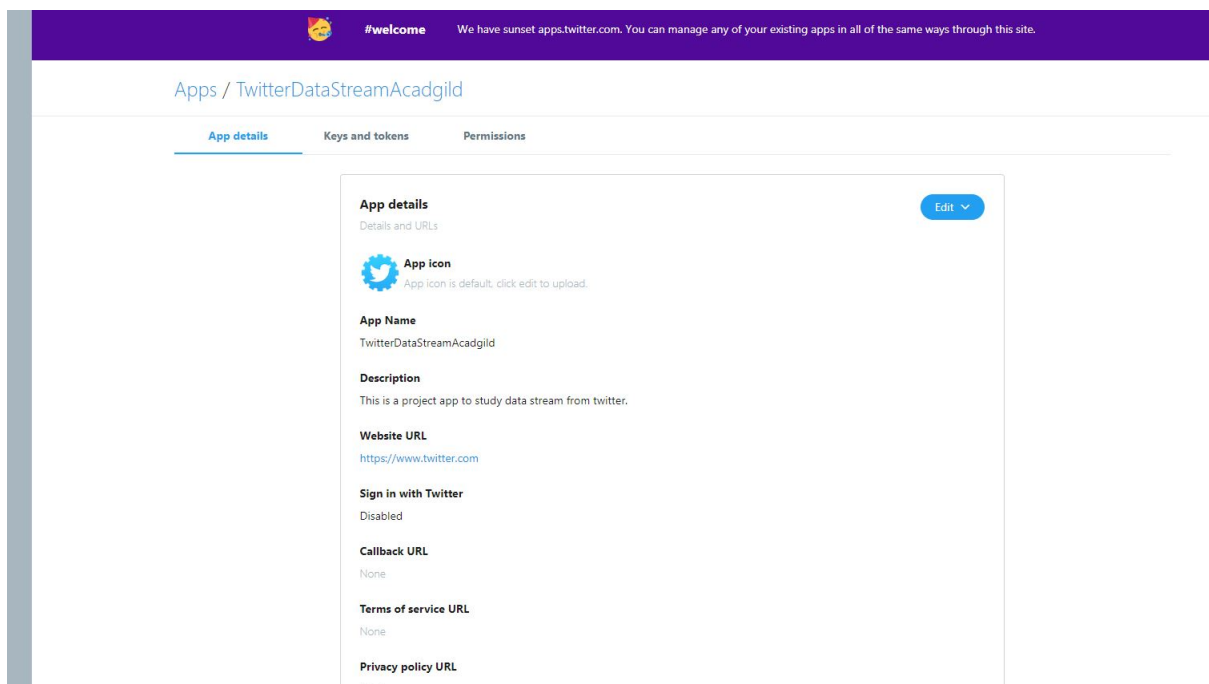


ASSIGNMENT 12.1

Create a flume agent that streams data from Twitter and stores in the HDFS.

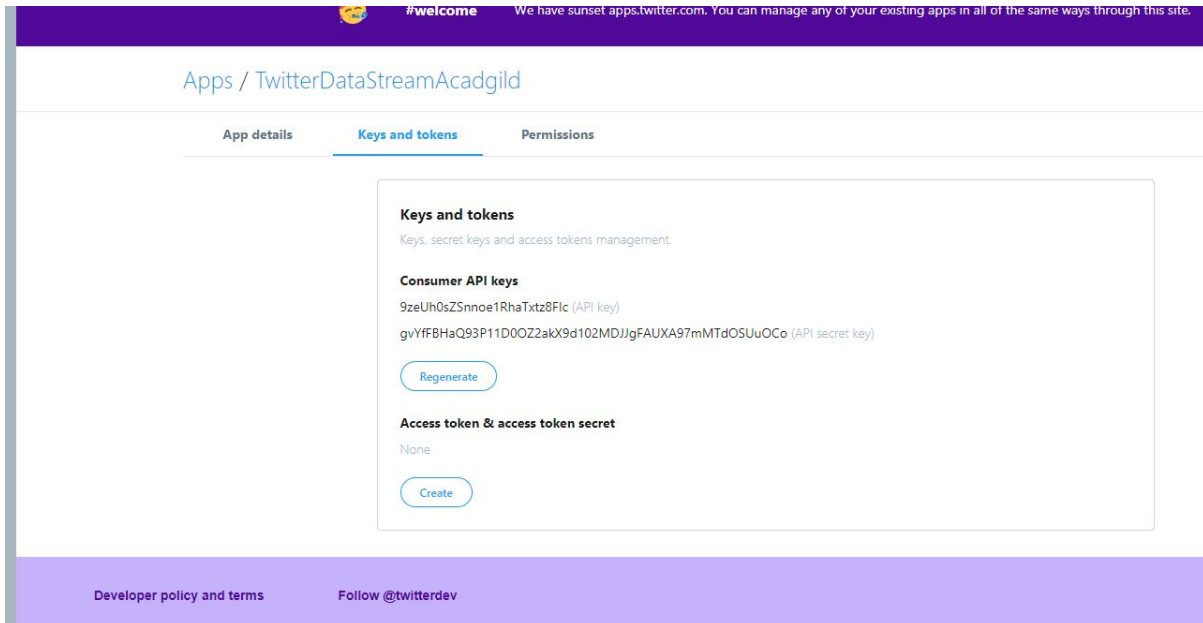
Login in to <https://apps.twitter.com/> and create a new app.



The screenshot shows the 'App details' page in the Twitter Developer Portal. The page has a purple header with a welcome message and a navigation bar with three tabs: 'App details' (selected), 'Keys and tokens', and 'Permissions'. The main content area displays the following information:

- App details** (with an 'Edit' button): Details and URLs
- App icon**: A Twitter logo icon with a note that the default icon is used and can be changed.
- App Name**: TwitterDataStreamAcadgild
- Description**: This is a project app to study data stream from twitter.
- Website URL**: <https://www.twitter.com>
- Sign in with Twitter**: Disabled
- Callback URL**: None
- Terms of service URL**: None
- Privacy policy URL**: None

Get the consumer key and consumer secret and generate the **access key** and **access token**.



The screenshot shows the Twitter Developer Portal interface. At the top, a purple banner contains the Twitter logo, a '#welcome' message, and a link to 'apps.twitter.com'. Below this, the breadcrumb 'Apps / TwitterDataStreamAcadgild' is visible. The main content area has three tabs: 'App details', 'Keys and tokens' (which is active), and 'Permissions'. The 'Keys and tokens' section is titled 'Keys and tokens' with a subtitle 'Keys, secret keys and access tokens management.' It contains two main sections: 'Consumer API keys' and 'Access token & access token secret'. The 'Consumer API keys' section displays two keys: '9zeUh0sZSnnoe1RhaTztz8Flc (API key)' and 'gvYfFBHaQ93P11D0OZ2akX9d102MDJjgFAUXA97mMTdOSUuOC0 (API secret key)', with a 'Regenerate' button below them. The 'Access token & access token secret' section shows 'None' and a 'Create' button. The footer of the page is purple and contains links for 'Developer policy and terms' and 'Follow @twitterdev'.

Make a directory for **tweetdata.conf** file in local file system

```
acdgild@localhost:~/tweetconf
[acdgild@localhost ~]$ mkdir tweetconf
[acdgild@localhost ~]$ cd tweetconf
You have new mail in /var/spool/mail/acdgild
[acdgild@localhost tweetconf]$ ls
tweetdata.conf
[acdgild@localhost tweetconf]$
```

Make an output directory **tweetdata** for output streaming data in hdfs

```
acadgild@localhost:~  
[acadgild@localhost ~]$ hdfs dfs -mkdir /tweetdata  
18/02/24 15:49:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...  
[acadgild@localhost ~]$ hdfs dfs -ls /tweetdata  
18/02/24 15:49:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...  
[acadgild@localhost ~]$ hdfs dfs -ls /  
18/02/24 15:49:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform...  
Found 5 items  
drwxr-xr-x - acadgild supergroup 0 2018-02-24 13:39 /hbase  
drwxr-xr-x - acadgild supergroup 0 2018-02-02 12:49 /sqoopout111  
drwxrwx--- - acadgild supergroup 0 2018-02-09 11:35 /tmp  
drwxr-xr-x - acadgild supergroup 0 2018-02-24 15:49 /tweetdata  
drwxr-xr-x - acadgild supergroup 0 2018-02-09 14:50 /user  
You have new mail in /var/spool/mail/acadgild  
[acadgild@localhost ~]$
```

Tweetdata.conf file

```
# Naming the components on the current agent.  
TwitterAgent.sources = Twitter  
TwitterAgent.channels = MemChannel  
TwitterAgent.sinks = HDFS  
  
# Describing/Configuring the source  
TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource  
  
TwitterAgent.sources.Twitter.consumerKey=nIa9JUBjTqLUWP2CnG2hzQOvm  
TwitterAgent.sources.Twitter.consumerSecret=mtOvwzkUP6pYdRkTMxs5fANMSjCjv2JqO2EPwg6f4QeziRoLHH  
TwitterAgent.sources.Twitter.accessToken=2245638606-taJAEc77EopDaedhPpEEwxElYOOQWRuWSWfJ47H  
TwitterAgent.sources.Twitter.accessTokenSecret=rYreOHPrjOFMstgHioImkZwbommwEoglUbP7Sw2C2H7To  
  
TwitterAgent.sources.Twitter.keywords = hadoop,bigdata,spark,flume,pig,MapReduce,Java,scala  
TwitterAgent.sources.Twitter.interceptors = il  
TwitterAgent.sources.Twitter.interceptors.il.type = timestamp  
  
# Describing/Configuring the sink  
  
TwitterAgent.sinks.HDFS.channel=MemChannel  
TwitterAgent.sinks.HDFS.type = hdfs  
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://localhost:54310/tweetdata  
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream  
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text  
TwitterAgent.sinks.HDFS.hdfs.batchSize = 100  
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0  
TwitterAgent.sinks.HDFS.hdfs.rollCount = 100  
TwitterAgent.sinks.HDFS.hdfs.rollInterval = 5  
  
# Describing/Configuring the channel  
TwitterAgent.channels.MemChannel.type = memory  
TwitterAgent.channels.MemChannel.capacity = 10000  
TwitterAgent.channels.MemChannel.transactionCapacity = 100  
  
# Binding the source and sink to the channel  
TwitterAgent.sources.Twitter.channels = MemChannel  
TwitterAgent.sinks.HDFS.channel = MemChannel  
acade@localhost:~/Desktop$
```

Below command in the screen shot will start fetching the data from twitter and streams it into HDFS given path.

```
@localhost:~$ flume-ng agent -n TwitterAgent -f /home/master/Desktop/tweetdata.conf
ag: No configuration directory set! Use --conf <dir> to override.
Including Hadoop libraries found via (/usr/local/hadoop/bin/hadoop) for HDFS access
Excluding /usr/local/hadoop/share/hadoop/common/lib/slf4j-api-1.7.10.jar from classpath
Excluding /usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar from classpath
Including HBASE libraries found via (/usr/local/hbase/bin/hbase) for HBASE access
Excluding /usr/local/hbase/lib/slf4j-api-1.7.7.jar from classpath
Excluding /usr/local/hadoop/share/hadoop/common/lib/slf4j-api-1.7.10.jar from classpath
Excluding /usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar from classpath
Including Hive libraries found via (/usr/local/hive) for Hive access
/usr/lib/jvm/java-8-openjdk-amd64/bin/java -Xmx20m -cp '/usr/local/flume/lib/*:/usr/local/hadoop/etc/hadoop:/usr
r:/usr/local/hadoop/share/hadoop/common/lib/apacheds-118n-2.0.0-M15.jar:/usr/local/hadoop/share/hadoop/common/li
p/share/hadoop/common/lib/api-asn1-api-1.0.0-M20.jar:/usr/local/hadoop/share/hadoop/common/lib/api-util-1.0.0-M2
jar:/usr/local/hadoop/share/hadoop/common/lib/avro-1.7.4.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-k
/lib/commons-beanutils-core-1.8.0.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-cli-1.2.jar:/usr/local/h
/local/hadoop/share/hadoop/common/lib/commons-collections-3.2.2.jar:/usr/local/hadoop/share/hadoop/common/lib/co
common/lib/commons-configuration-1.6.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-digester-1.8.jar:/usr
-3.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-io-2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/
on/lib/commons-logging-1.1.3.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/ha
local/hadoop/share/hadoop/common/lib/curator-client-2.7.1.jar:/usr/local/hadoop/share/hadoop/common/lib/curator-fr
lib/curator-recipes-2.7.1.jar:/usr/local/hadoop/share/hadoop/common/lib/gson-2.2.4.jar:/usr/local/hadoop/share/h
e/hadoop/common/lib/hadoop-annotations-2.7.2.jar:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.7.2.jar
-1.3.jar:/usr/local/hadoop/share/hadoop/common/lib/htrace-core-3.1.0-incubating.jar:/usr/local/hadoop/share/had
re/hadoop/common/lib/httpcore-4.2.5.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-core-asl-1.9.13.jar:/u
9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop
t/hadoop/common/lib/java-xmlbuilder-0.4.jar:/usr/local/hadoop/share/hadoop/common/lib/jaxb-api-2.2.2.jar:/usr/loc
/usr/local/hadoop/share/hadoop/common/lib/jersey-core-1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-j
ersey-server-1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/jets3t-0.9.0.jar:/usr/local/hadoop/share/hadoop/co
common/lib/jetty-6.1.26.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-util-6.1.26.jar:/usr/local/hadoop/s
/share/hadoop/common/lib/jsp-api-2.1.jar:/usr/local/hadoop/share/hadoop/common/lib/jsr305-3.0.0.jar:/usr/local/h
hadoop/share/hadoop/common/lib/log4j-1.2.17.jar:/usr/local/hadoop/share/hadoop/common/lib/mockito-all-1.8.5.jar:
nal.jar:/usr/local/hadoop/share/hadoop/common/lib/paranamer-2.3.jar:/usr/local/hadoop/share/hadoop/common/lib/pr
mon/lib/servlet-api-2.5.jar:/usr/local/hadoop/share/hadoop/common/lib/snappy-java-1.0.4.1.jar:/usr/local/hadoop
hadoop/share/hadoop/common/lib/xmlenc-0.52.jar:/usr/local/hadoop/share/hadoop/common/lib/xz-1.0.jar:/usr/local/had
al/hadoop/share/hadoop/common/hadoop-common-2.7.2.jar:/usr/local/hadoop/share/hadoop/common/hadoop-common-2.7.2-
fs-2.7.2.jar:/usr/local/hadoop/share/hadoop/common/jdiff:/usr/local/hadoop/share/hadoop/common/lib:/usr/local/h

reduce-examples-2.7.2.jar:/usr/local/hadoop/share/hadoop/mapreduce/lib:/usr/local/hadoop/share/hadoop/mapreduce/lib-examples:/usr/local/hadoop/share/hadoop/m
urces:/usr/local/hadoop/contrib/capacity-scheduler/*.jar:/usr/local/hbase/conf:/usr/local/hive/lib/*' -Djava.library.path=/usr/local/hadoop/lib/native:/usr/
p/lib/native org.apache.flume.node.Application -n TwitterAgent -f /home/master/Desktop/tweetdata.conf
18/03/09 14:57:53 INFO node.PollingPropertiesFileConfigurationProvider: Configuration provider starting
18/03/09 14:57:53 INFO node.PollingPropertiesFileConfigurationProvider: Reloading configuration file:/home/master/Desktop/tweetdata.conf
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Processing:HDFS
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Processing:HDFS
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Processing:HDFS
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Processing:HDFS
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Processing:HDFS
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Added sinks: HDFS Agent: TwitterAgent
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Processing:HDFS
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Processing:HDFS
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Processing:HDFS
18/03/09 14:57:53 INFO conf.FlumeConfiguration: Post-validation flume configuration contains configuration for agents: [TwitterAgent]
18/03/09 14:57:53 INFO node.AbstractConfigurationProvider: Creating channels
18/03/09 14:57:53 INFO channel.DefaultChannelFactory: Creating instance of channel MemChannel type memory
18/03/09 14:57:53 INFO node.AbstractConfigurationProvider: Created channel MemChannel
18/03/09 14:57:53 INFO source.DefaultSourceFactory: Creating instance of source Twitter, type org.apache.flume.source.twitter.TwitterSource
18/03/09 14:57:53 INFO twitter.TwitterSource: Consumer Key: 'nIa9JUBjTqLUWP2CnG2hzQ0vm'
18/03/09 14:57:53 INFO twitter.TwitterSource: Consumer Secret: 'mtOvzkUP6pYdRkTmx5fANMSjCjv2JqO2EPwg6f4Qez1RoLHh'
18/03/09 14:57:53 INFO twitter.TwitterSource: Access Token: '2245638606-taJAEc77Eop0aedhPPEEwxE1Y0OQWRuW5WfJ47H'
18/03/09 14:57:53 INFO twitter.TwitterSource: Access Token Secret: 'rYreOHP7rjOfMtqHicImK2wbomwEogLUbP7Sw2C2H7To'
18/03/09 14:57:53 INFO sink.DefaultSinkFactory: Creating instance of sink: HDFS, type: hdfs
18/03/09 14:57:53 INFO node.AbstractConfigurationProvider: Channel MemChannel connected to [Twitter, HDFS]
18/03/09 14:57:53 INFO node.Application: Starting new configuration: ( sourceRunners: [Twitter=EventDrivenSourceRunner: ( source:org.apache.flume.source.twitter
urce(name:Twitter,state:IDLE) ) ] sinkRunners: [HDFS=SinkRunner: ( policy:org.apache.flume.sink.DefaultSinkProcessor@2f5ca0cf counterGroup:( name:null counters
annels:(MemChannel=org.apache.flume.channel.MemoryChannel(name: MemChannel)) ) )
18/03/09 14:57:53 INFO node.Application: Starting Channel MemChannel
18/03/09 14:57:53 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: CHANNEL, name: MemChannel: Successfully registered new MBean.
18/03/09 14:57:53 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started
18/03/09 14:57:53 INFO node.Application: Starting Sink HDFS
18/03/09 14:57:53 INFO node.Application: Starting Source Twitter
18/03/09 14:57:53 INFO twitter.TwitterSource: Starting twitter source org.apache.flume.source.twitter.TwitterSource(name:Twitter,state:IDLE) ...
18/03/09 14:57:53 INFO instrumentation.MonitoredCounterGroup: Monitored counter group for type: SINK, name: HDFS: Successfully registered new MBean.
18/03/09 14:57:53 INFO instrumentation.MonitoredCounterGroup: Component type: SINK, name: HDFS started
18/03/09 14:57:53 INFO twitter.TwitterSource: Twitter source Twitter started.
18/03/09 14:57:53 INFO twitter4j.TwitterStreamImpl: Establishing connection.
18/03/09 14:57:56 INFO twitter4j.TwitterStreamImpl: Connection established.
18/03/09 14:57:56 INFO twitter4j.TwitterStreamImpl: Receiving status stream.
18/03/09 14:57:56 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/03/09 14:57:56 INFO hdfs.BucketWriter: Creating hdfs://localhost:54310/tweetdata/FlumeData.1520587676565.tmp
```

```

565
18/03/09 14:58:03 INFO hdfs.HDFSEventSink: Writer callback called.
18/03/09 14:58:04 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/03/09 14:58:04 INFO hdfs.BucketWriter: Creating hdfs://localhost:54310/tweetdata/FlumeData.1520587684467.tmp
18/03/09 14:58:05 INFO twitter.TwitterSource: Processed 300 docs
18/03/09 14:58:08 INFO twitter.TwitterSource: Processed 400 docs
18/03/09 14:58:09 INFO hdfs.BucketWriter: Closing hdfs://localhost:54310/tweetdata/FlumeData.1520587684467.tmp
18/03/09 14:58:09 INFO hdfs.BucketWriter: Renaming hdfs://localhost:54310/tweetdata/FlumeData.1520587684467.tmp to hdfs://localhost:54310/tweetdata/
467
18/03/09 14:58:09 INFO hdfs.HDFSEventSink: Writer callback called.
18/03/09 14:58:10 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/03/09 14:58:10 INFO hdfs.BucketWriter: Creating hdfs://localhost:54310/tweetdata/FlumeData.1520587690493.tmp
18/03/09 14:58:11 INFO twitter.TwitterSource: Processed 500 docs
18/03/09 14:58:14 INFO twitter.TwitterSource: Processed 600 docs
18/03/09 14:58:15 INFO hdfs.BucketWriter: Closing hdfs://localhost:54310/tweetdata/FlumeData.1520587690493.tmp
18/03/09 14:58:15 INFO hdfs.BucketWriter: Renaming hdfs://localhost:54310/tweetdata/FlumeData.1520587690493.tmp to hdfs://localhost:54310/tweetdata/
493
18/03/09 14:58:15 INFO hdfs.HDFSEventSink: Writer callback called.
18/03/09 14:58:16 INFO hdfs.HDFSDataStream: Serializer = TEXT, UseRawLocalFileSystem = false
18/03/09 14:58:16 INFO twitter.TwitterSource: Processed 700 docs
18/03/09 14:58:16 INFO hdfs.BucketWriter: Creating hdfs://localhost:54310/tweetdata/FlumeData.1520587696723.tmp
18/03/09 14:58:19 INFO twitter.TwitterSource: Processed 800 docs
18/03/09 14:58:21 INFO hdfs.BucketWriter: Closing hdfs://localhost:54310/tweetdata/FlumeData.1520587696723.tmp
18/03/09 14:58:22 INFO hdfs.BucketWriter: Renaming hdfs://localhost:54310/tweetdata/FlumeData.1520587696723.tmp to hdfs://localhost:54310/tweetdata/
723
18/03/09 14:58:22 INFO hdfs.HDFSEventSink: Writer callback called.

```

Output files of twitter data in output directory.



Browse Directory

/tweetdata								Go!
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
-rw-r--r--	master	supergroup	104.44 KB	3/9/2018, 2:58:03 PM	1	128 MB	FlumeData.1520587676565	
-rw-r--r--	master	supergroup	97.78 KB	3/9/2018, 2:58:09 PM	1	128 MB	FlumeData.1520587684467	
-rw-r--r--	master	supergroup	113.56 KB	3/9/2018, 2:58:15 PM	1	128 MB	FlumeData.1520587690493	
-rw-r--r--	master	supergroup	91.86 KB	3/9/2018, 2:58:21 PM	1	128 MB	FlumeData.1520587696723	
-rw-r--r--	master	supergroup	100.88 KB	3/9/2018, 2:58:27 PM	1	128 MB	FlumeData.1520587702435	
-rw-r--r--	master	supergroup	95.09 KB	3/9/2018, 2:58:32 PM	1	128 MB	FlumeData.1520587707723	
-rw-r--r--	master	supergroup	106.96 KB	3/9/2018, 2:58:38 PM	1	128 MB	FlumeData.1520587713445	
-rw-r--r--	master	supergroup	74.87 KB	3/9/2018, 2:58:44 PM	1	128 MB	FlumeData.1520587719489	

Output :-

[illegible]

