



YOTABITES

Big Data Summit KC 17

StreamSets WORKSHOP

**Yotabites Consulting
LLC**

**888-441-
8629**

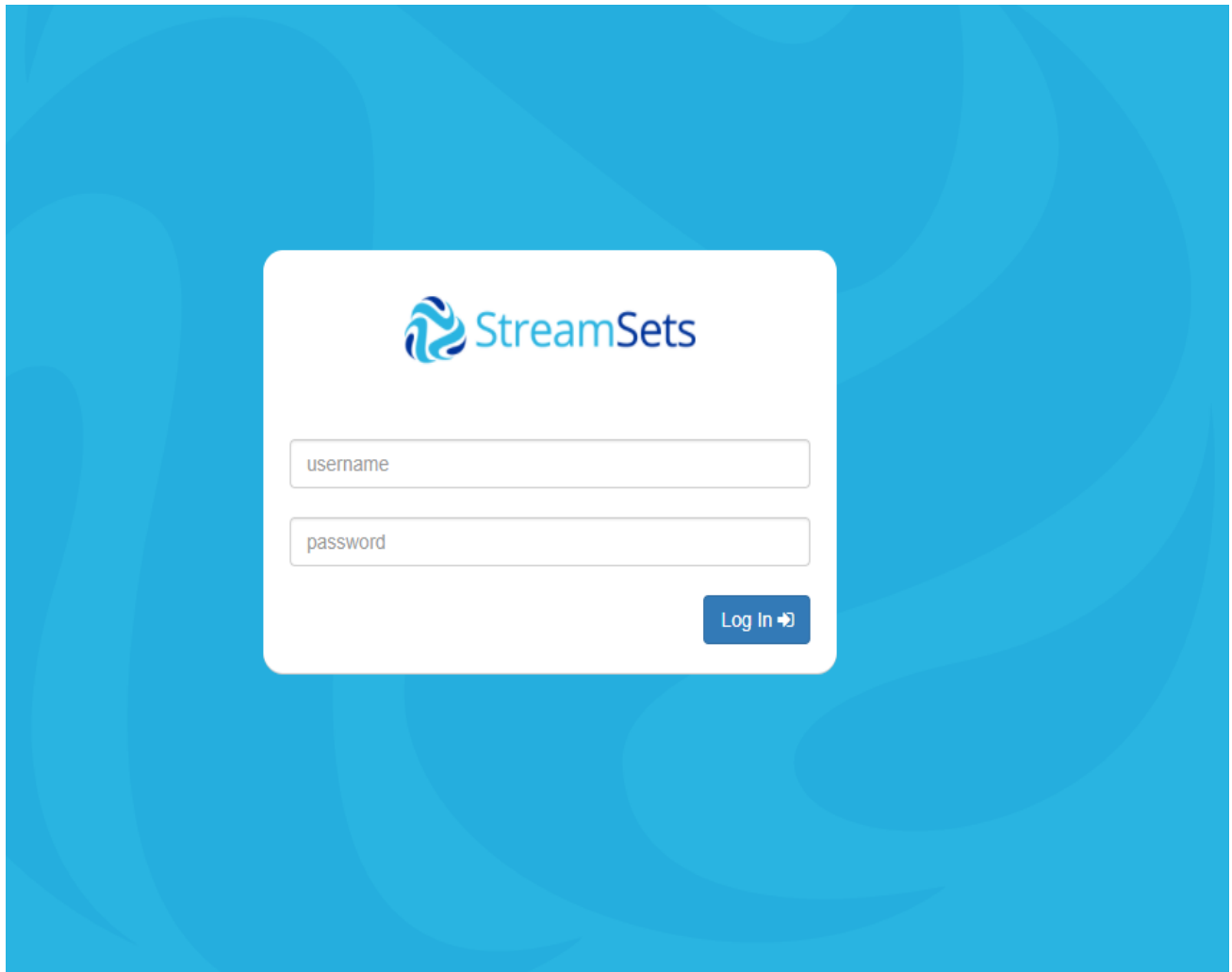
bigdata@yotabites.com



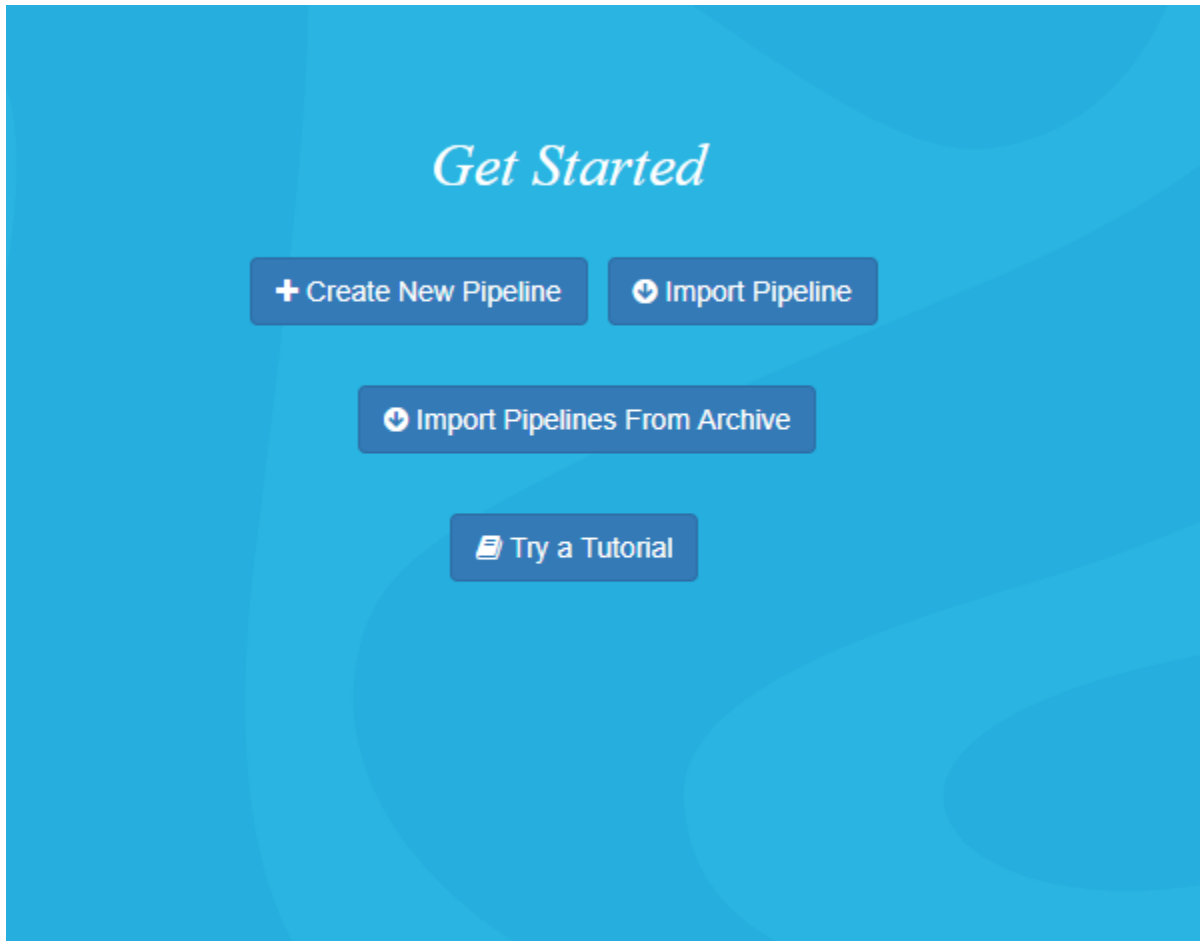
BUILDING PIPELINES USING STREAMSETS

StreamSets – Web UI

1.1	Login to StreamSets.
1.1.1	Go to http://ec2-34-203-42-12.compute-1.amazonaws.com:18630/ to access StreamSets Web UI in browser
1.1.2	Enter the username and password sent to your registered Email id and hit Log In.

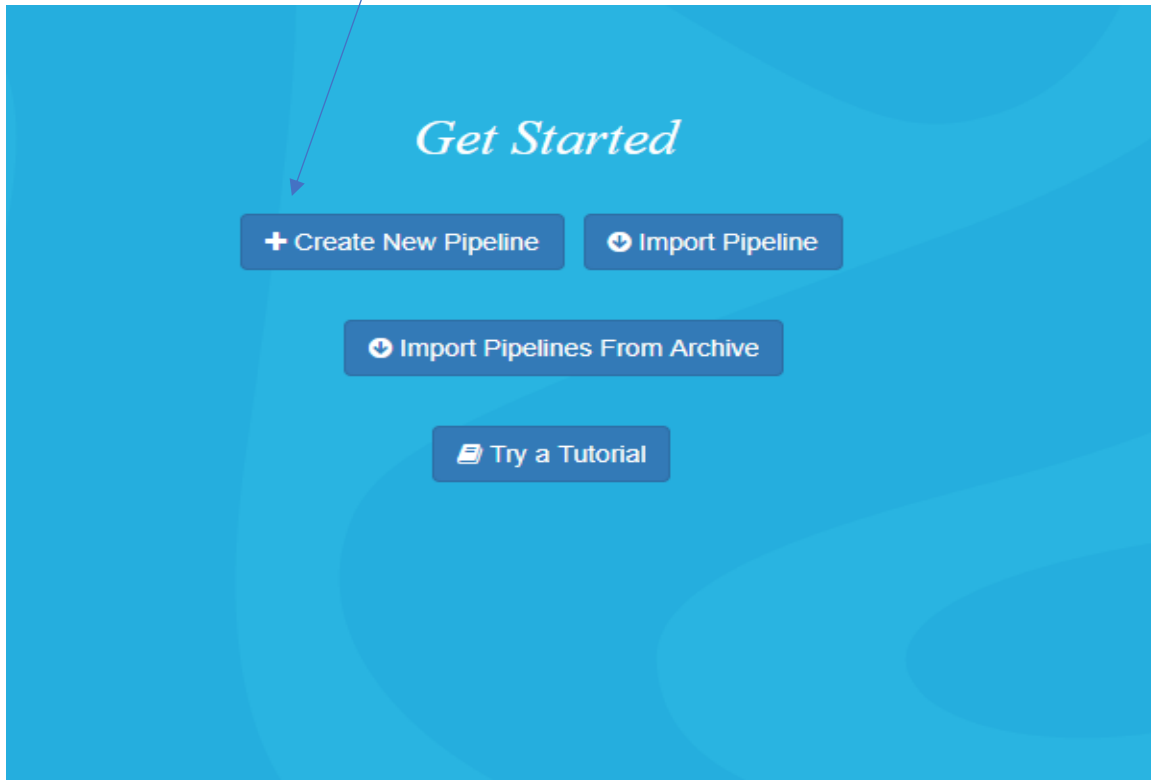


1.2	Once your login is successful, you will see the following landing page.
-----	---

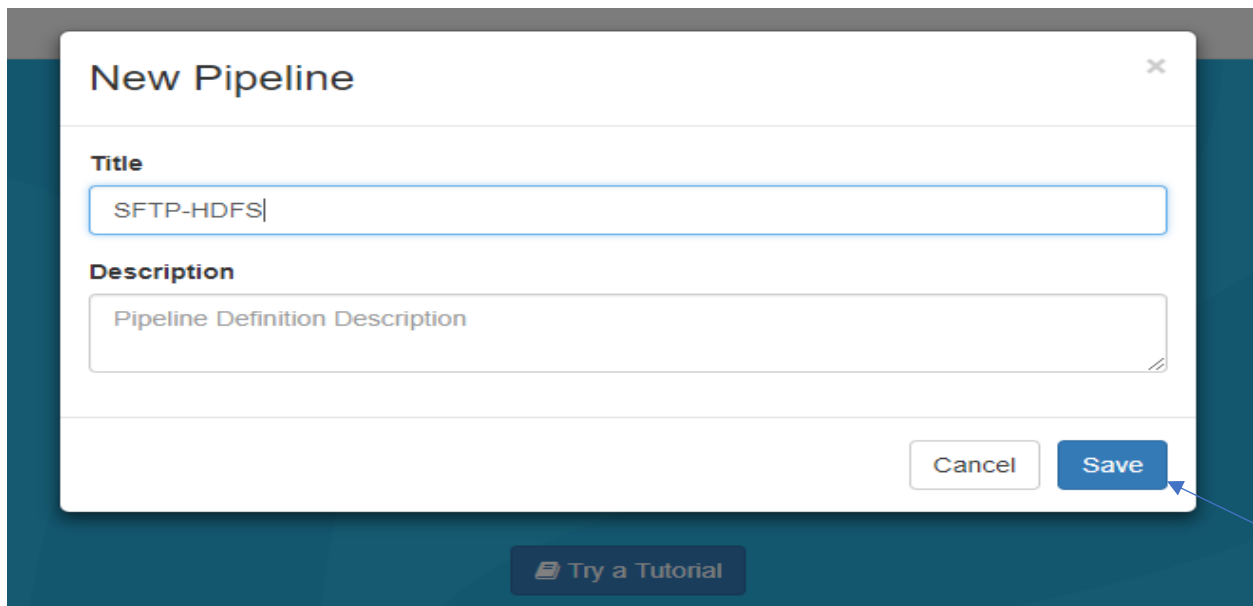


SFTP-HDFS

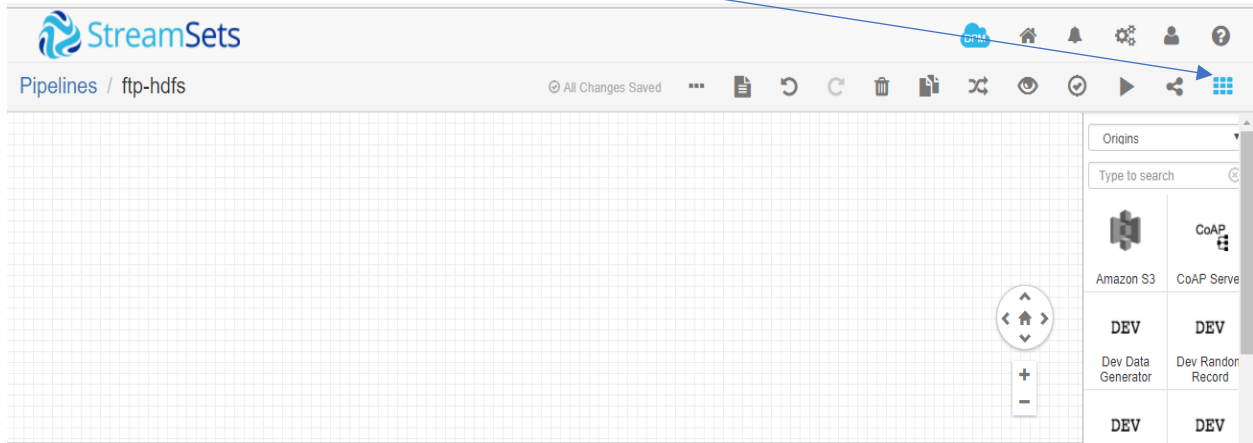
2.1 Click on **create new pipeline**



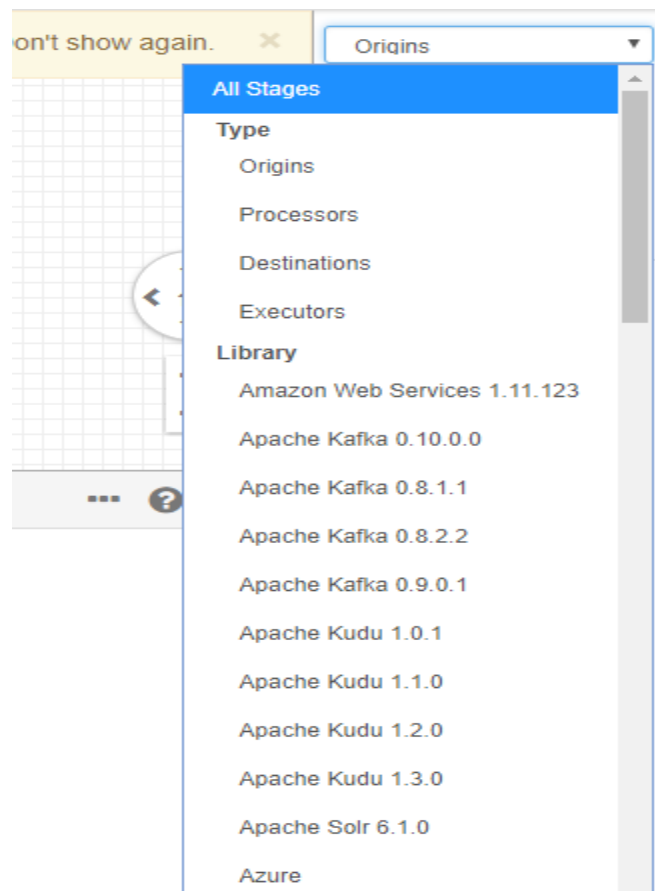
2.2 Input your **project title** and description in the **following window** and click save



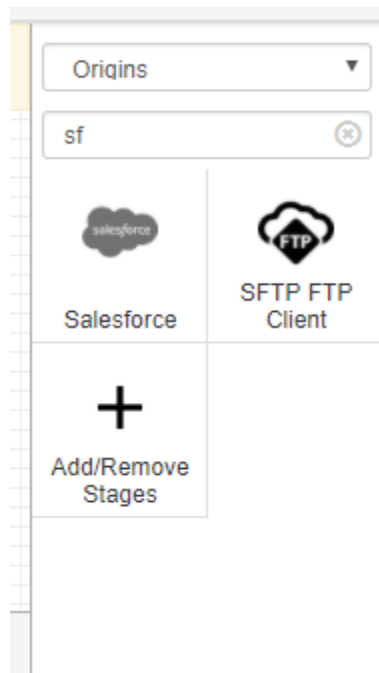
2.3 Now click on the **Stage library** option on the top right corner



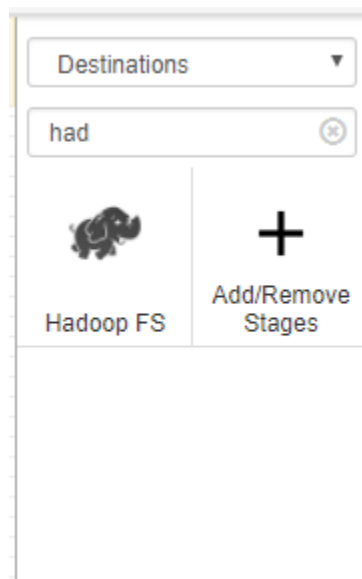
2.4 In the dropdown list select the **origins**



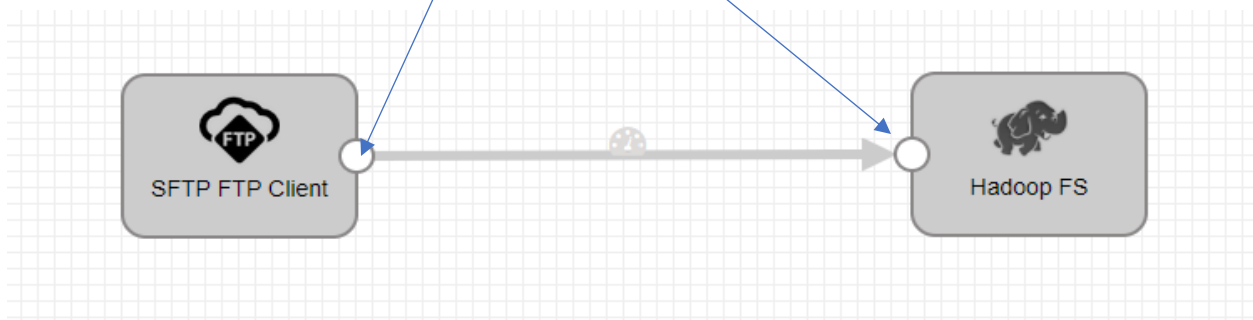
2.5	Now search for SFTP FTP Client and click once on the SFTP FTP Client option.
-----	---



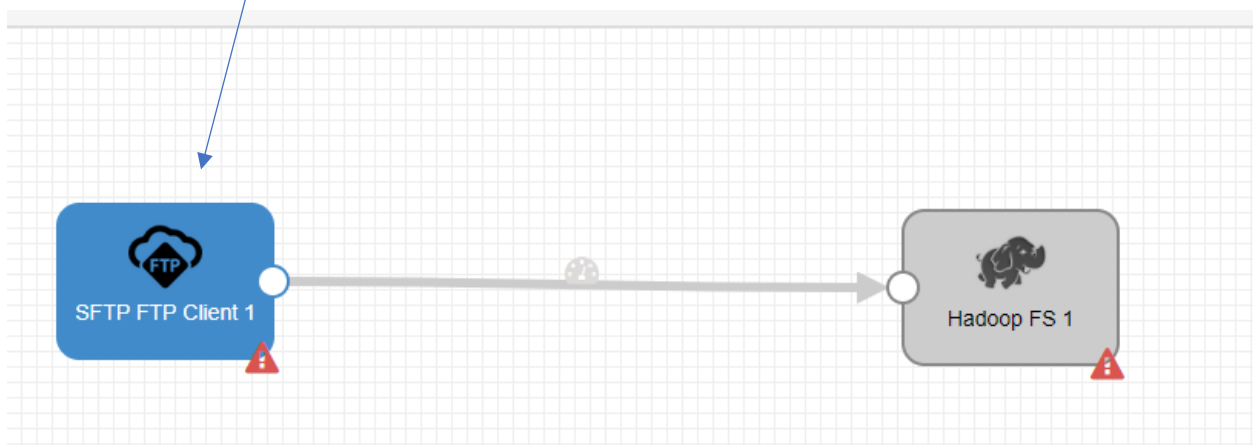
2.6	Now select Destinations in the dropdown list and select Hadoop Fs option and click once.
-----	--



2.7 Now connect the **origin** with **destination** by left clicking on the circle on the origin and drag the arrow to connect to circle on the destination



2.8 Now **click** on the **SFTP FTP Client** to provide the configuration to the origin



2.9 Enlarge the configuration window by clicking on the **enlarge** option

A screenshot of a configuration window for 'SFTP/FTP'. The window has a title bar with a question mark icon and a maximize icon. Below the title bar are tabs: 'General' (selected), 'SFTP/FTP', 'Credentials', 'Error Handling', and 'Data Format'. The 'General' tab is active, showing fields for 'Name' (SFTP FTP Client), 'Description' (empty), and 'On Record Error' (Send to Error). A blue arrow points from the text in step 2.9 to the maximize icon in the title bar.

2.10	Now click on the SFTP/FTP option and fill the Resource URL and File Name Pattern
------	---

General SFTP/FTP Credentials Error Handling Data Format

Resource URL ⓘ

Path Relative to User Home Directory ⓘ ☒

Process Subdirectories ⓘ ☐

File Name Pattern ⓘ

First File to Process ⓘ

Max Batch Size (records) ⓘ

Batch Wait Time (ms) ⓘ

2.11	Select Credentials window. In the Authentication drop down list select Password option and enter your username and password
------	---

General SFTP/FTP Credentials Error Handling Data Format

Authentication ⓘ

Username ⓘ

Password ⓘ

Strict Host Checking ⓘ ☐

2.12	Now select the Data Format and input the configuration as specified in the below picture.
------	--

General
SFTP/FTP
Credentials
Error Handling
Data Format

Data Format ⓘ Delimited ▼

Compression Format ⓘ None ▼

Delimiter Format Type ⓘ Default CSV (ignores empty lines) ▼

Header Line ⓘ With Header Line ▼

Max Record Length (chars) ⓘ 1024

Root Field Type ⓘ List-Map ▼

Lines to Skip ⓘ 0

Parse NULLs ⓘ ☐

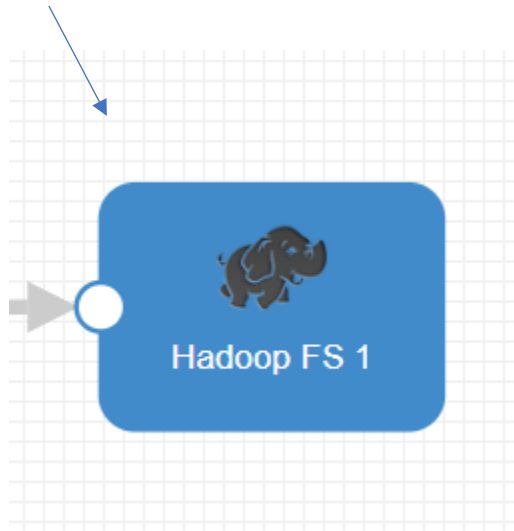
Charset ⓘ UTF-8 ▼

Ignore Control Characters ⓘ ☐

2.13	Now minimize the configuration window by clicking the minimize option on the top right of window
------	---



2.14 Now click on the **Hadoop FS** destination



2.15 Click on General and Select the **Stage library** as per your CDH version

General Hadoop FS Output Files Late Records Data Format

Name

Description

Stage Library

Produce Events ☐

Required Fields

[Select Fields Using Preview Data](#)

Preconditions

[Switch to bulk edit mode](#)

On Record Error

2.16	Now input the URI to connect to HDFS and Conf directory location(conf directory contain files like core-site.xml....etc)
------	---

General
Hadoop FS
Output Files
Late Records
Data Format

Hadoop FS URI ⓘ

HDFS User ⓘ

Kerberos Authentication ⓘ
☐

Hadoop FS Configuration Directory ⓘ

Hadoop FS Configuration ⓘ

[Switch to bulk edit mode](#)

2.17	Now click on output files and input the path where you want yours records to be stored.
------	--

General
Hadoop FS
Output Files
Late Records
Data Format

File Type ⓘ

Files Prefix ⓘ

Files Suffix ⓘ

Directory in Header ⓘ
☐

Directory Template ⓘ

Data Time Zone ⓘ

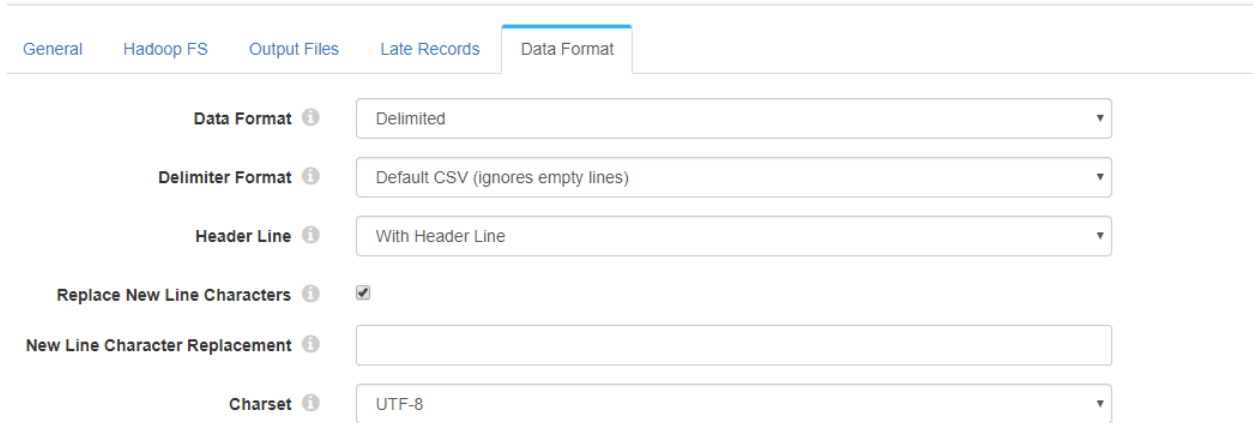
Time Basis ⓘ

Max Records in File ⓘ

Max File Size (MB) ⓘ

Idle Timeout ⓘ

2.18	Now click on the Data Format and select the format of data coming from origin. In this pipeline we are using CSV.
------	--



General Hadoop FS Output Files Late Records Data Format

Data Format *i* Delimited ▼

Delimiter Format *i* Default CSV (ignores empty lines) ▼

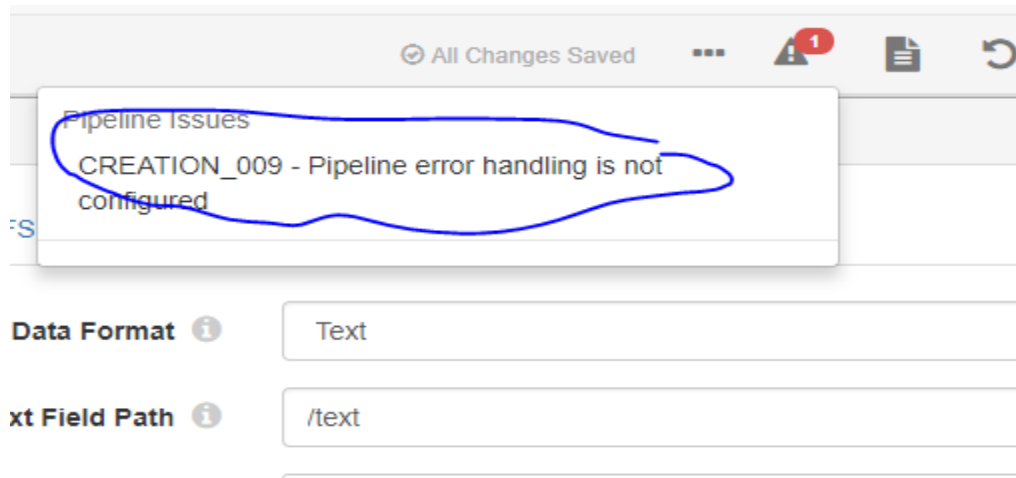
Header Line *i* With Header Line ▼

Replace New Line Characters *i* ☒

New Line Character Replacement *i*

Charset *i* UTF-8 ▼

2.19	Now minimize configuration window and correct the following error by clicking on the error .
------	---



All Changes Saved

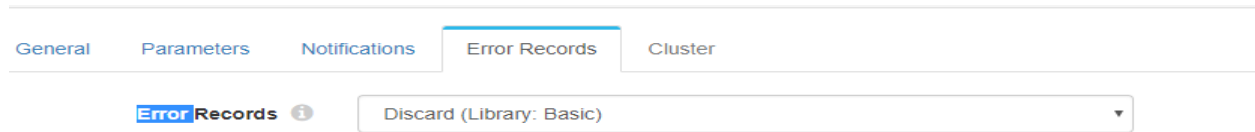
Pipeline Issues

CREATION_009 - Pipeline error handling is not configured

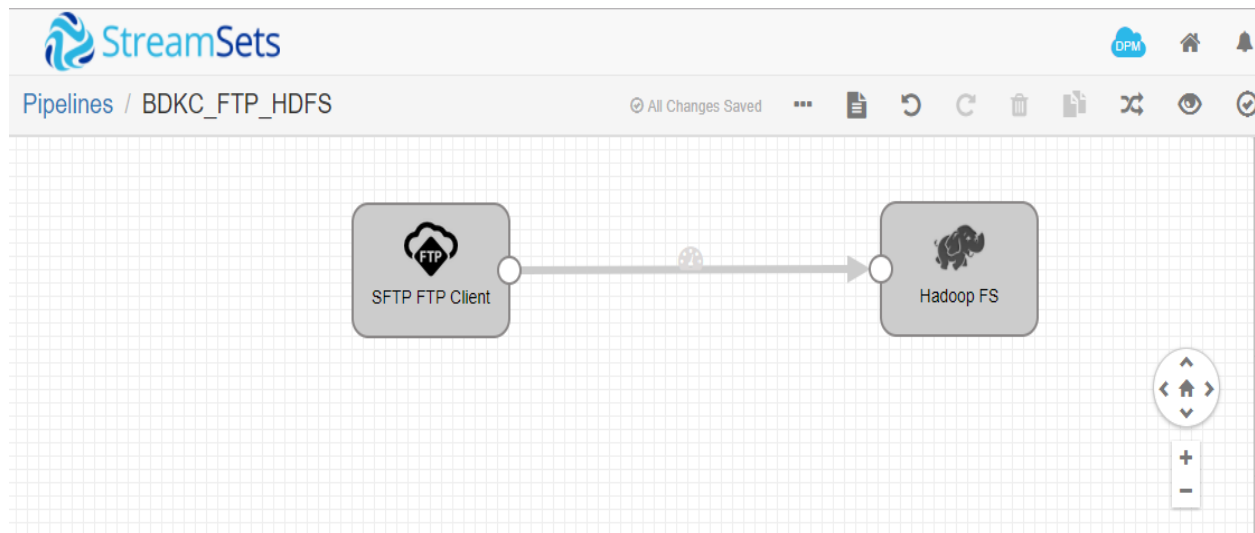
Data Format *i* Text

Text Field Path *i* /text

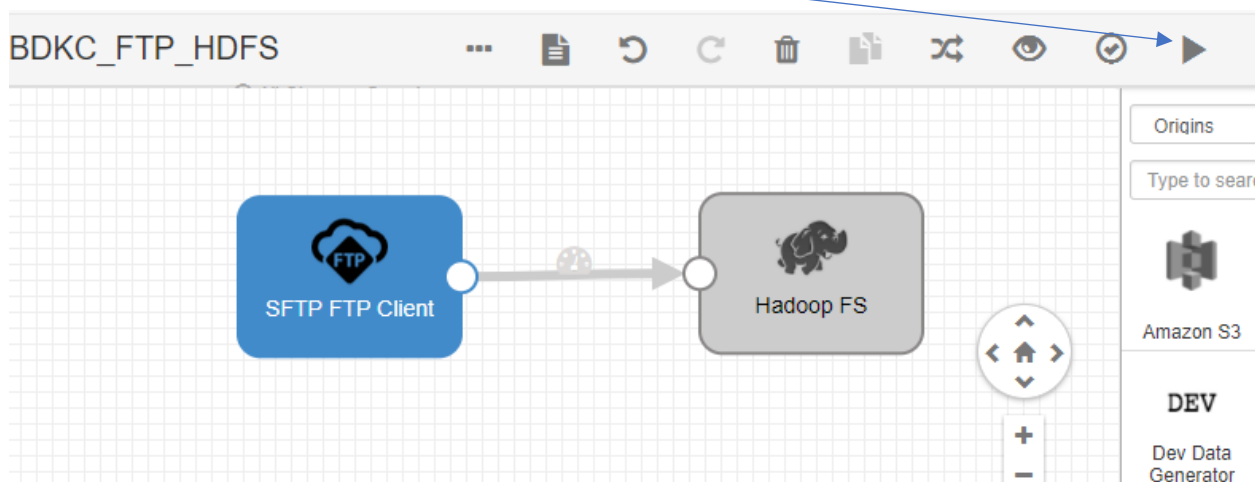
2.20 | Select **Discard** option from the drop-down list.



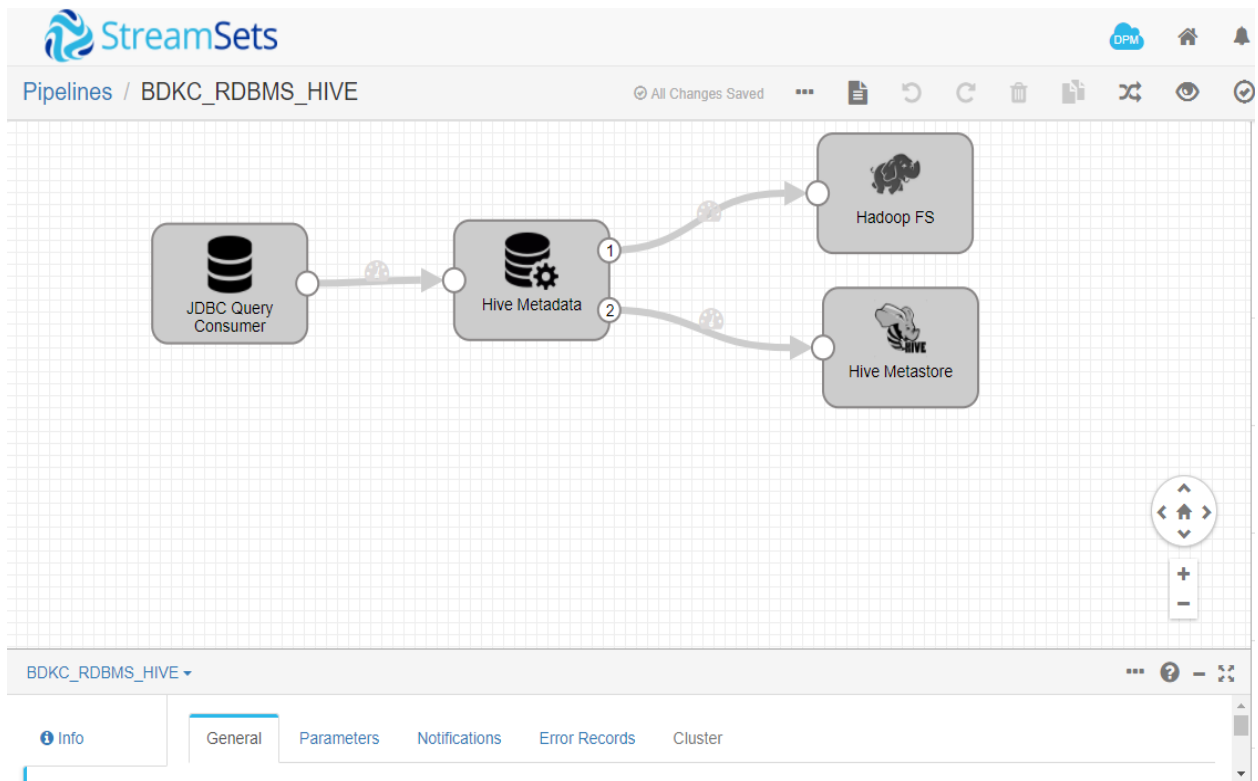
2.21 | Once all the configurations are successfully validated. We will the following pipeline.



2.22 | Now Click on the **start** option on the top left of the taskbar.



RDBMS-HIVE



	JDBC Query Consumer – Origin
Window Name	Configuration
JDBC	<p>JDBC Connection String = <i>jdbc:mysql://ip-172-31-89-20.ec2.internal:3306/employee</i></p> <p>Incremental Mode = <i>Check the box</i></p> <p>SQL Query = <i>select * from tbldesignations where ID > \${OFFSET} ORDER BY ID</i></p> <p>Initial Offset = 0</p> <p>Offset Column = <i>ID (needs to be unique and keeps incrementing)</i></p> <p>Root Field Type = <i>ListMap</i></p> <p>Note: <i>Any key which is incrementing in an order can be taken to compare the offsets. We have used incrementing column as primary key for our convenience.</i></p>
Credentials	<p>Username=<i>demo</i></p> <p>Password=<i>demo</i></p>
Advanced	<i>Create JDBC Namespace Headers = Check the box</i>

	Hive Metadata – Processor
Window Name	Configuration
General	Stage Library = <i>CDH 5.9.2</i>
Hive	<p>JDBC URL = <i>jdbc:hive2://ip-172-31-89-20.ec2.internal:10000/default</i></p> <p>JDBC Driver Name = <i>org.apache.hive.jdbc.HiveDriver</i></p> <p>Hadoop Conf Directory = <i>/etc/hive/conf</i></p>

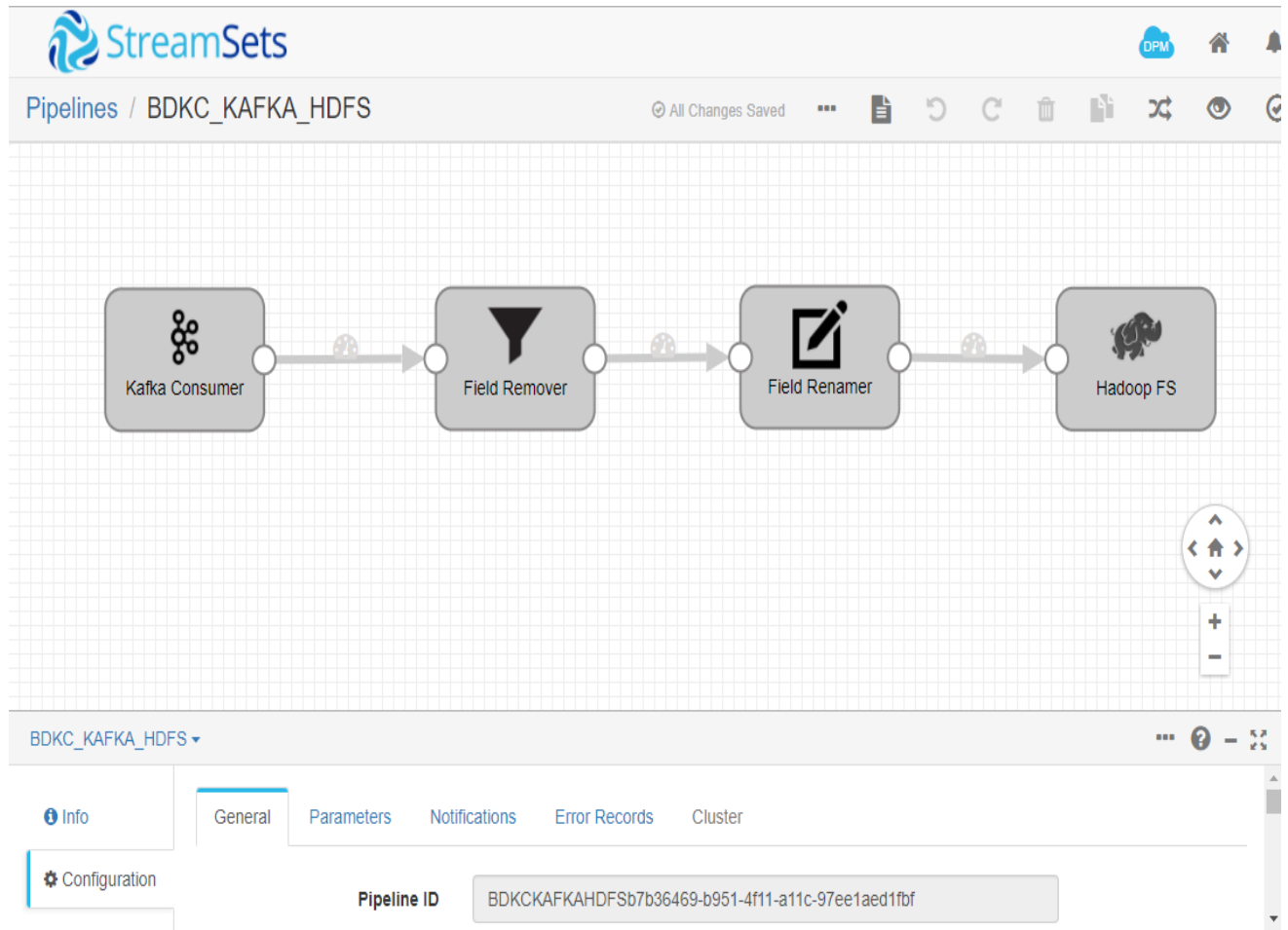
	Hive Metadata - Processor
Table	Database Expression = <i>\${record:attribute('default')}</i> Table Name = <i><your userid ></i>
Data Format	Data Format = <i>Avro</i>

	Hadoop FS – Destination
Window Name	Configuration
General	Stage Library = <i>Select the library with respect to CDH version</i>
Hadoop FS	HadoopFS URI = <i>hdfs:// ip-172-31-89-20.ec2.internal:8020</i> HadoopFS Configuration Directory = <i>/etc/hadoop/conf</i>
Output Files	File Type = <i>Text Files</i> Directory in Header = <i>Check the box</i> Max Records in File = <i>1</i> Use Roll Attribute = <i>Check the box</i> Roll Attribute Name = <i>roll</i>

Data Format	Data Format = Avro Avro Schema Location = In Record Header
--------------------	---

	Hive Metastore- Destination
Window Name	Configuration
Hive	JDBC URL = jdbc: hive2:// ip-172-31-89-20.ec2.internal:10000/default JDBC Driver Name = org.apache.hive.jdbc.HiveDriver Hadoop Conf Directory = /etc/hive/conf
Advanced	Stored as Avro = check the box

KAFKA – HDFS



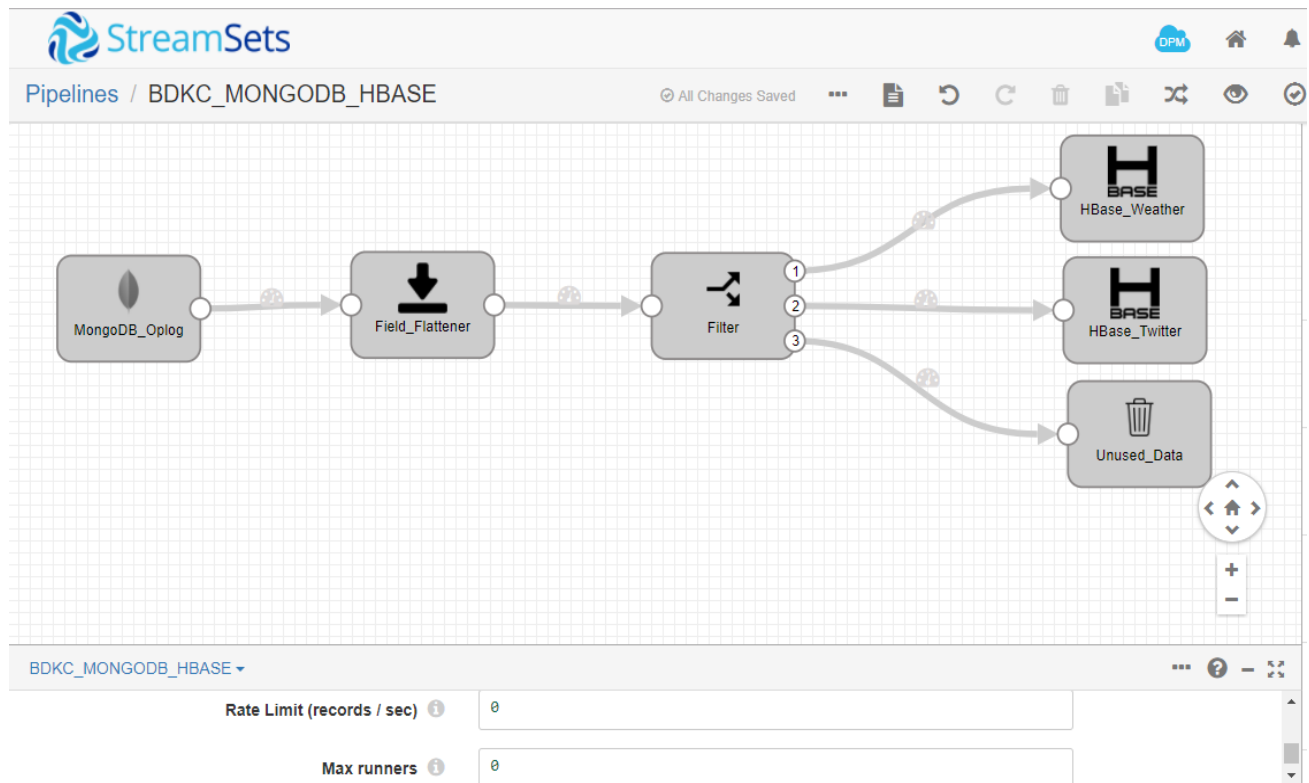
	Kafka Consumer – Origin
Window Name	Configuration
General	Stage Library = <i>Apache Kafka 0.10.0.0</i>
Kafka	Broker URI = <i>ip-172-31-89-20.ec2.internal:9092</i> Zookeeper URI = <i>ip-172-31-89-52.ec2.internal:2181/Kafka</i> Topic = <i>meetup</i>
Data Format	Data Format = <i>JSON</i> Json Content = <i>Multiple Json Objects</i>

	Field Remover – Processor
Window Name	Configuration
Remove/Keep	Fields = <i>/member/photo ; /event/event_url</i>

	Field Renamer – Processor
Window Name	Configuration
Rename	Source Field Expression = <i>/event/time</i> Target Field Expression = <i>/event/mtime</i>

	HDFS – Destination
Window Name	Configuration
General	Stage Library = <i>CDH 5.9.2</i>
Hadoop FS	HadoopFS URI = <i>hdfs:// ip-172-31-89-20.ec2.internal:8020</i> HadoopFS Configuration Directory = <i>/etc/hadoop/conf</i>
Data Format	Data Format = <i>Json</i> Json Content = <i>Multiple Json Object</i>

MongoDB – HBase



	MongoDB Oplog – Origin
Window Name	Configuration
MongoDB	Connection String = <i>mongodb://172.31.89.20:27017/streamset</i> Collection = <i>oplog.rs</i>
Credentials	Authentication Type = <i>None</i>

	Field Flattener – Processor
Window Name	Configuration
Flatten	Flatten = <i>Flatten entire record</i> Name separator = <i>.</i>

	Stream Selector – Processor
Window Name	Configuration
Conditions	<p>Condition = 1 <code>\${record:attribute('ns')}=="streamset.weather"</code></p> <p>2 <code>\${record:attribute('ns')}=="streamset.twitter"</code></p> <p>3 <code>default</code></p> <p>Note: We have taken weather and twitter data. It can be replaced by any data streams. Use “+” on the right to add more conditions.</p>

	HBase – Destination(Weather)
Window Name	Configuration
Hbase	<p>Zookeeper Quorum = <code>172.31.89.52,172.31.81.155,172.31.87.163</code></p> <p>Zookeeper Client Port = <code>2181</code></p> <p>Zookeeper Parent Znode = <code>/hbase</code></p> <p>Table Name = <code>streamset_weather</code></p> <p>Rowkey = <code>/o.place</code></p> <p>Storage Type = <code>Text</code></p> <p>Fields = <i>Map the Field Path to Column with a proper syntax</i></p> <p>Ex : Field path = <code>/o.time_zone</code></p> <p>Column = <code>Column_family:column_qualifier</code></p> <p>Storage = <code>Text</code></p>

	HBase – Destination(Twitter)
Window Name	Configuration
Hbase	Zookeeper Quorum = 172.31.89.52,172.31.81.155,172.31.87.163 Zookeeper Client Port = 2181 Zookeeper Parent Znode = /hbase Table Name = <i>streamset_twitter</i> Rowkey = /o.screen_name Storage Type =Text Fields = Map the Field Path to Column with a proper syntax Ex: Field path = /o. time_zone Column = Column_family: column_qualifier Storage = Text

	Trash- Destinations
Window Name	Configuration
General	Name = <i>Unused data</i> Description = <i>we send all the remaining data after filter to this pipeline.</i>