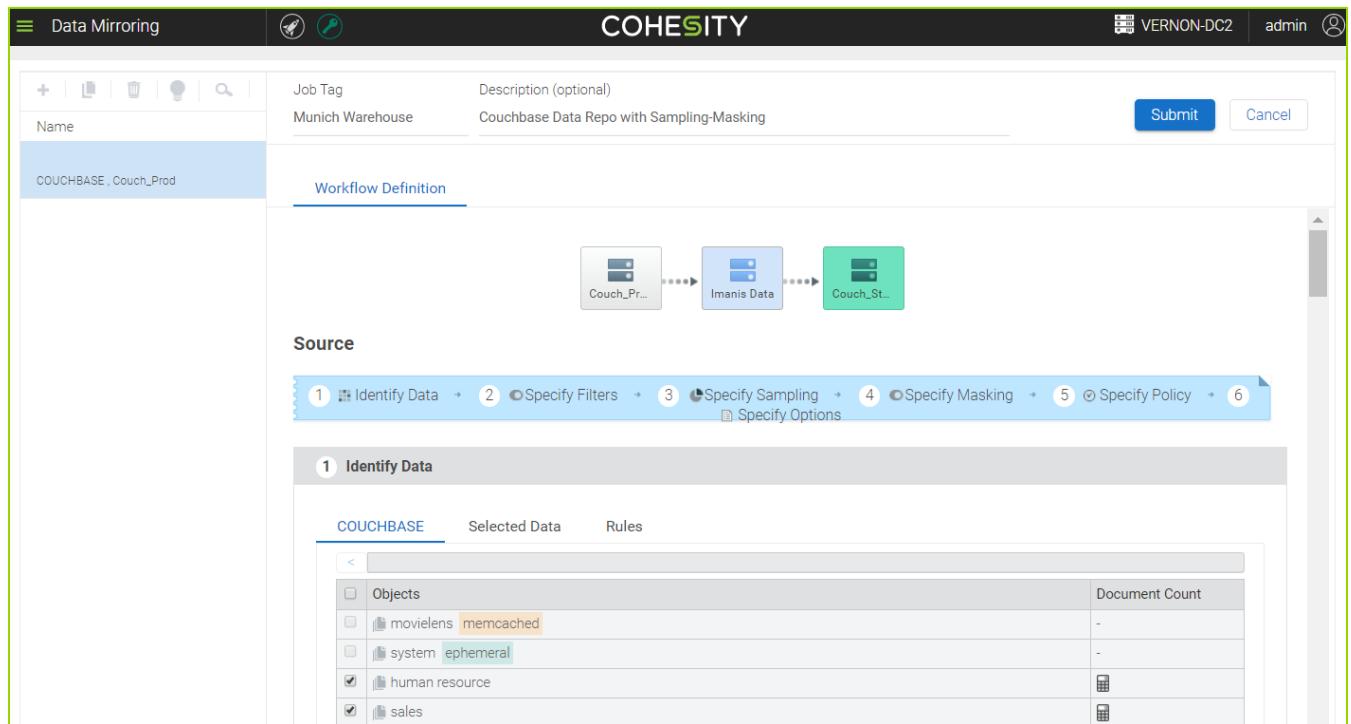
Imanis Data software supports data masking and sampling for Couchbase data sets at the jobtag and bucket level.

To start a data mirroring workflow for Couchbase, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the **Data Mirroring** page, click the  **+ Add New** button or the  icon. The **New Data Mirroring** dialog appears.



3. In the **New Data Mirroring Workflow** dialog, select a source data repository from the **Data Repository** drop-down menu, select **Couchbase** from the **Application** drop-down menu, and then click **OK**.
4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Source** section, in the **Identify Data** area, do the following:
 - a. In the **Couchbase** tab, identify the jobtag and bucket level data that you want to backup by selecting the corresponding check boxes. The following screenshot displays buckets:



- b. In the **Selected Data** tab, verify your selection or click the **X** icon to remove unwanted items.
- c. In the **Rules** tab, click **Yes** to include all buckets in the backup job. This will also include any buckets that get added in the future.



6. In the **Specify Filter** section, click the **Apply** link. For example, like the **Apply** link for **human resource** option.

2 Specify Filters	
Buckets	Filters
human resource	Apply
sales	Apply
payroll	Apply
legal	Apply

NOTE: Imanis Data software supports filtering only on document keys and document content.

In full data recovery, both document key and document content filtering is supported:

This option is visible only if "bucket" is selected on full data recovery.

In Couchbase Recovery module, under **Documents**, under **Select Filtering mechanism**, click **Document ID** and type a regular express to match document IDs.

In incremental data recovery, only document content filtering is supported:

All terms must be separated by one space

Comparisons with "<=", "<", ">=", ">", "==", "!=" are supported. &&, || for and, or respectively.

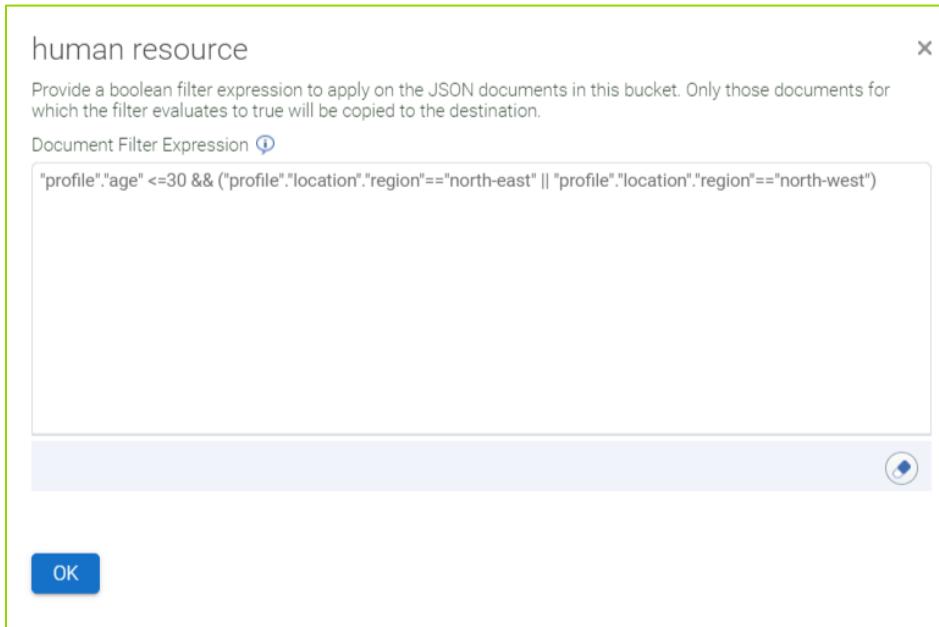
Filtering on nested items like {"i1" : {"i2" : 43}} can be specified as "i1".i2" > 21

Any condition where variables are not found in json is treated as false. Json docs are restored only if filter expression evaluates to true. Binary docs are always restored.

Parenthesis are also supported, however, whitespace are not permitted after opening brace and before closing brace. For example,

"ibu" > 0 && ("type" == "beer" || "abv" > 5) .

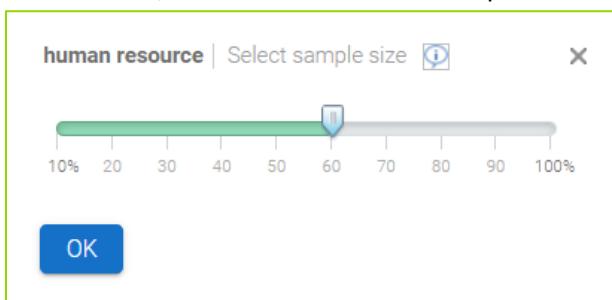
7. In the **Human Resource** dialog box, type a Boolean filter expression to apply on the JSON documents in this bucket. Only those documents for which the filter evaluates to true will be copied to the destination. For example, in the following screenshot you can select employee profiles under the age of 30 that belong to north-east or north-west regions.



8. In the **Specify Sampling** section, click the **100%** link.

3 Specify Sampling	
Objects	Sample Size
human resource	100%
sales	100%
payroll	100%
legal	100%

9. In the Sample Size dialog box for the bucket, for example, **human resource** dialog box for the bucket the **Warehouse1**, move the slider to set sample size:



10. In the **Specify Masking** section, click the **Apply** link.

Objects	Masks	Exclude binary Documents <small>i</small>
human resource	Apply	ON
sales	Apply	ON
payroll	Apply	ON
legal	Apply	ON

11. In the **human resource** dialog box, type attribute, set data type, and then set type of mask in the respective fields, and then click **OK**:

Attribute	Data Type	Mask
"property1" "property2"	String	Taxpayer Identification Number: 12-3456789 <small>▼</small> <small>X</small>

+ OK

12. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

Select a data backup policy
Backup to Cloud (s3)

Retention
Allow retention on cloud
Yes No

13. In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a configuration wizard. Under 'Cloud options', 'Data Repository' is set to 'S3 Cloud Storage' and 'Buckets' is set to 'hadoop_dr_archive'. A dropdown arrow is visible next to 'hadoop_dr_archive'.

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).

The screenshot shows the 'Specify Options' step of a configuration wizard. Under 'Cloud options', 'Data Repository' is set to 'Azure Cloud Storage' and 'Containers' is set to 'hadoop_dr_backup'. A dropdown arrow is visible next to 'hadoop_dr_backup'.

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a configuration wizard. Under 'Cloud options', 'Data Repository' is set to 'S3 Cloud Storage' and 'Buckets' has a text input field containing 'hadoop_dr_archive'. A dropdown arrow is visible next to the input field.

14. In the **Destination** section, select the destination data repository from the drop-down menu where you want to move the backed-up data.

The screenshot shows the 'Destination' section of a configuration wizard. The 'Data Repository' dropdown menu is open, showing 'Couch_Stage' as the selected option. A dropdown arrow is visible next to 'Couch_Stage'.

15. In the **Mirroring Options**, in the **Additional Options** area, under **Overwrite Behavior**, do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied
- Click **Append** to add new data to an existing bucket

The screenshot shows a user interface titled "1 Mirroring Options". Under the heading "Overwrite Behavior", there are three buttons: "Replace", "Keep" (which is highlighted in grey), and "Append".

16. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, 19092019.

The screenshot shows a user interface titled "2 More Options for Selected Data". Under the heading "Suffix", there is a text input field containing the value "19092019".

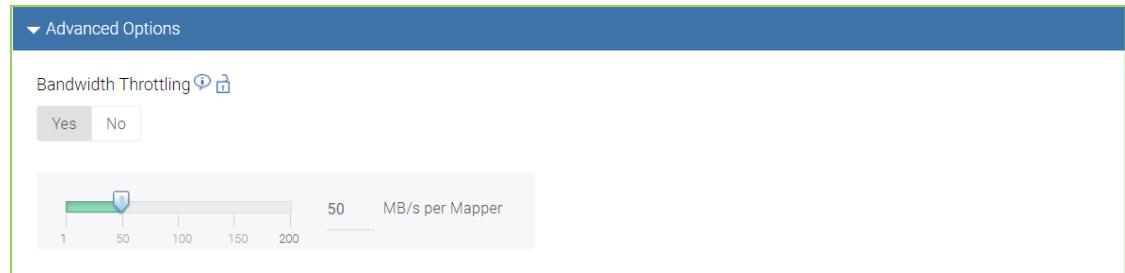
17. In the **More Options for Selected Data**, edit the object name and rename it, if needed.

The screenshot shows a user interface titled "2 More Options for Selected Data". It displays a table with two columns: "Objects" and "Recover As". The table contains five rows, each showing an object name and its corresponding renamed version with a suffix:

Objects	Recover As
human resource	human resource19092019
sales	sales19092019
payroll	payroll19092019
legal	legal19092019

18. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

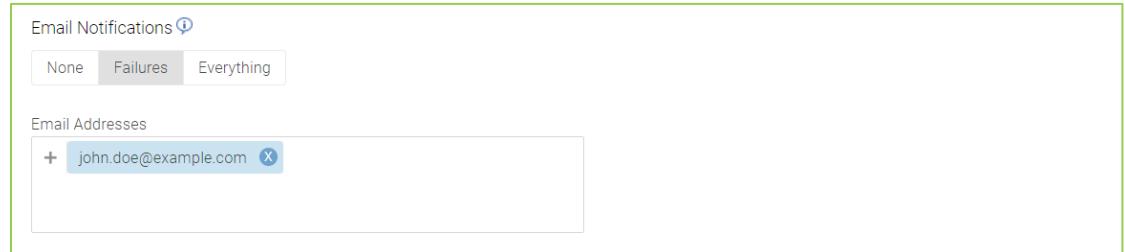
- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



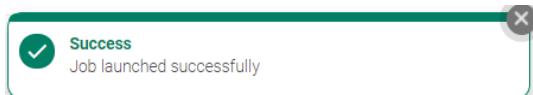
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



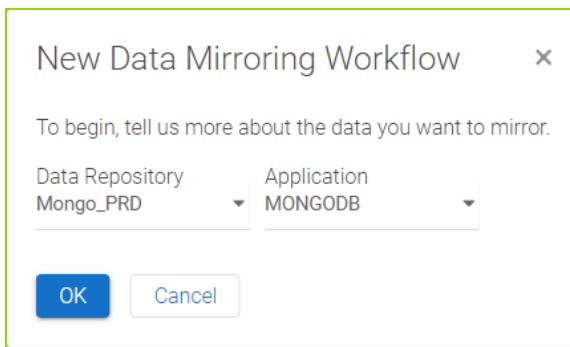
NOTE: Recovery of Ephemeral and Memcached buckets is not supported.

10.1.7 Data Mirroring for MongoDB

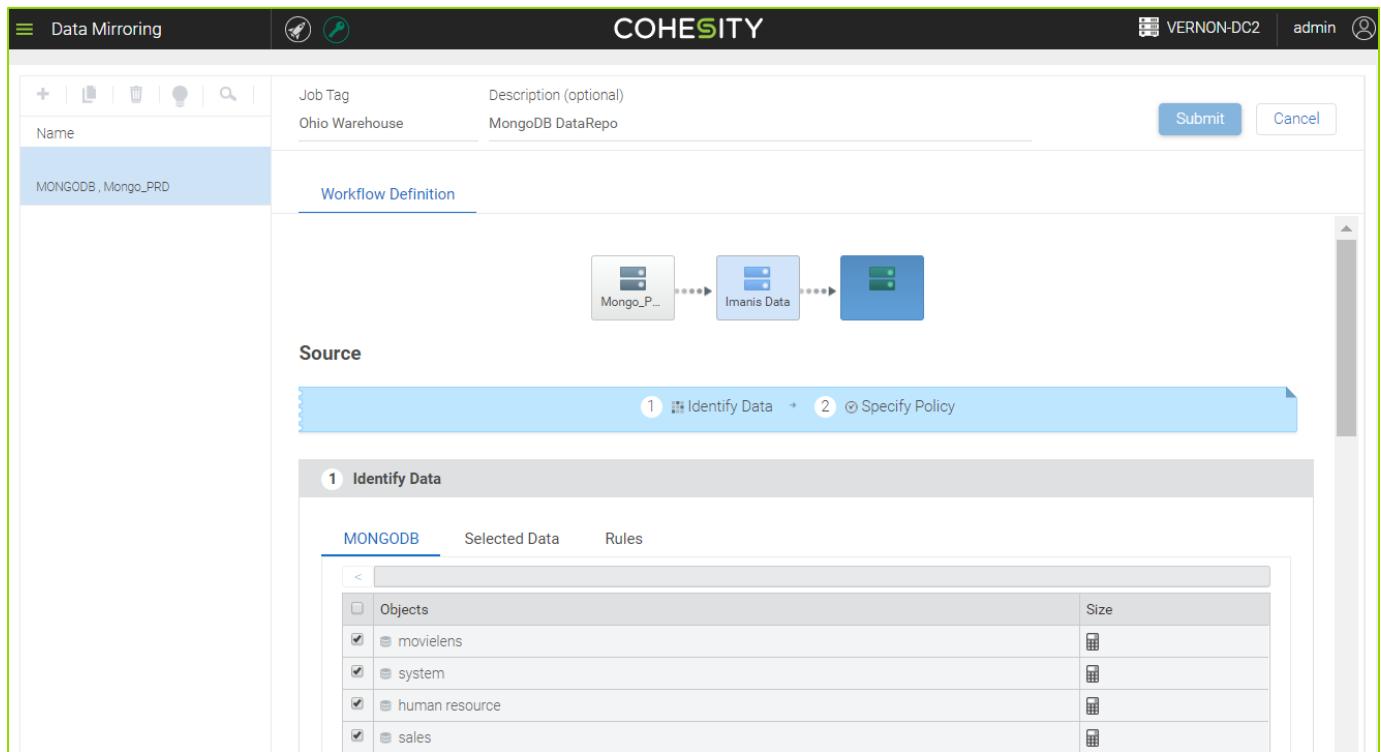
Imanis Data software supports data mirroring for MongoDB data sets at the database and collection level.

To start a data mirroring workflow for MongoDB, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the **Data Mirroring page**, click the  **+ Add New** button or the  icon. The **New Data Mirroring Workflow** dialog box appears.



3. In the **New Data Mirroring Workflow** dialog box, do the following:
 - a. Select a source data repository from the **Data Repository** drop-down menu.
 - b. Select **MongoDB** from the **Application** drop-down menu and click **OK**.
4. In the **Data Mirroring** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Source** section, in the **Identify Data** area, do the following:
 - a. In the **MONGODB** tab, select Database or Collection.
 - b. In the **Selected Data** tab, verify your selection or click the  icon to remove unwanted items.
 - c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.



To know how to specify Collection specific properties ('shardKeyJson' and 'unique'), refer to this section: More Options for Selected Data.

What exactly are the Collection specific properties shardKeyJson & unique?

* {{shardKeyJson}}: The index specification document to use as the shard key. The shard key determines how MongoDB distributes the documents among the shards. This should be specified as a JSON. Example: {{\{_id: "hashed"\}}}

* {{unique}}: When true, the unique option ensures that the underlying index enforces a unique constraint. Hashed shard keys do not support unique constraints. This should be specified as either {{true}} or {{false}}

For details please refer to MongoDB documentation at

<https://docs.mongodb.com/manual/reference/method/sh.shardCollection/>

6. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

The screenshot shows a configuration interface for specifying a data backup policy. At the top, a header reads "2 Specify Policy". Below it, a section titled "Select a data backup policy" has a dropdown menu set to "Backup to Cloud (s3)". Underneath, a "Retention" section includes the text "Allow retention on cloud" followed by two radio buttons: "Yes" (selected) and "No".

7. In the **Specify Options** section, under **Cloud Options**, do one of the following:

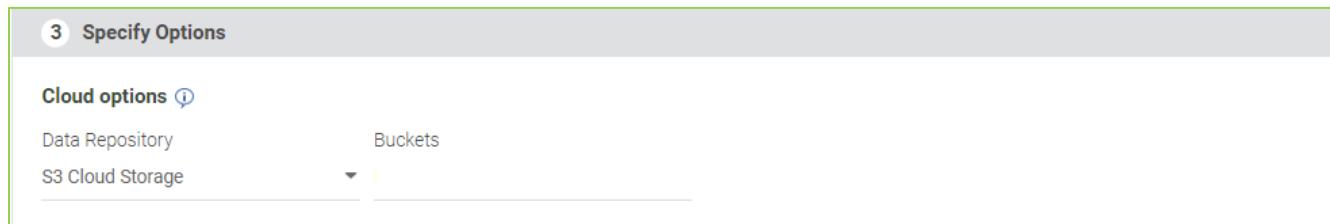
- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

The screenshot shows a configuration interface for specifying options. At the top, a header reads "3 Specify Options". Below it, a "Cloud options" section has a "Data Repository" dropdown set to "S3 Cloud Storage" and a "Buckets" dropdown set to "mongo_qa_backup".

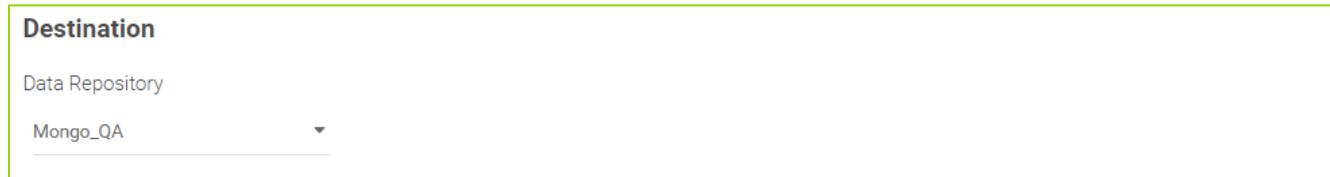
- Select an Azure cloud repository from the Data Repository drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).

The screenshot shows a configuration interface for specifying options. At the top, a header reads "3 Specify Options". Below it, a "Cloud options" section has a "Data Repository" dropdown set to "Azure Cloud Storage" and a "Containers" dropdown set to "mongo_qa_backup".

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

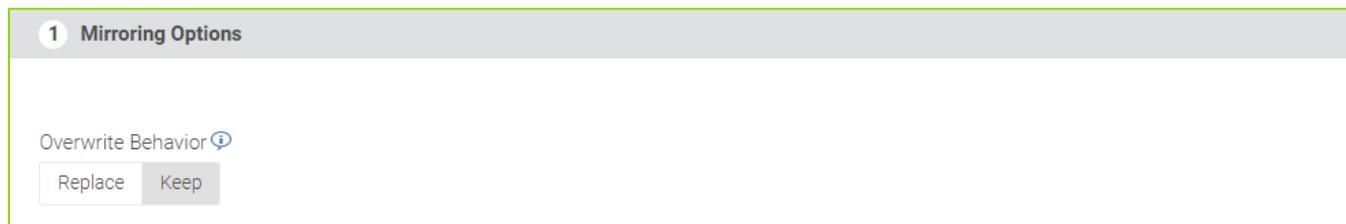


8. In the **Destination** section, select the destination data repository from the drop-down menu where you want to move the backed-up data.

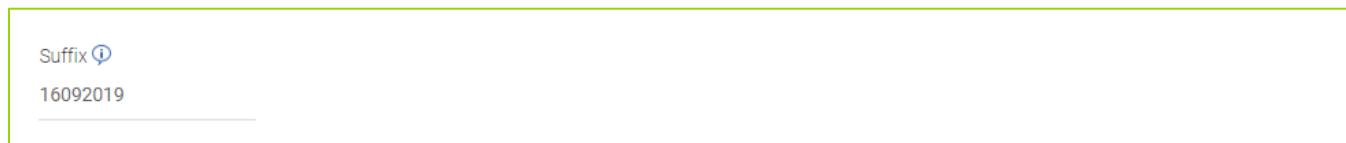


9. In the **Mirroring Options**, under **Overwrite Behavior** do one of the following:

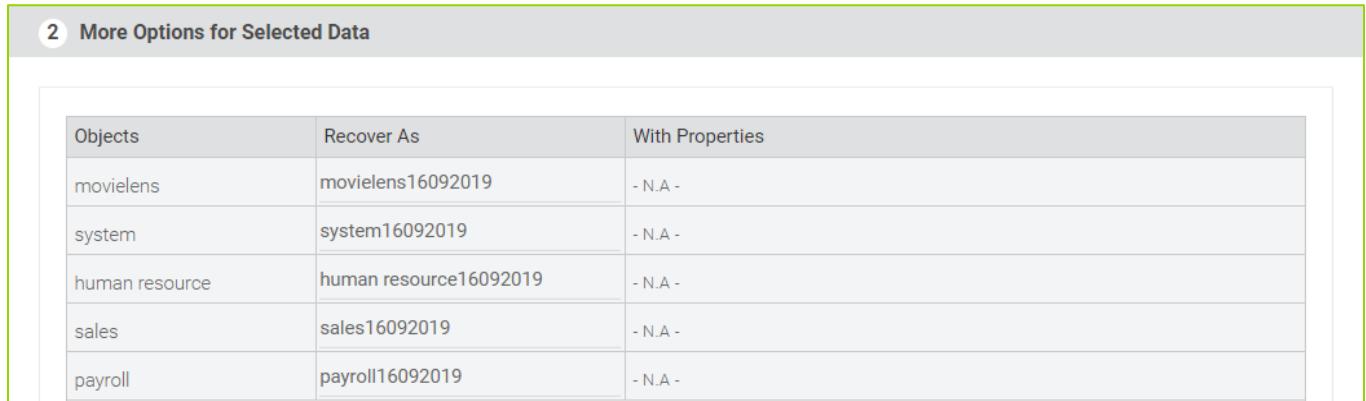
- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied



10. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _24012017.



11. In the **More Options for Selected Data**, you can specify Collection specific properties ('shardKeyJson' and 'unique') in the GUI. See the following screenshot:



The screenshot shows a table titled 'More Options for Selected Data' with the following data:

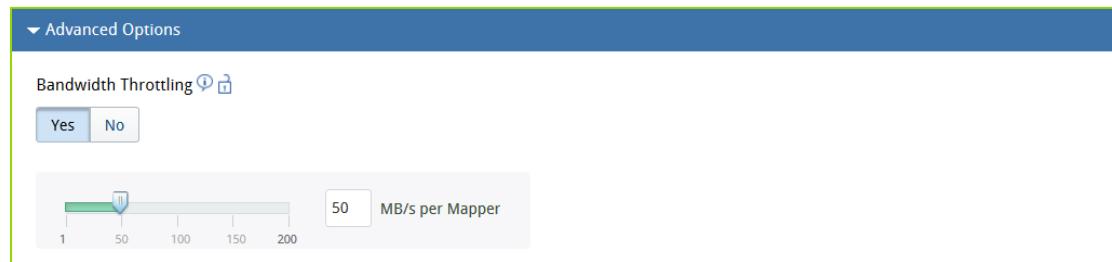
Objects	Recover As	With Properties
movielens	movielens16092019	- N.A -
system	system16092019	- N.A -
human resource	human resource16092019	- N.A -
sales	sales16092019	- N.A -
payroll	payroll16092019	- N.A -

IMPORTANT: The following information regarding limitations of specifying collection specific properties ('shardKeyJson' and 'unique') must be noted:

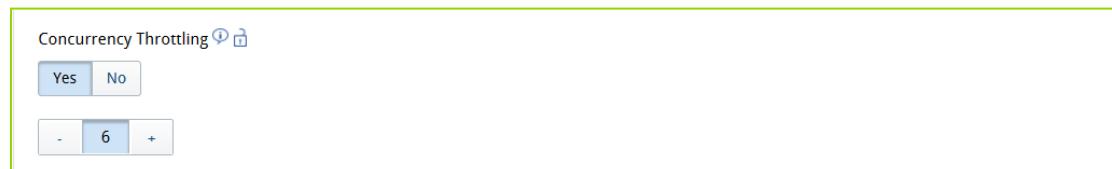
1. Currently, these properties are not auto detected from Primary in the case of Mirroring workflow. However, these properties can be specified in the 'Rename with Properties' section while creating a Mirroring job. These values will then override any existing configuration on the relevant collections.
2. In case of Restore, the existing values for these properties, if any, are displayed, that is, the values for these properties when the collections were backed up are displayed. You can choose to add, update, delete these property values.
3. The Imanis Data GUI does not enforce any relationship between these two properties while collecting this information from the user. The user is expected to input a valid combination of the two properties. If any of the parameters left blank, corresponding value from the source collection would be used. The Imanis Data GUI validates that the user has entered a valid Json as the value for shardKeyJson. However, the schema for this Json is not validated. Similarly, GUI enforces that as the value for 'unique' property, the only two values that are permitted to be entered are true or false.

12. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Backup** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

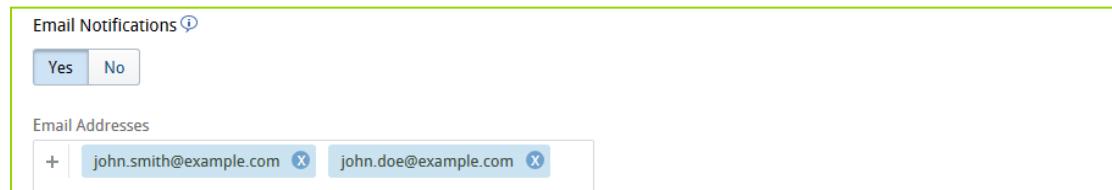
- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the in the MB/s per Mapper field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the Email Addresses field that will receive the job completion and/or failure notifications:



13. Click **Submit**.

10.2 Direct Replication for Hadoop

You can back up exact copy of data directly from the primary or source cluster to the destination cluster. However, you must create a direct replication policy before proceeding to create the direct replication workflow.

In the current release, direct replication can be executed for Hadoop (HDFS, Hive, and Hbase) only. Now the obvious question you may have is what is the difference between a Data Mirroring workflow and the Direct Replication workflow?

In the Data Mirroring workflow, an exact copy of data is replicated on the Imanis Data cluster as well as the destination cluster. The Data Mirroring workflow is a 2-step process:

1. Data is copied from the primary cluster to Imanis Data cluster

2. Data is copied from the Imanis Data cluster to destination cluster

However, in the Direct Replication workflow, an exact copy of data is directly replicated onto the destination cluster. In this process, the step of replicating an exact copy of data on to Imanis Data cluster is excluded.

10.2.1 Direct Replication for HDFS and HBase

For illustration purposes, let's get to know the procedure of executing the direct replication workflow for HDFS data objects in the following paragraph. However, the procedure remains the same for HBase too.

During data backup, HBase does not record the number of regions for a table. When a table is restored to a destination cluster, the table is created with a single region only. This single region may then get split into many regions depending on the size of the data and the configuration of the destination cluster.

However, there is a special configuration on Imanis Data cluster that will enable HBase to record the number of regions for a table during backup. Thus, during the data restore, the table will be created with the same number of regions as recorded during the backup.

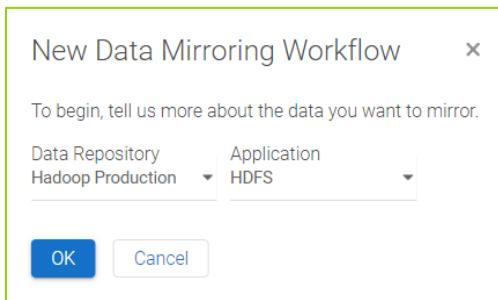
NOTE: In the current release, HDFS SOLR workflow is not supported in HDFS Direct Replication workflow.

To enable this configuration, add the following key to hdfs-site.xml on Imanis Data cluster:

```
<property>
<name>hbase.restore.table.pre.split</name>
<value>true</value>
</property>
```

To directly replicate data, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the **Data Mirroring** page, click the  **+ Add New** button or the  icon. The **New Data Mirroring Workflow** dialog appears:
3. In the **New Data Mirroring Workflow** dialog box, select a **Hadoop** source data repository from the **Data Repository** drop-down menu, select **HDFS** from the **Application** field, and then click **OK**.



4. In the **Data Mirroring** page, type a new job tag in the **Job Tag** field and a jog tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Source** section, in the **Identify Data** area, do the following:
 - a. In the **HDFS** tab, identify the file and directory level data that you want to backup by selecting the corresponding check boxes.
 - b. In the **Selected Data** tab, verify your selection or click the **X** icon to remove unwanted items.
 - c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

	Objects	Modified Time	Owner	Size
<input checked="" type="checkbox"/>	movielens	2019-09-18 07:48:52 ...	talena	2.8 GB
<input checked="" type="checkbox"/>	system	2019-09-18 07:48:52 ...	talena	29.7 GB
<input checked="" type="checkbox"/>	human resource	2019-09-18 07:48:52 ...	talena	937.6 MB

6. In the **Specify Policy** section, under the **Keep a copy on Cohesity Imanis DataCluster** option, do the following:
 - a. Click **No** to directly replicate data on destination. **Yes** option is selected by default for mirroring.
 - b. Select a direct replication policy from the **Select a Direct Replication policy** drop-down menu.

2 Specify Policy

Keep a copy on Cohesity Imanis Data cluster Yes No

Select a Direct Replication policy
Direct Replication Policy

Replication Schedule

Priority

NOTE: If the **Keep a copy on Imanis** option is set to **No**, then HDFS SOLR indexes will not be restored on the destination.

7. In the **Destination** section, select the destination **Hadoop** data repository from the **Data Repository** drop-down menu where you want to move the backed-up data.

Destination

Data Repository
Hadoop Production

8. In the **Mirroring Options** area, under the **Overwrite Behavior** option, do one of the following:
 - Click **Replace** to replace existing data with new data thus erasing any previously existing data
 - Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied

1 Mirroring Options

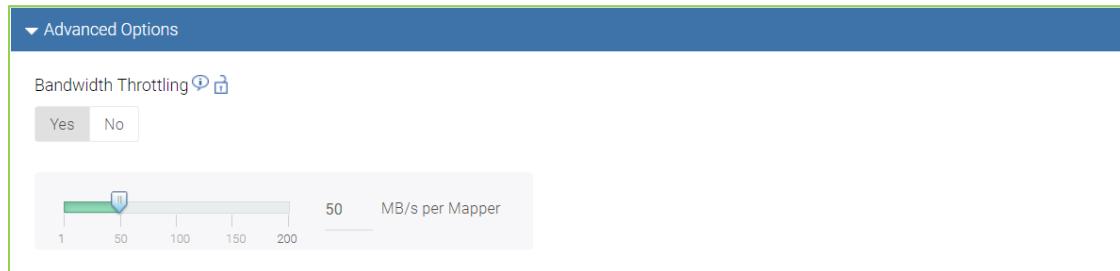
Overwrite Behavior Replace Keep

9. In the **Destination Directory** field, type a name for the directory. For example, /18092019. Imanis Data software creates a new directory with this name.

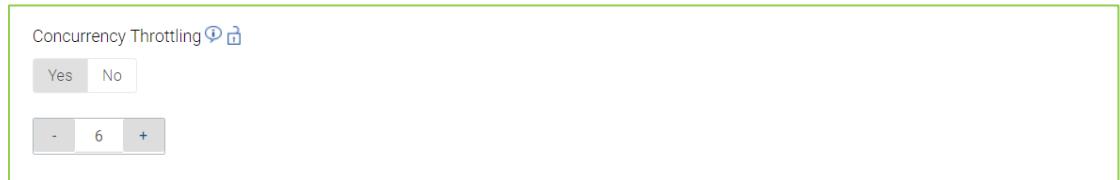
Destination Directory
/18092019

10. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

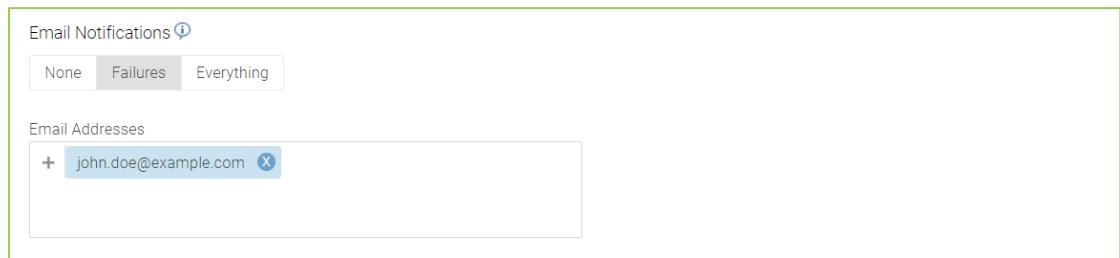
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



11. Click **Submit**.

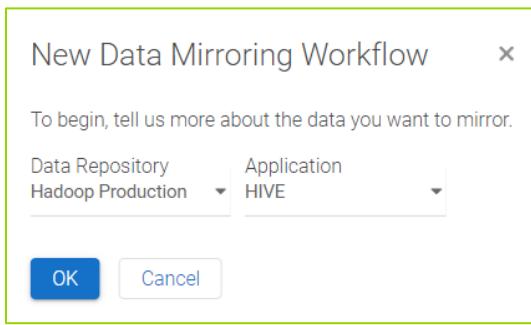
10.2.2 Direct Replication for HIVE

While adding Hive data repository if Hive Server (which is required only if sampling and/or masking data) is activated or enabled, then the Sampling and Masking sections will be displayed on the page. The sampling-masking sections is not supported in direct replication.

NOTE: If a Hive data repository added prior to 3.3.0 is being used for direct replication, then the data repository must be re-verified and saved. Hence, you must access the Data Repository page, identify the Hive data repository added before 3.3.0 release, click the edit button, verify the data repository, and then click the Save button.

To directly replicate data, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the Data Mirroring page, click the  **+ Add New** button or the  icon. The **New Data Mirroring Workflow** dialog appears.
3. In the **New Data Mirroring Workflow** dialog box, select a **Hadoop** source data repository from the **Data Repository** drop-down menu, select **HDFS** from the **Application** field, and then click **OK**.



4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Source** section, in the **Identify Data** area, do the following:
 - a. In the **HDFS** tab, identify the file and directory level data that you want to backup by selecting the corresponding check boxes.
 - b. In the **Selected Data** tab, verify your selection or click the  icon to remove unwanted items.
 - c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

The screenshot shows the 'Data Mirroring' section of the Cohesity interface. A job named 'Sydney Wareouse' is being configured with a 'Hive Data Repo with Direct Replication Policy'. The 'Workflow Definition' shows a flow from 'Hadoop P...' to 'Imanis Data'. The 'Source' section includes steps: 1. Identify Data, 2. Specify Masking, 3. Specify Sampling, and 4. Specify Policy.

6. In the **Specify Policy** section, under the **Keep a copy on Imanis** option, do the following:
 - a. Click **No** to directly replicate data on destination. By default, the **Yes** option is selected for the regular data mirroring workflow.
 - b. Select a direct replication policy from the **Select a Direct Replication policy** drop-down menu.

The 'Specify Policy' screen shows the following settings:

- Keep a copy on Cohesity Imanis Data cluster**: No (radio button selected)
- Select a Direct Replication policy**: Direct Replication Policy (dropdown menu)
- Replication Schedule**: One time, immediately (radio button selected)
- Priority**: High (radio button selected)

7. In the **Destination** section, select the destination **Hadoop** data repository from the drop-down menu where you want to move the backed up data.

Destination

Data Repository

Hadoop DR

8. In the **Mirroring Options** area, under the **Overwrite Behavior** option, do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied

1 Mirroring Options

Overwrite Behavior ⓘ

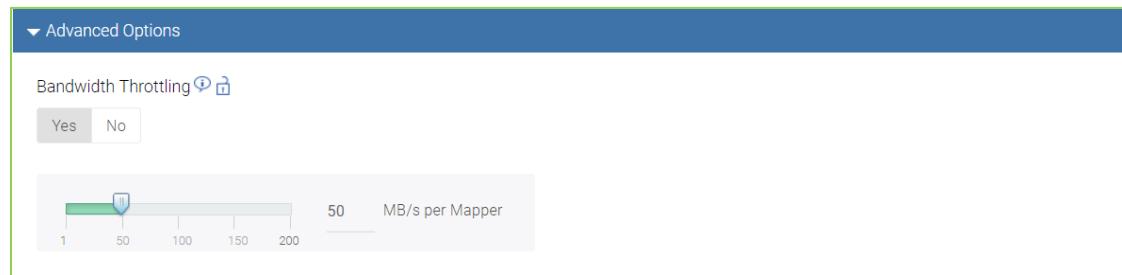
9. In the **Suffix** field, type a number and/or character to add a suffix to the data objects being recovered from the Imanis Data cluster. For example, _25052018.

Suffix ⓘ

18092019

10. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

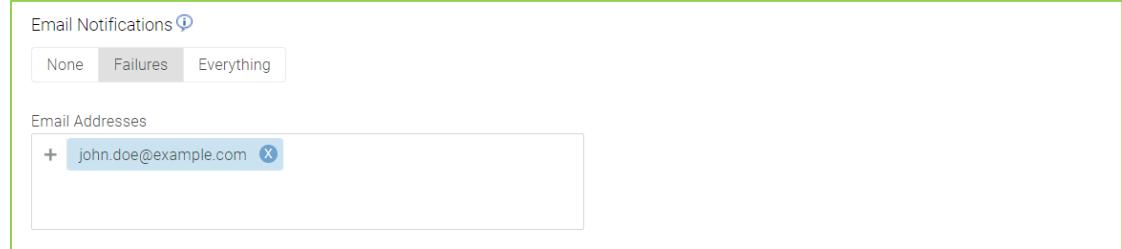
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



11. Click **Submit**.

11 Data Pipeline

This section describes the features of the Data Pipeline menu in Imanis Data software.

11.1 Overview

Data Pipeline is the process of moving data objects from a production cluster to one or more non-production clusters such as test, research, development, and so on.

11.2 Getting Started with Data Pipeline

Data is backed up from the primary data repository and moved to another cluster referred to as the destination. In the Data Pipeline user interface, Source represents a data repository from which the data is copied from and Destination represents a data repository where the data will be eventually copied.

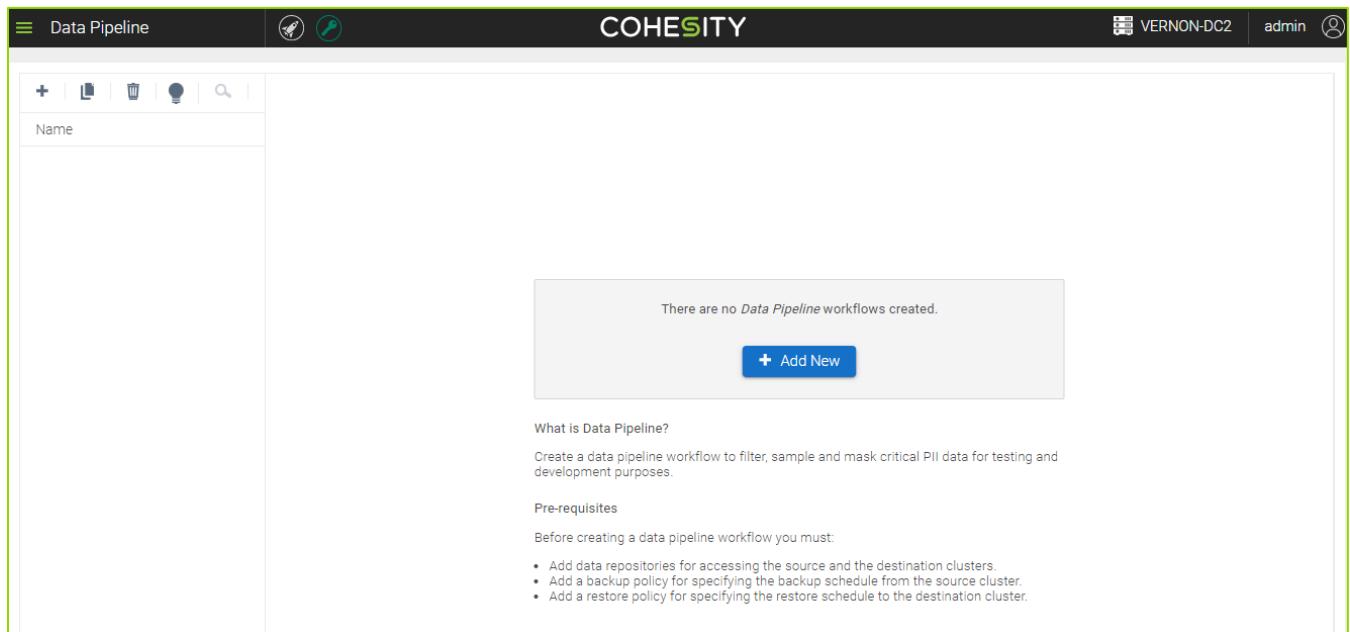
IMPORTANT: The administrator must create backup and restore policies prior to creating a Data Pipeline workflow.

11.2.1 Data Pipeline for HDFS or Hive

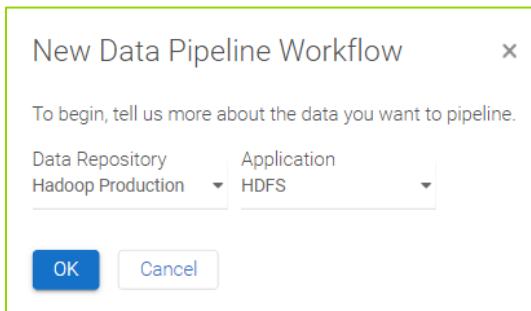
You can create data pipeline workflows for both HDFS and Hive in Imanis Data software.

To start data pipeline workflow for HDFS or Hive, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Pipeline**. The following page appears:



2. On the **Data Pipeline** page, click the **+ Add New** button or the icon to create data pipeline workflows. The **New Data Pipeline Workflow** dialog box appears.
3. In the **New Data Pipeline Workflow** dialog box, select the source data repository from the **Data Repository** drop-down menu, select **HDFS** or **Hive** from the **Application** drop-down menu, and then click **OK**.



4. In the **Backup Page**, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.

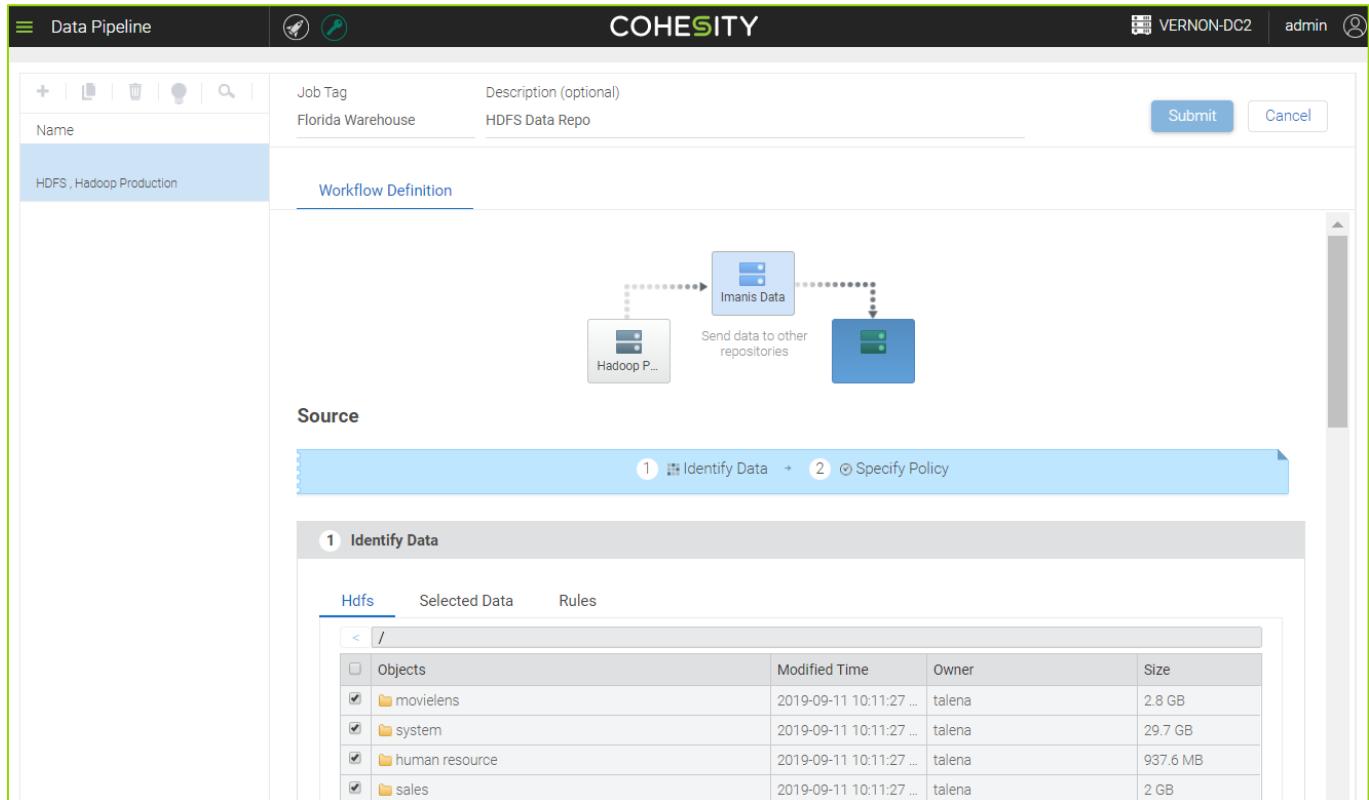
5. In the **Identify Data** area, do the following:

- In the **HDFS** tab, identify the files, tables, directories, databases, and partitions that you want to backup by selecting the corresponding check boxes. Use regular expressions (regex) for primary repository browsing.

IMPORTANT: The Imanis Data does not recommend having blank spaces in the inclusion or exclusion regex.

- In the **Selected Data** tab, verify your selection or click the **X** icons to remove unwanted items.
- In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

Refer to the section Appendix A: Rules for Data Inclusions and Exclusion.



NOTE: Imanis Data software enables administrators to do primary browsing and to specify glob expressions, a regex like expression, to find specific tables from a database. For example, db* retrieves all the databases starting with 'db' or database1.tb* and retrieves all the tables in database1 starting with "tb". However, Imanis Data software does not support db*.tbl*, where both the database and table is unknown. In this case, search results may not appear.

6. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

The screenshot shows the 'Specify Policy' screen. At the top, it says 'Select a data backup policy' with 'On Demand Backup' selected. Below that is a 'Retention' section with 'Allow retention on cloud' set to 'Yes'. A diagram shows a file icon connected by a line to a trash bin icon, labeled '20 days' below it. There are 'Yes' and 'No' buttons for the cloud retention setting.

7. In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' screen for S3. Under 'Cloud options', 'Data Repository' is set to 'S3 Cloud Storage' and 'Buckets' is set to 'hadoop_dr_archive'.

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).

The screenshot shows the 'Specify Options' screen for Azure. Under 'Cloud options', 'Data Repository' is set to 'Azure Cloud Storage' and 'Containers' is set to 'hadoop_dr_backup'.

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

3 Specify Options

Cloud options ⓘ

Data Repository Buckets

S3 Cloud Storage

8. In the **Destination 1** section, from the **Data Repository** drop-down menu, select a data repository where the data will be moved to, that is, the target or destination.

Destination 1

Data Repository

Hadoop QA

9. In the **Identify Data** area, do the following:

- In the **HDFS** tab, clear the check boxes for the corresponding files, tables, directories, databases, and partitions that you DO not want to recover to Destination 1.
- In the **Selected Data** tab, verify your selection or click the **X** icons to remove unwanted items.
- In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section Appendix A: Rules for Data Inclusions and Exclusions.

1 Identify Data

Hdfs Selected Data Rules

	Objects	Modified Time	Owner
<input checked="" type="checkbox"/>	movielens	2019-09-11 10:11:27 ...	talena
<input checked="" type="checkbox"/>	system	2019-09-11 10:11:27 ...	talena
<input checked="" type="checkbox"/>	human resource	2019-09-11 10:11:27 ...	talena
<input checked="" type="checkbox"/>	sales	2019-09-11 10:11:27 ...	talena
<input checked="" type="checkbox"/>	payroll	2019-09-11 10:11:27 ...	talena
<input checked="" type="checkbox"/>	legal	2019-09-11 10:11:27 ...	talena

10. In the **Specify Policy** area, select a recovery policy from the drop-down menu to recover data in the selected location. You will be able to see recovery policies that were created earlier.

2 Specify Policy

Select a recovery policy

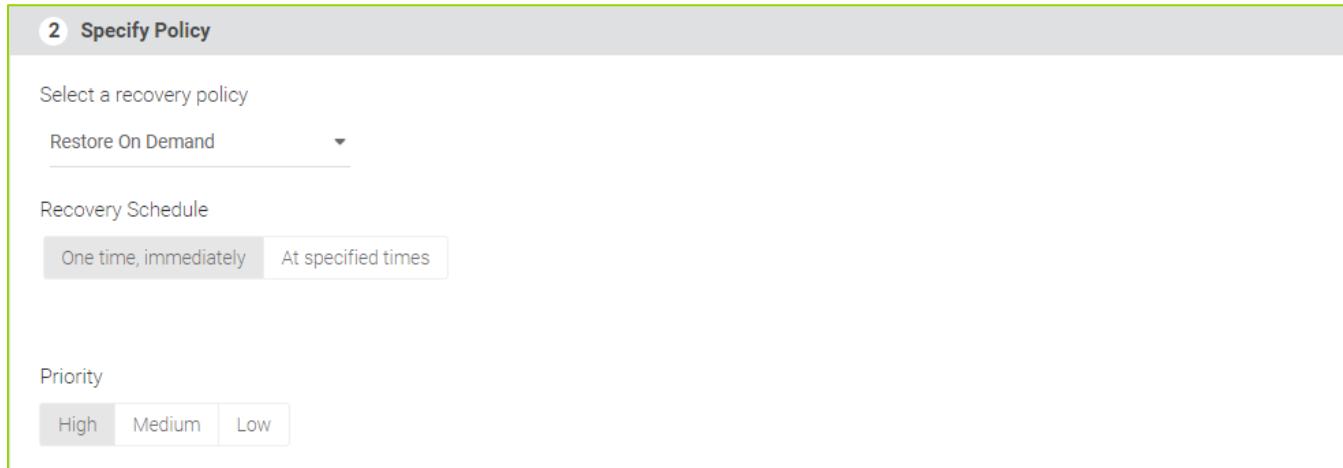
Restore On Demand

Recovery Schedule

One time, immediately At specified times

Priority

High Medium Low



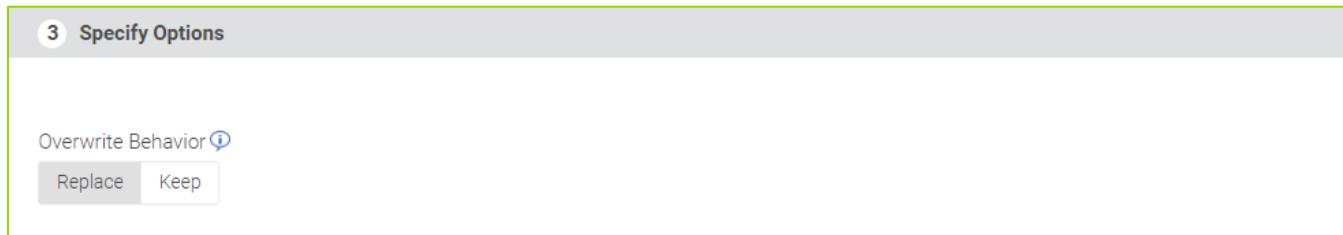
11. In the **Specify Options** area, under **Overwrite**, do one the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data

3 Specify Options

Overwrite Behavior ⓘ

Replace Keep

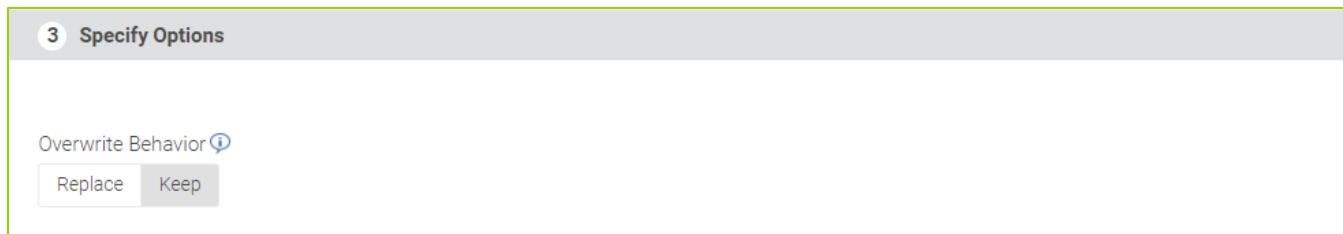


- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied

3 Specify Options

Overwrite Behavior ⓘ

Replace Keep



12. In the **Destination Directory** option, do one of the following:

- For **HDFS**: Type the name for the directory that will be created in the data repository that you selected earlier. For example, /example_directory. This is an optional step to restore data in the same directory or an alternate location. The data is restored in the same directory structure as the source
- For **Hive**, choose to add a suffix or overwrite the data



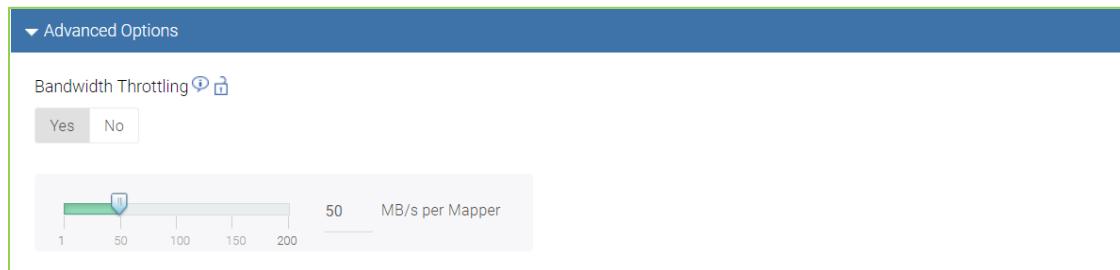
Destination Directory
/example_directory

NOTE: Imanis Data software supports the use of alphanumeric characters in the suffix field; however, the use of uppercase is not recommended.

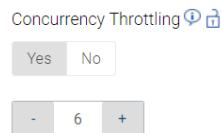
WARNING: In Hive, for example, if you create a data pipeline workflow and select three tables to be recovered with suffix as '_01', then the recovery will be successful. However, if you edit the workflow to add one row each in all three tables with suffix as '_02' and re-run the workflow, then only the new rows that were added will be restored. This is a known issue and will be fixed in future releases.

13. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

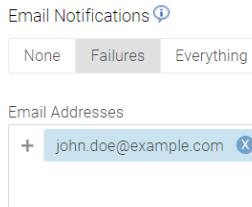
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



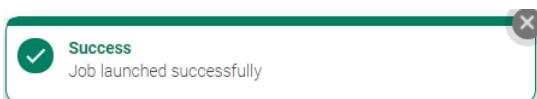
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



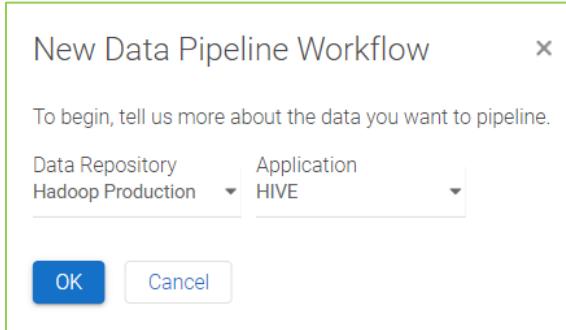
11.2.1.1 Data Masking & Sampling for Hive

Imanis Data software enables enterprises to conceal private and sensitive data such as credit card or social security numbers in testing or development environments with the data masking and sampling feature in the Data Pipeline process.

IMPORTANT: Sampling-Masking is not supported if the primary cluster is configured against JRE 1.7 or earlier versions.

To start data pipeline workflow for Hive, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Pipeline**.
2. On the **Data Pipeline** page, click the  **+ Add New** button or the  icon. The **New Data Pipeline Workflow** dialog box is displayed.



3. In the **New Data Pipeline Workflow** dialog box, select a **Hive** data repository from the **Data Repository** drop-down menu, and then click **OK**.
4. Type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Identify Data** area, do the following:
 - In the **Hive** tab, identify the database that includes private and sensitive data (within tables) that you want to conceal and/or downsample by selecting the corresponding check boxes. For example, click the **WareHouse_NYC1** as shown in the above screenshot
 - In the **Selected Data** tab, verify your selection or click the  icon to remove unwanted items
 - In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

6. In the **Specify Masking** area, click the **Apply** link to apply masking for the tables selected in the preceding step. For example, click **Apply** for the **schema_keyspaces** object table. A new window appears by the name of **schema_keyspaces**.

The screenshot shows a table titled 'Specify Masking' with two columns: 'Objects' and 'Masks'. The 'Objects' column lists several database components: 'schema_keyspaces', 'local', 'peers', 'schema_columns', and 'schema_columnfamilies'. The 'Masks' column contains the word 'Apply' next to each object name, indicating that masking has been applied.

Objects	Masks
schema_keyspaces	Apply
local	Apply
peers	Apply
schema_columns	Apply
schema_columnfamilies	Apply

7. In the **schema_keyspaces** dialog box, do the following:
- Select the corresponding check boxes for the parameters in the **Column Name** column.
 - Select the appropriate masks for the parameters listed in the **Mask** column.

For example, for '**keyspace_name**' select mask as **Full Name: John. E. Doe**. See the following screenshot:

The screenshot shows a dialog box titled 'schema_keyspaces' with a table for configuring masking. The table has three columns: 'Column Name', 'Data Type', and 'Mask'. There are five rows, each corresponding to a column name: 'keyspace_name', 'durable_writes', 'strategy_class', and 'strategy_options'. The 'keyspace_name' row has a 'Mask' value of 'Email: john.doe@somedomain.com'. All other rows have a 'Mask' value of 'None'. At the bottom left of the dialog is a blue 'OK' button.

Column Name	Data Type	Mask
keyspace_name	string	Email: john.doe@somedomain.com
durable_writes	boolean	
strategy_class	string	None
strategy_options	string	None

- c. Click **OK**.

NOTE: In the above screenshot, all the column names are pre-selected by default. You can manually clear the check boxes of column names that you do not want to mask.

8. Optionally to edit the masking, you may click the **Edit** link or **Remove All** link to clear all the masking that you may have applied (see the following screenshot):

2 Specify Masking

Objects	Masks
schema_keyspaces	Edit Remove All
local	Apply
peers	Apply
schema_columns	Apply
schema_columnfamilies	Apply

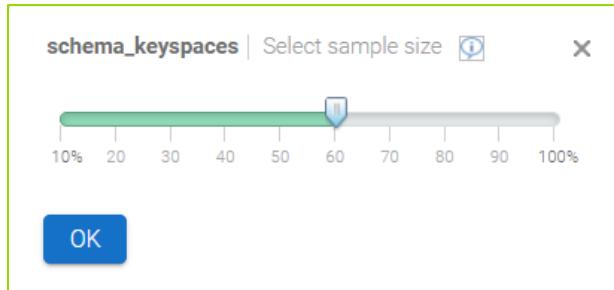
9. In the **Specify Sampling** area, click the link that denotes value **100%**.

3 Specify Sampling

Objects	Sample Size
schema_keyspaces	100%
local	100%
peers	100%
schema_columns	100%
schema_columnfamilies	100%

The following window appears. By default, the sample size is always set at 100%.

10. Move the slider to set a data sample size, and then click **OK**. For example, 60%.



11. Once you click **OK**, the table available in the **Specify Sampling** area displays a new link **On Entire Row**. See the following screenshot:

Objects	Sample Size
schema_keyspaces	60% On Entire Row
local	100%
peers	100%
schema_columns	100%
schema_columnfamilies	100%

12. Click the **On Entire Row** link displayed in the above screenshot. The following page appears: Imanis Data software offers **On Entire Row** and **On Selected Columns** as sampling options.

schema_keyspaces | Select sampling option

On Entire Row On Selected Columns

Sampling will be driven independently of the data and is likely to lead to an accurate sample size.

Available Column	Data Type
keyspace_name	string
durable_writes	boolean
strategy_class	string
strategy_options	string

Selected Column	Data Type

Column names annotated with a * indicate bucketing columns

OK

The ‘**On Entire Row**’ sampling option denotes that sampling is applied on the entire row instead of an individual column. While the ‘**On Selected Columns**’ sampling option is based on the value of the column and denotes the set of columns on which the table is hash-partitioned or clustered on.

NOTE: In the current release, some data may be repeated in case of Sampling when you select the On Entire Row option.

13. Do one of the following:

- Select the On Entire Row option, then click **OK**

OR

- a. Select the **Selected Columns** option
- b. Identify the columns to select and click the icon to transfer the selected column from the **Available Columns** section to the **Selected Columns** section. You can select maximum two columns only.
- c. Click **OK**.

14. In the **Specify Policy** area, under **Select a data backup policy**, select a backup policy with or without cloud retention.

4 Specify Policy

Select a data backup policy

Backup to Cloud (s3)

Retention

Allow retention on cloud

Yes No

15. In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

3 Specify Options

Cloud options ⓘ

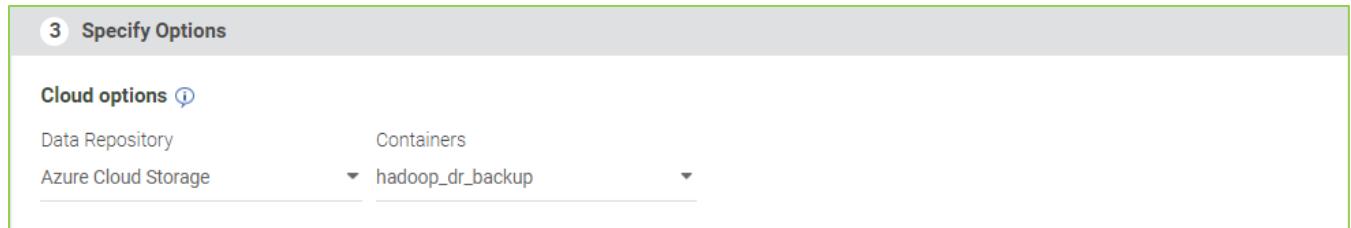
Data Repository

S3 Cloud Storage

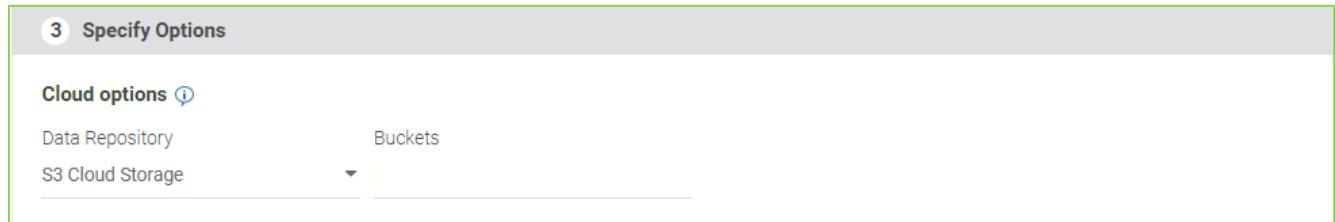
Buckets

hadoop_dr_archive

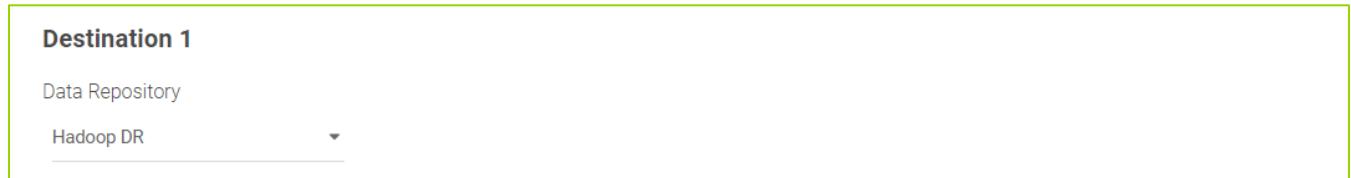
- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).



- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:



16. In the **Destination 1** area, from the **Data Repository** drop-down menu, select a data repository where the data will be moved to, that is, the target or destination.



17. In the **Destination 1** area, in the **Identify Data** area, do the following:

- In the **Hive** tab, clear the check boxes for the corresponding files, tables, directories, databases, and partitions that you DO not want to recover to Destination 1.
- In the **Selected Data** tab, verify your selection or click the **X** icon to remove unwanted items.
- In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

Objects	Owner	Table Type
schema_keyspaces	talena	
local	talena	
peers	talena	
schema_columns	talena	
schema_columnfamilies	talena	

18. In the **Specify Policy** area, select a recovery policy from the drop-down menu to recover data in the selected location.

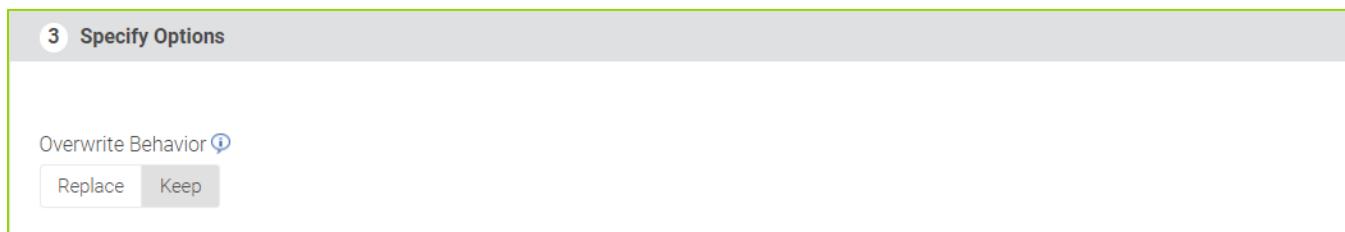
Select a recovery policy
Restore On Demand

Recovery Schedule
One time, immediately At specified times

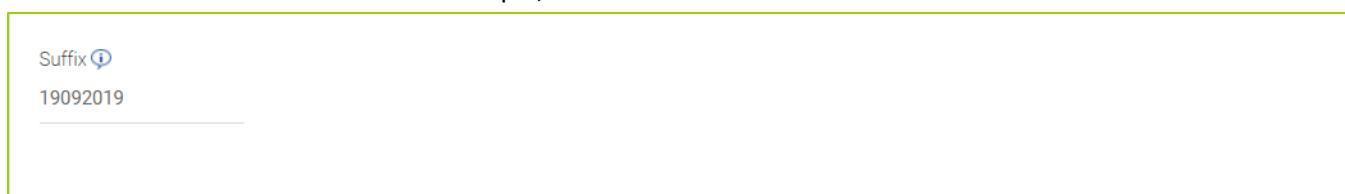
Priority
High Medium Low

19. In the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied



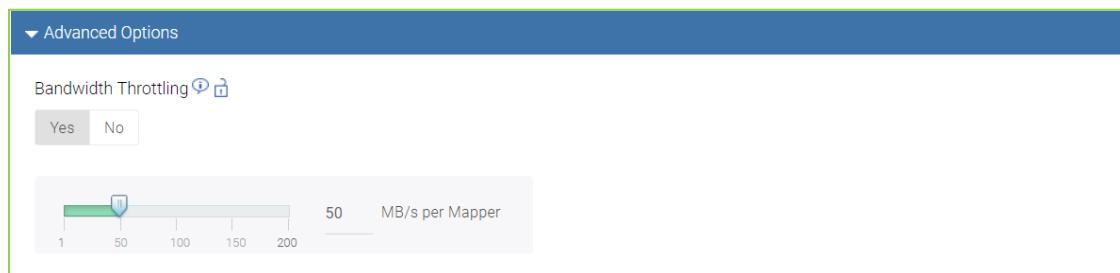
20. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, 19092019.



NOTE: Imanis Data software supports the use of alphanumeric characters in the suffix field; however, the use of uppercase is not recommended.

21. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:

Concurrency Throttling  

Yes No

 6 

- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:

Email Notifications 

None Failures Everything

Email Addresses

 john.doe@example.com 

IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

22. Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



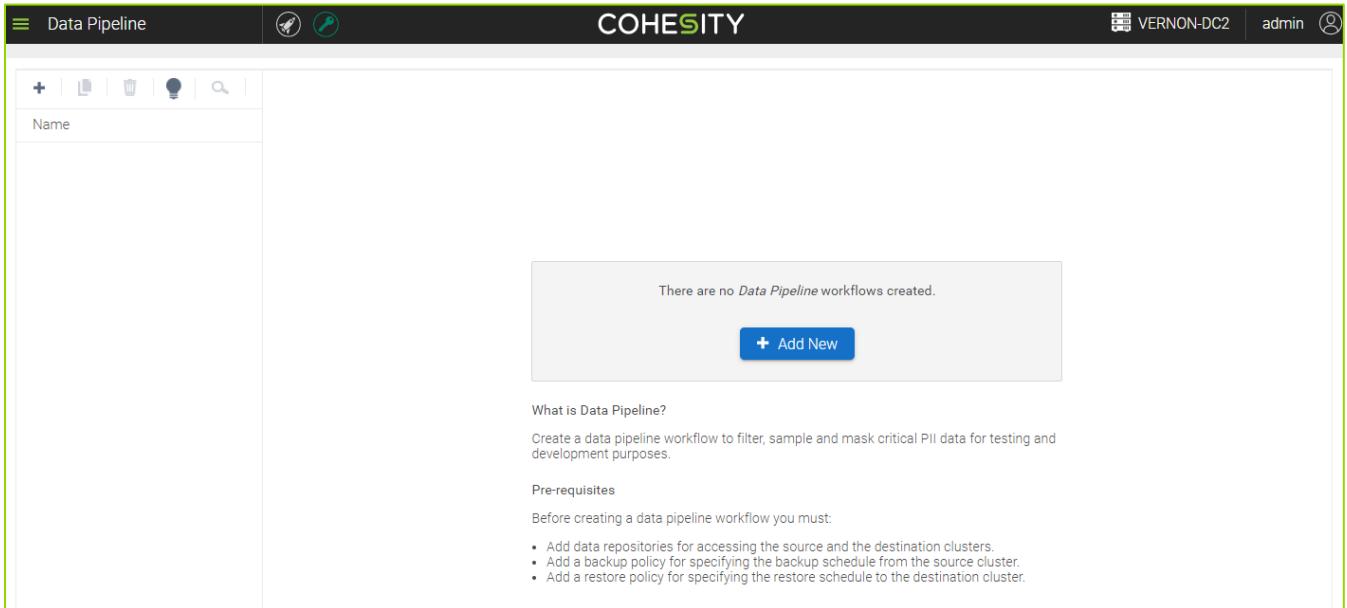
NOTE: Imanis Data software creates backup of your data objects in a non-incremental manner if masking and sampling are specified as part of a Data Pipeline or Data Mirroring workflow. For example, if masking and sampling is specified to a Hive table, the entire table is moved to Imanis Data each time the workflow runs.

11.2.2 Data Pipeline for HBase

You can create data pipeline workflows for HBase in Imanis Data software.

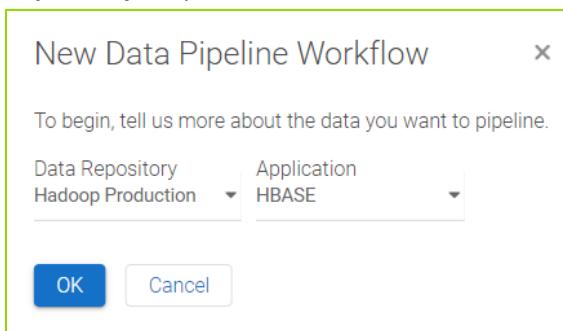
To start data pipeline workflow for HBase, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Pipeline**. The following page appears:



The screenshot shows the 'Data Pipeline' page in the Cohesity Imanis Data software. The main content area displays a message: 'There are no Data Pipeline workflows created.' with a blue '+ Add New' button. Below this, there is a 'What is Data Pipeline?' section with a brief description and a 'Pre-requisites' section with a bulleted list of requirements.

2. On the **Data Pipeline** page, click the  button or the  icon. The **New Data Pipeline Workflow** dialog box appears.
3. In the **New Data Pipeline Workflow** dialog box, select a **HBase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.



4. In the **Data Pipeline** screen, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Identify Data** area, do the following:
 - In the **HBase** tab, identify the namespaces and tables that you want to backup by selecting the corresponding check boxes. Use regular expressions (regex) for primary repository browsing.
 - In the **Selected Data** tab, verify your selection or click the **X** icons to remove unwanted items.
 - In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

The screenshot shows the 'Data Pipeline' screen. At the top, there are tabs for 'Data Pipeline' (selected), 'Workflow', and 'Policy'. Below the tabs, there's a form for defining a job tag:

Name	Job Tag: Dahanu Warehouse	Description (optional): Hbase Data Repo
------	---------------------------	---

Below the form is a 'Workflow Definition' section with a diagram showing data flow from 'Hadoop P...' to 'Imanis Data' and then to 'Send data to other repositories'.

The main area is divided into sections:

- Source:** Shows the connection between Hadoop and Imanis Data.
- Workflow Definition:** Shows the flow of data from Hadoop to Imanis Data and then to other repositories.
- Identify Data:** This section is currently active, showing the 'HBASE' tab. It lists objects: movieLens, system, and human resource, all of which are checked.
- Specify Policy:** This section is shown below the Identify Data section.

6. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

The screenshot shows the 'Specify Policy' section. It includes:

- A dropdown menu for 'Select a data backup policy' set to 'Backup to Cloud (s3)'.
- A 'Retention' section with the sub-instruction 'Allow retention on cloud' and two buttons: 'Yes' and 'No'.

7. In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a process. Under 'Cloud options', the 'Data Repository' dropdown is set to 'S3 Cloud Storage' and the 'Buckets' dropdown is set to 'hadoop_dr_archive'.

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).

The screenshot shows the 'Specify Options' step of a process. Under 'Cloud options', the 'Data Repository' dropdown is set to 'Azure Cloud Storage' and the 'Containers' dropdown is set to 'hadoop_dr_backup'.

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a process. Under 'Cloud options', the 'Data Repository' dropdown is set to 'S3 Cloud Storage' and the 'Buckets' dropdown is empty, indicating a manual entry field.

8. In the **Destination 1** section, from the Data Repository drop-down menu, select a data repository where the data will be moved to, that is, the target or destination.

The screenshot shows the 'Destination 1' step of a process. Under 'Destination 1', the 'Data Repository' dropdown is set to 'Hadoop DR'.

9. In the **Identify Data** area, do the following:

- In the **HBase** tab, clear the check boxes for the corresponding namespaces and tables that you DO not want to recover to Destination 1.
- In the **Selected Data** tab, verify your selection or click the icons to remove unwanted items.
- In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.
Refer to the section Appendix A: Rules for Data Inclusions and Exclusions.

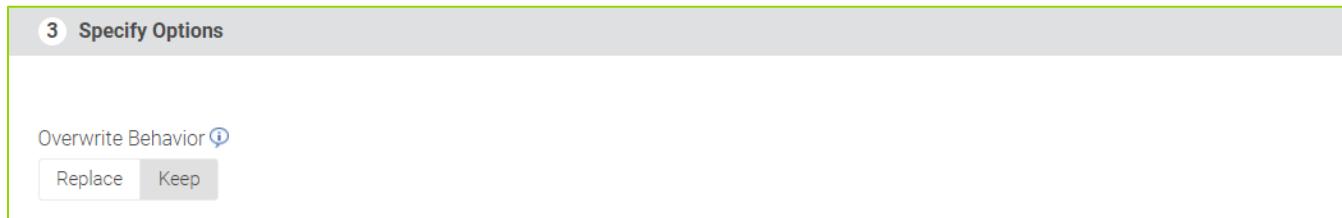
The screenshot shows the 'Identify Data' step of a process. At the top, there are three tabs: 'HBASE' (which is selected), 'Selected Data', and 'Rules'. The 'HBASE' tab displays a list of namespaces and tables. The 'Objects' checkbox is unchecked. The 'movielens', 'system', 'human resource', 'sales', and 'payroll' checkboxes are checked. There are also several empty rows below the checked items.

10. In the **Specify Policy** area, under **Select a recovery policy**, select a recovery policy from the drop-down menu to recover data in the selected location.

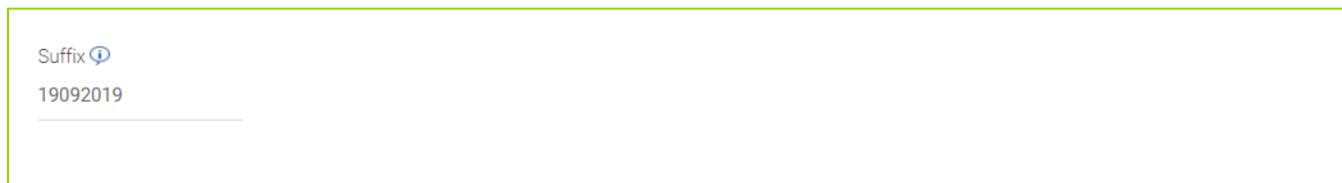
The screenshot shows the 'Specify Policy' step of a process. At the top, it says 'Select a recovery policy' and has a dropdown menu currently set to 'Restore On Demand'. Below that is a 'Recovery Schedule' section with two options: 'One time, immediately' (which is selected) and 'At specified times'. At the bottom is a 'Priority' section with three options: 'High' (which is selected), 'Medium', and 'Low'.

11. In the **Specify Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied

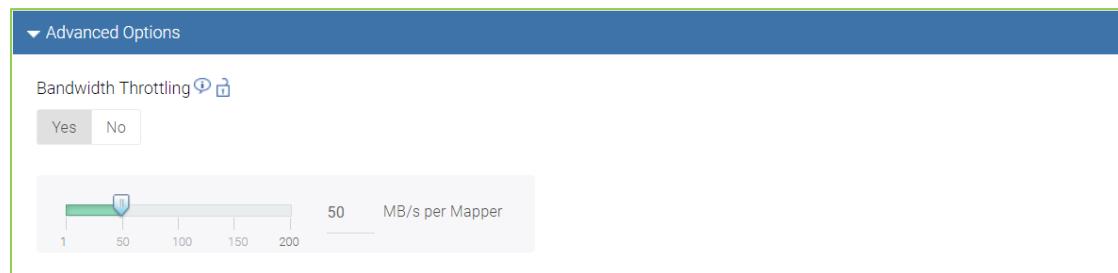


12. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.



13. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:

Concurrency Throttling   Yes No 6

- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:

Email Notifications  None Failures Everything

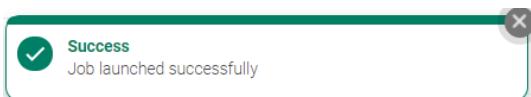
Email Addresses

john.doe@example.com 

IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



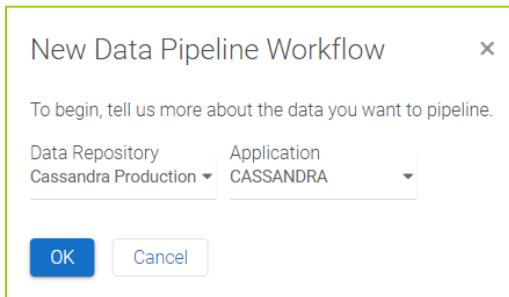
11.2.3 Data Pipeline for Cassandra

Imanis Data software enables you to create data pipeline workflows for Cassandra and Solr-enabled DataStax Enterprise (DSE) Cassandra. During the data pipeline process, Imanis Data software automatically identifies Solr-enabled Cassandra in both backup and recovery process.

To start data pipeline workflow for Cassandra, do the following:

- Click the **Main Menu**  > **Data Management** > **Data Pipeline**.

2. On the **Data Pipeline** page, click the **+ Add New** button or the **+**. The **New Data Pipeline Workflow** dialog box is displayed.
3. In the **New Data Pipeline Workflow** dialog box, select a Cassandra source data repository from the **Data Repository** drop-down menu, and then click **OK**.



4. In the **Data Pipeline** page, do the following:
 - Type a new job tag in the **Job Tag** field
 - Type a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters
5. In the **Identify Data** area, do the following:
 - In the **Cassandra** tab, identify the keyspaces and tables that you want to backup by selecting the corresponding check boxes. Use regular expressions (regex) for primary repository browsing.

The screenshot shows the Cohesity Data Pipeline interface. At the top, there's a navigation bar with 'Data Pipeline', a search icon, and user information ('admin'). The main area has tabs for 'Workflow Definition' and 'Identify Data'. Under 'Workflow Definition', there's a diagram showing a 'Source' (Cassandra) connected to a 'Target' (Imanis Data). A tooltip says 'Send data to other repositories'. Below this, a progress bar indicates steps 1 (Identify Data), 2 (Specify Policy), and 3 (Specify Options). The 'Identify Data' step is expanded, showing the 'Cassandra' tab with several keyspaces listed: 'Objects', 'movielens', 'system', and 'human resource'. Each item has a checkbox and a 'Size(approx)' column.

- In the **Selected Data** tab, verify your selection or click the **-** icons to remove unwanted items.

- In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.
6. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

The screenshot shows the 'Specify Policy' screen with the following details:

- Select a data backup policy:** Backup to Cloud (s3)
- Retention:**
 - Allow retention on cloud
 - Yes** (selected)
 - No
- On Imanis Data:** 1 day
- On Cloud:** 365 days
- Cloud Retention:** A trash can icon indicates data is deleted from the cloud after 365 days.

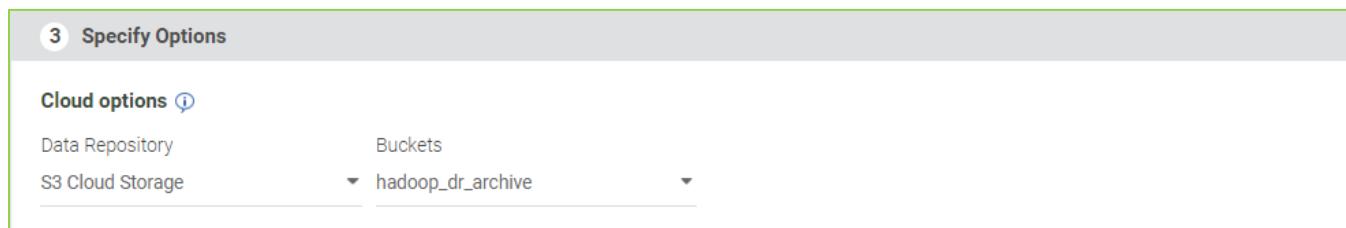
7. In the **Specify Options** section, under **Data Centers**, select a data center or data centers from where you want to backup the data.

The screenshot shows the 'Specify Options' screen with the following details:

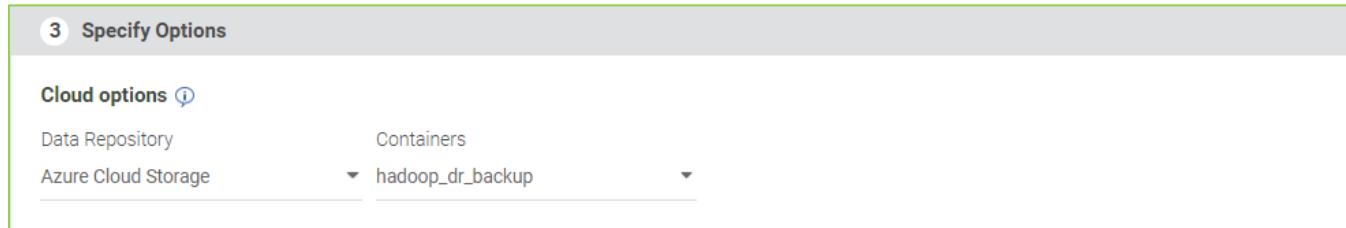
- Data Centers:**
 - Select All
 - DC1
 - DC2
- Cloud options:**
 - Data Repository:** S3 Cloud Storage
 - Buckets:** cassandra_dev_backup

In the preceding cases, the **Data Centers** and **Cloud options** are available ONLY if you have activated data centers in the Cassandra data repository and if you select a backup policy in which you have enabled the cloud retention feature respectively.

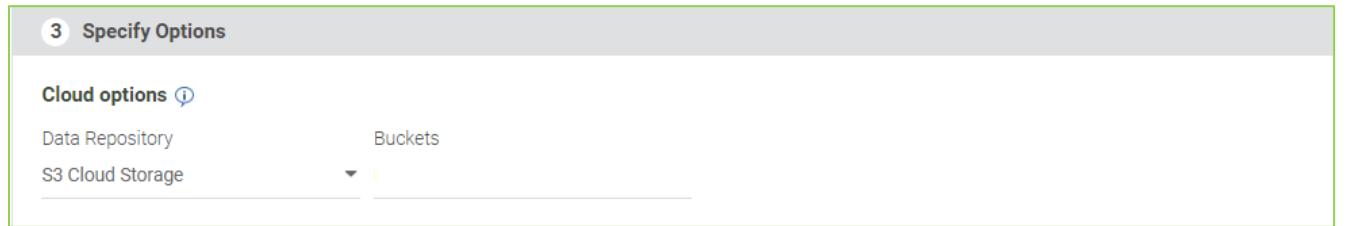
8. In the **Specify Options** section in the **Cloud Options** area, do one of the following:
- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:



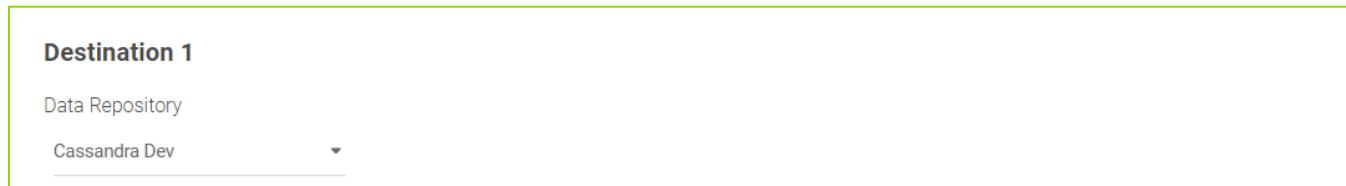
- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).



- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:



9. In the **Destination 1** section, from the **Data Repository** drop-down menu, select a data repository where the data will be moved to, that is, the destination cluster.



10. In the **Identify Data** area, do the following:

- a. In the **Cassandra** tab, clear the check boxes for the corresponding tables that you DO NOT want to recover to Destination 1.
- b. In the **Selected Data** tab, click the icons to remove unwanted items, if required.
- c. In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects.

Refer to the section Appendix A: Rules for Data Inclusions and Exclusions.

The screenshot shows the 'Identify Data' step of a process. At the top, there are three tabs: 'Cassandra' (which is selected and highlighted in blue), 'Selected Data', and 'Rules'. The 'Cassandra' tab displays a list of database objects. There is a checkbox next to each object name. The objects listed are: 'Objects' (unchecked), 'movielens' (checked), 'system' (checked), 'human resource' (checked), 'sales' (checked), and 'payroll' (checked). The background of the interface is light gray, and the overall layout is clean and modern.

11. In the **Specify Policy** area, under **Select a recovery policy**, select a recovery policy from the drop-down menu to recover data in the selected location.

The screenshot shows the 'Specify Policy' step of a process. At the top, it says 'Select a recovery policy' and has a dropdown menu set to 'Restore On Demand'. Below that is a 'Recovery Schedule' section with two options: 'One time, immediately' (which is selected) and 'At specified times'. Further down is a 'Priority' section with three radio buttons: 'High' (selected), 'Medium', and 'Low'. The interface uses a light gray background and standard form elements.

12. In the **Specify Options** area, under **Data Centers**, do the following:

- Select one or multiple data centers, in the **Data Centers** option, where you want to recover the data.

The screenshot shows a step titled "3 Specify Options". Under the "Data Centers" section, there is a checkbox labeled "Select All" which is checked. Below it are two other checkboxes: "DC1" and "DC2", both of which are also checked.

- In the **Additional Options** area, under **Overwrite Behavior**, do one of the following:

The screenshot shows a step titled "Overwrite Behavior". It contains two buttons: "Replace" and "Keep".

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied

13. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _10102017.

The screenshot shows a step titled "Suffix". A text input field contains the value "19092019".

14. For DSE 6.X, for **Recovery Staging Directory**, you can mention a temporary directory if you do not wish Imanis to use the Cassandra storage for staging. The Recovery Staging Directory field accepts a single directory or a comma separated list of directories. Ensure that the directories are present on all the nodes before executing the restore job.

The screenshot shows a step titled "Recovery Staging Directory". A text input field contains the value "/tmp/stage1,/tmp/stage2".

15. In the **More Options for Selected Data** section, do the following:

4 More Options for Selected Data		
Objects	Recover As	With Properties
movielens	movielens19092019	+
system	system19092019	+
human resource	human resource19092019	+
sales	sales19092019	+
payroll	payroll19092019	+

- a. To rename restored objects, type the new name in the **Recover As** column.
- b. To change property of the restored object, click the  icon, and type the values in the Key and Value field. Usually, the key is auto-completed by the UI. The Value field must contain the complete value of the property. Only following property changes are allowed: keyspace (replication) and table (compression and compaction).

For example, let's assume that a user wants to change replication for a source keyspace wherein the create query for the keyspace is as follows:

```
CREATE KEYSPACE tutorialspoint WITH replication = {'class':'SimpleStrategy', 'replication_factor' : 3};
```

To change the replication factor to 1, the key value pair would be as follows:

KEY: replication

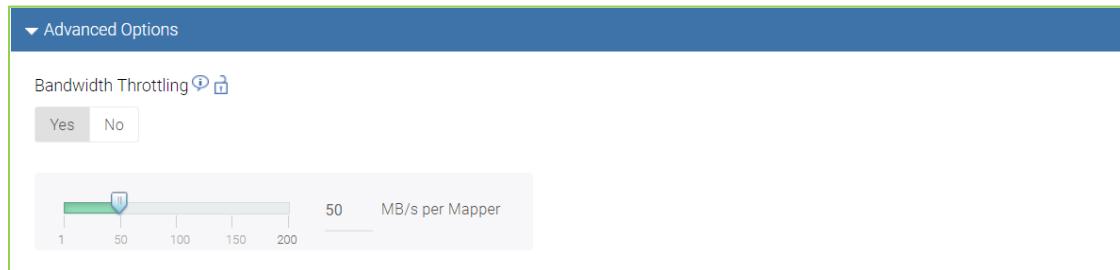
VALUE: {'class':'SimpleStrategy', 'replication_factor' : 1}

IMPORTANT: If the recovery process is executed to an '**Alternate Location**', Imanis Data software always uses 'SimpleStrategy' as the replication strategy unless it is overridden through the "**More Options for Selected Data**" section in the UI.

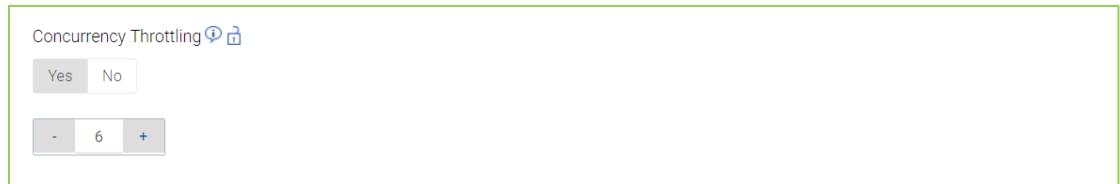
NOTE: In the current release, Imanis Data software does not support changing of compaction strategy to CFSCompactionStrategy from any other strategy. For example, Imanis Data software does not support changing of compaction strategy from Leveled Compaction Strategy, Size Tiered Compaction Strategy or any other strategy to CFSCompactionStrategy.

16. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

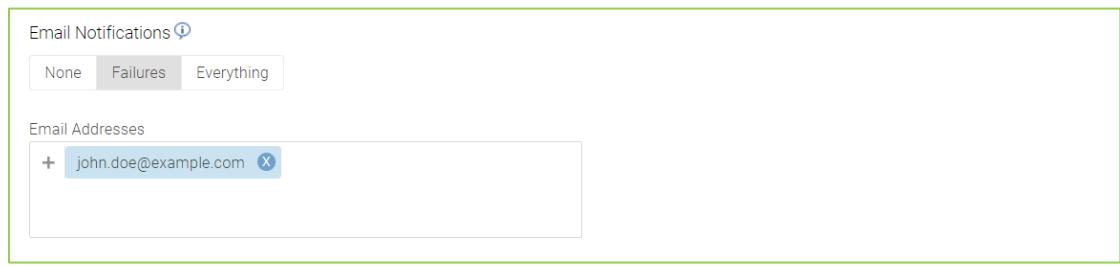
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

17. Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



IMPORTANT: Ensure that the system partitioner – a configurable parameter in the cassandra.yaml file that determines how data is distributed across the nodes in the cluster – at the source cluster matches with the system partitioner that is configured at the destination cluster. If you do not configure the system partitioner at the destination cluster, the data restore process will fail.

IMPORTANT: As part of the data recovery process, the keyspace/table schemas are created (if required) on the destination cluster. However, this schema creation may fail if a data center(s) is down in a multi-dc environment. In such cases, the Cassandra data recovery process will fail.

NOTE: In the current release, multiple schema alteration queries executed on the single column on the primary Cassandra cluster will not be propagated to the destination Cassandra cluster during a single invocation of data pipeline workflow. For example, if the following schema alteration commands are used on a table:

1. ALTER TABLE newyork.transit_map DROP events [where events column is dropped]
2. ALTER TABLE newyork.transit_map ADD events text [where events column of type text is added again to the schema]

Both these commands are executed on the primary Cassandra cluster and then the incremental data pipeline workflow is started to move the data to the destination Cassandra cluster. This process will result in an error with the incremental data pipeline workflow.

IMPORTANT: For DSE 5.0 and later releases, Imanis Data software does not restore permissions of tables or databases if the required set of roles are not present on destination restore cluster. Thus, the user must create the required set of roles on destination cluster before initiating the recovery workflow. The user must also ensure to create transitive role dependencies if any exists. For example, on the Source cluster, you have roles 'supervisor', 'admin', 'staff'.

The role 'admin' has "Select" privileges on database 'company'.

The role 'staff' has "Insert" and "Select" privileges on database 'company'.

While the role 'supervisor' has transitive dependency wherein it inherits all the privileges of both the roles 'admin' and 'staff' with the following command:

```
Grant admin to supervisor;
Grant staff to supervisor;
```

Thus, the user must ensure that such transitive dependent roles (like the role 'supervisor') are created on destination cluster before initiating recovery workflow.

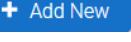
IMPORTANT: If Authorization is enabled on destination cluster, ensure that users present in Cassandra primary cluster are also present on destination cluster.

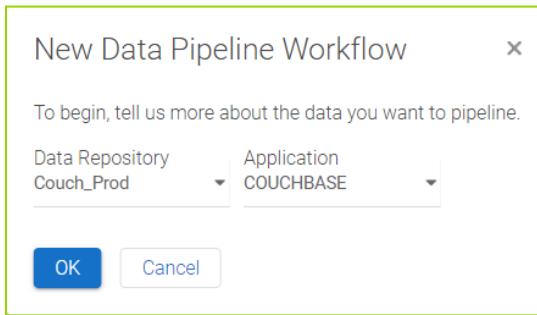
11.2.4 Data Pipeline for Couchbase

Imanis Data software supports data pipelining for Couchbase data sets at the jobtag and bucket level. Prior to starting the recovery process, you must first identify the bucket(s) that you want to recover at the source cluster and then manually create the bucket(s) with the same name(s) at the destination cluster.

For example, at the source cluster you identify test_bucket1, test_bucket2, and test_bucket3 that you want to recover. At the destination cluster, you must manually create test_bucket1, test_bucket2, and test_bucket3 before starting the recovery process.

To start a data pipeline workflow for Couchbase, do the following:

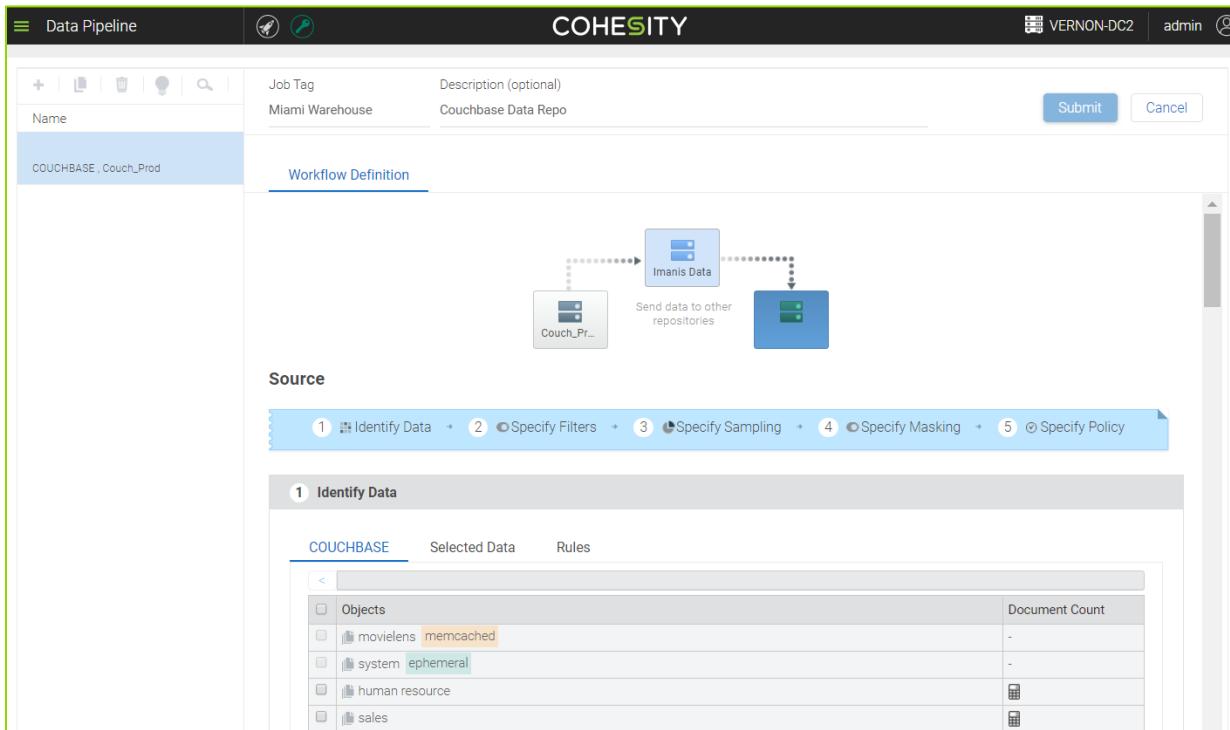
1. Click the **Main Menu**  > **Data Management** > **Data Pipeline**.
2. On the **Data Pipeline** page, click the  **+ Add New** button or the  icon. The **New Data Pipeline Workflow** dialog box appears.



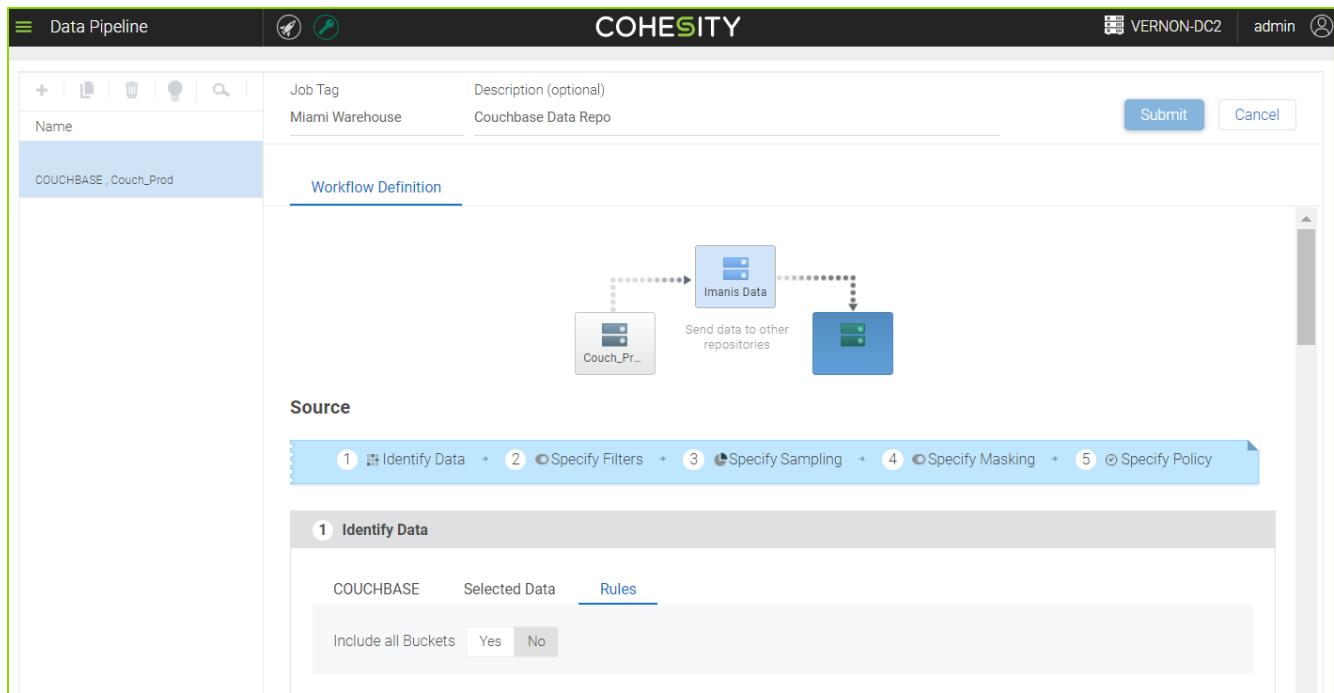
3. In the **New Data Pipeline Workflow** dialog, select a **Couchbase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.
4. In the **Data Pipeline Workflow**, type a new job tag in the **Job Tag field** and a job tag description in the **Description field**. The job tag name can include alphanumeric characters, numbers and/or special characters.

5. In the **Source** section, in the **Identify Data** area, do the following:

- In the **Couchbase** tab, identify the jobtag or bucket level data that you want to backup by selecting the corresponding check boxes.



- In the **Selected Data** tab, verify your selection or click the **X** icon to remove unwanted items.
- In the **Rules** tab, click Yes to include all buckets in the backup job. This will also include any buckets that get added in the future.

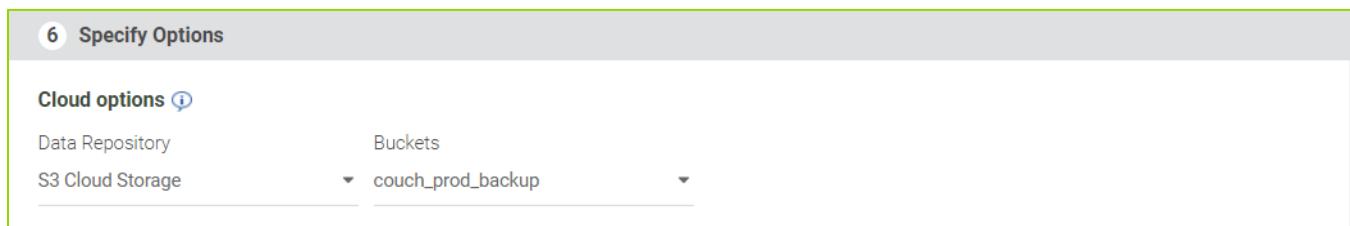


- In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

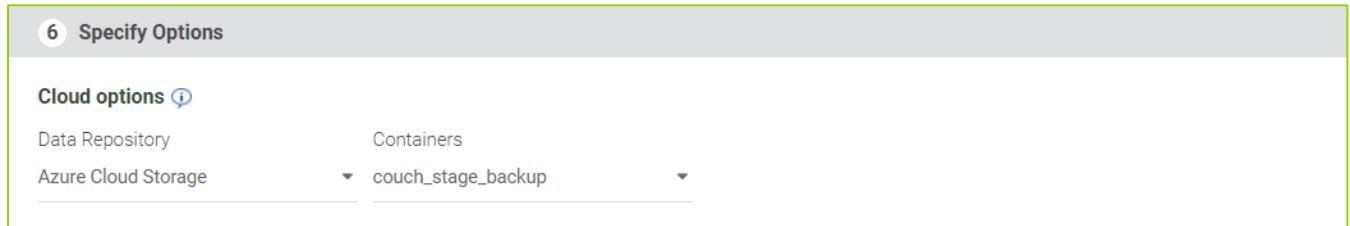
This screenshot shows the 'Specify Policy' section. It starts with a header '5 Specify Policy' and a sub-instruction 'Select a data backup policy'. A dropdown menu is open, showing 'Backup_Policy'. Below this is a 'Retention' section with the sub-instruction 'Allow retention on cloud'. A 'Yes' button is highlighted, while 'No' is unselected. Further down is a section 'On Imanis Data' which shows a circular icon with a flag and a circular icon with a trash can, connected by a line, with the text '60 days' indicating the retention period.

- In the **Specify Options** section, under **Cloud Options**, do one of the following:

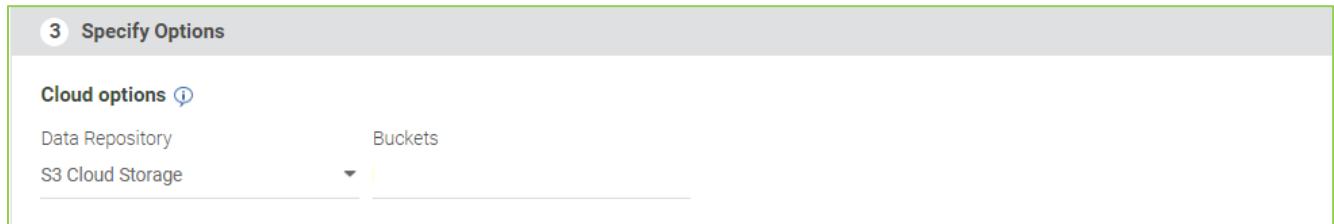
- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:



- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud).



- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:



NOTE: In Couchbase, Tombstones are records of expired or deleted items that include item keys and metadata. Couchbase deletes tombstones permanently after metadata purge interval has elapsed. Due to this process, Couchbase Database Change Protocol (DCP) issues rollbacks when a client requests for mutations (incremental) after tombstones are removed, which may lead to a full backup. If the metadata purge interval and job interval is configured such that incremental backup always runs after purge, every backup may end up getting a rollback. Refer to Couchbase documentation here for information:

<https://developer.couchbase.com/documentation/server/3.x/admin/Concepts/concept-tombstone.html>

NOTE: Customers should set the backup frequency less than their metadata purge interval. Typically, it is recommended to keep the metadata purge intervals to 7 days just to be on safer side.

8. In the **Destination 1** section, from the **Data Repository** drop-down menu, select a data repository where the data will be moved to, that is, the destination cluster.

Destination 1

Data Repository
Couch_Prod

9. In the **Identify Data** area, do the following:

- Clear the check boxes for the corresponding jobtag or bucket level data that you DO not want to recover to Destination 1.
- In the **Selected Data** tab, click the **X** icons to remove unwanted items, if required.
- In the **Rules** tab, edit the regex if needed.

1 Identify Data

COUCHBASE Selected Data Rules

<input type="checkbox"/> Objects
<input checked="" type="checkbox"/> human resource
<input checked="" type="checkbox"/> sales
<input checked="" type="checkbox"/> payroll

10. In the **Specify Policy** area, under **Select a recovery policy**, select a recovery policy from the drop-down menu to recover data in the selected location. You will be able to see recovery policies that were created earlier.

2 Specify Policy

Select a recovery policy
Restore On Demand

Recovery Schedule
 One time, immediately At specified times

Priority
 High Medium Low

11. In the **Specify Options**, in the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data

- Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied
- Click **Append** to add new data to an existing bucket

3 Specify Options

Overwrite Behavior ⓘ

Replace Keep Append

12. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _13092019.

Suffix ⓘ

13092019

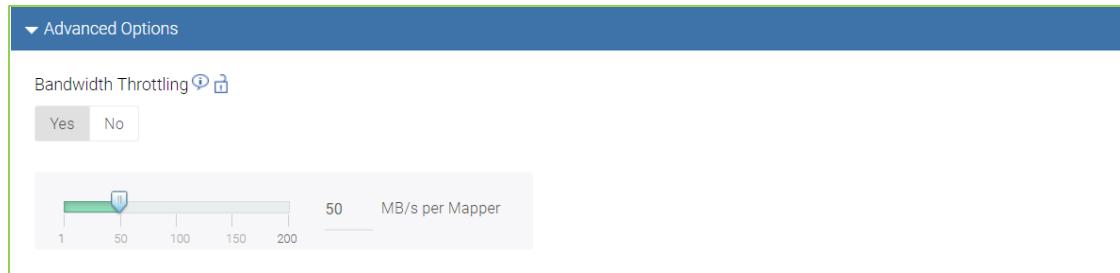
13. In the **More Options for Selected Data**, edit the object name and rename it, if needed.

4 More Options for Selected Data

Objects	Recover As
human resource	human resource13092019
sales	sales13092019
payroll	payroll13092019

14. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

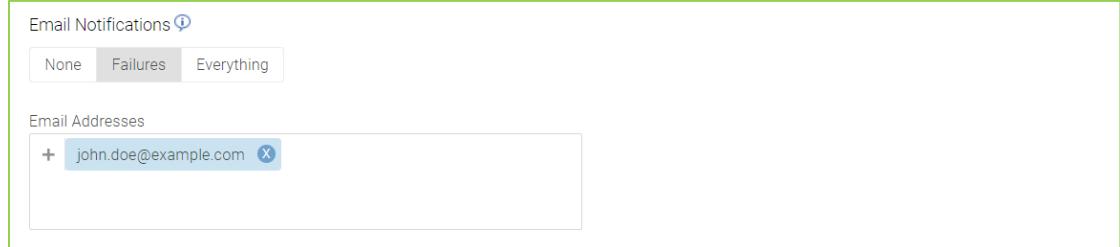
- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:



- In the **Concurrency Throttling** option, click the lock icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



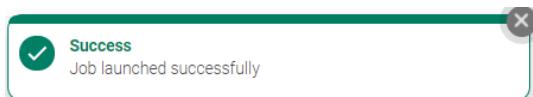
- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



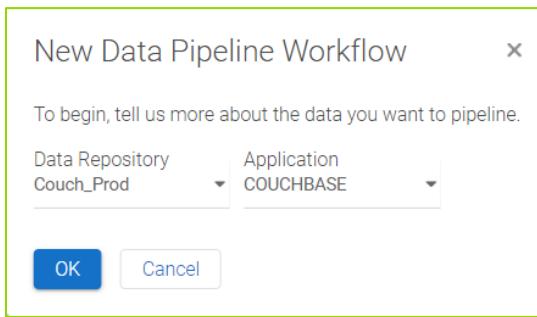
NOTE: Recovery of Ephemeral and Memcached buckets is not supported.

11.2.4.1 Data Masking & Sampling for Couchbase

Imanis Data software supports data masking and sampling for Couchbase data sets at the jobtag and bucket level.

To start a data pipeline workflow for Couchbase, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Pipeline**.
2. On the **Data Pipeline** page, click the  **+ Add New** button or the  icon. The **New Data Pipeline Workflow** dialog box appears.



3. In the **New Data Pipeline Workflow** dialog, select a **Couchbase** source data repository from the **Data Repository** drop-down menu, and then click **OK**.
4. In the **Data Pipeline Workflow**, type a new job tag in the **Job Tag field** and a job tag description in the **Description field**. The job tag name can include alphanumeric characters, numbers and/or special characters.
5. In the **Source** section, in the **Identify Data** area, do the following:
 - a. In the **Couchbase** tab, identify the jobtag or bucket level data that you want to backup by selecting the corresponding check boxes.

The screenshot shows the Cohesity Data Pipeline interface. At the top, there's a navigation bar with 'Data Pipeline', a search icon, and user information ('VERNON-DC2 admin'). Below the navigation is a job card with 'Name: COUCHBASE_Couch_Prod', 'Job Tag: Miami Warehouse', and 'Description (optional): Couchbase Data Repo'. There are 'Submit' and 'Cancel' buttons. The main area is titled 'Workflow Definition' and shows a flowchart: 'Couch_Pr...' (Source) → 'Imanis Data' (Target) → 'Send data to other repositories'. Below the flowchart is a 'Source' section with a numbered list: 1. Identify Data, 2. Specify Filters, 3. Specify Sampling, 4. Specify Masking, 5. Specify Policy. The 'Identify Data' step is expanded, showing the 'COUCHBASE' tab selected. Under 'Selected Data', there's a table with columns 'Document Count' and rows for 'Objects', 'movielens [memcached]' (selected), 'system [ephemeral]', 'human resource', and 'sales'. The 'Rules' tab is also visible.

- b. In the **Selected Data** tab, verify your selection or click the **X** icon to remove unwanted items.
- c. In the **Rules** tab, click **Yes** to include all buckets in the backup job. This will also include any buckets that get added in the future.

This screenshot is identical to the one above, but the 'Rules' tab is selected in the 'Identify Data' step. It shows the same interface elements: 'Data Pipeline' navigation, job card, workflow definition, and the expanded 'Identify Data' step. The 'Rules' tab is highlighted, and below it, there's a 'Include all Buckets' checkbox with 'Yes' selected.

6. In the **Specify Filter** section, click the **Apply** link.



Buckets	Filters
human resource	Apply
sales	Apply
payroll	Apply

NOTE: Imanis Data software supports filtering only on document keys and document content.

In full data recovery, only document key filtering is supported:

This option is visible only if "bucket" is selected on full data recovery.

Under Documents, under **Select Filtering mechanism**, click **Document ID** and type a regular express to match document IDs.

In incremental data recovery, only document content filtering is supported:

All terms must be separated by one space

Comparisons with "<=", "<", ">=", ">", "==", "!=" are supported. &&, || for and, or respectively.

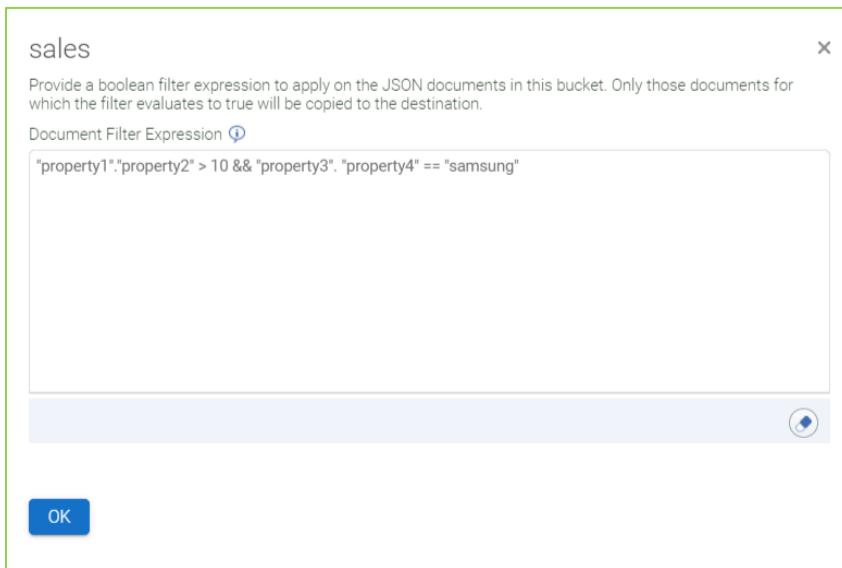
Filtering on nested items like {"i1" : {"i2" : 43}} can be specified as "i1"."i2" > 21

Any condition where variables are not found in json is treated as false. Json docs are restored only if filter expression evaluates to true. Binary docs are always restored.

Parenthesis are also supported, however, whitespace is not permitted after opening brace and before closing brace. For example,

"ibu" > 0 && ("type" == "beer" || "abv" > 5) .

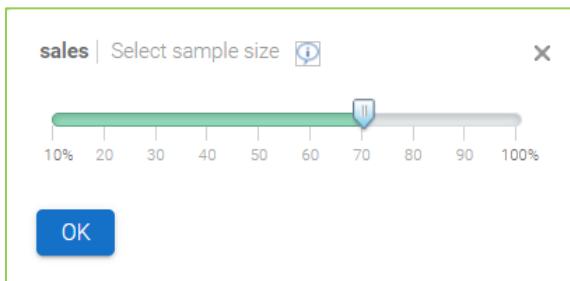
7. In the **sales** dialog box, type the boolean expression. See the example in the screenshot:



8. In the **Specify Sampling** section, click the **100%** link.

Objects	Sample Size
human resource	100%
sales	100%
payroll	100%

9. In the **Warehouse1** dialog box, move the slider to set sample size:



10. In the **Specify Masking** section, click the **Apply** link.

4 Specify Masking		
Objects	Masks	Exclude binary Documents <small>(i)</small>
human resource	Apply	ON
sales	Apply	ON
payroll	Apply	ON

11. In the **Warehouse1** dialog box, enter **attribute**, set **data type**, and then set **type of mask** in the respective fields:

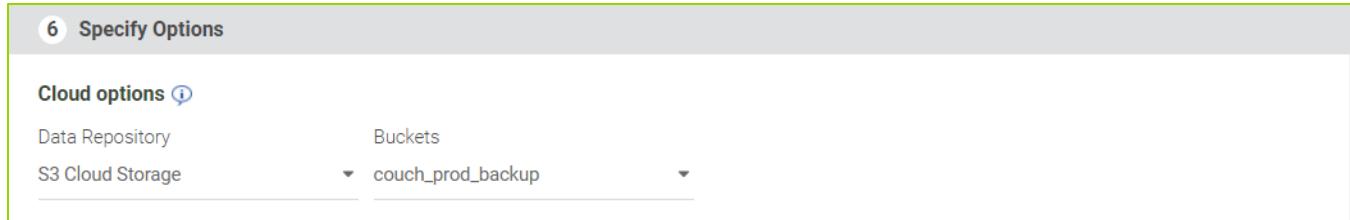
The screenshot shows a dialog box titled "sales". It contains a table with three columns: "Attribute", "Data Type", and "Mask". The "Attribute" column has a single entry: "[property1].[property2]". The "Data Type" column shows "String". The "Mask" column displays "Employee Identification Number: 12-3456789" with a red "X" icon to its right, indicating it can be deleted. Below the table is a "+" button and an "OK" button.

12. In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

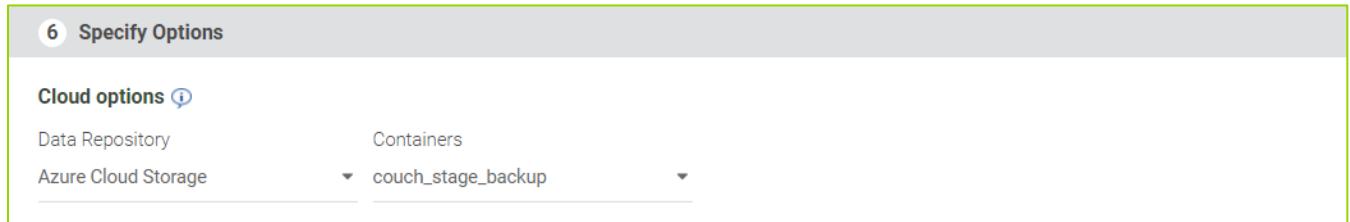
The screenshot shows the "Specify Policy" dialog box. Under "Select a data backup policy", a dropdown menu is open with "Backup_Policy" selected. In the "Retention" section, the "Allow retention on cloud" checkbox is checked, with "Yes" selected. Below this, the "On Imanis Data" section shows a timeline starting with a trash bin icon and ending with a "60 days" label.

13. In the **Specify Options** section, under **Cloud Options**, do one of the following:

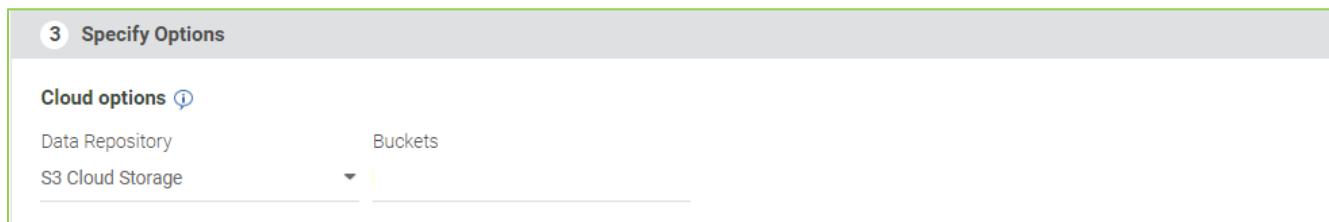
- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:



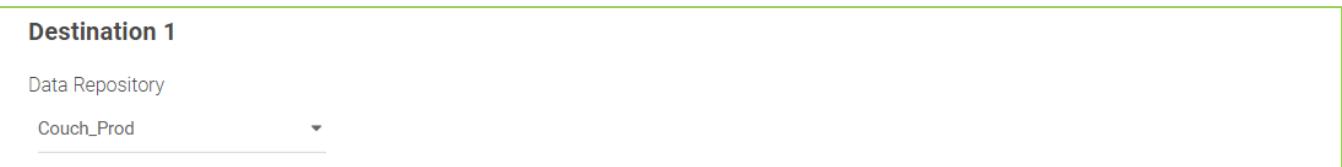
- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud.



- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:



14. In the **Destination 1** section, from the **Data Repository** drop-down menu, select a data repository where the data will be moved to, that is, the destination cluster.



15. In the **Identify Data** area, do the following:

- Clear the check boxes for the corresponding jobtag or bucket level data that you DO not want to recover to Destination 1.
- In the **Selected Data** tab, click the **X** icons to remove unwanted items, if required.

c. In the **Rules** tab, edit the regex if needed.

The screenshot shows the 'Identify Data' step of a process. At the top, there are tabs for 'COUCHBASE', 'Selected Data', and 'Rules'. Under 'COUCHBASE', there is a list of objects: 'Objects' (unchecked), 'human resource' (checked), 'sales' (checked), and 'payroll' (checked). The 'Selected Data' and 'Rules' tabs are also visible.

16. In the **Specify Policy** area, under **Select a recovery policy**, select a recovery policy from the drop-down menu to recover data in the selected location. You will be able to see recovery policies that were created earlier.

The screenshot shows the 'Specify Policy' step. It includes a 'Select a recovery policy' dropdown set to 'Restore On Demand', a 'Recovery Schedule' section with 'One time, immediately' selected, and a 'Priority' section with 'High' selected.

17. In the **Specify Options**, in the **Additional Options** area, under **Overwrite Behavior** do one of the following:

- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied
- Click **Append** to add new data to an existing bucket

The screenshot shows the 'Specify Options' step. It includes an 'Overwrite Behavior' section with three buttons: 'Replace' (selected), 'Keep', and 'Append'.

18. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _13092019.

Suffix 
13092019

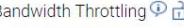
19. In the **More Options for Selected Data**, edit the object name and rename it, if needed.

4 More Options for Selected Data

Objects	Recover As
human resource	human resource13092019
sales	sales13092019
payroll	payroll13092019

20. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click Yes to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the in the MB/s per Mapper field:

Bandwidth Throttling 

Yes No

 50 MB/s per Mapper

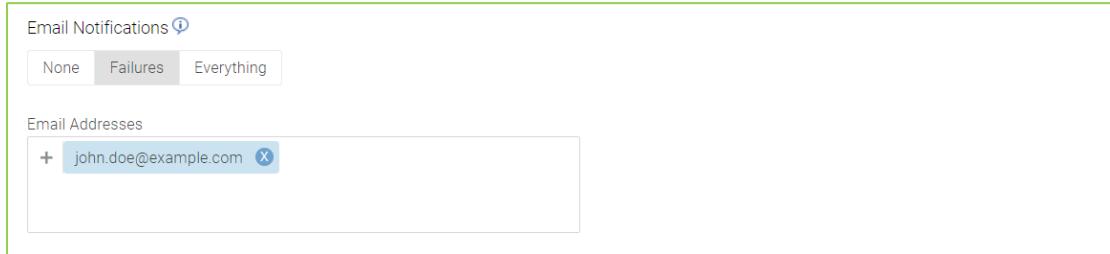
- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click Yes to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:

Concurrency Throttling 

Yes No

 - 6 +

- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the Email Addresses field that will receive the job completion and/or failure notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

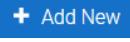
```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

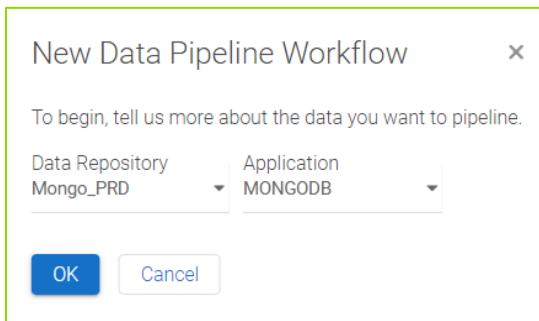
- Click **Submit**.

11.2.5 Data Pipeline for MongoDB

Imanis Data software supports data pipelining for MongoDB data sets at the database and collection level.

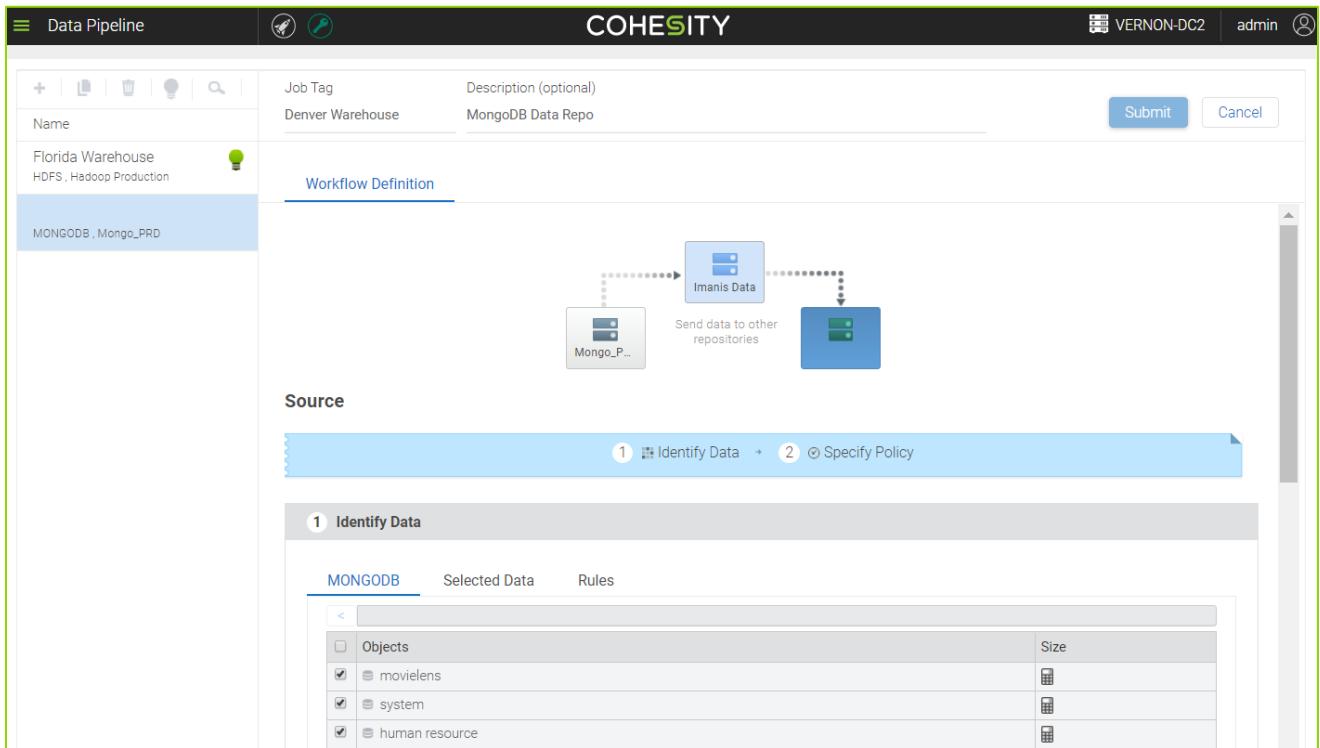
To start a data pipeline workflow for Couchbase, do the following:

- Click the **Main Menu**  > **Data Management** > **Data Pipeline**.
- On the Data Pipeline page, click the  button or the  icon. The **New Data Pipeline Workflow** dialog box appears.
- In the **New Data Pipeline Workflow** dialog box, select a **MongoDB** source data repository from the **Data Repository** drop-down menu, and then click **OK**.



- In the **Data Pipeline** page, type a new job tag in the **Job Tag** field and a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters.
- In the **Source** section, in the **Identify Data** area, do the following:

- In the **MongoDB** tab, identify the database and/or collection level data that you want to back up by selecting the corresponding check boxes.
- In the **Selected Data** tab, verify your selection or click the X icon to remove unwanted items.
- In the **Rules** tab, type regular expressions (regex) to exclude or include specific data objects. Refer to the section Appendix A: Rules for Data Inclusions and Exclusions.



- In the **Specify Policy** section, under **Select a data backup policy**, select a backup policy with or without cloud retention.

The screenshot shows the "Specify Policy" section of the Cohesity interface. It includes a dropdown menu for "Backup_Policy" and a "Retention" section. The "Retention" section contains a toggle for "Allow retention on cloud" (set to "No") and a timeline for "On Imanis Data" ranging from "60 days" to "1 year".

- In the **Specify Options** section, under **Cloud Options**, do one of the following:

- Select an S3 from the Data Repository drop-down menu and select a bucket (S3) from the Buckets drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a backup job setup. Under 'Cloud options', the 'Data Repository' dropdown is set to 'S3 Cloud Storage' and the 'Buckets' dropdown is set to 'mongo_prd_backup'.

- Select an Azure cloud repository from the **Data Repository** drop-down menu and select a container (Azure from the Containers drop-down menu to retain data in the cloud.

The screenshot shows the 'Specify Options' step of a backup job setup. Under 'Cloud options', the 'Data Repository' dropdown is set to 'Azure Cloud Storage' and the 'Containers' dropdown is set to 'mongo_qa_backup'.

- As a user, if you do not have list permissions on S3 buckets or Azure containers, then you must type the name of the bucket (S3) or the container (Azure) for which you have been granted access by your organization in the **Buckets / Containers** field. The following screenshot displays an example of S3 data repository:

The screenshot shows the 'Specify Options' step of a backup job setup. Under 'Cloud options', the 'Data Repository' dropdown is set to 'S3 Cloud Storage' and the 'Buckets' dropdown is empty, indicated by a small vertical line.

8. In the **Destination 1** section, from the Data Repository drop-down menu, select a data repository where the data will be moved to, that is, the destination cluster.

The screenshot shows the 'Destination 1' section of a backup job setup. The 'Data Repository' dropdown is set to 'Mongo_PRD'.

9. In the **Identify Data** area, do the following:

- a. Clear the check boxes for the corresponding jobtag or bucket level data that you DO not want to recover to Destination 1.
- b. In the **Selected Data** tab, click the **X** icons to remove unwanted items, if required.

c. In the **Rules** tab, edit the regex if needed.

1 Identify Data

MONGODB Selected Data Rules

	Objects
<input type="checkbox"/>	Objects
<input checked="" type="checkbox"/>	movielens
<input checked="" type="checkbox"/>	system
<input checked="" type="checkbox"/>	human resource
<input checked="" type="checkbox"/>	sales

10. In the **Specify Policy** area, under **Select a recovery policy**, select a recovery policy from the drop-down menu to recover data in the selected location. You will be able to see recovery policies that were created earlier.

2 Specify Policy

Select a recovery policy

Restore On Demand

Recovery Schedule

One time, immediately At specified times

Priority

High Medium Low

11. In the **Specify Options**, under **Overwrite Behavior**, do one of the following:

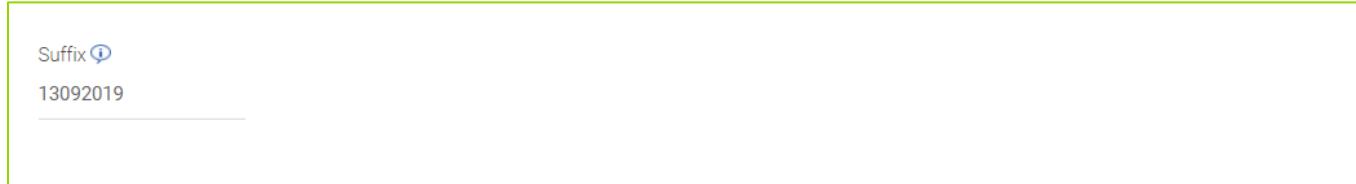
- Click **Replace** to replace existing data with new data thus erasing any previously existing data
- Click **Keep** to retain existing data (if any). However, if there is no existing data then the new data will be copied

3 Specify Options

Overwrite Behavior ⓘ

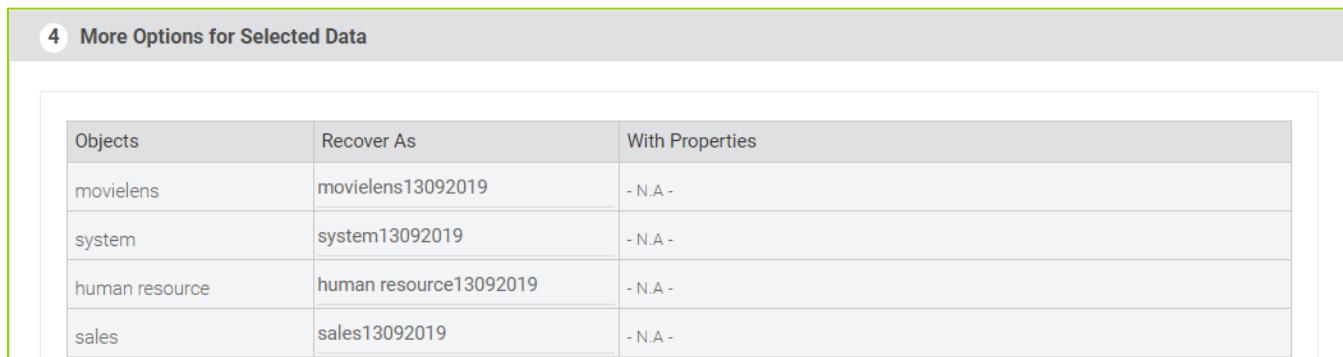
Replace Keep

12. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _recovered. In the current release, . (dot) is not supported in Suffix field.



The screenshot shows a single-line input field labeled "Suffix" with a blue info icon. The value "13092019" is typed into the field. The entire input area is enclosed in a light gray border.

13. In the **More Options for Selected Data**, you can rename MongoDB data objects (database and collection) with properties. A few of the properties like shardCollectionOptions, createCollectionOptions have values which are JSONs. These JSON values are validated if they are well formed and correct through the UI.

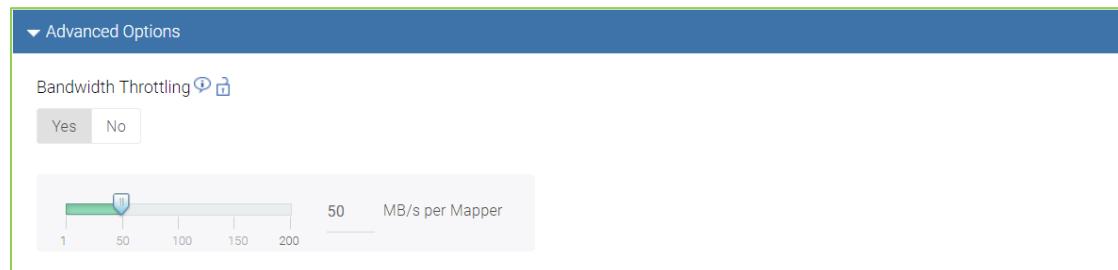


The screenshot shows a table titled "More Options for Selected Data". The table has three columns: "Objects", "Recover As", and "With Properties". The rows show the following data:

Objects	Recover As	With Properties
movielens	movielens13092019	- N.A -
system	system13092019	- N.A -
human resource	human resource13092019	- N.A -
sales	sales13092019	- N.A -

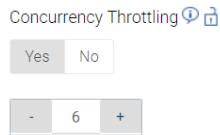
14. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications** for both **Data Capture** and **Data Restore**. You can decrease congestion over the network and primary cluster and reduce the number of concurrent jobs through the Bandwidth and Concurrency throttling feature. If you do not specify throttling parameters, default values are applied from mapred-site.xml:

- In the **Bandwidth Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then pick a value on the slider at which the bandwidth consumed by each individual mapper will be capped. The desired bandwidth cap may also be directly entered in the **MB/s per Mapper** field:

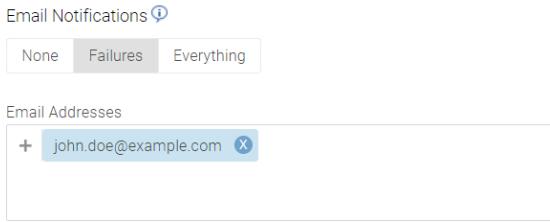


The screenshot shows the "Advanced Options" pane expanded. It contains a section for "Bandwidth Throttling" with a "Yes" button (which is selected) and a "No" button. Below this is a slider with a scale from 1 to 200, currently set to 50, with the label "MB/s per Mapper".

- In the **Concurrency Throttling** option, click the lock  icon to use the feature, click **Yes** to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the concurrency for the job:



- In the **Email Notifications** option, click **Yes** to confirm, click the plus sign (+), and then type one or more the email addresses in the **Email Addresses** field that will receive the job status notifications:



IMPORTANT: Make sure the following command works from all the Imanis Data nodes:

```
#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>
```

- Click **Submit**. A confirmation message will be displayed on the page indicating the job is successfully launched.



12 Managing Jobs

This section describes how you can manage jobs in Imanis Data software. For example, you can manage jobs in Imanis Data by rerunning the ‘run now’ jobs, stopping the jobs, editing a job, deactivating a job, and so on.

12.1 Rerunning ‘Run Now’ jobs

Once a Run Now — a one-time job — has completed running, administrators can re-run that job. A one-time job is a job that has a backup schedule of **One time, immediately**.

While creating a workflow, if you select a policy that has the **One time, immediately** option enabled (see screenshot below), you can re-run the job once it completes its data backup or data recovery cycle.

A **Run Now** job is easy to identify: the job WITHOUT the bulb is a Run Now job. You can execute one-time job runs for Data Backup, Data Pipeline, and Data Mirroring as long as the policy that you select has the ‘**One time, immediately**’ option enabled, and the job has completed running.

The following example illustrates re-running a one-time data backup job.

To run a one-time job, do the following:

1. On the **Data Backup** page, identify the workflow that has the **Run Now** button.

The screenshot shows the 'Data Backup' page. At the top, there is a form for a job named 'NY Warehouse' with a description 'Cassandra DataRepo of NyWarehouse'. To the right is a 'Run Now' button. Below this, there are tabs for 'Workflow Definition', 'Stats', and 'Advanced Stats', with 'Workflow Definition' selected. A 'ThreatSense' toggle switch is also present. A blue progress bar at the bottom indicates steps 1 through 3: 'Identify Data', 'Specify Policy', and 'Specify Options'. The 'Identify Data' step is expanded, showing a table of selected data objects: 'movielens', 'system', 'human resource', and 'sales'. The 'Size(approx)' column shows values of 1.00 TB, 1.00 TB, 1.00 TB, and 1.00 TB respectively.

2. Do one of the following:

- Click the **Run Now** button to re-run the data backup job without making any changes
- Click the  icon and identify new files and directories from which you want to backup data and select the corresponding check boxes and then click the **Run Now** button. Remember that you are not permitted to edit the data repositories or policy of the workflow

NOTE: As Imanis Data software recovers all the data in one single, unified flow the **Run Now** button and the **edit** icon is not available in **Data Recovery** workflows. Due to the same reason the **Number of Job Runs** tile and the **Job Runs** expandable pane is also not available.

IMPORTANT: For **Data Pipeline** workflows, both the **Data Backup** and **Data Recovery** policies must be scheduled to run **One time, immediately**.

12.2 Cloning a job

Imanis Data software enables administrators to clone any job available in the system.

All you have to do is identify the job that you want to clone and then select the clone button. Imanis Data software clones the name with –CLONE suffix to the original name. The job cloning feature is available in backup, recovery, data lifecycle management, data mirroring, and data pipelining workflows.

The following example illustrates how the cloning feature works in a typical backup workflow.

To clone a job, do the following:

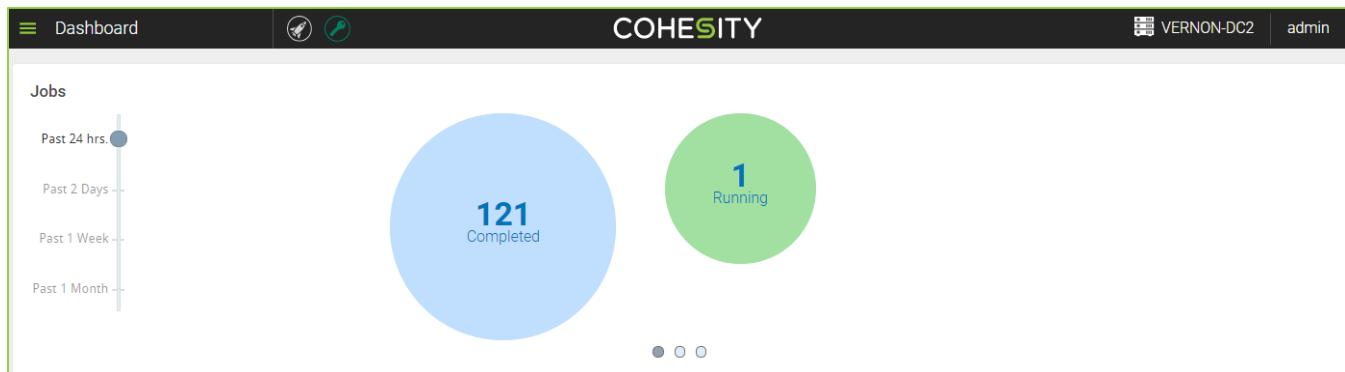
1. On the **Dashboard**, click **Backup**.
2. On the **Data Backup** page, identify the job that you need to clone and then click the clone button . A new copy of the job is created with suffix -CLONE. For example, if you clone a job or workflow named as MiamiWarehouse, then the cloned job will be renamed as MiamiWarehouse-CLONE.

12.3 Stopping a job

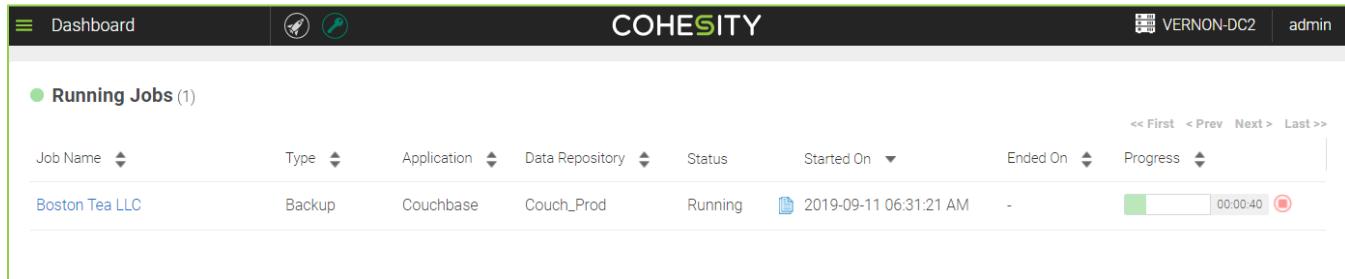
Imanis Data software enables you to stop any running job. This is a hard stop, that is, once the job is stopped it cannot be restarted again. Details of a job that is stopped can be viewed upon clicking the ‘Failed’ bubble on the Dashboard. In case the job is in an advanced stage of execution, the job cannot be stopped.

To stop a job, do the following:

1. On the **Dashboard**, click the **Running** bubble. The **Running Jobs** page is displayed.



2. In the **Running Jobs** page, identify the job that you want to stop and click the **Stop** button under the **Progress** column.



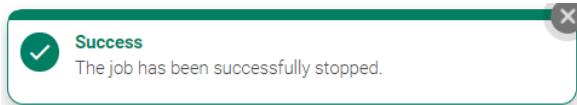
3. Click **Yes** to confirm that you are stopping the job.

The job may not be stopped if it is in an advanced state of execution. Do you want to attempt stopping it?

Yes

No

4. A confirmation message will be displayed on the page indicating the data repository is successfully stopped.



12.4 Editing a job

Go ahead and edit any job initiated through the Data Backup, Data Pipeline, and Data Mirroring menu.

What you can edit?

You can edit the job tag, data repository (as long as it refers to the same logical structure) and a few parameters of policy such as the following:

- **Type of frequency:** For the At specified times frequency option, you can increase or decrease the value under the At hours of the day option and/or increase or decrease the value under the Repeat on days of the week option
- **Backup Policy:** You can change the increase or decrease the value of Preserve recovery points for ___ days option
- **DLM Policy:** You can increase or decrease the Glacier value or change the value to Forever

What you cannot edit?

You are not permitted to edit the following:

- **Type of policy:** You cannot change a ‘backup policy’ to a ‘recovery policy’ or a DLM policy to a backup policy.
- **Type of frequency:** You cannot change **One time, immediately** to **At Specified times** and vice versa
- **DLM Policy:** You cannot change the Skip value for the Glacier step to a numeric value and vice versa
- **Suffix:** You cannot change the value in the suffix field. This restriction is by design and applicable across all the applications where ever the suffix field is enabled

For illustration purposes, the following steps show how to edit a job in the Data Backup menu:

To edit a job, do the following:

1. On the **Data Backup** page, identify the job that you want to edit from the left pane, and select it.
2. Click the  edit icon to make the required changes and click **Submit**.

12.5 Deactivating a job

The process of deactivating temporarily suspends the next upcoming runs of the job. Deactivating a job makes sense when the storage is full or if there is ongoing maintenance to the data repository.

All the jobs initiated through Data Backup, Data Pipeline, and Data Mirroring, which can be deactivated by using the following steps.

To deactivate a job, do the following:

1. Identify the job that you want to deactivate and click the  icon.
2. Click the **Yes** button on the confirmation message **Do you really want to de-activate this job?** Your job is de-activated temporarily.

NOTE: If you de-activate jobs that are active and currently underway, the job de-activates only after it completes its current running cycle.

12.6 Activating & Deactivating multiple jobs

You can activate or deactivate all jobs in one go. However, you can always identify the job that you need to activate or deactivate and click the respective icon. This feature is available in backup, data lifecycle management, data pipeline, and data mirroring module.

When you activate a previously deactivated job, its next run will be as per the applied policy. In case, the job has a 'one time policy', the job can now be run using the 'run now' button.

To activate or deactivate a job, do the following:

1. In the menu, click the  icon to activate or deactivate to activate or de-activate multiple jobs at one go.
2. Click the Activate all backup jobs  **Activate all Data Backup jobs** option or the Deactivate all backup jobs option  **Deactivate all Data Backup jobs**.
3. Click the **Yes** button on the confirmation message to activate or deactivate all jobs in one go.

12.7 Deleting a job

Imanis Data enables you to delete a job even if it is currently in a running state. All the data and metadata associated with the workflow is permanently deleted from the Imanis Data cluster and not from the primary data repository or the data pipeline destination. Data continues to reside in the destination cluster and will have to be manually deleted by the administrator.

All the workflows initiated through Data Backup, Data Recovery, Data Lifecycle Management, Data Pipeline, and Data Mirroring can be deleted using the following steps. However, it is recommended that you review the information below to understand the consequences of deleting a job.

To delete a job:

1. Identify the job that you want to delete and click the  icon.
2. Type the name of the job to confirm your decision.
3. Click the **I understand the consequences; delete this job** button to delete your job permanently.

What happens when you delete a running job?

- **Data Backup job:** If a running Data Backup job is deleted, all the data on the Imanis Data cluster is deleted automatically. However, the data from Amazon Glacier is not deleted. Administrators need to manually delete this data or create a recovery workflow to restore the data

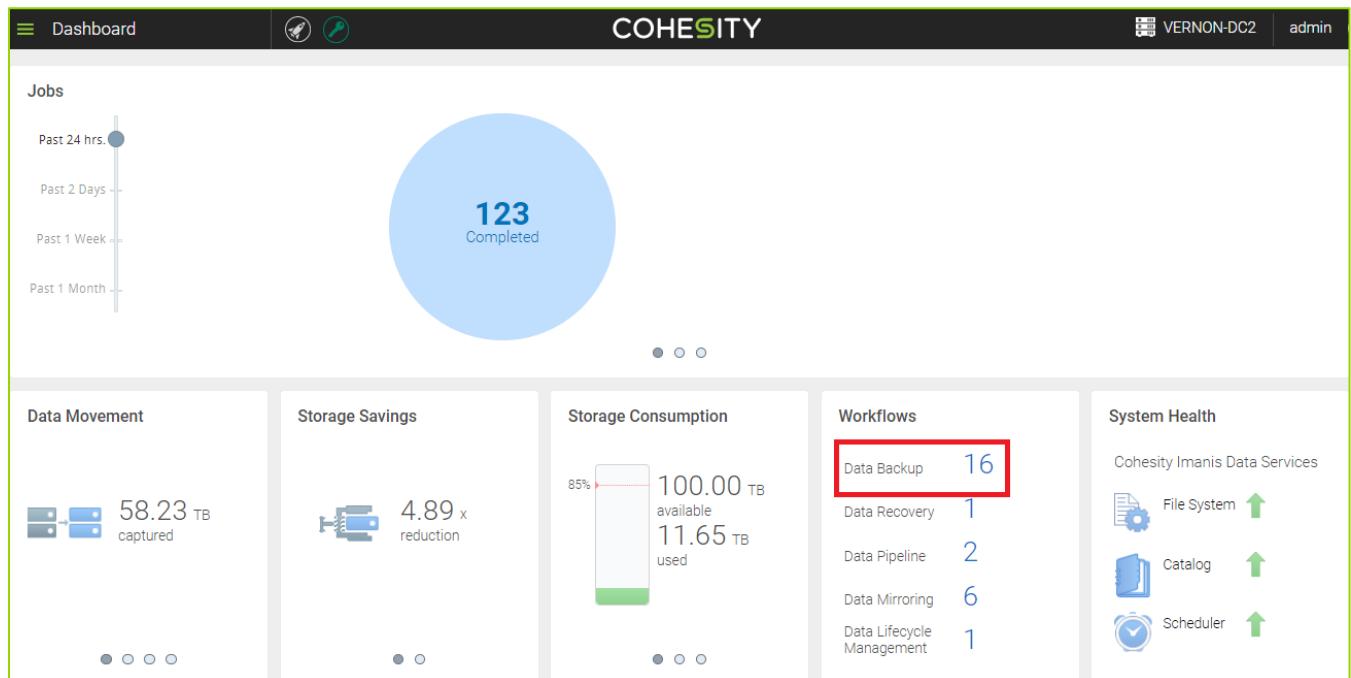
- **Data Recovery job:** If a running Data Recovery job is deleted, all the data that was available prior to the job deletion remains on the destination cluster and will need to be manually deleted
- **Data Pipeline and Data Mirroring job:** If a running Data Pipeline or Data Mirroring job is deleted, the job is terminated immediately, all the related data and metadata from the Imanis Data cluster is deleted

12.8 Monitoring Jobs

Imanis Data software enables you to view workflow details and job run statistics for data backup, data recovery, data pipeline, and data mirroring.

12.8.1 Viewing Workflow Details

- On the **Dashboard**, under **Workflows**, click the number corresponding to **Data Backup** or **Data Recovery**, **Data Pipeline**, and **Data Mirroring**.



- In the **Data Backup** page, click on the job in the left pane whose workflow details you want to view. For example, the screenshot below displays the details of a workflow named ADLS_Hourly



12.8.2 Viewing Job Run Statistics

You can view the job and data-related statistics. The set of statistics displayed within the Stats tab depends on the type of workflow.

1. On the **Dashboard**, under **Workflows**, click the number corresponding to **Data Mirroring**. The Data Mirroring page appears.
2. In the **Data Mirroring** page, click on the job in the left pane whose statistics you want to view.
3. Click the **Stats** tab to do the following:
 - a. View the Last run date and time of the workflow, the next scheduled workflow run, and the current status of the workflow
 - b. View the number of job runs, the average data backup time, average data backup size, and average number of mutations
 - c. Click the **Job Runs** pane to expand and view individual job run statistics

See the following screenshots for each of the workflows:

- **Data Backup Workflow**

The screenshot shows the Cohesity interface for a Data Backup Workflow. The left sidebar lists various backup jobs, including ADLS_Hourly, Boston Tea LLC, Boston Tea Party LLC, CASS_Backup_Now, Cass_Hourly2, Cassandra Hourly Backup, Couch Daily Backup, HBase Hourly Backup, HD Insight Prod Daily Back..., Hdfs Hourly Backup, and Hive Hourly Backup. The main area displays the 'Stats' tab for the selected 'ADLS_Hourly' job. It shows the last run was on 2019-08-09 at 06:00:28 AM, and the next scheduled run is indicated. Below this are four summary cards: 'Number of Job Runs' (324), 'Avg. Data Backup Time' (5.0 Minutes), 'Avg. Data Backup Size' (35.5 GB), and 'Avg. Number of Mutations' (1 Creations, 49 Modifications, 24 Deletions). At the bottom, the 'Job Runs' section is expanded, showing three recent runs with details like Date, Data Backup Time, Size, and Mutation counts.

Date	Data Backup Time	Size	Creations	Modifications	Deletions
2019-08-09 06:00:28 AM	5.3 Minutes	42.1 GB	0	80	40
2019-08-09 05:00:27 AM	5.3 Minutes	34.6 GB	0	76	38
2019-08-09 04:00:26 AM	5.2 Minutes	26.9 GB	0	77	38

- **Data Lifecycle Management Workflow**

The screenshot shows the Cohesity Data Lifecycle Management interface. On the left, a sidebar lists 'Data Lifecycle Management' and various job types. A 'DLM Job' entry is selected, showing its details. The main area displays a 'Workflow Definition' card with a 'Last Run' of 2019-09-11 at 07:29:22 AM. Below this are four summary cards: 'Number of Job Runs' (1), 'Avg. Data Recovery Time' (5.5 Minutes), 'Avg. Data Recovery Size' (47.3 GB), and 'Avg. Number of Mutations' (202 Creations, 0 Modifications, 0 Deletions). At the bottom is a table of 'Job Runs' with one entry: Date 2019-09-11 07:29:22 AM, Data Recovery Time 5.5 Minutes, Size 47.3 GB, Creations 202, Modifications 0, and Deletions 0.

- Data Pipeline or Data Mirroring Workflows: In the **Capture** section

The screenshot shows the Cohesity Data Mirroring interface. On the left, a sidebar lists various mirroring jobs. A 'Mongo Mirroring' entry is selected, showing its details. The main area displays a 'Capture' card with a 'Last Run' of 2019-08-09 at 06:00:28 AM. Below this are four summary cards: 'Number of Job Runs' (772), 'Avg. Data Backup Time' (5.0 Minutes), 'Avg. Data Backup Size' (713.9 MB), and 'Avg. Number of Mutations' (0 Creations, 1 Modifications, 1 Deletions). At the bottom is a table of 'Job Runs' with four entries: 2019-08-09 06:00:28 AM (5.1 Minutes, 642.6 MB, 0 Creations, 2 Modifications, 1 Deletions), 2019-08-09 05:00:27 AM (4.5 Minutes, 644.0 MB, 0 Creations, 1 Modifications, 0 Deletions), 2019-08-09 04:00:26 AM (4.9 Minutes, 904.9 MB, 0 Creations, 1 Modifications, 0 Deletions), and 2019-08-09 03:00:25 AM (4.8 Minutes, 730.7 MB, 0 Creations, 1 Modifications, 0 Deletions).

- Data Pipeline or Data Mirroring Workflows: In the **Recovery** section

The screenshot shows the Cohesity Imanis Data interface. On the left, a sidebar lists various mirroring jobs: Cassandra Mirroring, Couchbase_Mirroring, HBase Mirroring, Hive Mirroring, Mongo Mirroring (selected), and Vertical Prod to Analytics. The main area is titled 'Advanced Stats' under a 'Workflow Definition' section. It displays 'Recovery' statistics: Last Run (2019-08-09 06:04:58 AM), Total Data Recovery Time (5.0 Minutes), Total Data Recovery Size (717.4 MB), and Total Number of Mutations (0 Creations, 1 Modifications, 1 Deletions). Below this is a table titled 'Job Runs' with three entries:

Date	Data Recovery Time	Size	Creations	Modifications	Deletions
2019-08-09 06:04:58 AM	4.6 Minutes	742.5 MB	0	2	1
2019-08-09 05:04:57 AM	5.2 Minutes	796.0 MB	0	2	1
2019-08-09 04:04:56 AM	5.1 Minutes	838.3 MB	0	2	1

12.8.3 Viewing Advanced Job Run Statistics

You can view advanced job run and data-related statistics in the Advanced Stats tab. The Advanced Stats tab consists of the Deduplication Stats and the S3 Migration Stats.

12.8.3.1 Deduplication Stats

Imanis Data software uses intelligent data compression techniques to eliminate duplicate copies of data available on Imanis Data cluster. The Deduplication Stats tab displays statistics of the deduplication process such as the number of runs, average time taken for deduplication by Imanis Data software, and the average storage savings on the Imanis Data cluster.

You can also refer to the Deduplication Runs table for a more granular view of the deduplication statistics.

To view Deduplication Stats, do the following:

- On the **Dashboard**, under **Workflows**, click the number corresponding to **Data Backup**.
The Data Backup page appears.
- In the **Data Backup** page, identify a workflow in the left pane whose deduplication statistics you want to view.
- Click the **Advanced Stats** tab to do the following:
 - View the tiles of number of runs, average deduplication time taken by Imanis Data software, and average savings you made on the Imanis Data cluster.

- b. Click the Deduplication Runs bar to individual job run statistics such as date of the job run, deduplication time, input data (before deduplication), de-duplicated data (after deduplication), and storage savings done for each job run (see the following screenshot):

The screenshot shows the Cohesity Dashboard interface. At the top, there are three tabs: 'Workflow Definition', 'Stats', and 'Advanced Stats', with 'Advanced Stats' being the active tab. Below the tabs, the title 'Deduplication Stats' is displayed. Under this title, there are three cards: 'Number of Runs' (1289), 'Avg. Deduplication Time' (5.0 Minutes), and 'Avg. Storage Savings' (39.96x). Below these cards is a table titled '▼ Deduplication Runs'. The table has columns for Date, Deduplication Time, Input Data, Deduplicated Data, and Storage Savings. The data in the table is as follows:

Date	Deduplication Time	Input Data	Deduplicated Data	Storage Savings
2019-09-11 06:00:21 AM	4.8 Minutes	7.7 GB	1.7 GB	4.59x
2019-09-11 05:00:20 AM	5.1 Minutes	10.6 GB	2.3 GB	4.56x
2019-09-11 04:00:20 AM	5.5 Minutes	15.3 GB	2.7 GB	5.68x
2019-09-11 03:00:19 AM	5.5 Minutes	13.9 GB	2.5 GB	5.55x

12.8.3.2 Cloud Migration Stats

Imanis Data software manages the migration of duplicated data from the Imanis Data cluster to Cloud (repository meant for global deduplication purposes) repository. The Cloud Migration Stats tab displays the statistics of the migration process such as number of migrations, average migration time, average migration size, and total data size.

You can also refer to the Cloud Migrations table for a more granular view of the de-duplicated data migration statistics.

The Cloud Migration Stats are displayed only if a particular backup workflow uses a backup policy that is scheduled to copy data on to the Cloud (data repository meant for global deduplication purposes).

To view Cloud Migration Stats, do the following:

1. On the **Dashboard**, under **Workflows**, click the number corresponding to **Data Backup**. The **Data Backup** page appears.

2. In the **Data Backup** page, identify a data backup workflow whose cloud migration statistics you want to view.
3. Click the **Advanced Stats** tab to do the following:
 - a. View the **Deduplication Runs** stats such as **Number Of Runs**, **Average Deduplication Time** (taken by Imanis Data software), and **Average Storage Savings** on both Imanis Data cluster and Cloud (data repository meant for global deduplication purposes).
 - b. Click the **Cloud Migrations** bar to open and view individual migration statistics such as **Data Backup Timestamp**, **Cloud Migration Timestamp**, **Cloud Migration Time**, **Cloud Migration Size** (of the data), and **Storage Savings** (on Imanis Data).

13 Appendix A: Rules for Data Inclusions and Exclusions

Imanis Data software supports Unix-like regular expressions (regex) for primary repository browsing and “inclusions and exclusions” under the Rules tab for all the applications. However, Imanis Data software does not support regex for the parent and child level.

How does regex work?

The regular expressions (regex) in the Exclusions and Inclusions field have to be separated by a new line. Regular expressions work as follows:

- The inclusion filter generates a list of data objects on the selected data repository that match the regular expression. This list is in addition to the data objects that you have already selected in the Identify Data section
- The exclusion filter is applied to the list of data objects generated by the inclusion filter and the selected objects during the Identify Data step. The filter intelligently excludes objects that match the regex

For example, a user has selected keyspace named “keyspace_abc” in the Identify Data section:

- In the Inclusion field, if the user specifies an inclusion filter as keyspace_*, then the user has selected the keyspace named “keyspace_abc” and also all the keyspaces from the data repository starting with name prefix 'keyspace_'
- In the Exclusion field, if the user specifies an exclusion filter as *table_tmp*, then the user has selected tables containing string table_tmp, which will be excluded from the list of data objects that have been selected using the Identify Data section and inclusion filter

IMPORTANT: Avoid blank spaces at the start or end in the inclusion or exclusion regex. In Imanis Data software, a space is considered as a character. When spaces are inserted, the data backup process is executed successfully; however, the actual data is not backed up because the source list contains a space which does not match the source data.

13.1 HDFS Regular Expressions

This section describes the properties of HDFS object names.

13.1.1 Inclusion for HDFS

You must specify paths beginning with root of the file system (“/”) until the required directories or files are encountered.

ACTION TO TAKE	CORRECT USAGE
To include all text files from a directory	/dir1/*.txt
To include file(s) from set of directories	/dir*/file*
To include a range of directories or files	/dir[1-5]/reports.txt.
To include files or directories which are present at 3rd level in file system hierarchy	/*/*/dir*

13.1.2 Exclusions for HDFS

You must specify paths beginning with Root of the File System (“/”) until required directories or files are encountered.

ACTION TO TAKE	CORRECT USAGE
To exclude all log files from backup path	*.log
To exclude log files from a specific directory	/dir1/*.log

WARNING: Imanis Data software does not support “,” and “-” in regular expressions. For example, expressions like /dir{1,2}/*.log and /dir[1-3]/file1 are not supported currently.

13.2 Hive Regular Expressions

The following is supported for backup, restore, and data orchestration (DO):

- Hive Backup is supported at the database and table level
- Hive Restore is supported at the database, table, and partition level
- Hive DO is supported at the database and table level

The following section describes the properties of Hive object names:

- Hive object names have to be complete names
- Complete name consists of three parts separated by ':': First part: database name, Second part table name, Third part: partition name (same as which appears in the HDFS directory corresponding to the partition, in format 'key1=value/key2=value/key3=value...')

According to this rule:

- * Database complete name consists of one part
- * Table complete name consists of two parts separated by '.' where the first part is the database to which table belongs
- * Partition complete name consists of three parts separated by '.' where first two parts make complete table name to which that partition belongs

13.2.1 Inclusions for Hive

Hive object name present in the inclusion list should be a complete Hive object name. Regex is applicable separately on different parts of complete Hive object name. For example: If the reflex is to be applied in partition name, then it should contain 'database name/regex' followed by 'table name/regex'.

'.' has special meaning and cannot be skipped by defining as part of regex. For example: If user needs to specify regex for all tables ending with _suffix from database db_1, db_2, db_3. Then, user must specify regex db_[1-3].*_suffix and NOTdb_[1-3]*_suffix. Here the first expression will be correctly considered as regex for table. However, the second expression will be incorrectly considered as regex for database name.

ACTION TO TAKE	CORRECT USAGE
To include all databases starting with prefix 'db'	db*
To include databases with any one character following prefix database_	database_?
To include all tables starting with prefix 'table' and belonging to any of databases (database1, database2, database3, database4, database5)	database[1-5].table*

13.2.2 Exclusions for Hive

These can be a complete Hive object names or glob patterns. All rules specified in the inclusion list to apply here. Additionally exclude regex does not consider '.' as a special character and gives freedom to consider '.' as part of regular expression. For example: If a user needs to exclude tables ending with '_suffix' from databases db_1, db_2, db_3. Then user can specify db_[1-3].*_suffix or db_[1-3]*_suffix, both are allowed.

To exclude all tables containing string 'meta':

Correct: *meta*

14 Appendix B: EC2 Installation Cheat Sheet

In this section you will learn how to set up multi-dc configuration and verify the data source of EC2 Cassandra cluster.

14.1 Setting Up Multi-DC Configuration

In EC2, when a Cassandra application is set up in a multi-data center (DC) configuration setup, there are certain configuration parameters that must be set correctly. Refer to the following steps to set the configuration parameters:

To set configuration parameters, do the following:

1. Access and open the `cassandra.yaml` file.
2. In the `cassandra.yaml` file, set `rpc_address` to `0.0.0.0`, `broadcast_address` to `<public_ip>`, `listen_address` to `<private_ip>`.
3. In Cassandra 2.1 onward, configure parameter `broadcast_rpc_address` to `<public_ip>`. This is used by drivers to connect to Cassandra. Here `<public_ip>` refers to the IP address in `a.b.c.d` format and not FQDN.
4. Set `JVM_OPTS="$JVM_OPTS -Djava.rmi.server.hostname=<public_ip>"` in `cassandra-env.sh`.
5. After making the configuration changes, run the `nodetool status` to verify if the configuration is correct. Then run `cqlsh` (or `cqlsh <host_ip>`) to see if it works.

Please be aware that the verification process may pass without executing the preceding steps, however; the connector could fail because the verification process uses JMX only. However, if the issue persists, check the `system.peers` table by executing the following query in `cqlsh`:

```
SELECT peer, data_center, host_id, preferred_ip, rack, rpc_address, release_version  
from system.peers;
```

In the output of the above query, the `rpc_address` column should return public IPs in a multi-DC setup. If it contains private IP or `0.0.0.0` value, Imanis Data software may not be able to connect from outside the AWS region. Also ensure that there is no firewall blocking JMX and RPC ports. If the Imanis Data cluster and Cassandra are in the same region private IP access will work.

NOTE: If you do not set RMI server hostname, it may not cause any issues when running `cqlsh` or `nodetool`, however; it may cause the Imanis Data Cassandra JMX connector to fail. This behavior was observed in Cassandra 2.0.10 version. Configuration files on all nodes must be changed as indicated in the preceding section and Cassandra must be restarted, if required.

14.2 Verifying the Data Source of EC2 Cassandra Cluster

If the Imanis Data cluster and Cassandra data source are set in different AWS region, the EC2 Cassandra cluster data source verification may fail. Refer to the following steps to set the configuration parameters:

To set configuration parameters, do the following:

1. Access and open the `cassandra-env.sh` file.
2. In the `cassandra-env.sh` file, set the following configuration parameter:
`JVM_OPTS="$JVM_OPTS -Djava.rmi.server.hostname=<public_ip>"`.
Here `<public_ip>` refers to means the IP address in a.b.c.d format and not public DNS.

NOTE: To be able to carry data center-specific data movement, names of all data centers must be entered at the time of data repository discovery because currently automatic discovery is not supported. Use the nodetool status to get the list of data centers and enter them in the textbox labeled Data Centers each separated by a comma. The Data Centers field will not be populated in case of EC2 specific snitches. In Amazon EC2MultiregionSnitch, data center name is a concatenation of region followed by data center suffix. For example, in us-east_dc1 where 'us-east' is the region and '_dc1' is the data center suffix.

15 Appendix C: Quick Troubleshooting

In this section you will learn about some basic steps to help you quickly fix some issues with Imanis Data software.

15.1 Job Log Collection Tool

The Job Log Collection tool is used to collect various logs and statistics associated with a job run in Imanis Data software. The tool collects information that is required by the Imanis Data Support team to understand and debug the issue when reported by customers.

To run the job collection tool, do the following:

Run the following Job Collection tool script as \$SERVICE_USER for Imanis Data (usually hdfs by default) talena-joblog.sh:

```
$INSTALL_DIR/bin/talena-joblog.sh --action stats|logs|systemjoblogs|all --jobname <jobname> --outdir <dirname> [--limit <n>]

stats : Collect job stats and job history for given jobname.
logs : Collect yarn/agent logs for given jobname.
systemjoblogs : Collect yarn/agent logs of system jobs corresponding to given jobname
all : Collect job stats/history/agent logs/yarnlogs/system job logs
--limit : n days old stats/history/logs will be collected.
If --limit value is not provided then last 7 runs log will be collected.
--outdir : Absolute path of directory where logs will be collected. Please make sure that directory exists and has write permissions for hdfs user.
```

To collect job logs and stats for global compaction system job, do the following:

```
opt/talena/bin/talena-joblog.sh --action globaljoblogs --outdir <dirname> [--limit <n>]
```

To collect job logs and stats for given jobname for given job run, do the following:

```
/opt/talena/bin/talena-joblog.sh --action runlogs --jobname <jobname> --runid <jobRunID> --outdir <dirname>
```

--runid can be obtained from the following screen:

The screenshot shows a logs interface with the following details:

- Logs** tab is selected.
- Hive Mirroring_from_Imanis** is the job name.
- Status**: Completed
- Start time**: 2019-03-05 03:05:50 PM
- End time**: 2019-03-05 03:10:58 PM
- Job Run ID**: 2019_02_26T1423

The log content is as follows:

```

2015-11-27 00:14:11 INFO mapreduce.Job:Running job: job_1448610330496_0030
2015-11-27 00:14:18 INFO mapreduce.Job:Job job_1448610330496_0030 running in uber mode : false
2015-11-27 00:14:18 INFO mapreduce.Job: map 0% reduce 0%
2015-11-27 00:14:27 INFO mapreduce.Job: map 100% reduce 0%
2015-11-27 00:14:27 INFO mapreduce.Job: Job job_1448610330496_0030 completed successfully
2015-11-27 00:14:28 INFO mapreduce.Job: Counters: 30 File System Counters FILE: Number of bytes read=0 FILE: Number of bytes written=139769 FILE: Number of read operations=0 FILE: Number of large read operations=0 FILE: Number of write operations=0 HDFS: Number of bytes read=7476 HDFS: Number of bytes written=0 HDFS: Number of read operations=3 HDFS: Number of large read operations=0 HDFS: Number of write operations=0 Job Counters Launched map tasks=1 Data-local map tasks=1 Total time spent by all maps in occupied slots (ms)=6560 Total time spent by all reduces in occupied slots (ms)=0 Total time spent by all map tasks (ms)=6560 Total vcore-seconds taken by all map tasks=6560 Total megabyte-seconds taken by all map tasks=26869760 Map-Reduce Framework Map input records=0 Map output records=0 Input split bytes=242 Spilled Records=0 Failed Shuffles=0 Merged Map outputs=0 GC time elapsed (ms)=0 CPU time spent (ms)=2690 Physical memory (bytes) snapshot=476782592 Virtual memory (bytes) snapshot=4331425792 Total committed heap usage (bytes)=2026373120 File Input Format Counters Bytes Read=7234 File Output Format Counters Bytes Written=0
2015-11-27 00:14:28 INFO zookeeper.ZooKeeper: Initiating client connection, connectString=talena-33:2181,talena-41:2181,talena-42:2181 sessionTimeout=60000 watcher=org.kiji.schema.layout.impl.ZooKeeperClient$SessionWatcher@4c380929
2015-11-27 00:14:28 INFO zookeeper.ClientCnxn: Opening socket connection to server talena-33/10.1.10.23:2181. Will not attempt to authenticate using SASL (unknown error)
2015-11-27 00:14:28 INFO zookeeper.ClientCnxn: Socket connection established to talena-33/10.1.10.23:2181, initiating session

```

15.2 Masking dialog box does not respond

Problem

While using the Masking-Sampling feature, if a user loads the data, applies masking by clicking the Apply link in the Masking dialog box, closes the dialog box, and again clicks the Apply link the Imanis Data software GUI stops responding.

This problem is caused by the Skype add-on (version 8.1.x.xxxx or lower) present in the Mozilla Firefox Web browser. If you have enabled the ‘auto-update’ feature in the Mozilla Firefox version 46, then this version of the Skype add-on will be automatically added to the browser without asking the user's permission. You can disable the Skype add-on however; you cannot uninstall or remove the add-on from this version of the Firefox browser as there is no button or link to do it.

Resolution

1. Disabling the Skype add-on

The Skype add-on in the Firefox browser searches for phone numbers available on the Masking dialog box. This process slows down the performance of Imanis Data software. In some cases, the Imanis Data software UI stops responding completely.

To resolve this problem, you must disable the Skype add-on. Refer to the Mozilla Firefox documentation for information on how to disable or remove add-ons: <https://support.mozilla.org/en-US/kb/disable-or-remove-add-ons>

2. Reinstalling the Mozilla Firefox Web browser

In case you frequently use the Skype add-on and want to continue using it, you must first uninstall Mozilla Firefox and install it again. Then you can proceed to add the Skype add-on to the Firefox browser.

The Mozilla Firefox team has resolved this Skype add-on issue in the latest release (46.0.1). In the new version of Firefox 46.0.1, the Skype add-on will not be automatically added to the browser. You would also be able to disable or remove. Make sure that the Skype version is 8.2.x.xxxx or above.

15.3 Valid revisions for objects not available

Problem

When using an exact or partial search, the results may contain data objects that are being currently loaded (as part of a catalog load operation) or deleted (as part of a restore point delete operation).

Such data objects are unrecoverable since their respective restore points are in flux — either they're partially loaded or are being deleted.

The following message is displayed when an attempt is made to select such objects for recovery: "No valid revisions available for this object at this time".

Resolution

As the process of loading the catalog is in progress, you can retry accessing these data objects after some time.

15.4 Email not received after job is complete or failed

Problem

The 'Email Notifications' feature sends an automated email to the user upon completing or failing of a particular job. However, there is a possibility that this email is not received in the inbox of the user.

Resolution

- Check the Spam folder. There is a possibility that some email service providers send Imanis Data email in the spam folder
- On all the Imanis Data nodes, check if mailx is installed using the following mailx command:
`#echo "TestMail" | mailx -v -s "TestMail from Imanis Data Cluster" <email-address>`

- On all the Imanis Data nodes, check if service sendmail is active using command "service sendmail status"

15.5 Hive row count does not match after recovery

Problem

Select count (*) after recovery does not match with the count of backup.

Resolution

1. On the HIVE CLI set the following property to false: `hive.compute.query.using.stats=false`
2. Then run `count (*)`

16 Appendix D: Mirroring Workflow in TDE/Encrypted Environment in Cassandra

In this section you will learn how to set up a Mirroring workflow between a Cassandra (TDE/Encrypted environment) and Cassandra destination cluster where you must edit the 'table' properties for 'compression strategy'.

However, before you begin, you must ensure the destination cluster has both the encryption keys or Truststores (external TDE) from the source Cassandra cluster and the Destination Cassandra clusters configured.

To start a data mirroring workflow in TDE/Encrypted environment, do the following:

1. Click the **Main Menu**  > **Data Management** > **Data Mirroring**.
2. On the **Data Mirroring** screen, click the  **+ Add New** button or the  icon to create a data mirroring workflow. The **New Data Mirroring Workflow** dialog box appears.
3. In the **New Data Mirroring Workflow** dialog box, select a Cassandra data repository from the **Data Repository** drop-down menu. Imanis Data software auto-populates the **Application** field with **Cassandra** as it recognizes the type of data repository that is selected.
4. Click **OK**.
5. Do the following:
 - Type a new job tag in the **Job Tag** field
 - Type a job tag description in the **Description** field. The job tag name can include alphanumeric characters, numbers and/or special characters
6. In the **Source** area, in the **Identify Data** area, do the following:
 - In the **Cassandra** tab, identify the keyspaces and tables that you want to backup by selecting the corresponding check boxes. Use regular expressions (regex) for primary repository browsing.
 - In the **Selected Data** tab, verify your selection or click the  icon to remove unwanted items.
 - In the **Rules** tab, type regular expressions (regex) or search term to exclude or include specific data objects in the Rules tab.
7. In the **Specify Policy** section, under Select a data backup policy, select a backup policy with or without cloud retention.
8. In the **Cloud** options area, do one of the following. This option is displayed if you selected a backup policy with a cloud retention option.
 - Select an S3 or Azure cloud repository from the **Data Repository** drop-down menu and select a bucket (S3) or a container (Azure) from the **Buckets/Containers** drop-down menu to retain data in the cloud. The following screenshot displays an example of S3 data repository:
 - Select an S3 or Azure cloud repository from the **Data Repository** drop-down menu and type the name of the bucket (S3) or the container (Azure) for which you have been granted access by

your organization in the **Buckets/Containers** field. The following screenshot displays an example of S3 data repository:

9. In the **Destination** section, select the destination data repository from the drop-down menu where you want to move the backed up data.
10. In the **Destination** area, in the **Mirroring Options**, do the following:
 - Click **Replace** to replace existing data with new data thus erasing any previously existing data
 - Click **Keep** to retain existing data (if any). However, if there is not existing data then the new data will be copied
11. In the **Suffix** field, type a number and/or character as a suffix to the data objects being recovered from the Imanis Data cluster. For example, _example.
12. In the **More Options for Selected Data** section, do the following:
 - To rename restored objects, type the new name in the **Recover As** column.
 - To change property of the restored object, click the icon, and type the values in the **Key** and **Value** field. Usually, the key is auto-completed by the UI as soon as you start to type. The **Value** field must contain the complete value of the property.

Only the following property changes are allowed:

- keyspace -- replication strategy
- table -- compression strategy and compaction strategy

The following screenshot displays an example where table properties are edited for compression strategy:

The screenshot shows a table titled "2 More Options for Selected Data". The table has three columns: "Objects", "Recover As", and "With Properties".

Objects	Recover As	With Properties
movies_directory	movies_directory	+ compression = {'class': 'Encryptor', 'key_provider': 'KmipKeyProviderFactory', 'kmip_host': 'kmip_group_name' ['key_namespace' = 'kmip_namespace'], 'cipher_algorithm': 'AES/ECB/PKCS5Padding', 'secret_key_strength': 128}
ratings_directory	ratings_directory	+
genre_directory	genre_directory	+

NOTE: You must ensure that the `kmip_host` refers to the `kmip_host` used at the destination cluster. The destination cluster must have `kmip_host` from the source Cassandra cluster if `kmip_host` at the destination cluster is different from the source cluster.

13. Click the **Advanced Options** pane to expand and set **Bandwidth Throttling**, **Concurrency Throttling**, and **Email Notifications**. Imanis Data software can handle several jobs easily. However, during certain times of the day, you can decrease congestion over the network and primary cluster and reduce the number of concurrent job through the throttling feature. If you do not specify throttling parameters, default values are applied from `mapred-site.xml`.
 - In the Bandwidth Throttling option, click the lock icon to use the feature, click Yes to confirm, and then set the number of MBs you want to request for each individual mapper by moving the slider or typing a number in the MB/s per Mapper field.
 - In the Concurrency Throttling option, click the lock icon to use the feature, click Yes to confirm, and then click the plus sign (+) or the minus sign (-) to increase or decrease the request of number of mappers for each job.
It may be possible that the requested number of MBs per mapper or number of mappers per job may not be granted but you will not receive more throughputs (in MBPs) per mapper or number of mappers per job than what you have requested.
 - In the Email Notifications option, click Yes to confirm, click the plus sign (+), and then type the email id where you want to receive the job completion or failure email in the Email Addresses field.
14. Click **Submit**.

17 Appendix D: Mirroring Workflow in TDE/Encrypted Environment in Hadoop

17.1 HDFS

Before taking a restore in Hadoop TDE/encrypted environment, you must ensure the following:

- The 'hdfs' service user of primary Hadoop cluster must be removed from Key Management Server Access Control's blacklist as WEBHDFS internally runs as HDFS user. HDFS user must have 'Decrypt_EEK' permission. HDFS service user must be removed from the following properties of KMS server:
 - kms_blacklist_users
 - hadoop.kms.blacklist.DECRYPT_EEK
- The restore destination directory specified must have been previously created and must be an encryption zone to ensure that the restored data is encrypted.
- If restoring to root directory ("/"), the encryption zone must be at least created at the parent directory level where data must be restored.
For example, if the directory to be restored is '/dir0/dir1/dir2' at destination cluster, then encryption zone must be created on '/dir0' or '/dir0/dir1' and not at restore directory level '/dir0/dir1/dir2' as HDFS restore operation will overwrite the directory '/dir0/dir1/dir2' at destination cluster if already present, but will not affect its parent directories and zones will be preserved.

17.2 HBase

- TDE encryption for HBase is supported at parent directory of HBase root directory.
For example, if hbase.rootdir = '/apps/hbase/data', then the encryption zone must be created at '/apps/hbase', as HBase internally uses '/apps/hbase/' as staging directory which is supposed to be within the zone.
- Ensure Hbase staging directory is within the encryption zone directory.

17.3 HIVE

- TDE encryption for HIVE is only supported at database level and hive default warehouse directory level.

- For database level encryption zone, before running restore job, user must perform the following steps:
 1. create database directory at destination cluster and create a zone on database directory.
 2. Run the restore job. It will restore the data to destination database directory that user has created above.

For example, if database '**xyz**' must be restored and the destination directory is under hive warehouse location. If the destination warehouse location for managed and external is **/user/hive/managed/** and **/user/hive/external/** respectively, then the user must create the following directories:

 - **/user/hive/managed/xyz.db**
 - **/user/hive/external/xyz.db**

A zone must be created on the above directories. The restore job will copy hive data files to these directories under encryption zone.

If the database 'db1' uses external location such as '**/data/databases/db1**', then user must create directory and zone at '**/data/databases/db1**' before running the restore job.

- If the table or partition uses external location which is outside hive warehouse directory, then the encryption zone must be created at the parent directory of the object's locations.

For example, if the external table 'table1' with location '**/data/tables/table1**' to be restored, then the encryption zone can be at '**/data/tables**' or '**/data**' but not at table location, that is , '**/data/tables/table1**'
For example, if external partition 'p=1/p=2' with location '**/data/partitions/p=1/p=2**' which is not under table's location, then encryption zone can be at '**/data/partitions**' or '**/data**' but not at partition location i.e. '**/data/partitions/p=1/p=2**'.

18 Appendix E: Couchbase Point-in-Time (PIT) Recovery Limitation

In this section you will learn about an limitation of Couchbase Point-in-Time (PIT) recovery functionality.

Couchbase backups can only get the latest mutation for a specific key. Thus, if a key has been edited or mutated multiple times between two backups, only the second backup will get the latest copy of the data. Point-in-Time (PIT) Recovery runs get only the metadata and rely on next backup to get the actual mutations.

Couchbase PIT recovery is described in the following table where **kXvY implies version Y of Key kX**:

		1 ST BACKUP RUN		1 ST PITR RUN		2 ND PITR RUN	2 ND BACKUP RUN
Keys							
k1	k1v1						
k2			k2v1				
k3					k3v1		
k4	k4v1		k4v2				
k5	k5v1				k5v2		
k6	k6v1		k6v2		k6v3		

	RESTORE FOR PITR1	RESTORE FOR PITR2
Keys		
k1	k1v1	k1v1
k2	k2v1	k2v1
k3	Not restored	k3v1
k4	k4v2	k4v2

	RESTORE FOR PITR1	RESTORE FOR PITR2
k5	k5v1	k5v2
k6	k6v1	k6v3

NOTE: For key k6, the 2nd Backup run gets the latest copy k6v3. Thus, the PIT Recovery for 1st run of PITR cannot restore k6v2.

Your Feedback

Was this document helpful? [Send us your feedback!](#)

ABOUT COHESITY

Cohesity makes your data work for you by consolidating secondary storage silos onto a hyperconverged, web-scale data platform that spans both private and public clouds. Enterprise customers begin by radically streamlining their backup and data protection, then converge file and object services, test/dev instances, and analytic functions to provide a global data store. Cohesity counts many Global 1000 companies and federal agencies among its rapidly growing customer base and was named to Forbes' "Next Billion-Dollar Startups 2017," LinkedIn's "Startups: The 50 Industry Disruptors You Need to Know Now," and CRN's "2017 Emerging Vendors in Storage" lists.

For more information, visit our [website](#) and [blog](#), follow us on [Twitter](#) and [LinkedIn](#) and like us on [Facebook](#).

© 2019. Cohesity, Inc. Confidential & Proprietary. For Internal Distribution Only.

Cohesity, the Cohesity logo, SnapFS, SnapTree, SpanFS, and SpanOS, are registered trademarks, and DataPlatform, DataProtect, and Helios are trademarks of Cohesity, Inc. All rights reserved.