How to Guide

Data Readers (Data Pipeline)

Version: Release 1.1

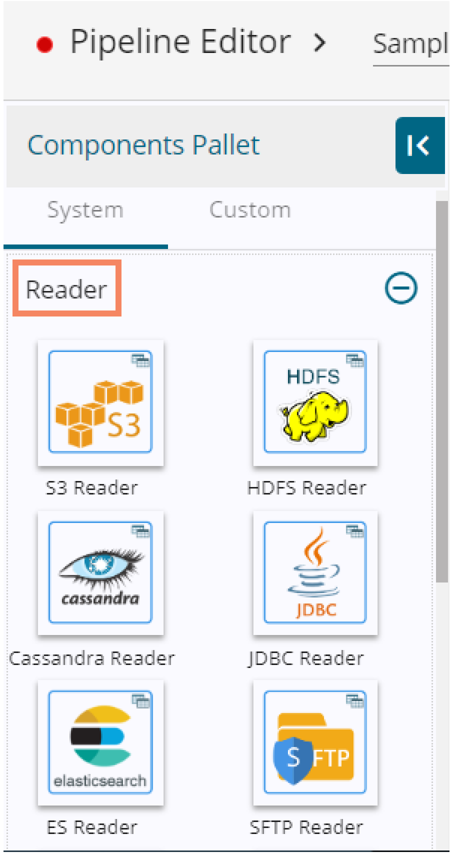
Contents

1. Components Pallet

The Components Pallet of the BDB Data Pipeline contains data readers, data writers, Transformations, ML, Ingestion, and WebSocket to create various pipeline workflows based on the user's need.

1.1 Reader

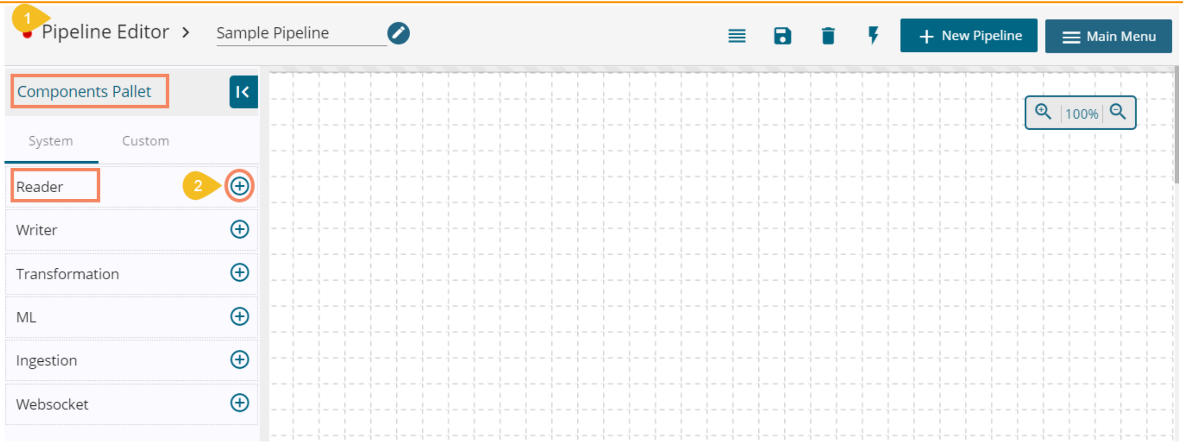
Data Reader components help to read data from different sources and ingest that data into the Pipeline for further processing.
BDB Data Pipeline contains the following Data Readers under the Component Pallet



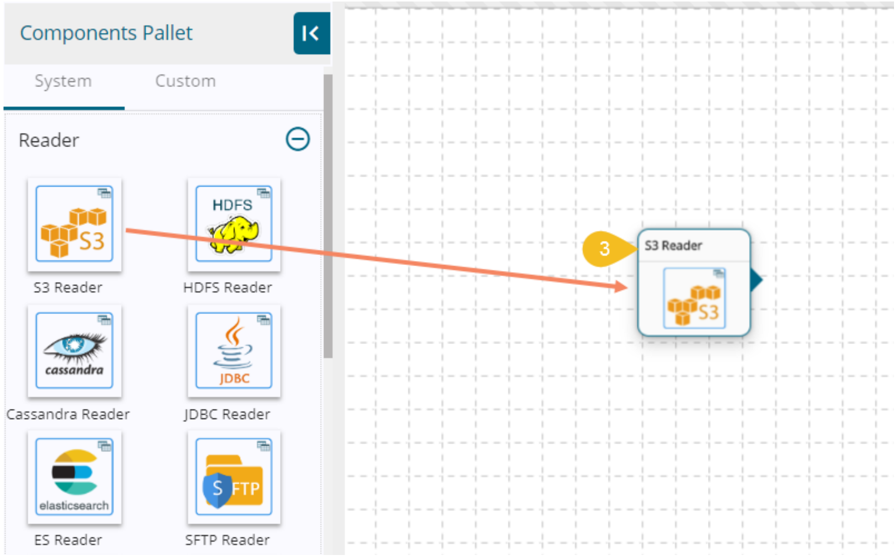
The document aims to describe all the readers with configuration details.

1.1.1 S3 Reader

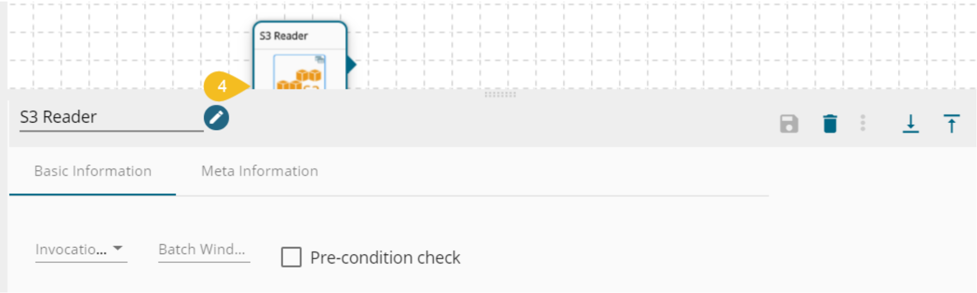
1. Navigate to the Data Pipeline Editor
2. Expand the Reader section provided under the Component Pallet



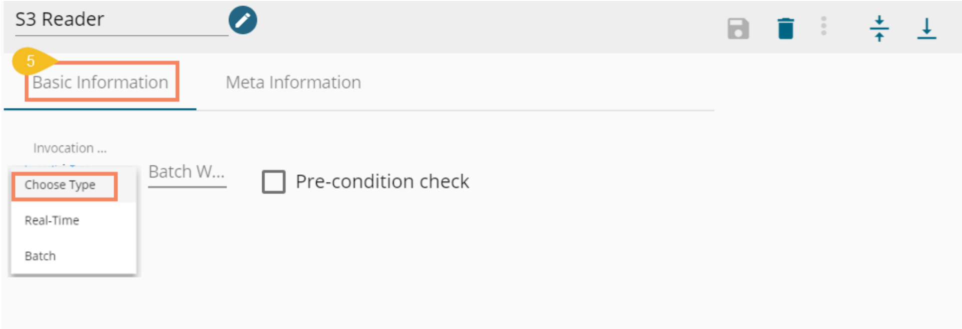
3. Drag and drop the S3 Reader component to the Workflow-editor



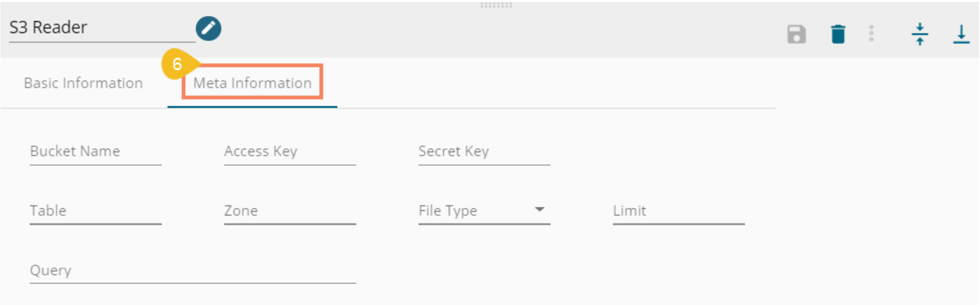
4. Click on the dragged S3 Reader to get the Configuration panel



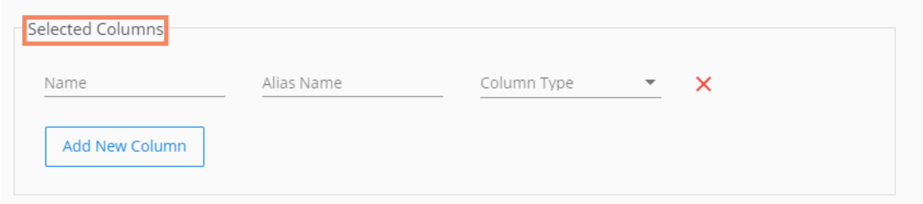
5. S3 Reader requires some configuration for reading data from S3. There are 2 sections in Information panel:
   a. The first section is for **Basic Information**, in which the user needs to select the running mode for component out of ‘Batch’ or ‘Real-Time’)



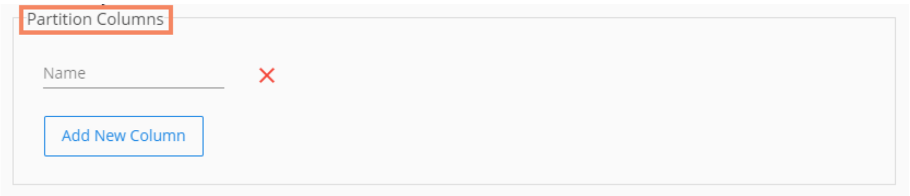
   b. The second section is for Meta Information.
      i. Configure the following fields:
         1. Bucket Name(*): Folder Name
         2. Access Key(*): Access key shared by AWS to login
         3. Secret Key(*): Secret key shared by AWS to login
         4. Table(*): Table or object name which has to be read
         5. Zone(*): S3 Zone location
         6. File Type(*): Select a file type from the drop-down menu (CSV, JSON, PARQUET, AVRO are the supported file types)
         7. Limit: Set limit for the number of records
         8. Query: Insert an SQL query (it takes query containing a Join statement as well)



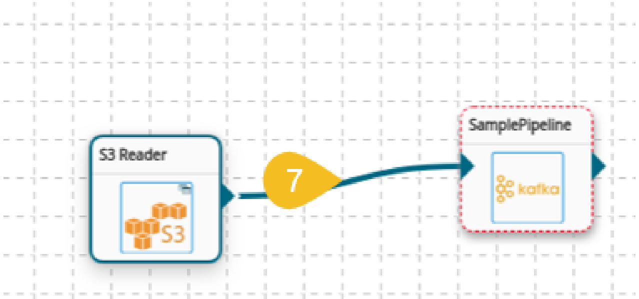
      ii. There is also a section for the selected columns in the Meta Information tab if the user can select some specific columns from the table to read data instead of selecting a complete table so this can be achieved by using the ‘Selected Columns’ section. Select the columns which you want to read and if you want to change the name of the column, then put that name in alias name section otherwise **keep alias name same as of column name** and then select a Column Type from the drop-down menu.



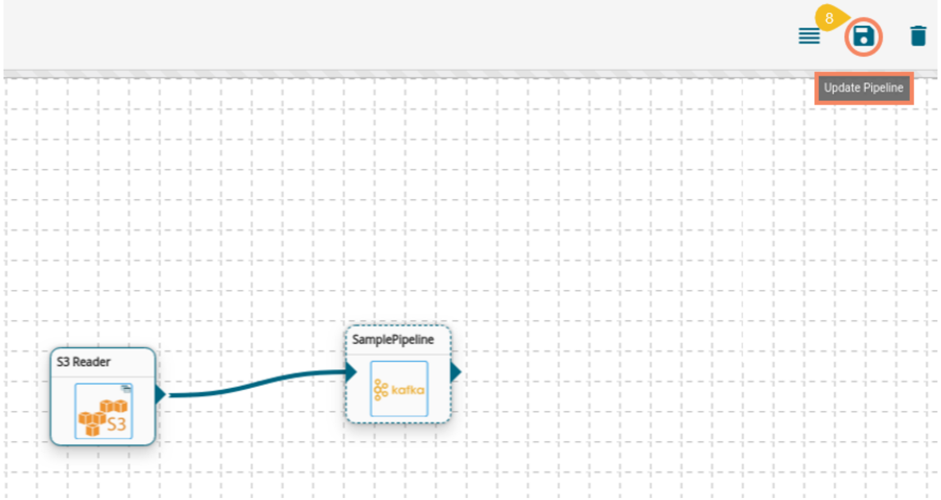
iii. Partition Column- Provide a unique Key column name to partition data in Spark.



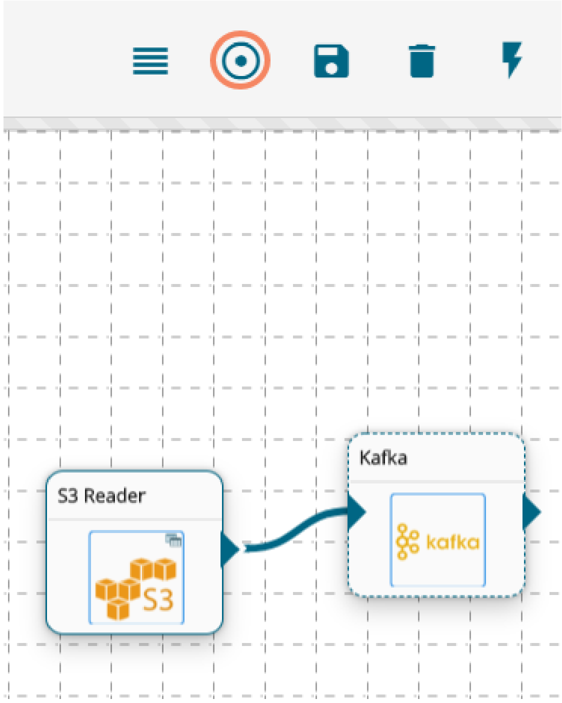
6. After doing all the configurations click the 'Save' icon provide in the reader configuration panel to save the component.
7. Connect the S3 output node to an event to send data after reading it from the S3 reader.



8. Click the 'Update Pipeline' icon on the Pipeline Editor.



9. A success message appears to confirm the pipeline update.
10. An icon to Activate the Pipeline appears on the header panel.

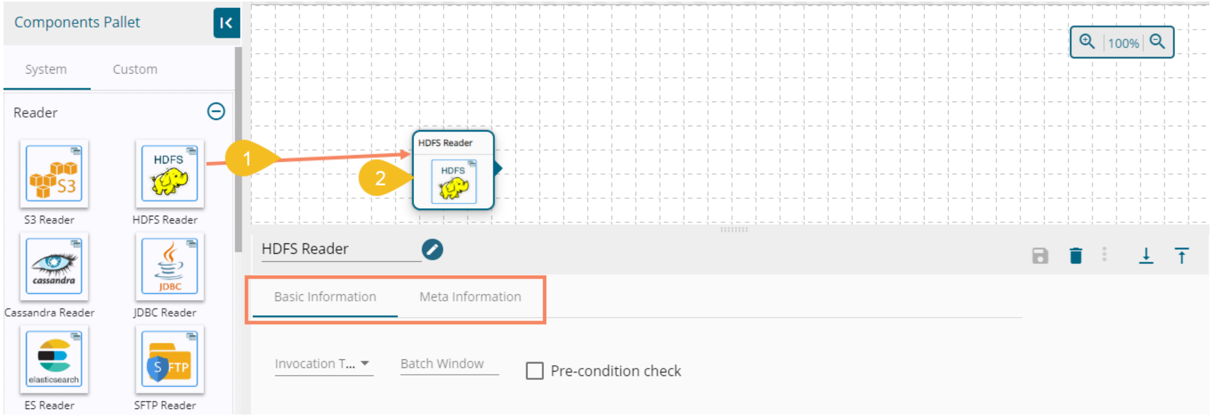


Note:

a. (*) symbol indicates that the field is mandatory.
b. Either table or query must be specified for the data readers except for SFTP Reader.
c. Selected Columns- There should not be data type mismatch in the Column Type for all the Reader components.

1.1.2 HDFS Reader

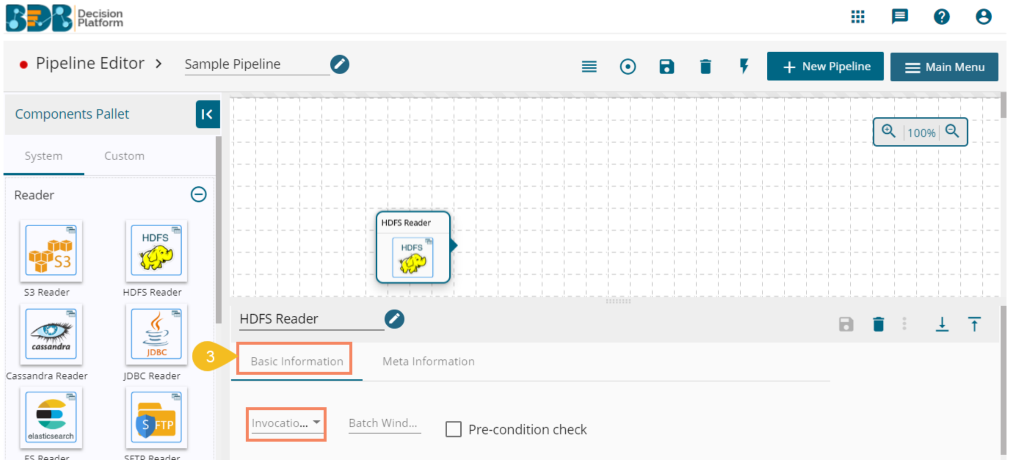
1. Drag & Drop the HDFS Reader component on the Work-flow editor.
2. Click on the dragged reader component to open the configuration tabs below.



3. Configure the following information for an HDFS Reader:

   a. Basic Information

      Select an Invocation type from the drop-down menu to confirm the running mode of the reader component (Currently, 'Real-time' option is supported)

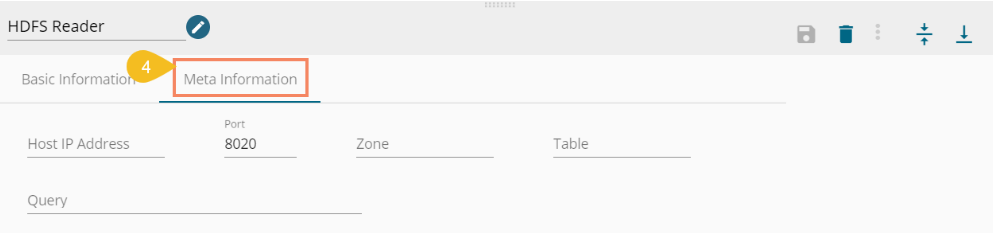


   b. Meta Information

      Fill all the connection specific details of HDFS, and there is a section of Selected Columns in Metadata.

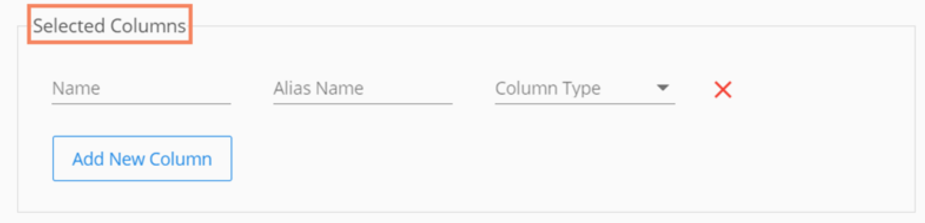
      i. Configure the following fields:

         1. Host IP Address(*): Hadoop IP address of the host
         2. Port(*): Port number
         3. Zone(*): HDFS Zone location
         4. Table(*): Table or object name which must be read
         5. Query: Insert an SQL query (it takes query containing a Join statement as well)

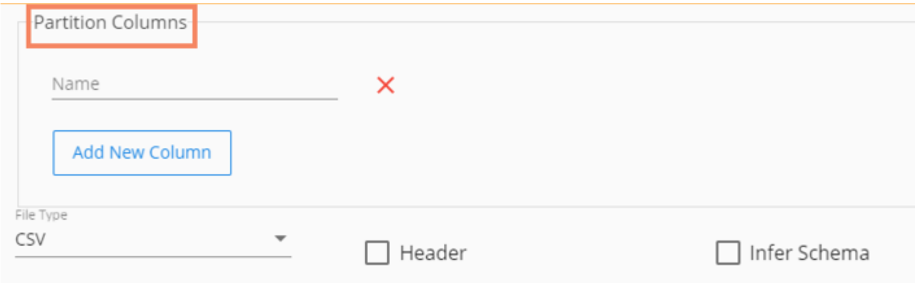


Port: 8020

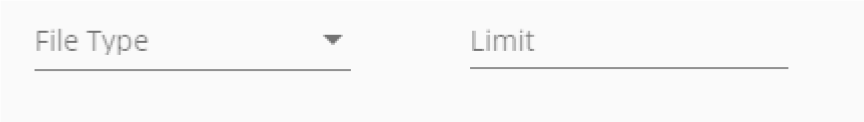
      ii. The users can select some specific columns from the table to read data instead of selecting a complete table; this can be achieved via the 'Selected Columns' section. Select the columns which you want to read and if you want to change the name of the column, then put that name in alias name section otherwise **keep alias name same as of column name** and then select a Column Type from the drop-down menu.



iii. Partition Column- Provide a unique Key column name to partition data in Spark.

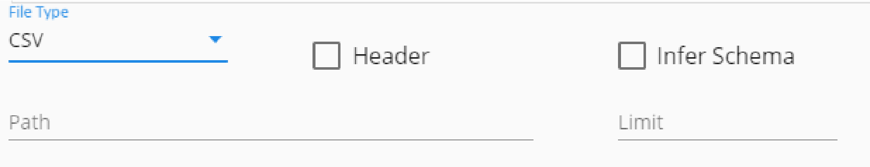


iv. File Type
The users get the File Type and Limit fields at the end. The File Type selection is a mandatory field for the HDFS reader.



Currently, we support CSV, JSON, PARQUET, AVRO file types. The configuration fields may vary based on the selection of the file type.

1. CSV
   Configure the following fields when the selected file type is CSV:
   a. Header- Enable or Disable the Header option to confirm whether the stored data has header or not
   b. Infer Schema- Enable or disable the Infer Schema option to confirm whether the schema is provided or not
   c. Path- Provide a specific path for the file
   d. Limit- Set limit for the number of rows



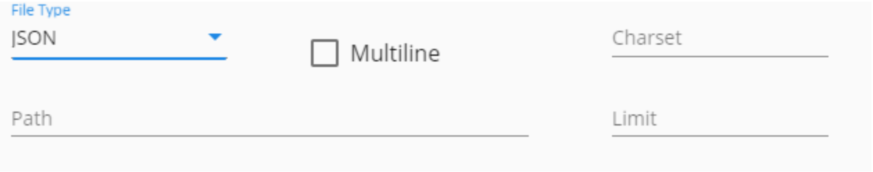
2. JSON
   Configure the following fields when the selected file type is JSON:

a. Multiline-Enable this option, if the stored JSON has line breaks in-between

b. Charset-Character set encoding, E.g., UTF-8

c. Path-provide a specific path for the file

d. Limit-set limit for the number of rows



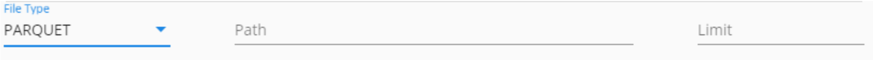
File Type: JSON

3. PARQUET

   Configure the following fields when the selected file type is PARQUET:

   a. Path-provide a specific path for the file

   b. Limit- set limit for the number of rows



File Type: PARQUET

4. AVRO

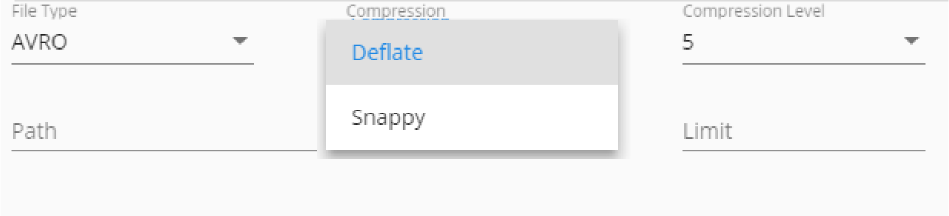
   Configure the following fields when the selected file type is AVRO:

   a. Compression: Select a compression option from the drop-down menu out of ‘Deflate’ and ‘Snappy’ options.

   b. Compression Level- If the selected Compression option is ‘Deflate’ then select compression level from 0-9 (where 0 and 9 are included)

   c. Path-provide the specific path for the file

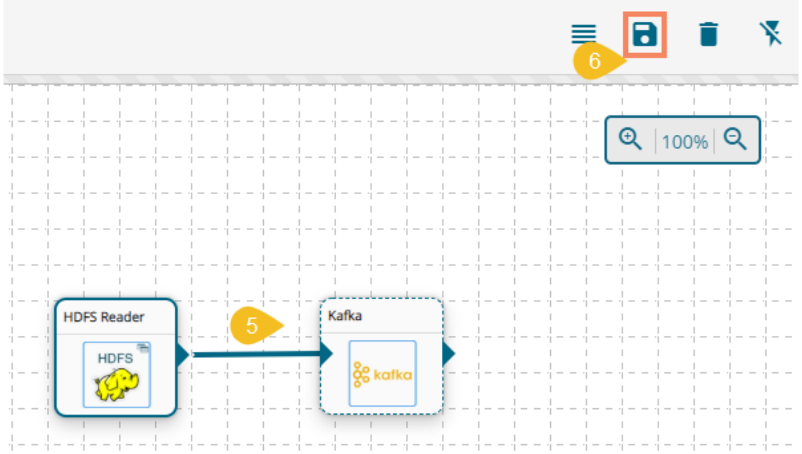
   d. Limit- set limit for the number of rows



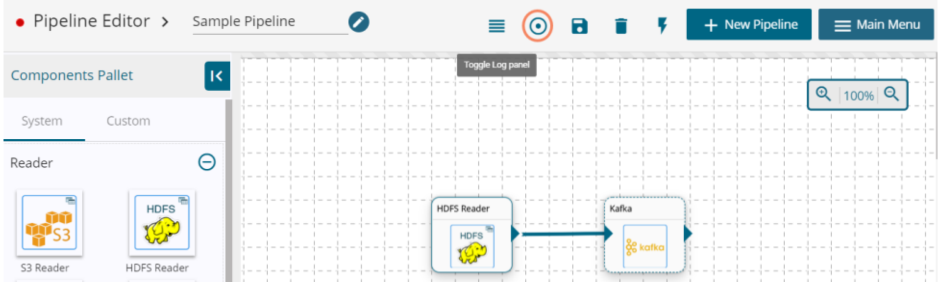
File Type: AVRO

Compression Level: 5

4. After doing all the configurations click the ‘Save’ icon provide in the reader configuration panel to save the component.
5. Connect the HDFS reader’s output node to an event to send data to this event.
6. Click the ‘Update Pipeline’ icon on the Pipeline Editor.

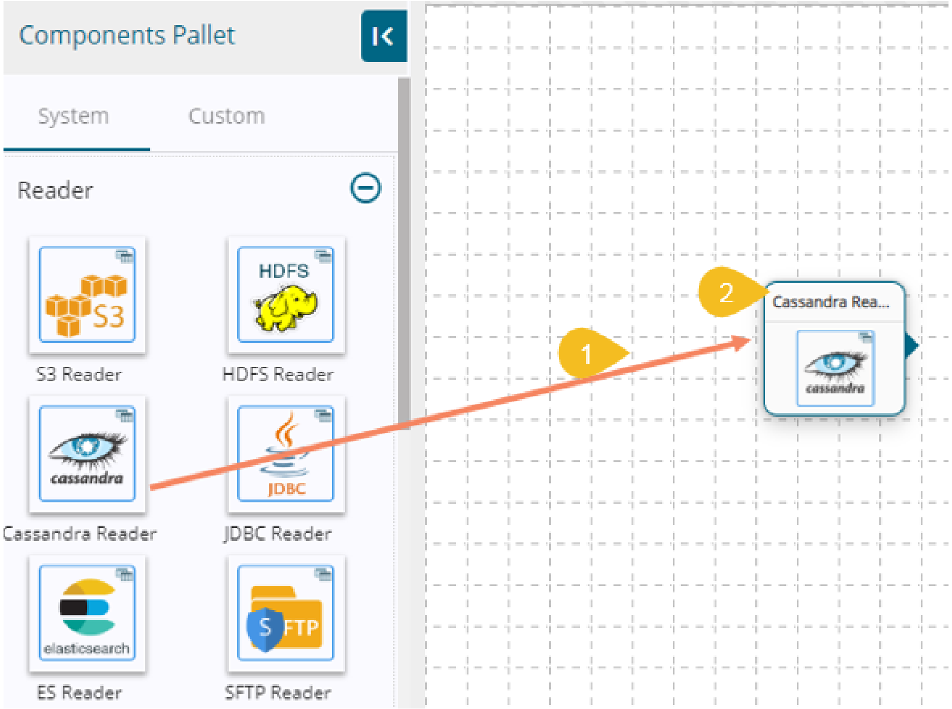


7. A success message appears to confirm the pipeline update.
8. An icon to Activate the Pipeline appears on the header panel.



1.1.3 Cassandra Reader

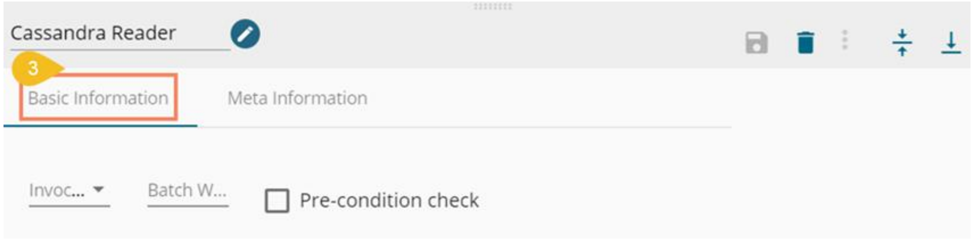
1. Drag & Drop the Cassandra Reader component on the Work-flow editor
2. Click on the dragged reader component to open the configuration tabs below.



3. Configure the following information for an HDFS Reader:

   a. Basic Information

      Select an Invocation type from the drop-down menu to confirm the running mode of the reader component (Currently, 'Real-time' option is supported)

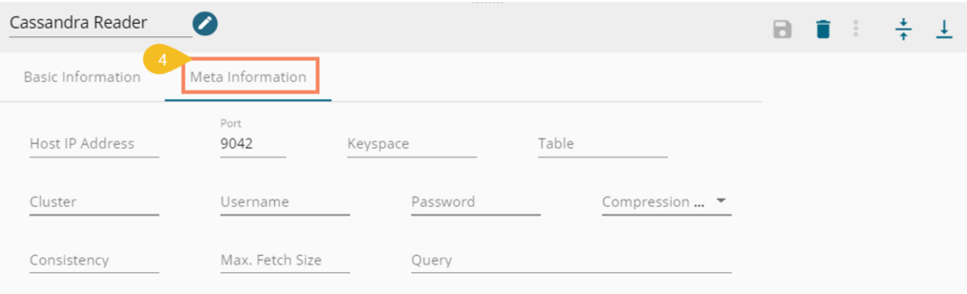


   b. Meta Information

      Fill all the connection specific details of Cassandra reader, and there is a section of Selected Columns in Metadata.

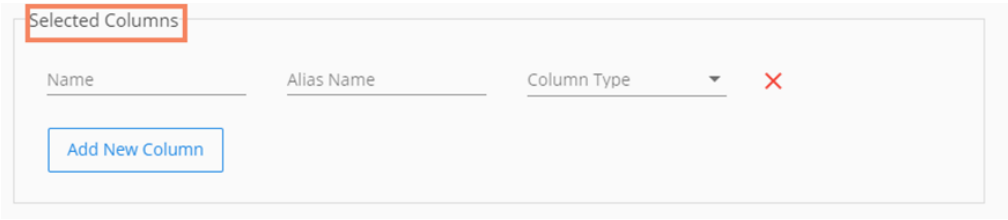
      i. Configure the following fields:

         1. Host IP Address (*): IP Address
         2. Port(*): Host server port number
         3. Keyspace(*): Key for table
         4. Table(*): Table Name to read data
         5. Cluster: Name of the cluster
         6. Username(*): Provide username for login
         7. Password(*): provide a valid password for login
         8. Compression Method: Select a compression method from the drop-down
         9. Consistency: the minimum number of Cassandra nodes that must acknowledge the operation
         10. Max. Fetch Size: Max number of records per batch
         11. Query: Insert an SQL query (it takes query containing a Join statement as well)



Port: 9042

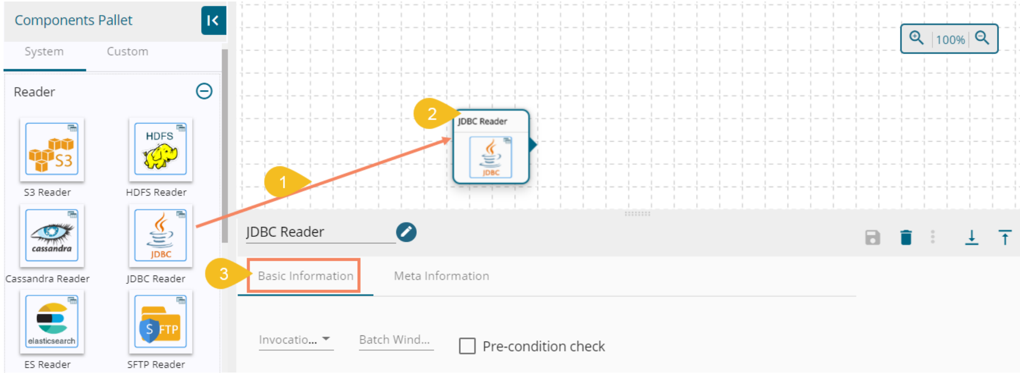
      ii. The users can select some specific columns from the table to read data instead of selecting a complete table; this can be achieved via the 'Selected Columns' section. Select the columns which you want to read and if you want to change the name of the column, then put that name in alias name section otherwise **keep alias name same as of column name** and then select a Column Type from the drop-down menu.



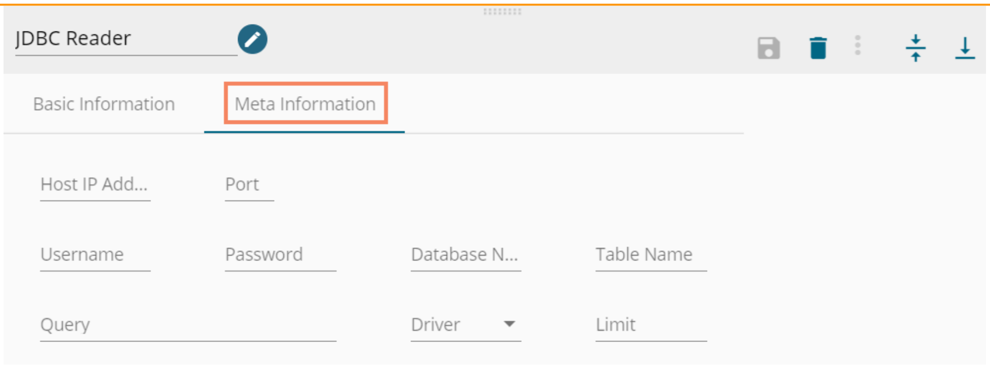
4. After filling the Meta-data, Save the reader Component. Now You can update the Pipeline. As you Update the Pipeline, an icon appears to Activate the Pipeline.

1.1.4 JDBC Reader

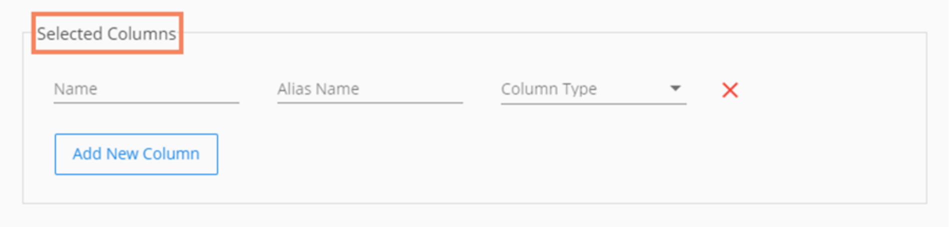
1. Drag & Drop the JDBC Reader component on the Work-flow editor
2. Click on the dragged reader component to open the configuration tabs below
3. Configure the following information for a JDBC Reader:
   a. Basic Information
      Select an Invocation type from the drop-down menu to confirm the running mode of the reader component (Currently, 'Real-time' option is supported)



   b. Meta Information
      Fill all the connection specific details of Cassandra reader, and there is a section of Selected Columns in Metadata.
      i. Configure the following fields:
         1. Host IP Address (*): IP Address
         2. Port(*): Host server port number
         3. Username(*): Username for login
         4. Password(*): Password for login
         5. Database Name(*): Provide the Database name
         6. Table Name: Provide the table name to read data
         7. Query: Insert an SQL query (it takes query containing a Join statement as well)
         8. Driver(*): Select Database type (MYSQL, MSSQL, Oracle, Postgres)
         9. Limit: Set limit for the number of records



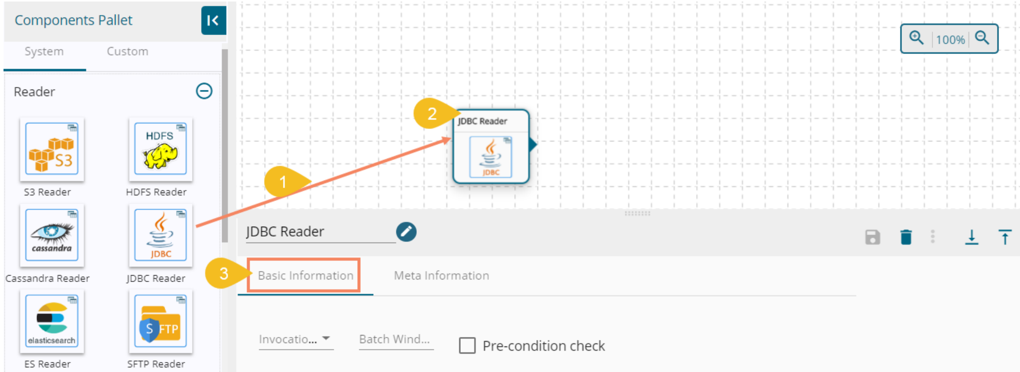
ii. The users can select some specific columns from the table to read data instead of selecting a complete table; this can be achieved via the ‘Selected Columns’ section. Select the columns which you want to read and if you want to change the name of the column, then put that name in alias name section otherwise **keep alias name same as of column name** and then select a Column Type from the drop-down menu.



4. After filling the Meta-data, Save the reader Component. Now You can update the Pipeline. As you Update the Pipeline, an icon appears to Activate the Pipeline.

1.1.5 ES Reader

1. Drag & drop the ES Reader component on the Work-flow editor
2. Click on the dragged reader component to open the configuration tabs below
3. Configure the following information for an ES Reader:
   a. Basic Information
   Select an Invocation type from the drop-down menu to confirm the running mode of the reader component (Currently, ‘Real-time’ option is supported)

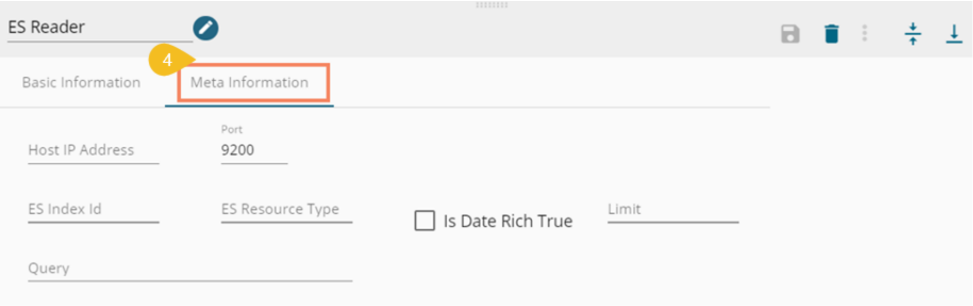


b. Meta Information

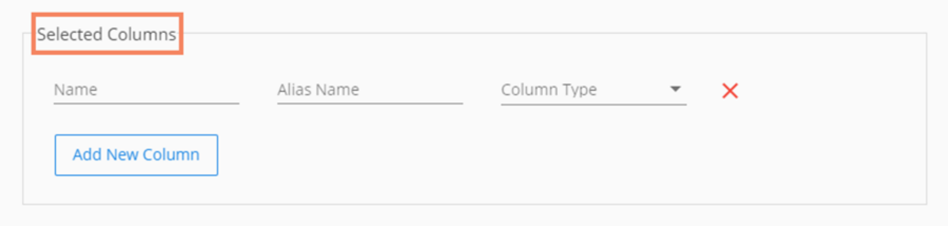
Fill all the connection specific details of Cassandra reader, and there is a section of Selected Columns in Metadata.

i. Configure the following fields:

1. Host IP Address (*): Database Host Address
2. Port(*): Database Host port
3. ES Index Id(*): Id of the elastic search index
4. ES Resource Type: Type of the elastic search index (can be given same as Id)
5. Is Data Rich True: Required while reading data from indexes having dates
6. Limit: Set limit for the number of records



ii. The users can select some specific columns from the table to read data instead of selecting a complete table; this can be achieved via the ‘Selected Columns’ section. Select the columns which you want to read and if you want to change the name of the column, then put that name in alias name section otherwise **keep alias name same as of column name** and then select a Column Type from the drop-down menu.

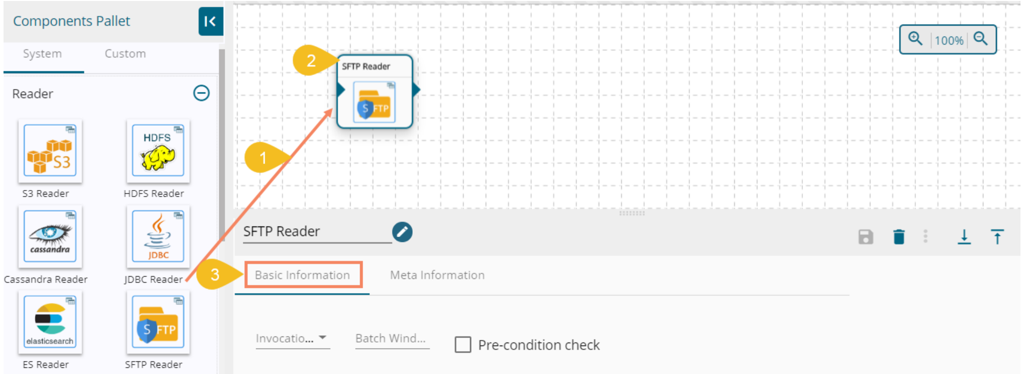


4. After filling the Meta-data, Save the reader Component. Now You can update the Pipeline. As you Update the Pipeline, an icon appears to Activate the Pipeline.

1.1.6 SFTP Reader

1. Drag & Drop SFTP Reader component on the Work-flow editor.
2. Click on the dragged reader component to open the configuration tabs below
3. Configure the following information for an SFTP Reader:
   a. Basic Information

Select an Invocation type from the drop-down menu to confirm the running mode of the reader component (Currently, 'Real-time' option is supported)



b. Meta Information

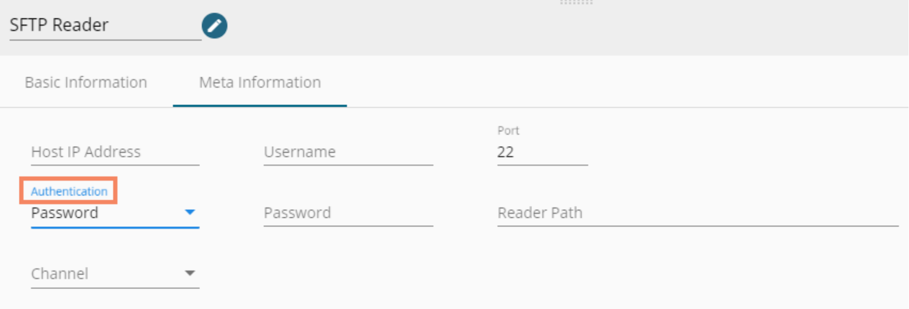
Fill all the connection specific details of Cassandra reader, and there is a section of Selected Columns in Metadata.

i. Configure the following fields:

1. Host IP Address (*): IP address of SFTP location
2. Username(*): Username of SFTP location
3. Port(*): Port of SFTP server (default:22)
4. Authentication(*): Select an authentication option for login to SFTP server using the drop-down menu

   a. Password: configure the following fields that appear after selecting the 'Password' option as authentication

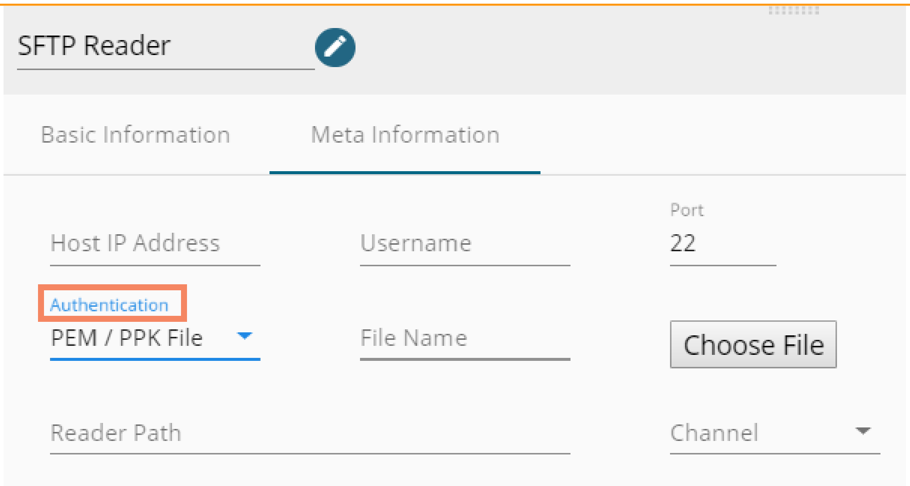
      i. Password: Provide a valid password to login



   b. PEM/PPK (file): Select PEM/PPK file database host port for login to SFTP server

      i. File Name: Provide a file name

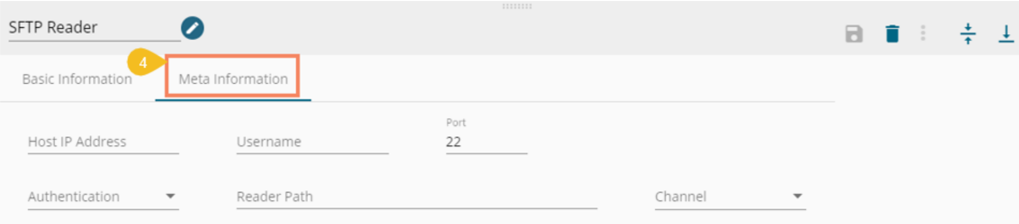
      ii. Choose File: select a PEM/PPK file using this option



Port: 22

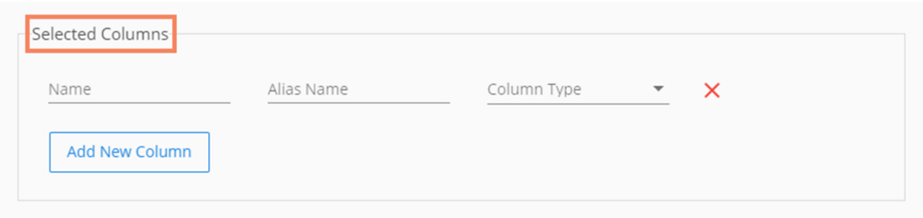
Authentication: PEM / PPK File

5. Reader Path(*) – complete path (absolute path) for the folder, from where new files have to be read
6. Channel(*) – Currently supporting SFTP



Port: 22

ii. The users can select some specific columns from the table to read data instead of selecting a complete table; this can be achieved via the 'Selected Columns' section. Select the columns which you want to read and if you want to change the name of the column, then put that name in alias name section otherwise **keep alias name same as of column name** and then select a Column Type from the drop-down menu.



4. After filling the Meta-data, Save the reader Component. Now, you can update the Pipeline. As you Update the Pipeline, an icon appears to Activate the Pipeline.