Attunity Compose for Data Lakes 6.4 Release Notes - August 2019

The new version of Compose for Data Lakes introduces new features and enhancements including support for provisioning to Google BigQuery.

In these release notes:

Post-Upgrade Procedures for Hive Project Optimization

In previous versions, when working with Hive projects, Compose would insert a new record to the **attrep_apply_batches** and **attrep_ddl_history** Control Tables every time a change occurred. As Hive creates a new file for each newly inserted record, this would sometimes lead to the accumulation of numerous small files which significantly degraded performance. This issue has been resolved in Attunity Compose for Data Lakes 6.4.

After upgrading, you need to run a procedure that generates a script for each of your Hive projects. Running the script in the project database will create new Control Tables based on the existing tables, but with a different structure. Records in the new Control Tables will be updated for each change (instead of being inserted), thereby preventing the accumulation of large numbers of small files.

The table names before and after the script is run are shown in the following table:

Current Table Name	Current Table Name after Upgrade	New Table Name (6.4)
attrep_apply_batches	OBSOLETE_attrep_apply_batches	attcmps_apply_cmps_batches
attrep_ddl_history	OBSOLETE_attrep_ddl_history	attcmps_ddl_history

Once the scripts have been run, you need to run an additional command to regenerate all of the task ETLs so that they are compatible with the new Control Table structure. Both of the required procedures are described below.

To generate and run the scripts:

1. From the Windows **Start** menu, open **Attunity Compose for Data Lakes** > **Compose for Data Lakes Command Line** and run the following command:

   ```
   composecli generate_upgrade_scripts
   ```

   This will create a script with the following name (for each Hive project) under **<install_dir>data\projects\<project name>\ddl-scripts**:

   ```
   ComposeUpgradeFrom6.3To6.4_<hive_project_name>__<timestamp>.sql
   ```

2. Leave the Compose for Data Lakes Command Line open as you will need it to generate the task ETLs (described below)
3. In the project database, run the script for each of your Hive projects.

> **Note** The new prefix for the Compose control tables - `attcmps` - allows the same database to be shared between the Landing Zone and the Storage Zone.

Once you are sure that the new Control Tables are working without issue, you can delete the old Control Tables (which can be identified by the OBSOLETE prefix).

To generate the task ETLs:

1. Run the following command:

   ```
   ComposeCli.exe connect
   ```

2. Generate all ETLs on Compose for Data Lakes 6.4 by running the following command:

   ```
   ComposeCli.exe generate_etls
   ```

   Any invalid tasks will be skipped and an appropriate error will be printed to the output.

   > **Notes**
   >
   > - The ETL generation process may take a while (depending on the number of tasks and projects) as Compose for Data Lakes needs to connect to each of the relevant databases.
   > - If you prefer, you can regenerate the ETL instructions manually for each task. Note however that a task will fail to run until its ETL instructions are regenerated.

Attunity Product Compatibility

Attunity Compose for Data Lakes 6.4 is compatible with the following Attunity products only:

- **Attunity Replicate**: Versions 5.5, 6.3, 6.4, and 6.2 during the grace period
- **Attunity Enterprise Manager (AEM)**: Version 6.4 SP03 and above

Provisioning to Google BigQuery

Compose for Data Lakes 6.4 introduces support for Google BigQuery. The new distribution is supported as a Provisioning Target in Compose for Data Lakes with Apache Spark projects.

Enhancements

- Cloudera 6.1 is now supported as a distribution platform for Spark and Hive projects.
- Support added for the following Replicate target endpoints: Hortonworks Data Platform, Amazon EMR, Microsoft Azure HDInsight, and Google Dataproc.
- When defining provisioning tasks (in Spark projects), there is now an option to create a Snapshot or HDS identical to the source tables, without any additional header columns.
- In Hive projects, the Compose Control Tables were renamed to allow the same database to be used for Landing, Storage and Provisioning Zones.
- When managing metadata, you can now select multiple entities for deletion.

Resolved Issues and Customer Requested Enhancements

The following are the resolved issues and customer requested enhancements in this release.

Component/Process	Description	Ref #
Change Processing	With some Hive versions, the partition name would be reset if the CDC task completed without any changes.	195234
Incrementally Updated ODS - Provisioning Task	The task will now create a current view (for each provisioned table) without header columns and with the Primary Key columns in the same position as the metadata.	193921
Platform support	Added support for Hortonworks 3.1.x as a distribution platform for Hive projects. **Note**: For Apache Spark project types, the Hortonworks cluster should be configured to set the Spark metadata store to the Hive metadata store.	191634
Discovery	In some cases, when discovering source tables, tables that were replicated by Attunity Replicate would not be included in the resulting table list. The tables would appear as Views and the number of instances of each view was equal to the number of times the Search button was pressed.	190753
Schema Evolution	In some scenarios, when using schema evolution, the Compose task would fail when writing rows. The following error would be displayed: `org.apache.spark.SparkException: Task failed while writing rows.`	190160
Provisioning	An error would occur when an Attribute name was different to the Attribute Domain Name.	189885

Component/Process	Description	Ref #
Storage	When inserting new records for Change Data tasks, Hive would create a new file in the table for each record, thus creating many small files and causing significant performance degradation.	183306 190120
Discovery	When the columns with the same name appeared in different tables, the attribute names were not preserved during discovery.	187832 161075 188383
Provisioning	When provisioning databases in a Spark project, if the specified database did not exist, the **Test Connection** would succeed without checking the existence and validity of the database.	160838
Table naming	Added support for adding a prefix to tables names in the Provisioning Zone.	CMPS-8112
Platform support	Added support for ADLS GEN2 HDInsight.	CMPS-8090
Project Type	Added support for the Databricks Delta project type. Supported with Databricks 5.3 only.	CMPS-8088
AEM Integration	Metadata integration on a Google platform was not supported by AEM.	CMPS-7561
Provisioning	An error would occur when trying to delete incremental HDS provisioning task data.	CMPS-7494
Model	Attributes with very long names would appear incorrectly in the confirmation UI dialog box.	CMPS-7344
Provisioning	When importing a project, the existing Landing Connection would be overridden.	CMPS-7106
Change Processing	When a Hive project was set up with **Create tables with ACID transactions** and an **Operational Data Store** project type, the Change Processing task would fail if there was a derived attribute that used a renamed column.	CMPS-6726

Component/Process	Description	Ref #
Drop and Create	After dropping and recreating entities, Full Load and Change Processing tasks already in the Change Processing stage would remain in the Storage area.	CMPS-6579
Provisioning Tasks	When saving an incremental task after editing (by clicking **Finish**), a message was displayed telling users to drop and recreate their data. The message would be displayed even if no changes had been made and no data existed (i.e. the task had not yet been run).	CMPS-6550
Provisioning Tasks	Records with null and duplicated values would be added to the storage and provisioning zones. After the fix, any duplicate records inserted into the storage zone will eventually be filtered out in the provisioning zone.	CMPS-6216

Known Issues

The following are the known issues in this release.

Component/Process	Description	Ref #
Provisioning Tasks	If you are using Amazon EMR Hive distribution version 5.20.0 or higher, the value for the `spark.sql.parquet.fs.optimized.committer.optimization-enabled` parameter is set by default to be **True**. Before running a Spark provisioning task, you must configure this parameter's value to be **False** - i.e. `spark.sql.parquet.fs.optimized.committer.optimization-enabled=false`	191640
Metadata	In an Apache Spark project, when deleting an attribute from the Metadata and then adding back the same attribute to the Metadata, the affected tables need to be dropped and recreated.	CMPS-7862
Provisioning Tasks	Mapping from multiple sources is not supported, in both Spark and Hive projects.	CMPS-7770
Provisioning Tasks	In Spark projects, Full Load overwrites previous data, so when running two tasks on the same target, the second task overwrites the first task results. Splitting the Full Load into several tasks also doesn't work for the same reason.	CMPS-7457
Schema Evolution	When a table is added to the source, the new table is added to the task (the mapping is created correctly and associated with the task), and the table is populated with data. However, when opening the mapping, there is an error that the source table does not exist in the database. **Workaround:** Click **Clear Cache** in the **Manage Data Storage Tasks** window (and then open the mapping again).	CMPS-7210

Component/Process	Description	Ref #
Knox	In the **Knox gateway path** field, Compose automatically appends "/hive" to the specified path. As users are unaware of this, they will also specify the path with "/hive" (resulting in "/hive/hive") causing the connection to fail.	CMPS-6626
Provisioning Tasks - Mapping	Two Data Storage Change Processing tasks with different sources that have mapping to the same entity, will result in incorrect data in incremental provisioning to ODS or HDS.	CMPS-6545
Identical task names across Compose projects	Multiple Compose projects with tasks that have the same name are currently not supported in AEM Metadata.	CMPS-6524
Generate	After performing a non-supported change in the metadata, regenerating the task instructions will appear to succeed without errors or warnings, but the task will fail if run later.	CMPS-6437
Hard delete	Hard delete performed on records with expressions, lookup, or derived attributes on Primary Keys does not work.	CMPS-5480
UI	When displaying an entity in the **Physical Metadata** tab and going back to **Logical Metadata** tab, the entity initially appears without any attributes.	CMPS-5421
Column Renaming	Renaming a column in Parquet or Avro format will cause loss of all data in that column.	CMPS-5416
Schema Change	After running a Compose Change Processing task and a schema change occurs, if you run the Compose Full Load task again, the task will fail. **Workaround:** Run the Replicate Full Load again.	CMPS-5394
Replicate Control Tables	If Replicate's **attrep_ddl_history** and **attrep_history** Control Tables are not in the same schema, Compose fails at runtime.	CMPS-5346

Component/Process	Description	Ref #
Discover and Generate	Discover and Generate read all the tables in the database, regardless of the selection. This may take a while if the database contains numerous tables.	CMPS-5332
Derived Attribute	A statement error occurs when changing the name of an attribute that is included in a derived attribute expression.	CMPS-5178
Multiple Landing Zones	Tasks that ingest data from several Landing Zones fail during generation of task instructions. Workaround: Create several tasks - one for each Landing Zone.	CMPS-5159
Validation	Validation of the Data Lake does not detect Compose columns (e.g. FROM_DATE) that have been renamed.	CMPS-4963
Spark Provisioning	When a Spark project is defined with a Microsoft Azure Data Lake Storage Gen1 data store, defining HDFS as a provisioning target is not currently supported.	CMPS-7859
Log File Download	The Compose Server log file cannot currently be downloaded via the UI. As a temporary workaround, the log file can be copied from the following directory: <PRODUCT_DIR>\data\logs	CMPS-7902