Trending movie analyzer

USING TWITTER FEEDS

Overview

1. What movies to track ?
2. How to retrieve relevant data from twitter?
3. What data structures to use for holding data? (In memory vs Disk)
4. How to perform sentiment analysis?
5. What are the features to store in order to answer future queries?
6. How does the overall architecture look like?

What movies to track ?

1. Manual or automatic discovery of new movies?
2. A data source to hold all the movies that we were planning to track. Periodical updates required!

How to retrieve relevant data from twitter?

1. Twitter has excellent support for Streaming API.

2. Repeatedly polling the data by making connection is expensive, can we do better?

3. HBC (Hosebird client)

https://github.com/twitter/hbc

What data structures to use for holding data?

In memory queue **vs** Disk File. Pros and cons? (consider hive's nature of batch processing here)

TODO: Use storms spout for stream processing and bolt for next stage(s).

How to perform sentiment analysis?

Difficult task.

Made even more difficult with twitter's unstructured data.

Typical ideas: Use polarity for each word/adjective. Classify incoming message into one of the several classes.

Data source:

http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#datasets

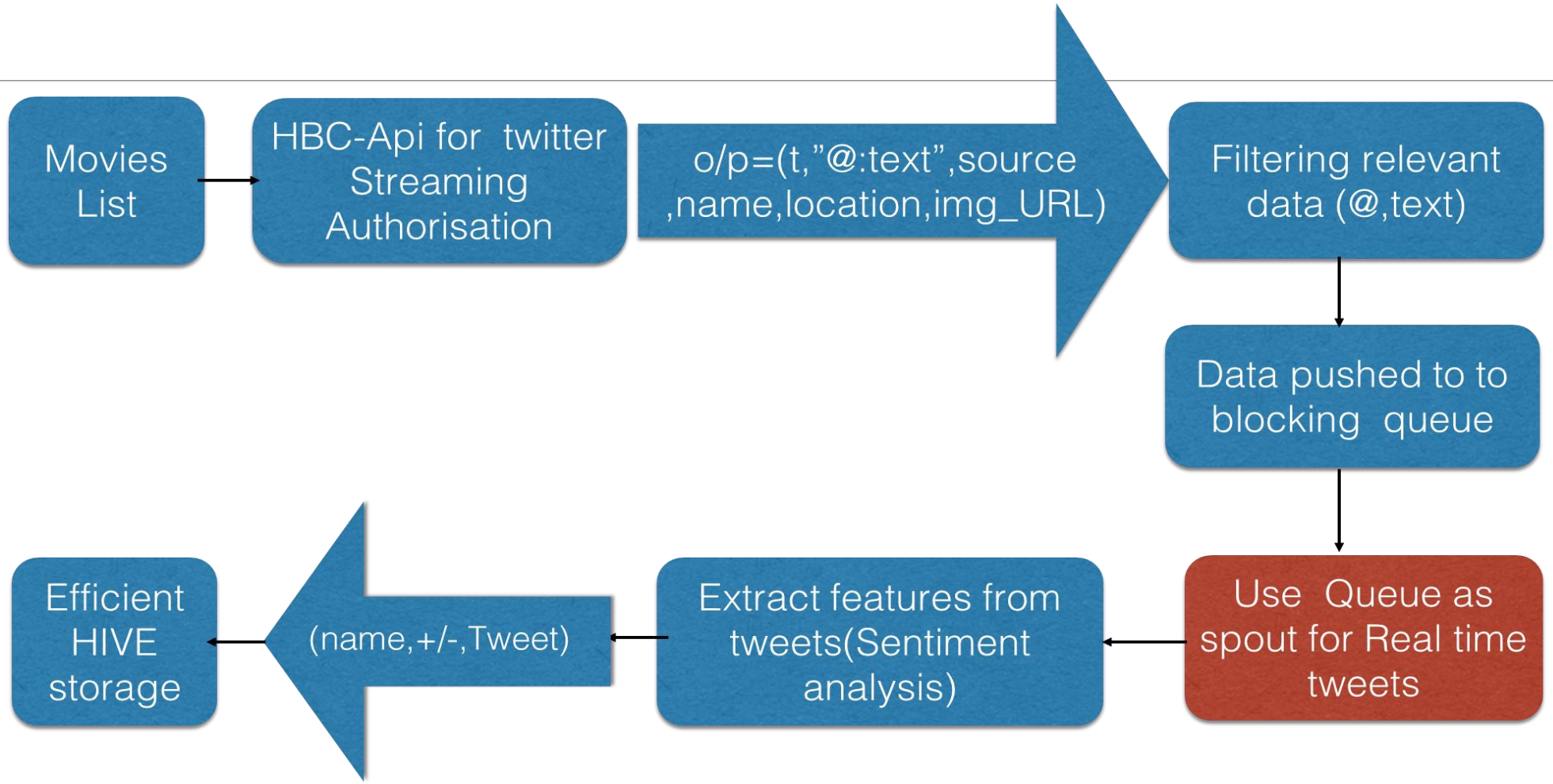
Stored vector

<Timestamp , Movie name, Sentiment, Original tweet>

Significance of each attribute?

Where to store? Advantages of each approach?

Big picture


Movies List
HBC-Api for twitter Streaming Authorisation
o/p=(t,"@:text",source ,name,location,img_URL)
Filtering relevant data (@,text)
Data pushed to to blocking queue
Use Queue as spout for Real time tweets
Extract features from tweets(Sentiment analysis)
(name,+/-,Tweet)
Efficient HIVE storage