**BigData And Hadoop**

**Project1.2**

**Problem Statement:**

Copy the data set into HDFS using Flume and send the screen shot of that with the project solution.

**Solution:**

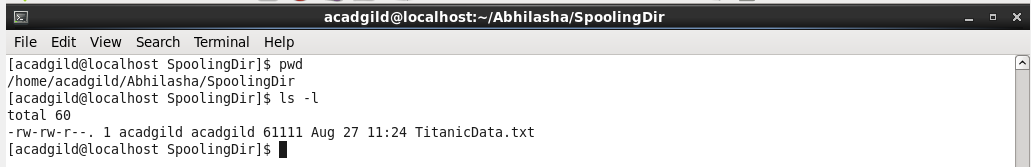
In order to load data into HDFS, the agents used are as follows:

1. Source : Spooling Directory Source
2. Channel : Memory channel
3. Sink : HDFS

**Source: Spooling Directory**

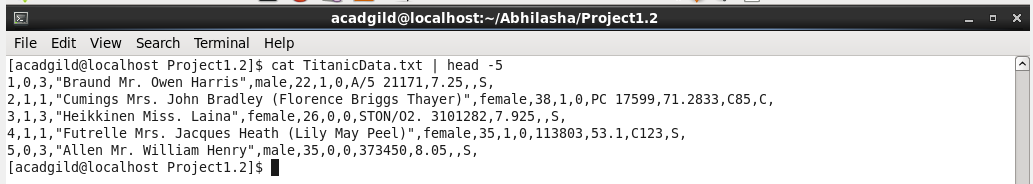
This source lets us ingest data by placing files to be ingested into a “spooling” directory on disk. This source will watch the specified directory for new files, and will parse events out of new files as they appear. After a given file has been fully read into the channel, it is renamed to indicate completion (or optionally deleted).

For this project, the spooling directory used is ‘/home/acadgild/Abhilasha/SpoolingDir’. The dataset file which is to be placed in HDFS is placed in this directory as follows:



Ls command is used to list that document in the spooling directory.

The few records from this file are printed on console using cat command as follows:



In order to place dataset into HDFS using flume, the configuration file named flume.conf is used. This file contains the properties of each of the source, sink and channel.

The properties defined in this file are as follows:

1. **Properties of Source**
2. type = spooldir

spooldir is the keyword that specifies that source is spooling directory

1. spoolDir = /home/acadgild/Abhilasha/SpoolingDir

This is the full qualified path of the folder used as spooling directory

1. fileHeader = false

This is to decide whether to add a header storing the absolute path filename.

1. fileSuffix = .COMPLETED

This is the suffix to append to completely ingested files

1. **Properties of channel:**
2. type = memory

This is to specify the we are using memory channel

1. capacity = 100  
   The maximum number of events stored in the channel
2. transactionCapacity = 100

The maximum number of events the channel will take from a source or give to a sink per transaction

1. **Properties of sink**
2. type = hdfs

This is to specify that hdfs is the sink

1. path = /abhilasha/FlumeTitanicData

The path on HDFS where the data will be transferred

1. batchSize = 1000

Number of events written to file before it is flushed to HDFS

1. rollSize = 268435456

File size to trigger roll, in bytes

1. rollInterval = 0

Number of seconds to wait before rolling current file

1. rollCount = 50000000

Number of events written to file before it rolled

1. writeFormat=Text

Format for sequence file records. One of “Text” or “Writable”

1. fileType = DataStream

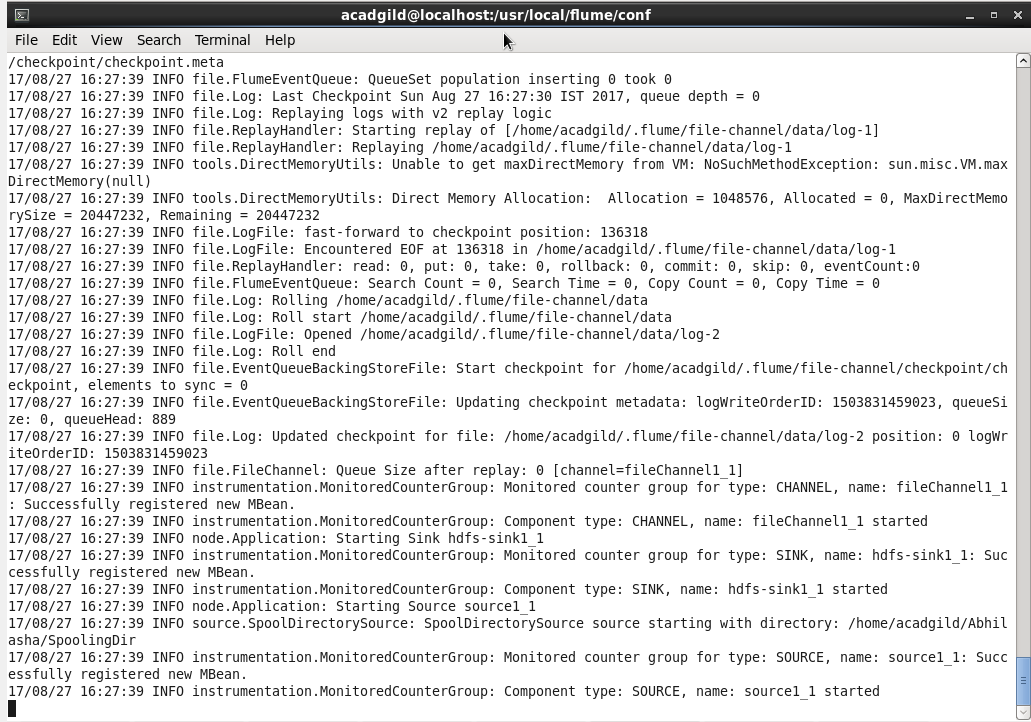
File format: currently SequenceFile, DataStream or CompressedStream

The conf file is executed as follows:

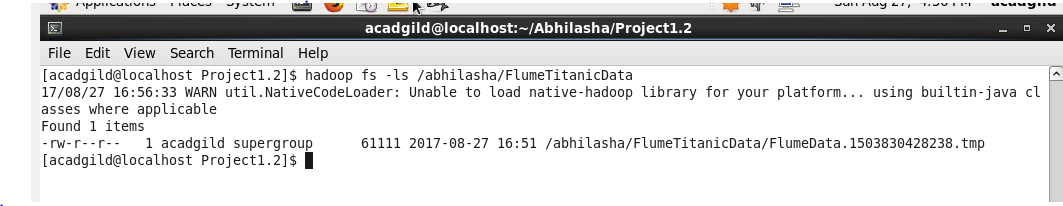


The path at which the file flume.conf is placed is ‘/usr/local/flume/conf’

The below screenshot mentions that flume is all set to ingest data from spooling directory.



Then we placed the test data in the spooling directory. This file got transferred to the destination HDFS path as follows:



This is how the dataset file is placed in hdfs using flume.