Information Builders

iWay Big Data Integrator New Features Bulletin and Release Notes

Version 1.5.2

DN3502232.0717

Chapter 1

iWay Big Data Integrator Version 1.5.2

This document describes new features and provides release notes for iWay Big Data Integrator (iBDI) version 1.5.2. It is intended for all levels of users, including system integrators, application developers, and administrators.

In this chapter:

New Features

This section provides a summary of the new features for iWay Big Data Integrator (iBDI) version 1.5.2.

- **Smaller Cluster Sizes** (BDI-504)

  iBDI can now run in smaller cluster sizes.

- **Non-Pipeline Objects** (BDI-513)

  Non-pipeline objects can now specify host name and port values.

Release Notes

This section provides release notes for iWay Big Data Integrator (iBDI) version 1.5.2 and version 1.5.1.

Important Considerations

This section describes important considerations that you should be aware when using iWay Big Data Integrator (iBDI) version 1.5.2.

JDBC Drivers

JDBC drivers are not shipped with iBDI and must be obtained from a relevant JDBC vendor (for example, Cloudera and Simba). Depending on your usage requirements, there are several open source Hive JDBC drivers available online. For additional security and availability options, commercial drivers are also available. Each driver includes a list of dependencies for the specific driver. Consult the driver provider for the list of .jar files that must be used with the specific driver version.

Data Definition Language (DDL)

Data Definition Language (DDL) generated by iBDI is specific to the JDBC driver being used. DDL generated for one database type (for example, MySQL) cannot be applied in another domain (for example, Hive).

Transformer

In a pipeline, the Transformer is only supported for Hive table objects.

Known Issues in Version 1.5.2

This section describes known issues for iWay Big Data Integrator (iBDI) in version 1.5.2.

Defining Run Configurations for a Flume

An endless loop occurs when using the *iWay Big Data Integrator Deploy* option for a Flume run configuration.

Resolution: Always use the *iWay Big Data Integrator Publish* option when creating run configurations for streaming and event options.

Thread Creation Error

When running the iBDI Client (Eclipse design time framework/workbench) in a memory constrained environment (for example, a vendor *sandbox*), the following error message may appear:

```
cannot create new native thread
```

Do not close the iBDI workbench, but click through the error dialog and continue.

Resolution: Increase the virtual machine memory limits.

Hive Connection Object

Entries cannot be edited directly when using a Hive connection object.

Resolution: Create another Hive target entry and access the profile through the new entry.

XML and JSON Deserializers

The current implementation requires flat input only.

Expression Field or a Constant Field in a Transform or Map

An expression field or a constant field in a transform or map cannot be renamed, even though the field appears to be renamed in the user interface.

Pipelines

The known issues described in this section apply to pipelines in iBDI version 1.5.2.

Pipeline Transformer Tool: Define Mappings Dialog (BDI-554)

The Define Mappings dialog currently allows the following incorrect behavior to occur:

- Right-clicking and selecting *Add Column* from the context menu.

  A NULL column is created in the Hive HQL, which invalidates the transform. Do not add these columns.

- Clicking on a column to rename its current label/title.

  Although you may specify a new name for a column using this action, the renamed value for the column is ignored.

Compute Type: Save

The pipeline Compute type *Save* generates a *class not found* exception.

Resolution: If you must use *Save* as a Compute type, then compose a Hive Query Language Statement using syntax similar to the following:

```
CREATE TABLE [IF NOT EXISTS] [db_name.]table_name
   LIKE existing_table_
      [LOCATION hdfs_path];
```

The complete Hive Data Definition Language (DDL) syntax can be found in the *Apache Hive Language/DDL* documentation.

Compute Type: Coalesce

The pipeline Compute type *Coalesce* can only be used when the number of partitions is greater than the value entered for Coalesce. For example, if you have 20 partitions, and enter a value of 30 for Coalesce, then the Spark runtime will generate an exception.

Hive Source to Hive Compute (BDI-553)

A pipeline using a Hive source to Hive compute generates a null pointer exception (*Failed to create the part's controls*).

Resolution: Do not use back to back transforms in a pipeline. Use Hive and an HQL object.

Data Serialization in Pipeline Components

Data format types in a pipeline source do not require a schema if they are of type *text*. Otherwise, a schema is required to deserialize the source. If a CSV file is used where the first line is a header describing the fields, then the *infer schema* can be used. However, the preference is still to have an Avro schema describing the structure, fields, and type of data.

Some pipeline components require deserialization from specific formats (for example, pipeline source of type *HDFS*, *Kafka*, *NiFi*, or *Stream*) or to specific formats, known as serialization (for example, a pipeline compute type of *Save* or a pipeline target type of *HDFS*). These components provide a drop-down list allowing you to select the data format type and an edit field that allows you to specify a schema name.

Source requirements: If the data format type is a streaming type (for example, Kafka, Kafka Direct, NiFi, or Stream), then a schema must be supplied in Avro format describing the stream. The schema should have a flat structure without any nested objects.

Kafka Topics

Kafka topics must be pre-configured for use as pipeline target producers.

Pipeline Target for HDFS Output

Do not select *CSV* as an option when specifying a pipeline target for HDFS output. Select the *HDFS CSV* option to write CSV to HDFS.

Pipeline Target Data File Formats

Avro will be saved as *Avro*, and *Orc* will be saved as *orc*. By default, all other format types will be saved as Parquet files. Use *HDFS CSV* to save to HDFS CSV.

Resolved Issues in Version 1.5.2

This section describes resolved issues for iWay Big Data Integrator (iBDI) in version 1.5.2.

Sending Files to the Edge Node Host Generates an Error Message on CentOS or Ubuntu (BDI-552)

An error in the Secure Copy command when a timestamp is created generates an error message in the Bash shell of the host (*mtime.sec not present*) when using the following operating systems:

- Ubuntu version 16 and higher
- CentOS version 7 and higher

Resolution: If you are using an earlier version of iBDI (1.5.1 or 1.5.0), then you must install iBDI version 1.5.2 Client (Eclipse design time framework/workbench) and provision a server edge node with iBDI version 1.5.2.

Known Issues in Version 1.5.1

This section describes known issues for iWay Big Data Integrator (iBDI) in version 1.5.1.

Issue Number	Description	Resolution
BDI-528	Deploy DDL may not create tables for some database types.	Database syntax is driver dependent, and the generic syntax created by the Eclipse data tools may need editing before submission.
BDI-532	Sqoop failure to create field when type is *tinyint*.	For more information, see section 27.2.5 in the *Sqoop User's Guide*.
BDI-535	Transformer functions may lack descriptions.	For more information, click *here* to view related information on functions.
BDI-536	Nifi missing .jar files exception.	Nifi use is considered experimental: Advanced deployment necessary, see add on document for Nifi.
BDI-539	Hive Source in Pipeline generates a *No Such Method* error.	Do not copy JDBC drivers to the JDBC folder of the iBDI Edge node location.

Issue Number	Description	Resolution
BDI-540	Data Source Explorer - Hive JDBC cannot edit driver after creation.	Create a new driver as a copy of the existing driver and modify the properties.
BDI-541	Pipeline Hive to HDFS *noSuchElementException:key not found:header*.	Do not use the *HDFS with CSV* selection option as the pipeline target. Use *HDFS CSV* as the pipeline target. CSV file format as source in pipeline check the field *header* as *on* or *off* depending on the presence of a header line in the source file.
BDI-542	Pipeline Rename Null Pointer Exception in run configurations.	When renaming a pipeline, any run configurations for that pipeline must be removed before renaming. Otherwise, when opening the Run Configurations dialog, a *Null Pointer Exception* will be generated for trying to read a non-existing pipeline.
BDI-543	DDL Create Table fails to create table under Hive database Schema.	Do not use upper case characters and quoted identifiers. For more information, see the Create Table topic in the *Apache Hive DDL Language Manual*.
BDI-544	DDL - Create Schema does not create database schema for mySQL.	The JDBC driver for mySQL is database dependent, so the database schema must be created in the mySQL workbench and not in the DDL execution engine in iBDI.
BDI-545	Pipeline Hive Transformer.	Rename target column fails to rename column. As a workaround, use the Big Data Mapper to rename table columns.

Issue Number	Description	Resolution
BDI-547	Flume used with Run Configuration Deploy option results in an endless loop.	Always use *Publish* for a Flume.

The following table lists and describes known issues that are related to **Virtual Machines** when using iWay Big Data Integrator (iBDI) version 1.5.1.

Issue Number	Description	Resolution
BDI-534	CDH Hadoop Explorer - NullPointer Exception cannot create thread.	Do not close the workbench, click through dialog boxes, thread model related.
BDI-537	CDH Sqoop import cannot load Postgres driver in Sqoop.	Postgres is not installed on CDH.
BDI-538	CDH Sqoop Export IoException when destination database is not on the Virtual Machine (VM).	CDH limits outbound connections. Do not use this option.
BDI-523	HDW JDBC to Hive cannot open connection after close and reopen Workbench.	This is related to the JDBC driver.
BDI-524	HDW Copying JDBC drivers results in no class definition found.	Do not copy JDBC drivers to the JDBC folder of the iBDI Edge node location.

Note: In this table:

- CDH - Cloudera
- HDW - HortonWorks

Run Configurations

The following types of run configurations are available:

- iWay Big Data Deploy
- iWay Big Data Pipeline
- IWay Big Data Publish

The *Pipeline* run configuration is for all pipelines. You can set the *Deploy* or *Publish* option for each pipeline inside the configuration.

The non-Pipeline objects have explicit *Deploy* or *Publish* configurations.

The *Deploy* run configuration uploads the objects to the Edge node and runs the configuration immediately, returning the results to the console in the iBDI workbench. Use *Deploy* when running simple objects with smaller amounts of data that can be run in the iBDI console

The *Publish* run configuration uploads the objects to the Edge node. A user logs in to the Edge node via SSH or another tool, navigates to the folder where the configuration is located, and runs the shell script that triggers the job. The *Publish* run configuration is mandatory for Flume, Kafka, Kafka Direct, and Stream. It is recommended when using HDFS with large results.

Warning: Do not use the *Deploy* run configuration with Flume. For more information, see BDI 547.

Using iWay Big Data Integrator on Vendor Provided Sandbox Virtual Machines

iWay Big Data Integrator (iBDI) version 1.5.2 supports the following versions of common Hadoop software:

- Sqoop version 1.4.6
- Flume version 1.6.0
- Spark version 1.6.0
- Kafka version 2.10 - 8.2.1

iBDI Client requires Java version 8 to support the Pipeline and Transformer features.

iBDI is designed to run on Hadoop Clusters, not single node machines. However, running in a small Virtual Machine (VM) footprint is possible by adhering to the following guidelines:

Please check the terms of service (TOS) of the VM vendor to see if installing third-party software is permitted. Most do not permit this without explicit permission. Installing other software or modifying the VM can violate the terms with the vendor. Check before you install.

The sandbox VMs (Cloudera, HortonWorks) have statically assigned host names and will re-shuffle assigned IP addresses to match their internal configuration. If these are tampered with, then the sandbox will usually not run. Check the vendor of the VM hostware (Oracle VirtualBox, VMware, or other) on their networking rules. The usual network setup for a sandbox VM will have one network port as a Network Address Translation (NAT) port, and you define a second port as *host only*. The iBDI communication between the local computer and the VM is through the *host only* port. The VMs usually have restrictions on outbound communication, so these modes should be sufficient.

Assign sufficient memory to the VM. 8GB of memory is the smallest configuration possible under most circumstances. However, 10GB, 12GB, or higher provides better results.

Cloudera Sandbox does not have Java version 8 pre-installed. If you intend to install iBDI on the Cloudera VM, then Java version 8 must be installed. Cloudera Sandbox does not have Kafka pre-installed. If you intend to test Kafka, then you will need to install Kafka. The default implementation of Spark does not have a Hive configuration, and the Spark memory defaults must be adjusted. Install iBDI on the image and run the product from the image. Doing so may void your Cloudera terms of service.

HortonWorks Sandbox runs inside a Docker image. The usual SSH port number 22 opens the Docker image, and port 2222 opens the actual sandbox. Before using SSH, you must set the password through the Hortonworks console or web console. If you are using a database such as mySQL, then you must allow Docker to open the default mySQL port number 3306 (or other database port numbers). There may also be permission issues if using mySQL or other databases on the local machine. Changing these options may void your HortonWorks terms of service. Install iBDI on your local machine and communicate with Hortonworks through the network address and IP port.

MapR

MapR is not supported with iBDI at this time. This is currently under research.

Feedback

Customer success is our top priority. Connect with us today!

Information Builders Technical Content Management team is comprised of many talented individuals who work together to design and deliver quality technical documentation products. Your feedback supports our ongoing efforts!

You can also preview new innovations to get an early look at new content products and services. Your participation helps us create great experiences for every customer.

To send us feedback or make a connection, contact Sarah Buccellato, Technical Editor, Technical Content Management at *Sarah_Buccellato@ibi.com.*

To request permission to repurpose copyrighted material, please contact Frances Gambino, Vice President, Technical Content Management at *Frances_Gambino@ibi.com.*

Information Builders

iWay Big Data Integrator New Features Bulletin and Release Notes

Version 1.5.2

Information Builders
Two Penn Plaza
New York, NY 10121-2898

Printed on recycled paper in the U.S.A.