

**Data Pipeline**

# Data Pipeline

Data pipeline is an event based serverless architecture, deployed on kubernetes cluster and Kafka based communication to handle real-time and batch data. Gain seamless insights from massive amount of structured, semi-structured, and unstructured data. Drag & Drop based interface where pipeline can be created, and models generated from ‘BDB Predictive Workbench’ & ‘Data Preparation’ can be embedded inside the workflow.

BDB Data Pipeline is a web service that you can use to automate the End-to-End movement and transformation of data. With BDB Data Pipeline, you can define data-driven workflows in order to get great insights and visualization for better decision making.

# Features of Pipeline

BDB data pipeline is an event based serverless architecture for smooth orchestration of data flow and processing of real time and batch data. Some of its features are –

* Runs on kubernetes container
* Provides easy scalability and fault tolerance.
* Build using Kafka for event management and streaming, and spark for distributed data processing.
* Using serverless architecture and just in time invocation of process, user can effectively utilize the computational resources.
* Provides seamless integration with Data preparation and Predictive Analytics workbench.
* Can consume any ML models from the predictive work bench as well as data preparation steps build using Data preparation module.
* Provides ability to run custom scrips using python, SSH and Perl
* Logging and monitoring facilities
* Drag and drop panel for the user to easily configure and build pipeline.

# Components of Pipeline

Data Pipeline has various components which are designed from scratch, with system and custom component. In case of custom component, it will list all the components designed by the user.

## System

## Reader

It helps to read data from different sources and ingest that data into the Pipeline for further processing

## Writer

This component is used to write input data to the required place

## Transformation

Allows user to transform data

## ML Components

Import the various predictive models to the Data Pipeline plugin and use them to create different Pipeline workflows for streaming data

## Ingestion

Allow the users to ingest data in the pipeline from outside the pipeline

## Websocket

It contains Websocket Listener and Websocket Writer components to produce and consume the live streaming data using socket

## Custom

Custom components are designed by user based on their own requirements.

# Description of Each Components of Pipeline

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.

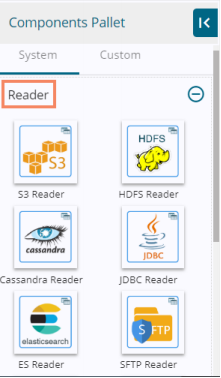
# Data Reader

It helps to read the 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.

* 1. **S3 Reader**
  2. **HDFS Reader**
  3. **Cassandra Reader**
  4. **JDBC Reader**
  5. **ES Reader**
  6. **SFTP Reader**

Drag and drop the component to the Workflow-editor and do the configuration for reading the data from reader component. While doing configuration user needs to select the running mode for component as ‘Batch’ or ‘Real-time’ in basic information while user also need to fill the details in meta information. After filling the Meta-data information, save the reader component.

Once configuration is done, you can update the Pipeline.



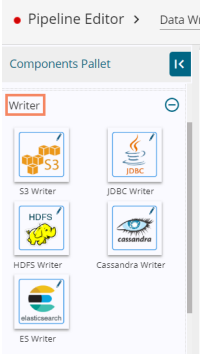
# Data Writers

BDB Data Pipeline provides Writer components to write input data to the required place. BDB Data Pipeline contains the following Data Writers under the Component Pallet.

* 1. **S3 Writer**
  2. **JDBC Writer**
  3. **HDFS Writer**
  4. **Cassandra Writer**
  5. **ES Writer**

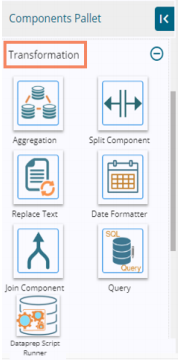
Drag and drop the component to the Workflow-editor and do the configuration for writing the data from reader component. While doing configuration user needs to select the running mode for component as ‘Batch’ or ‘Real-time’ in basic information while user also need to fill the details in meta information. After filling the Meta-data information, save the writer component.

Once configuration is done, you can update the Pipeline and activate it.



# Transformations

Transformation components allow the users to transform the data. BDB Data Pipeline provides the following Transformations components to transform a variety of data.



* 1. **Aggregation**

It allows the users to perform the following actions on the selected input columns:

1. If the input data type is number (E.g., Integer or double) then, the output column takes the data after performing the addition aggregation function on the selected columns.
2. If the input data type is String, then after applying the Aggregation transform the data gets concatenated from the selected columns
   1. **Split component**

The Split component helps users to split the selected column(s) from the input data set based on the given regular expression.

* 1. **Replace Text**

The Replace Text transform component allows users to replace the searched data in the selected columns with the user-defined replacement text.

* 1. **Date Formatter**

Users can alter the Date format by using this transform component

* 1. **Join Component**

Join transform allows users to join two or more input data sets as per the user-defined join conditions

* 1. **Query**

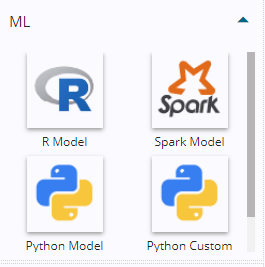
This component helps users to get data as per entered query.

* 1. **Dataprep Script Runner**

This transform component enables the users to use the exported Dataprep scripts in the Data Pipeline.

# ML

ML Model Runner Components allow us to use the models created on R, Spark and Python Workspaces predictive workbench inside the pipeline. There are 3 ML model (runner) components available in this section:

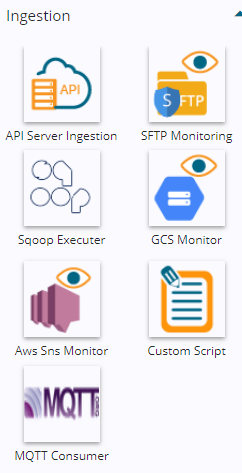


* 1. **R Model Runner**
  2. **Spark Model**
  3. **Python Model**
  4. **Python Custom Script**

All the runner component reads the data coming to an input event, runs the model, and gives the output data with predicted columns to the output event

# Ingestion

Ingestion components allow the users to ingest data in the pipeline from outside the pipeline as in files from some SFTP location or manual ingestion of data using a service in real-time.



* 1. **API Server Ingestion**

As we save the pipeline, an auto-generated Component Ingestion URL appears inside the Meta Information tab to ingest the data. Users can use the Component Ingestion URL with the following format in the program or add anywhere in the third-party portal.

* 1. **SFTP Monitor**

The SFTP Monitoring component monitors the file coming to the monitored path and copies the file in Copy Path location for SFTP Reader to read.

* 1. **Sqoop Executor**

Sqoop component in BDB Pipeline helps users to transfer tables and databases from RDBMS (supported by Sqoop) to HDFS (Hadoop System Distributed File).

* 1. **GCS Monitor**
  2. **Aws Sns Monitor**
  3. **Custom Script**
  4. **MQTT Consumer**

# WebSocket

The Websocket Component provided in the Components Pallet of the Data pipeline contains Websocket Listener and Websocket Writer components to produce and consume the live streaming data using socket.

* 1. **Websocket Listener**

Websocket Listener component listens to and accepts the incoming Websocket connection requests

* 1. **Websocket Producer**

Websocket Producer helps users to get the message received by the Kafka channel.

***To connect any component in Data Pipeline, user need to connect it via event. And toggle log panel gives the information that which component is up.***

A close up of a map

Description automatically generated