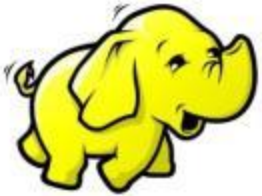
edureka!

Big Data & Hadoop



Connect with us

edureka!

- ✓ **24x7 Support** on Skype, Email & Phone
- ✓ **Skype ID – edureka.hadoop**
- ✓ **Email – hadoop@edureka.in**
- ✓ **Call us – +91 88808 62004**

Course Topics

edureka!

- ✓ Week 1
  - Introduction to HDFS
- ✓ Week 2
  - Setting Up Hadoop Cluster
- ✓ Week 3
  - Map-Reduce Basics, types and formats
- ✓ Week 4
  - PIG
- ✓ Week 5
  - HIVE
- ✓ Week 6
  - HBASE
- ✓ Week 7
  - ZOOKEEPER & SQOOP
- ✓ Week 8
  - Project Discussion

Solution Characteristics

edureka!

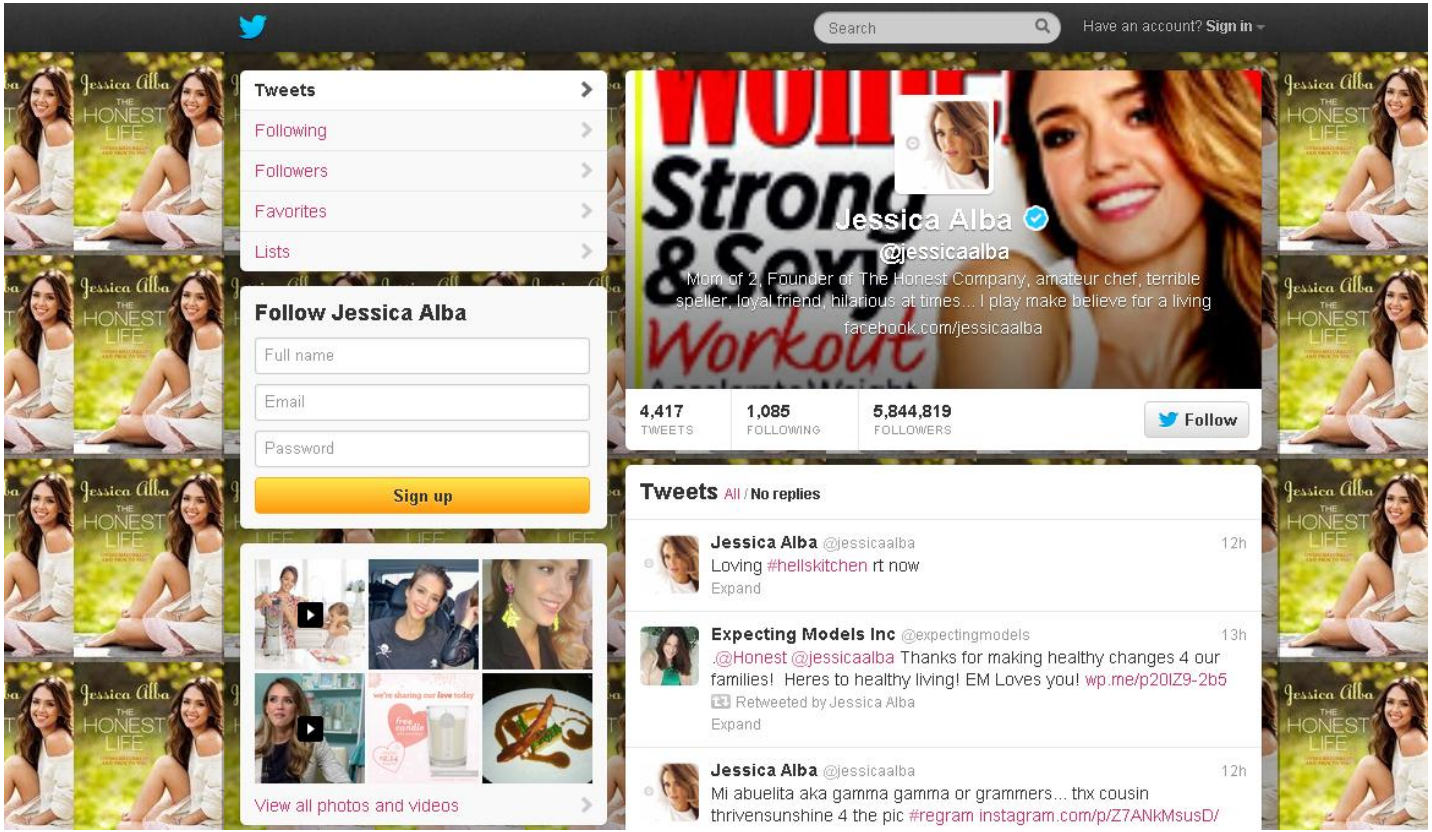
- ✓ Handle Big data
- ✓ Handle JSON data
- ✓ System should be able to fetch data from twitter
- ✓ Quick analytical capabilities

Who Is The Most Influential Celebrity?

edureka!

Problem Definition

edureka!

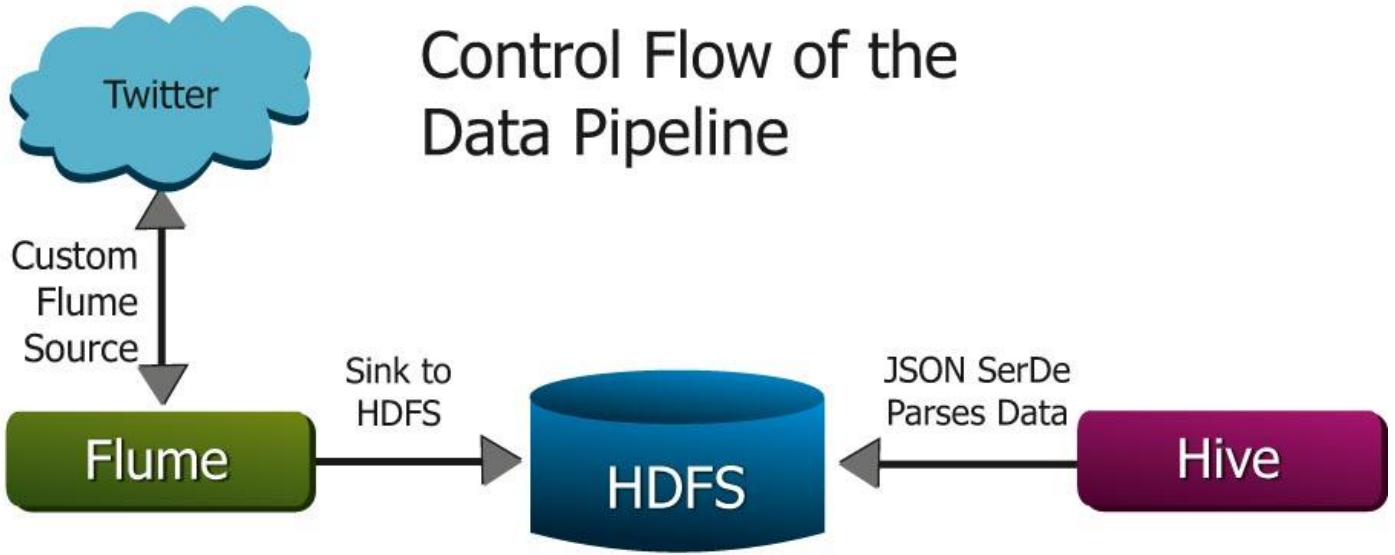


Problem Definition

edureka!

Social media has gained immense popularity with marketing teams, and Twitter is an effective tool for a company to get people excited about its products. Twitter makes it easy to engage users and communicate directly with them, and in turn, users can provide word-of-mouth marketing for companies by discussing the products. Given limited resources, and knowing we may not be able to talk to everyone we want to target directly, marketing departments can be more efficient by being selective about whom we reach out to.

Solution



Step By Step Guide

edureka!

Download Flume from http://flume.apache.org/download.html

Download

Apache Flume is distributed under the Apache License, version 2.0

The link in the Mirrors column should display a list of available mirrors with a default selection based on your inferred location. If you do not see that page different browser. The checksum and signature are links to the originals on the main distribution server.

	Mirrors	Checksum	Signature
Apache Flume binary (tar.gz)	apache-flume-1.3.1-bin.tar.gz	apache-flume-1.3.1-bin.tar.gz.md5	apache-flume-1.3.1-bin.tar.gz.asc
Apache Flume source (tar.gz)	apache-flume-1.3.1-src.tar.gz	apache-flume-1.3.1-src.tar.gz.md5	apache-flume-1.3.1-src.tar.gz.asc

Step 2

edureka!

- Unzip the flume contents in /usr/lib/flume
- Create flume-env.sh from template file

  cp /etc/flume-ng/conf/flume-env.sh.template /etc/flume-ng/conf/flume-env.sh

Step 3

edureka!

- Create flume.conf in /path-to-flume/conf/ directory and copy following file

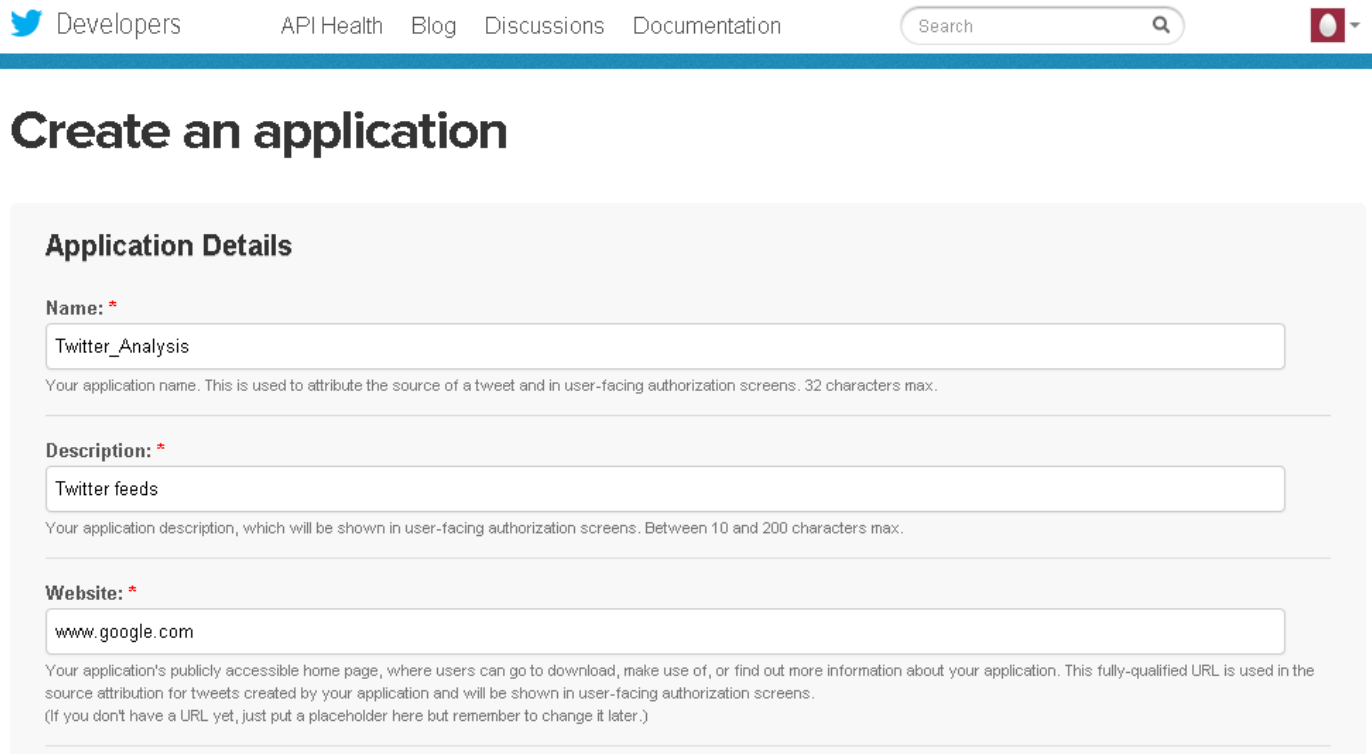
```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey = [required]
TwitterAgent.sources.Twitter.consumerSecret = [required]
TwitterAgent.sources.Twitter.accessToken = [required]
TwitterAgent.sources.Twitter.accessTokenSecret = [required]
TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientist, business intelligence, mapreduce, data
warehouse, data warehousing, mahout, hbase, nosql, newsql, businessintelligence, cloudcomputing

TwitterAgent.sinks.HDFS.channel = MemChannel
TwitterAgent.sinks.HDFS.type = hdfs
TwitterAgent.sinks.HDFS.hdfs.path = hdfs://hadoop1:8020/user/flume/tweets/%Y/%m/%d/%H/
TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream
TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text
TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000
TwitterAgent.sinks.HDFS.hdfs.rollSize = 0
TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000
TwitterAgent.sinks.HDFS.hdfs.rollInterval = 600

TwitterAgent.channels.MemChannel.type = memory
TwitterAgent.channels.MemChannel.capacity = 10000
TwitterAgent.channels.MemChannel.transactionCapacity = 100
```

Step 4.Register with Twitter



Step 5. Create Acess token

edureka!

Consumer key	nrObOHI1ErKL6saFKDhVZQ
Consumer secret	uc91pLc1vqCLu3avmrd45gsW4gh72w2vh13C7dW9E
Request token URL	https://api.twitter.com/oauth/request_token
Authorize URL	https://api.twitter.com/oauth/authorize
Access token URL	https://api.twitter.com/oauth/access_token
Callback URL	None
Sign in with Twitter	No

Your access token

It looks like you haven't authorized this application for your own Twitter account yet. For your convenience, we give you the opportunity to create your OAuth access token here, so you can start signing your requests right away. The access token generated will reflect your application's current permission level.

Create my access token

Step 6

edureka!

- Copy access keys information in Configuration file
- Add Flume jar in Flume classpath variable
- $ hadoop fs -mkdir /user/flume/tweets
- $ hadoop fs -chown -R flume:flume /user/flume
- $ hadoop fs -chmod -R 770 /user/flume
- Run Flume using

```
bin/flume-ng agent -n TwitterAgent -c conf -f conf/flume.conf
```

Data will start accumulating in HDFS

Hive Setup

edureka!

```
$ sudo -u <your-hdfs-user> hadoop fs -mkdir /user/hive/warehouse

$ sudo -u <your-hdfs-user> hadoop fs -chown -R hive:hive /user/hive

$ sudo -u <your-hdfs-user> hadoop fs -chmod 750 /user/hive

$ sudo -u <your-hdfs-user> hadoop fs -chmod 770 /user/hive/warehouse
```

Problem Definition

edureka!

```
ADD JAR <path-to-hive-serdes-jar>;

CREATE EXTERNAL TABLE tweets ( id BIGINT, created_at STRING, source STRING, favorited BOOLEAN,
retweeted_status STRUCT< text:STRING, user:STRUCT<screen_name:STRING,name:STRING>,
retweet_count:INT>, entities STRUCT< urls:ARRAY<STRUCT<expanded_url:STRING>>,
user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
hashtags:ARRAY<STRUCT<text:STRING>>>, text STRING, user STRUCT< screen_name:STRING,
name:STRING, friends_count:INT, followers_count:INT, statuses_count:INT, verified:BOOLEAN,
utc_offset:INT, time_zone:STRING>, in_reply_to_screen_name STRING )
ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe‘
LOCATION '/user/flume/tweets';
```

Load Data in Hive

edureka!

```
LOAD DATA INPATH '/user/flume/tweets/2013/02/25/16' INTO TABLE `default.tweets`
```

Problem Definition

edureka!

```
SELECT   t.retweeted_screen_name,
  sum(retweets) AS total_retweets,
  count(*) AS tweet_count FROM    (
            SELECT  retweeted_status.user.screen_name as retweeted_screen_name ,
            retweeted_status.text,
            max(retweet_count) as retweets
            FROM tweets
            GROUP BY
            retweeted_status.user.screen_name,  retweeted_status.text) t
GROUP BY t.retweeted_screen_name
ORDER BY total_retweets DESC
LIMIT 10;
```

Clarifications

edureka!

Q & A..?

edureka!

Thank You