# Design doc for HDFS output module (tuple based)

## Objective

This module is responsible for writing incoming data into HDFS.

## Functionality

This module assumes that all incoming tuples are to be written to the same destination file path. After max file size limit is reached; output will be rolled over to the new file with incremental id suffixed to it.

For example, if module is configured to write to filePath \user\username\path\to\destination\filename.txt and max file size is set to 100 MB. Then, output will be rolled over to filename.txt.1, filename.txt.2 …

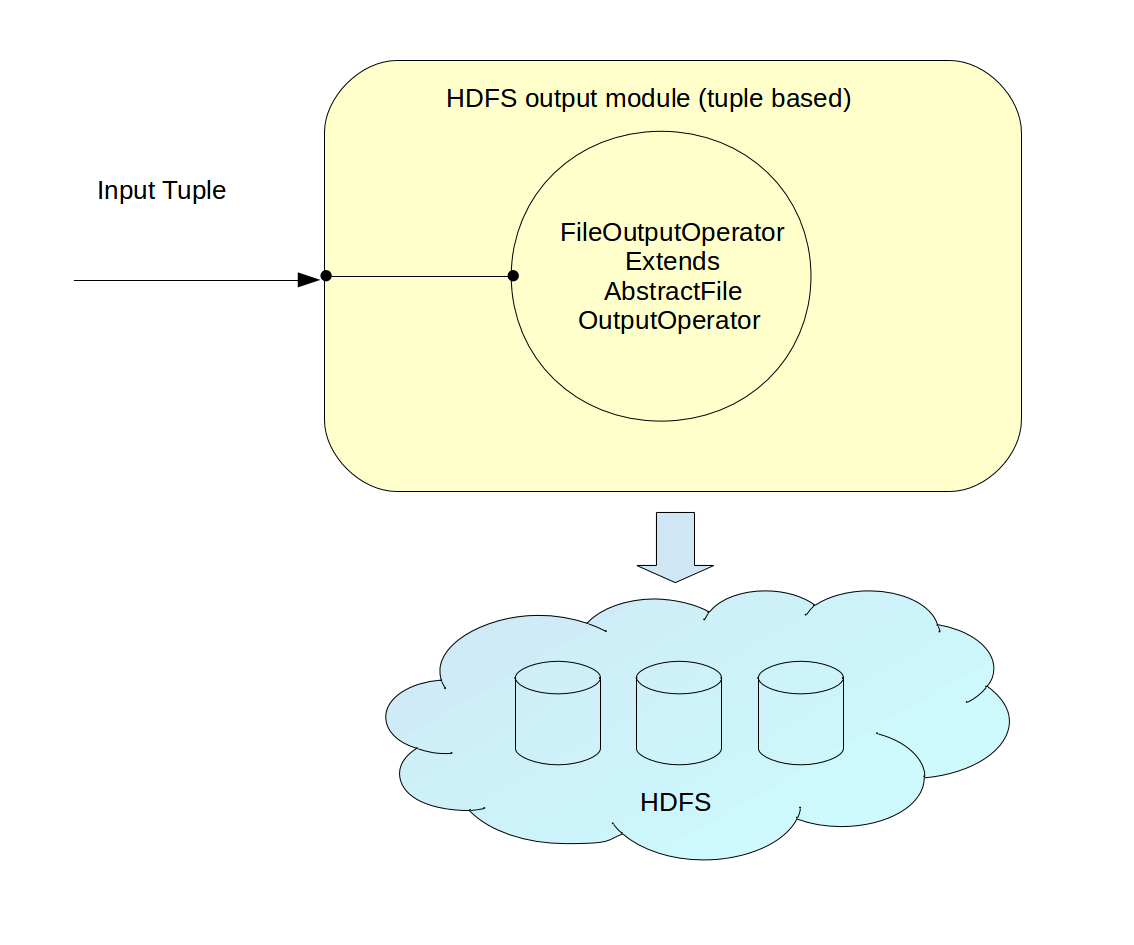
## Dependencies

This module will have a concrete implementation which extends com.datatorrent.lib.io.fs.AbstractFileOutputOperator under malhar library. Thus, it is heavily dependent on the functionality provided by that operator.

## Risks

Any changes in the functionality or side-effects of code changes done in com.datatorrent.lib.io.fs.AbstractFileOutputOperator under malhar library will have an impact on the behavior of this module.

## DAG



HDFS output module (tuple based) DAG

## Properties

```java //Path for the output file String filePath

//The maximum length in bytes of a rolling file. Long maxLength = Long.MAX\_VALUE

```

## Attributes

//No. of static partitions to be used for this module  
int partitionCount

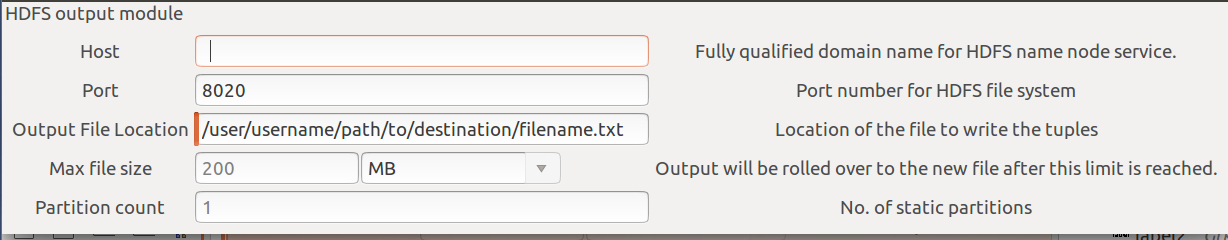
## Input ports

* **input**: Input port to receive incoming data to be written to HDFS.

## Output ports

Since this is output module; it will not have any output ports.

## Config UI



HDFS output module (tuple based) Config UI

## Counters

* Total number of bytes written
* Total number of messages written
* Number of bytes written per second
* Number of messages written per second

## Dashboard Widgets

* Total number of bytes written : Single value widget
* Total number of messages written : Single value widget
* Number of bytes written per second : Line chart
* Number of messages written per second : Line chart

## Usage cases

* dtingest: Copying messages from Kafka to HDFS
* writing output of Dedup to HDFS

## Pathological cases

* More than one instances of the HDFS output module writing to the same filePath. This case will not be handled in the first cut implementation.

## Future work

Adding support for writing to multiple files. One way to achieve this is as follows:

Each incoming tuple must implement following interface

```java public interface FileOutputData { public String getFileName(); public byte[] getBytesForTuple(); }

```

Thus, each tuple will have information about which file to write and what data to write.