How to Run the Big Data Management Utility Update for 10.1

Abstract

Install and run the EBF containing updates to the Big Data Management Utility version 10.1.

Supported Versions

- Informatica Big Data Management 10.1

Table of Contents

Overview

EBF 17557 contains updates to the Big Data Management Utility version 10.1. You can use the Big Data Management Configuration Utility to automate part of the configuration for Big Data Management. This article tells how to download and install the EBF to run the updated utilty.

The Big Data Management Configuration Utility assists with the following tasks:

- Creates configuration files on the machine where the Data Integration Service runs.
- Creates connections between the cluster and the Data Integration Service.
- Updates Data Integration Service properties in preparation for running mappings on the cluster.

After you run the utility, complete the configuration process for Big Data Management.

Note: The utility does not support Big Data Management for the following distributions:

- Highly available BigInsights clusters
- Amazon EMR
- Azure HDInsight
- MapR

To apply the EBF and use the Big Data Management utility, perform the following tasks:

1. Verify prerequisites.
2. Download and install the software.
3. Run the updated Big Data Management utility.

Verify Prerequisites

Verify the following prerequisites before you run the Big Data Management utility

1. Big Data Management 10.1 installed in a Hadoop cluster environment.
2. Administrator privileges on the Big Data Management instance.

Download and Install the Update to the Big Data Management Utility

To use the EBF update to the Big Data Management Configuration Utility, download EBF 17557 and run the clients update file.

1. Download EBF 17557 from the TS FTP site.
2. Uncompress the EBF archive.
3. Uncompress the client update archive file EBF17557_Server_Installer_linux_em64t.tar to a directory.
4. Follow these steps to update the Big Data Management Utility:
   a. Edit the `input.properties` file with the following information:
      - DEST_DIR - destination directory on the on the machine where the Data Integration Service runs.

   b. Type `installEBF.sh` to run the installer.

The installer updates servers and the Big Data Management Configuration Utility.

Run the Big Data Management Utility

The Big Data Management Configuration Utility edits Hadoop cluster properties on the machine where the Data Integration Service runs.

To automate part of the configuration process, perform the following steps:

1. On the machine where the Data Integration Service runs, open the command line.
2. Go to the following directory: `<Informatica installation directory>/tools/BDMUtil`.
3. Run `BDMConfig.sh`.
4. Press `Enter`.
5. Choose the Hadoop distribution that you want to use to configure Big Data Management:

Option	Description
1	Cloudera CDH
2	Hortonworks HDP
3	MapR
4	IBM BigInsights

Note: Select only 1 for Cloudera or 2 for Hortonworks. At this time, the utility does not support configuration for MapR or BigInsights.

6. Based on the option you selected in step 5, see the corresponding topic to continue with the configuration process:
    - “Use Cloudera Manager” on page 4
    - “Use Apache Ambari” on page 6
    - “Use SSH” on page 8

Use Cloudera Manager

If you choose Cloudera Manager, perform the following steps to configure Big Data Management:

1. Select the version of Cloudera CDH to configure.
2. Choose the method to access files on the Hadoop cluster:

    The following options appear:

Option	Description
1	Cloudera Manager. Select this option to use the Cloudera Manager API to access files on the Hadoop cluster.
2	Secure Shell (SSH). Select this option to use SSH to access files on the Hadoop cluster. This option requires SSH connections to the machines that host the name node, JobTracker, and Hive client. If you select this option, Informatica recommends that you use an SSH connection without a password or have sshpass or Expect installed.

    Note: Informatica recommends the Cloudera Manager option.
3. Select whether to use HiveServer 2 to run mappings.

    Note: If you select no, Big Data Management uses the default Hive driver to run mappings.
4. Verify the location of the Informatica Big Data Management installation directory on the cluster.

    The default location appears, along with the following options:

Option	Description
1	OK. Select this option to change the directory location.
2	Continue. Select this option to accept the default directory location.

5. Configure the connection to the Cloudera Manager.
    a. Enter the Cloudera Manager host.
    b. Enter the Cloudera user ID.
    c. Enter the password for the user ID.
    d. Enter the port for Cloudera Manager.
    e. Select whether to use Tez as the execution engine type.
        - 1 - No
        - 2 - Yes

    The Big Data Management Configuration Utility retrieves the required information from the Hadoop cluster.
6. Select whether to exit the utility, or update the Data Integration Service and create connections.

Select from the following options:

Option	Description
1	No. Select this option to exit the utility.
2	Yes. Select this option to continue.

7. Select whether to update Data Integration Service properties.

   Select from the following options:

Option	Description
1	No. Select this option to update Data Integration Service properties later.
2	Yes. Select this option to update Data Integration Service properties now.

8. Choose if you want to create connections to the Hadoop cluster:

Option	Description
1. Hive	Create a Hive connection.
2. HDFS	Create an HDFS connection.
3. Hadoop	Create a Hadoop connection.
4. HBase	Create an HBase connection.
5. Select all	Create all four types of connection.

   Press the number that corresponds to your choice.

9. Supply information about the Informatica domain.
   a. Enter the domain name.
   b. Enter the node name.
   c. Enter the domain user name.
   d. Enter the domain password.
   e. Enter the Data Integration Service name.
   f. If the cluster is Kerberos-enabled, enter the following additional information:
      - Hadoop kerberos service principal name
      - Hadoop kerberos keytab location -- Location of the keytab on the Data Integration Service machine.

   Note: After you enter the Data Integration Service name, the utility tests the domain connection, and then recycles the Data Integration Service.

10. In the **Connection Details** section, provide the connection properties.

    Based on the type of connection you choose to create, the utility requires different properties. For more information about the connection properties, see the *Big Data Management User Guide*.

    Note: When you specify a directory path for the Blaze working directory or the Spark staging directory, you must specify existing directories. The Big Data Management utility does not validate the directory paths that you specify.

The utility creates the connections that you configured.

11. The utility reports a summary of its operations, including whether connection creation succeeded, and the location of utilty log files.
12. Complete the manual configuration steps for Big Data Management.

The utility creates the following files in the `<Informatica installation directory>/tools/BDMUtil` directory:

ClusterConfig.properties

Contains details about the properties fetched from the Hadoop cluster and connection creation commands that can be used to create connections to the Hadoop cluster.

Note: Edit the connection name, domain username and password to use the generated commands.

HiveServer2_EnvInfa.txt

Contains the list of environment variables and values that need to be copied to the HiveServer2 environment on the Hadoop cluster. This file is created only if you choose HiveServer2.

Use Apache Ambari

If you choose Ambari, perform the following steps to configure Big Data Management:

1. Select the Hadoop distribution directory to configure.
2. Choose the method to access files on the Hadoop cluster:

   The following options appear:

Option	Description
1	Apache Ambari. Select this option to use the Ambari REST API to access files on the Hadoop cluster.
2	Secure Shell (SSH). Select this option to use SSH to access files on the Hadoop cluster. This option requires SSH connections to the machines that host the name node, JobTracker, and Hive client. If you select this option, Informatica recommends that you use an SSH connection without a password or have sshpass or Expect installed.

   Note: Informatica recommends the Apache Ambari option.

3. Select whether to use HiveServer 2 to run mappings.

   Note: If you select no, Big Data Management uses the default Hive Command Line Interface to run mappings.

4. Verify the location of the Informatica Big Data Management installation directory on the cluster.

   The default location appears, along with the following options:

Option	Description
1	OK. Select this option to change the directory location.
2	Continue. Select this option to accept the default directory location.

5. Configure the connection to the Ambari Manager.
   a. Enter the Ambari Manager host.
   b. Enter the Ambari user ID.
   c. Enter the password for the user ID.

d. Enter the port for Ambari Manager.

e. Select whether to use Tez as the execution engine type.

- 1 - No
- 2 - Yes

The Big Data Management Configuration Utility retrieves the required information from the Hadoop cluster.

6. Select whether to exit the utility, or update the Data Integration Service and create connections.

Select from the following options:

Option	Description
1	No. Select this option to exit the utility.
2	Yes. Select this option to continue.

7. Select whether to update Data Integration Service properties.

Select from the following options:

Option	Description
1	No. Select this option to update Data Integration Service properties later.
2	Yes. Select this option to update Data Integration Service properties now.

8. Choose if you want to create connections to the Hadoop cluster:

Option	Description
1. Hive	Create a Hive connection.
2. HDFS	Create an HDFS connection.
3. Hadoop	Create a Hadoop connection.
4. HBase	Create an HBase connection.
5. Select all	Create all four types of connection.

Press the number that corresponds to your choice.

9. Supply information about the Informatica domain.

a. Enter the domain name.

b. Enter the node name.

c. Enter the domain user name.

d. Enter the domain password.

e. Enter the Data Integration Service name.

f. If the cluster is Kerberos-enabled, enter the following additional information:

- Hadoop kerberos service principal name
- Hadoop kerberos keytab location -- Location of the keytab on the Data Integration Service machine.

Note: After you enter the Data Integration Service name, the utility tests the domain connection, and then recycles the Data Integration Service.

10. In the **Connection Details** section, provide the connection properties.

    Based on the type of connection you choose to create, the utility requires different properties. For more information about the connection properties, see the *Big Data Management User Guide*.

    Note: When you specify a directory path for the Blaze working directory or the Spark staging directory, you must specify existing directories. The Big Data Management utility does not validate the directory paths that you specify.

    The utility creates the connections that you configured.
11. The utility reports a summary of its operations, including whether connection creation succeeded, and the location of utilty log files.
12. Complete the manual configuration steps for Big Data Management.

The utility creates the following files in the `<Informatica installation directory>/tools/BDMUtil` directory:

ClusterConfig.properties

Contains details about the properties fetched from the Hadoop cluster and connection creation commands that can be used to create connections to the Hadoop cluster.

Note: Edit the connection name, domain username and password to use the generated commands.

HiveServer2_EnvInfa.txt

Contains the list of environment variables and values that need to be copied to the HiveServer2 environment on the Hadoop cluster. This file is created only if you choose HiveServer2.

Use SSH

If you choose SSH, you must provide host names and Hadoop configuration file locations.

Note: Informatica recommends that you use an SSH connection without a password or have sshpass or Expect installed. If you do not use one of these methods, you must enter the password each time the utility downloads a file from the Hadoop cluster.

Verify the following host names: name node, JobTracker, and Hive client. Additionally, verify the locations for the following files on the Hadoop cluster:

- `hdfs-site.xml`
- `core-site.xml`
- `mapred-site.xml`
- `yarn-site.xml`
- `hive-site.xml`

Perform the following steps to configure Big Data Management:

1. Enter the name node host name.
2. Enter the SSH user ID.
3. Enter the password for the SSH user ID.

   If you use an SSH connection without a password, leave this field blank and press enter.
4. Enter the location for the `hdfs-site.xml` file on the Hadoop cluster.
5. Enter the location for the `core-site.xml` file on the Hadoop cluster.

The Big Data Management Configuration Utility connects to the name node and downloads the following files: `hdfs-site.xml` and `core-site.xml`.

6. Enter the Yarn resource manager host name.

   Note: Yarn resource manager was formerly known as JobTracker.

7. Enter the SSH user ID.
8. Enter the password for the SSH user ID.

   If you use an SSH connection without a password, leave this field blank and press enter.

9. Enter the directory for the `mapred-site.xml` file on the Hadoop cluster.
10. Enter the directory for the `yarn-site.xml` file on the Hadoop cluster.

    The utility connects to the JobTracker and downloads the following files: `mapred-site.xml` and `yarn-site.xml`.

11. Enter the Hive client host name.
12. Enter the SSH user ID.
13. Enter the password for the SSH user ID.

    If you use an SSH connection without a password, leave this field blank and press enter.

14. Enter the directory for the `hive-site.xml` file on the Hadoop cluster.

    The utility connects to the Hive client and downloads the following file: `hive-site.xml`.

15. Choose if you want to create connections to the Hadoop cluster:

Option	Description
1. Hadoop	Create a Hadoop connection.
2. Hive	Create a Hive connection.
3. HDFS	Create an HDFS connection.
4. HBase	Create an HBase connection.
5. Select all	Create all four types of connection.

    Press the number that corresponds to your choice.

16. Supply information about the Informatica domain.
    a. Enter the domain name.
    b. Enter the node name.
    c. Enter the domain user name.
    d. Enter the domain password.
    e. Enter the Data Integration Service name.
    f. If the cluster is Kerberos-enabled, enter the following additional information:
       - Hadoop kerberos service principal name
       - Hadoop kerberos keytab location -- Location of the keytab on the Data Integration Service machine.

    Note: After you enter the Data Integration Service name, the utility tests the domain connection, and then recycles the Data Integration Service.

17. In the **Connection Details** section, provide the connection properties.

Based on the type of connection you choose to create, the utility requires different properties. For more information about the connection properties, see the *Big Data Management User Guide*.

Note: When you specify a directory path for the Blaze working directory or the Spark staging directory, you must specify existing directories. The Big Data management utility does not validate the directory paths that you specify.

The utility creates the connections that you configured.

18. The utility reports a summary of its operations, including whether connection creation succeeded, and the location of utilty log files.
19. Complete the manual configuration steps for Big Data Management.

The utility creates the following files in the `<Informatica installation directory>/tools/BDMUtil` directory:

ClusterConfig.properties

Contains details about the properties fetched from the Hadoop cluster and connection creation commands that can be used to create connections to the Hadoop cluster.

Note: Edit the connection name, domain username and password to use the generated commands.

HiveServer2_EnvInfa.txt

Contains the list of environment variables and values that need to be copied to the HiveServer2 environment on the Hadoop cluster. This file is created only if you choose HiveServer2.

Verify Data Integration Service Properties

After you complete configuration of Big Data Management with the Big Data Management utility, the utility automatically restarts the Data Integration Service. Before you proceed with additional configuration tasks, verify that the utility correctly set Data Integration Service properties.

Note: If you configured the utility not to automatically restart the Data Integration Service, restart it before proceeding.

1. Log in to the Administrator tool.
2. In the Domain Navigator, select the Data Integration Service. Verify that the following properties are correct:
   - Hadoop Kerberos service principal name
   - Hadoop Kerberos keytab
   - Informatica installation directory on Hadoop
   - Hadoop distribution directory
   - Data integration service Hadoop distribution directory

Author

Big Data Management Team