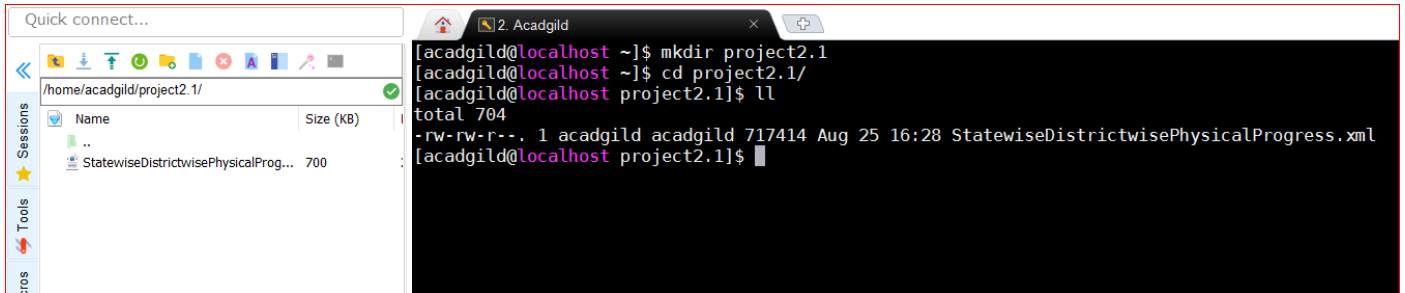


****Copying Dataset using Flume****

Download the dataset using the below link:

<https://drive.google.com/file/d/0Bxr27gVaXO5sUjd2RWFQS3hQQUE/view?usp=sharing>



Start Hadoop daemons and clear /flume sink if exists:

```
$ start-all.sh
```

```
$ hadoop fs -ls /
```

```
$ hadoop fs -rm -r /flume_sink
```

```
[acadgild@localhost ~]$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
17/08/25 16:31:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-acadgild-namenode-localhost.localdomain.out
localhost: starting datanode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-acadgild-datanode-localhost.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop-2.6.0/logs/hadoop-acadgild-secondarynamenode-localhost.localdomain.out
17/08/25 16:32:00 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-acadgild-resourcemanager-localhost.localdomain.out
localhost: starting nodemanager, logging to /usr/local/hadoop-2.6.0/logs/yarn-acadgild-nodemanager-localhost.localdomain.out
[acadgild@localhost ~]$ hadoop fs -ls /
17/08/25 16:34:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 8 items
drwxr-xr-x - acadgild supergroup          0 2017-07-16 01:28 /flume_sink
drwxr-xr-x - acadgild supergroup          0 2017-06-03 16:59 /hadoop
drwxr-xr-x - acadgild supergroup          0 2017-08-19 13:53 /hbase_output_dir
drwxr-xr-x - acadgild supergroup          0 2017-08-20 13:39 /hbasestorage
drwxr-xr-x - acadgild supergroup          0 2017-08-24 01:16 /sqoop
drwxrwxr-x - acadgild supergroup          0 2017-08-20 16:24 /tmp
drwxr-xr-x - acadgild supergroup          0 2017-07-11 15:57 /user
drwxr-xr-x - acadgild supergroup          0 2015-11-05 12:56 /zookeeper
[acadgild@localhost ~]$ hadoop fs -rm -r /flume_sink
17/08/25 16:34:51 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/08/25 16:34:52 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /flume_sink
[acadgild@localhost ~]$
```

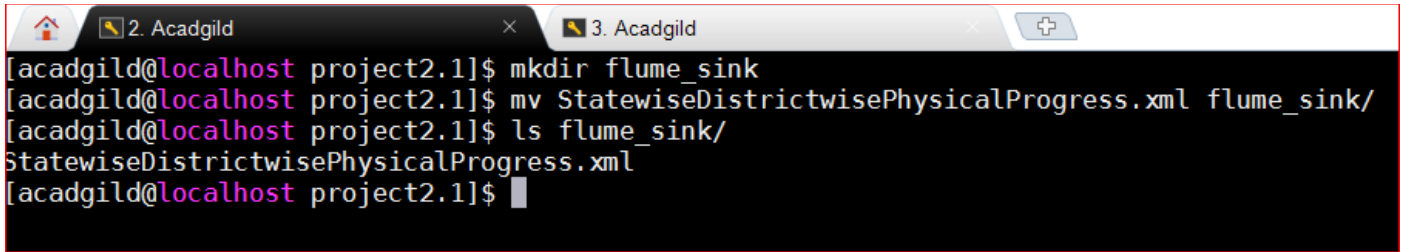
Copy dataset from local file system to HDFS using flume:

Step 1: create source directory and move dataset file into it:

```
$ mkdir flume_sink
```

```
$ mv StatewiseDistrictwisePhysicalProgress.xml flume_sink/
```

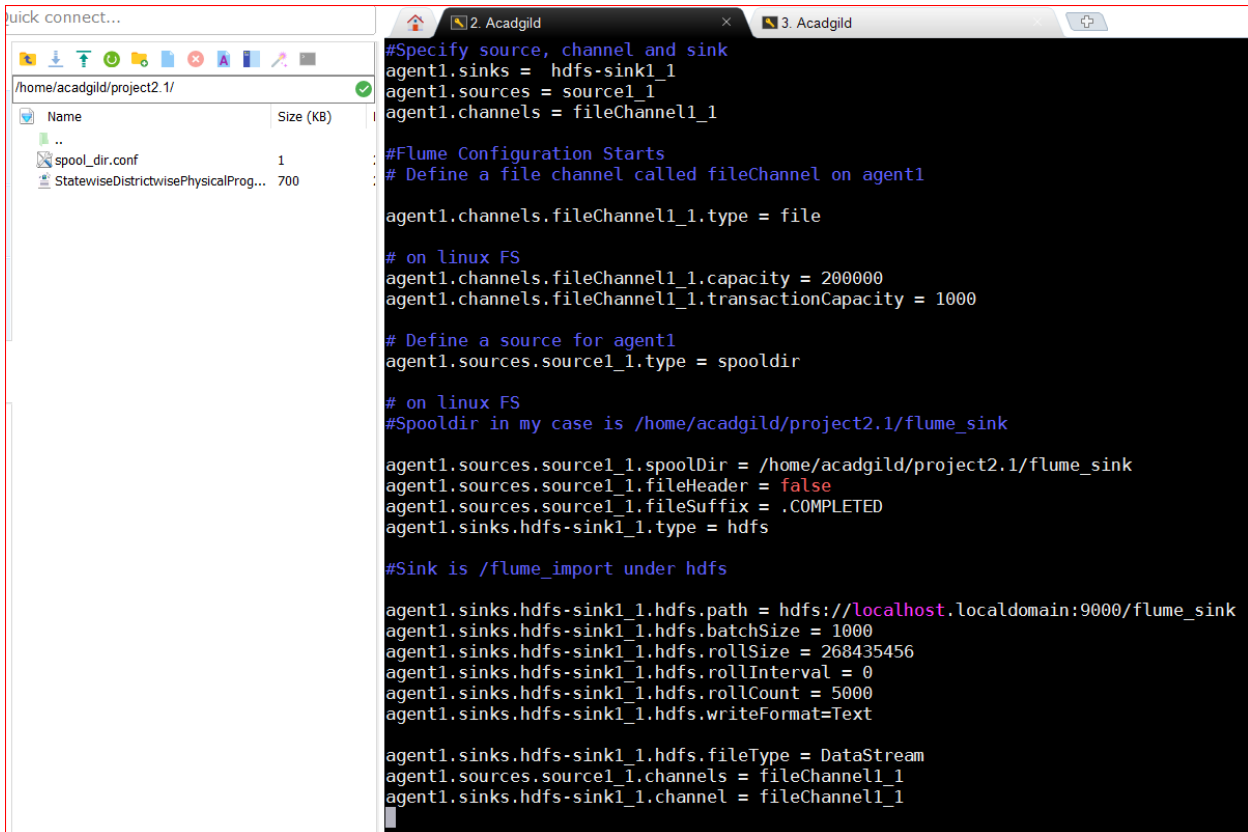
```
$ ls flume_sink/
```

A terminal window with two tabs labeled '2. Acadgild' and '3. Acadgild'. The terminal shows the execution of three commands: 'mkdir flume_sink', 'mv StatewiseDistrictwisePhysicalProgress.xml flume_sink/', and 'ls flume_sink/'. The output of the last command shows 'StatewiseDistrictwisePhysicalProgress.xml'.

Step 2: create flume config file which uses below source and target directory:

Source Directory: */home/acadgild/project2.1/flume_sink*

Target Directory: *hdfs://localhost.localdomain:9000/flume_sink*

A screenshot showing a file explorer on the left and a terminal window on the right. The file explorer shows a directory with files 'spool_dir.conf' (1 KB) and 'StatewiseDistrictwisePhysicalProg...' (700 KB). The terminal window shows the configuration for a flume agent named 'agent1'. The configuration includes settings for a file channel, a spool directory source, and an HDFS sink. The sink is configured to write to 'hdfs://localhost.localdomain:9000/flume_sink' in a DataStream format with a batch size of 1000 and a roll count of 5000.

Step 3: run flume agent agent1 from spool_dir.config file:

```
$ flume-ng agent -n agent1 -f spool_dir.conf
```

```

17/08/25 16:48:25 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: fileChannel1_1 started
17/08/25 16:48:25 INFO node.Application: Starting Sink hdfs-sink1_1
17/08/25 16:48:25 INFO node.Application: Starting Source source1_1
17/08/25 16:48:25 INFO source.SpoolDirectorySource: SpoolDirectorySource source starting with directory: /home/acadgild/project2.1/flume_sink
17/08/25 16:48:25 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: hdfs-sink1_1. Successfully registered new MBean.
17/08/25 16:48:25 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfs-sink1_1 started
17/08/25 16:48:25 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SOURCE, name: source1_1. Successfully registered new MBean.
17/08/25 16:48:25 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: source1_1 started
17/08/25 16:48:26 INFO avro.ReliableSpoolingFileEventReader: Last read took us just up to a file boundary. Rolling to the next file, if there is one.
17/08/25 16:48:26 INFO avro.ReliableSpoolingFileEventReader: Preparing to move file /home/acadgild/project2.1/flume_sink/StatewiseDistrictwisePhysicalProgress.xml to /home/acadgild/project2.1/flume_sink/StatewiseDistrictwisePhysicalProgress.xml.COMPLETED
17/08/25 16:48:26 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
17/08/25 16:48:26 INFO hdfs.BucketWriter: Creating hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906192.tmp
17/08/25 16:48:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
17/08/25 16:48:28 INFO hdfs.BucketWriter: Closing hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906192.tmp
17/08/25 16:48:29 INFO hdfs.BucketWriter: Renaming hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906192.tmp to hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906192
17/08/25 16:48:29 INFO hdfs.BucketWriter: Creating hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906193.tmp
17/08/25 16:48:29 INFO hdfs.BucketWriter: Closing hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906193.tmp
17/08/25 16:48:29 INFO hdfs.BucketWriter: Renaming hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906193.tmp to hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906193
17/08/25 16:48:29 INFO hdfs.BucketWriter: Creating hdfs://localhost.localdomain:9000/flume_sink/FlumeData.1503659906194.tmp
17/08/25 16:48:54 INFO file.EventQueueBackingStoreFile: Start checkpoint for /home/acadgild/.flume/file-channel/checkpoint/checkpoint, elements to sync = 12142
17/08/25 16:48:54 INFO file.EventQueueBackingStoreFile: Updating checkpoint metadata: logWriteOrderID: 1503659929134, queueSize: 0, queueHead: 168849
17/08/25 16:48:54 INFO file.Log: Updated checkpoint for file: /home/acadgild/.flume/file-channel/data/log-9 position: 1742203 logWriteOrderID: 1503659929134
17/08/25 16:48:54 INFO file.LogFile: Closing RandomReader /home/acadgild/.flume/file-channel/data/log-4

```

Source directory and file has been highlighted by green box.

Target directory and files have been highlighted by purple box.

Press ctrl+c for killing the flume agent since file have been copied as per the logs. However in real scenario if xml files are being generated or streamed then flume agent keeps on running to get the data into HDFS.

```

17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: hdfs-sink1_1 stopped
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.start.time == 1503659905179
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.stop.time == 1503660267913
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.batch.complete == 12
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.batch.empty == 75
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.batch.underflow == 1
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.connection.closed.count == 3
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.connection.creation.count == 3
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.connection.failed.count == 0
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.event.drain.attempt == 12142
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SINK, name: hdfs-sink1_1. sink.event.drain.sucess == 12142
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Component type: SOURCE, name: source1_1 stopped
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. source.start.time == 1503659905193
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. source.stop.time == 1503660267914
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. src.append.batch.accepted == 122
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. src.append.batch.received == 122
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. src.append.accepted == 0
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. src.append.received == 0
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. src.events.accepted == 12142
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. src.events.received == 12142
17/08/25 16:54:27 INFO instrumentation.MonitoredCounterGroup: Shutdown Metric for type: SOURCE, name: source1_1. src.open-connection.count == 0
17/08/25 16:54:27 INFO source.SpoolDirectorySource: SpoolDir source source1_1 stopped. Metrics: SOURCE:source1_1[src.events.accepted=12142, src.open-connection.count=0, src.append.received=0, src.append.batch.received=122, src.append.batch.accepted=122, src.append.accepted=0, src.events.received=12142]
[acadgild@localhost project2.1]$ █

```

When agent is killed it shows summary of the agent in our case it shows 12142 events have been captured along with start time, stop time & batch details.

Step 4: check source directory in local file system:

```
$ ll /home/acadgild/project2.1/flume_sink
```

```

[acadgild@localhost project2.1]$ ll /home/acadgild/project2.1/flume_sink
total 704
-rw-rw-r--. 1 acadgild acadgild 717414 Aug 25 16:28 StatewiseDistrictwisePhysicalProgress.xml.COMPLETED
[acadgild@localhost project2.1]$ █

```

File name has been appended with .COMPLETED as mentioned in flume configuration. It means file have been completely moved to HDFS.

Step 5: check files in hdfs:

```
$ hadoop fs -ls /flume_sink
```

```
[acadgild@localhost project2.1]$ hadoop fs -ls /flume_sink
17/08/25 16:59:56 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
-rw-r--r--  1 acadgild supergroup      295196 2017-08-25 16:48 /flume_sink/FlumeData.1503659906192
-rw-r--r--  1 acadgild supergroup      295013 2017-08-25 16:48 /flume_sink/FlumeData.1503659906193
-rw-r--r--  1 acadgild supergroup      127206 2017-08-25 16:54 /flume_sink/FlumeData.1503659906194
[acadgild@localhost project2.1]$
```

As we can see there are three files created as per the flume configuration. Dataset has been imported to HDFS successfully.