## Hadoop, Pig, and Twitter

Kevin Weil -- @kevinweil Analytics Lead, Twitter



#### Introduction

- Hadoop Overview
- Why Pig?
- Evolution of Data Processing at Twitter
- Pig for Counting
- Pig for Correlating
- Pig for Research and Data Mining
- Conclusions and Next Steps

## My Background

- Studied Mathematics and Physics at Harvard, Physics at Stanford
- Tropos Networks (city-wide wireless): mesh routing algorithms, GBs of data
- Cooliris (web media): Hadoop and Pig for analytics, TBs of data
- Twitter: Hadoop, Pig, machine learning, visualization, social graph analysis, ??? of data

#### Introduction

- Hadoop Overview
- Why Pig?
- Evolution of Data Processing at Twitter
- Pig for Counting
- Pig for Correlating
- Pig for Research and Data Mining
- Conclusions and Next Steps

## Data is Getting Big

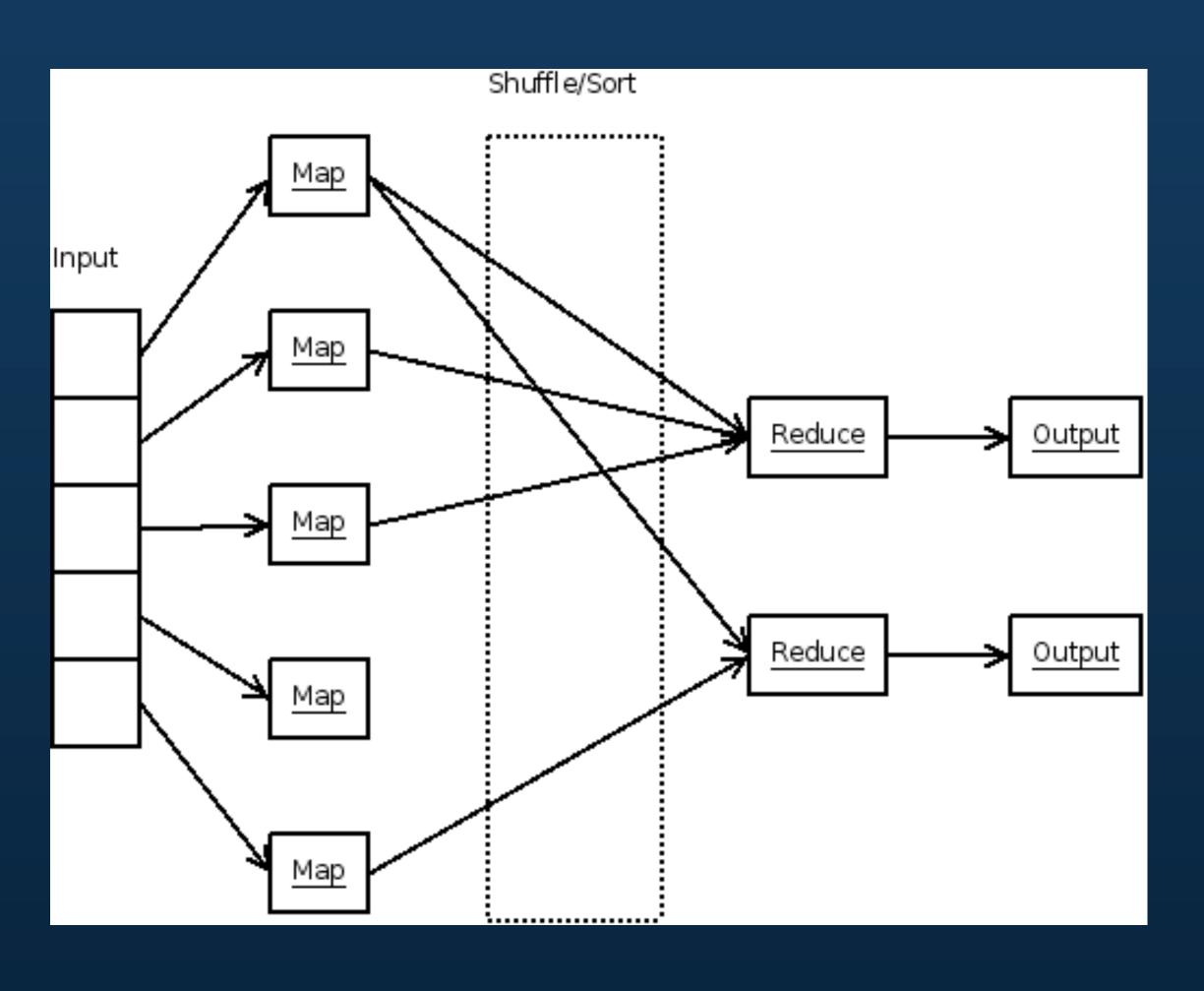
- NYSE: 1 TB/day
- Facebook: 20+ TB compressed/day
- CERN/LHC: 40 TB/day (15 PB/year!)
- And growth is accelerating
- Need multiple machines, horizontal scalability



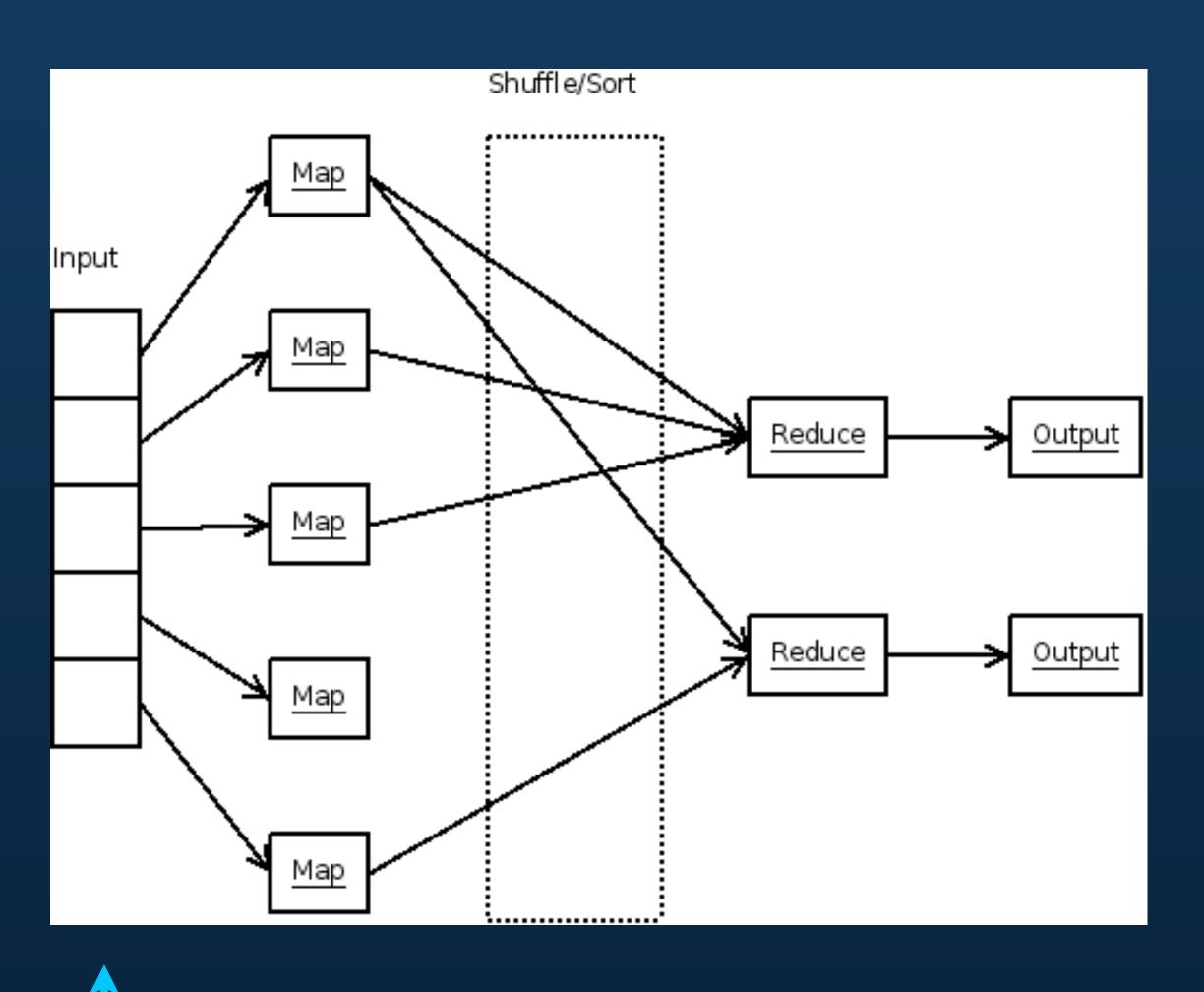
#### Hadoop



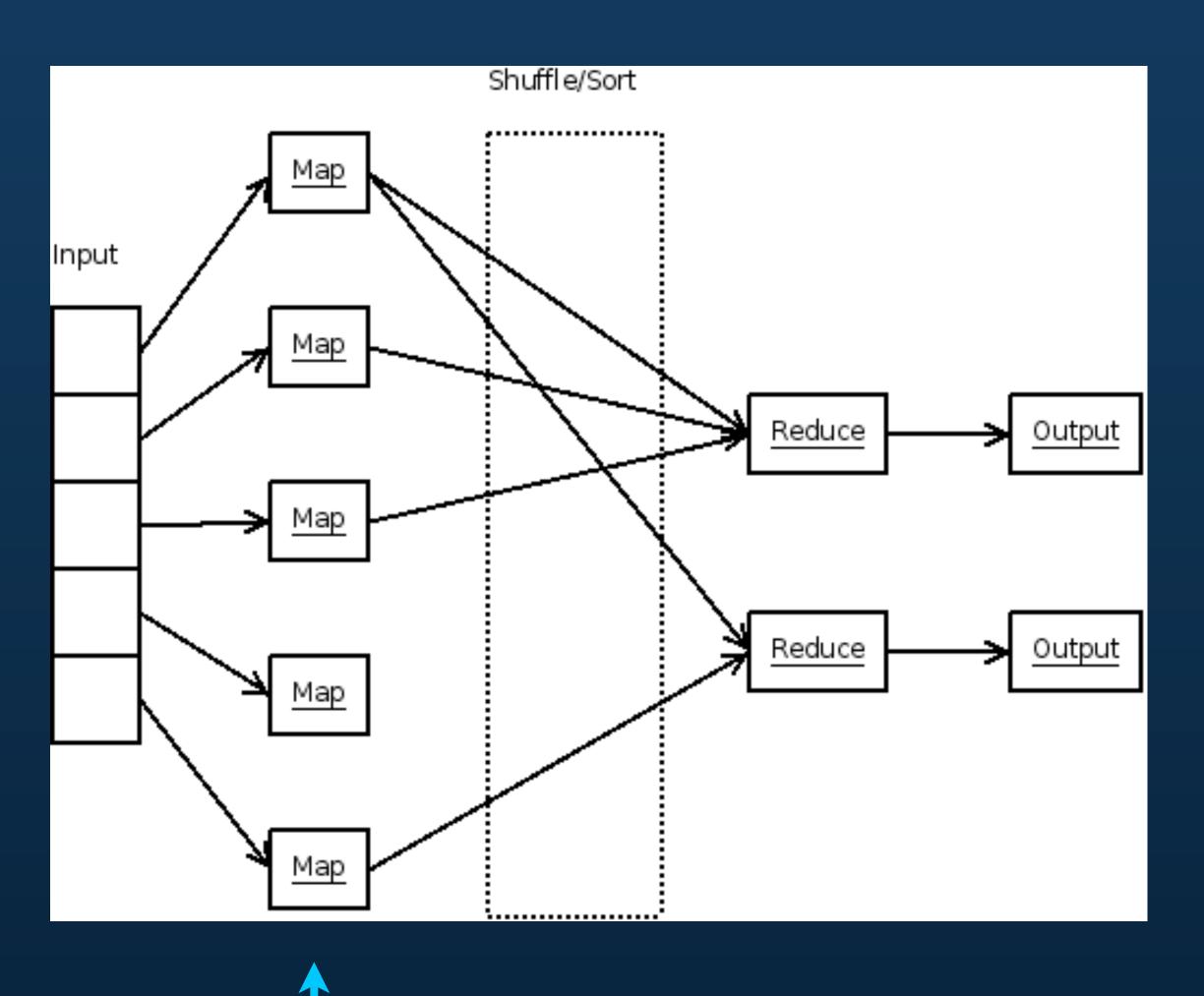
- Distributed file system (hard to store a PB)
- Fault-tolerant, handles replication, node failure, etc
- MapReduce-based parallel computation (even harder to process a PB)
- Generic key-value based computation interface allows for wide applicability
- Open source, top-level Apache project
- Scalable: Y! has a 4000-node cluster
- Powerful: sorted a TB of random integers in 62 seconds



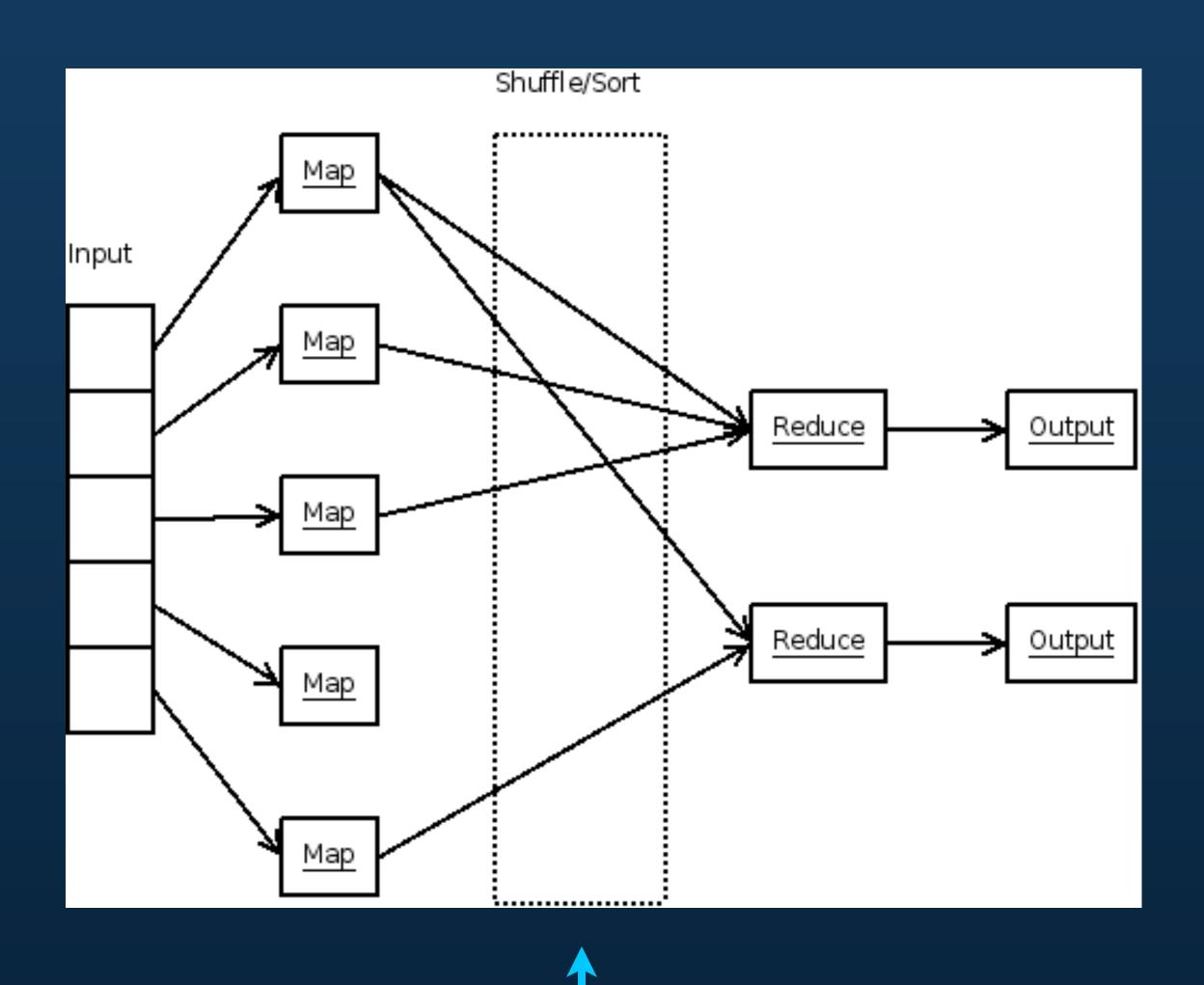
- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user\_id, value=1
- Shuffle: sort by user\_id
- Reduce: for each user\_id, sum
- Output: user\_id, tweet count
- With 2x machines, runs close to 2x faster.



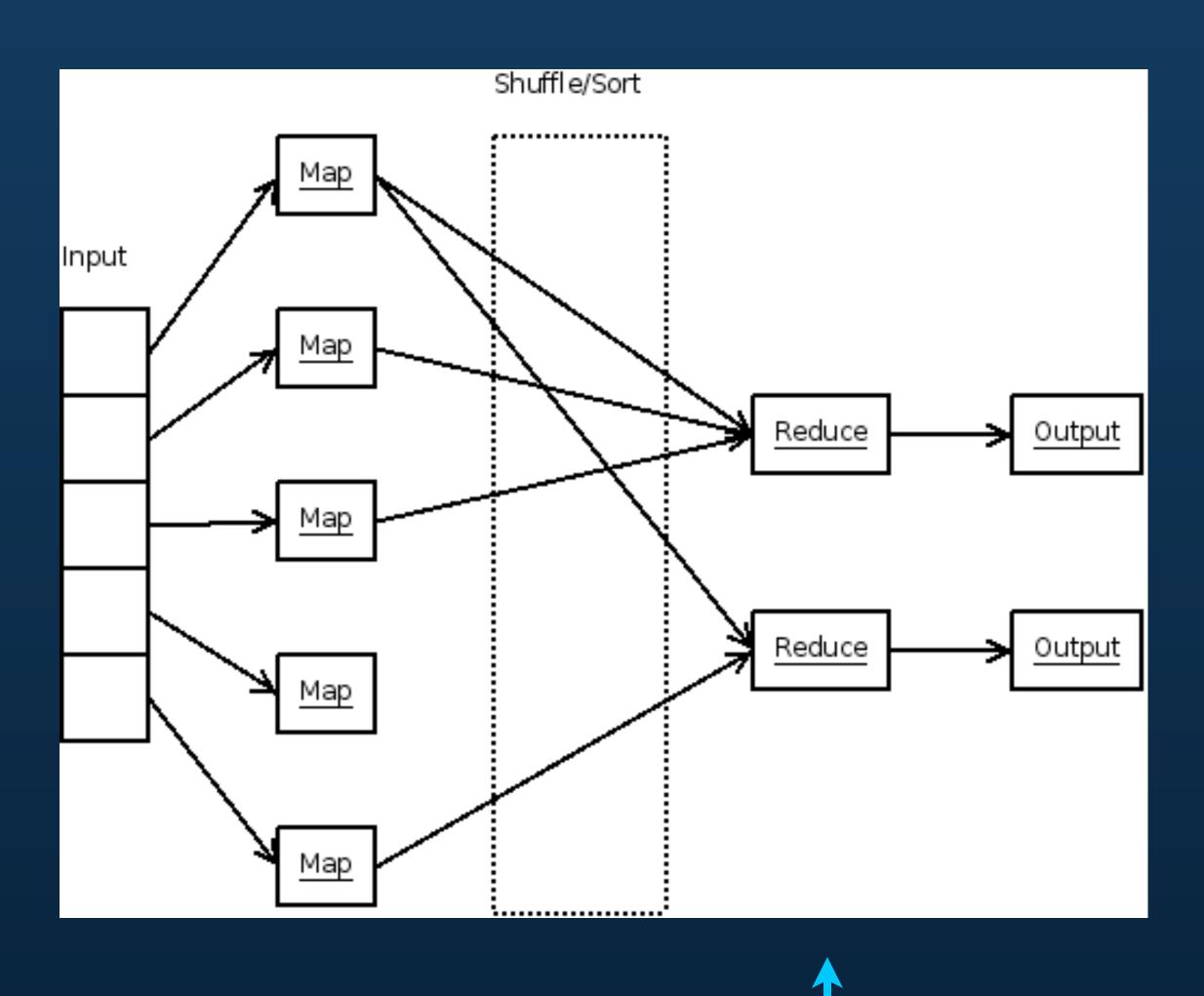
- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user\_id, value=1
- Shuffle: sort by user\_id
- Reduce: for each user\_id, sum
- Output: user\_id, tweet count
- With 2x machines, runs close to 2x faster.



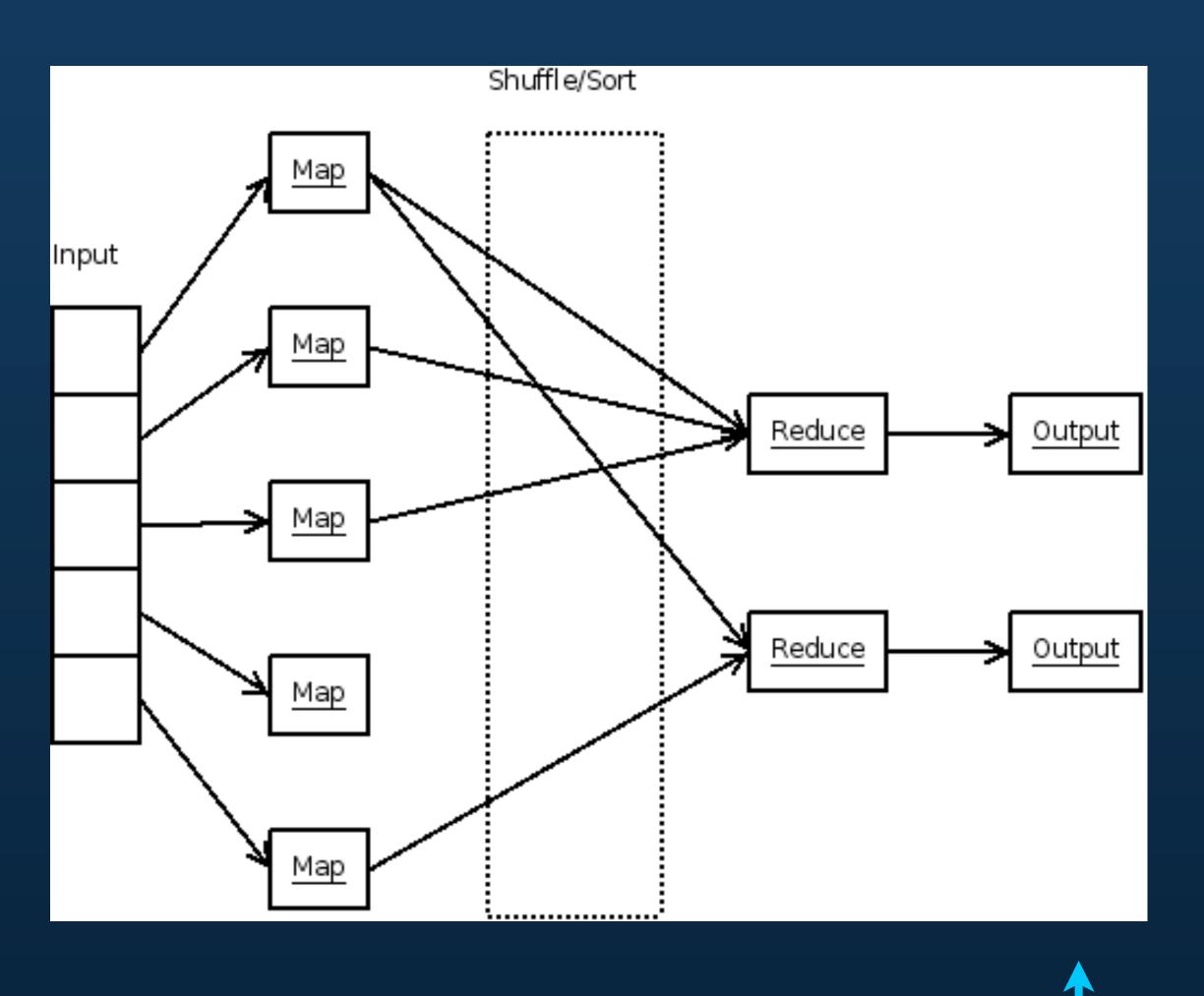
- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user\_id, value=1
- Shuffle: sort by user\_id
- Reduce: for each user\_id, sum
- Output: user\_id, tweet count
- With 2x machines, runs close to 2x faster.



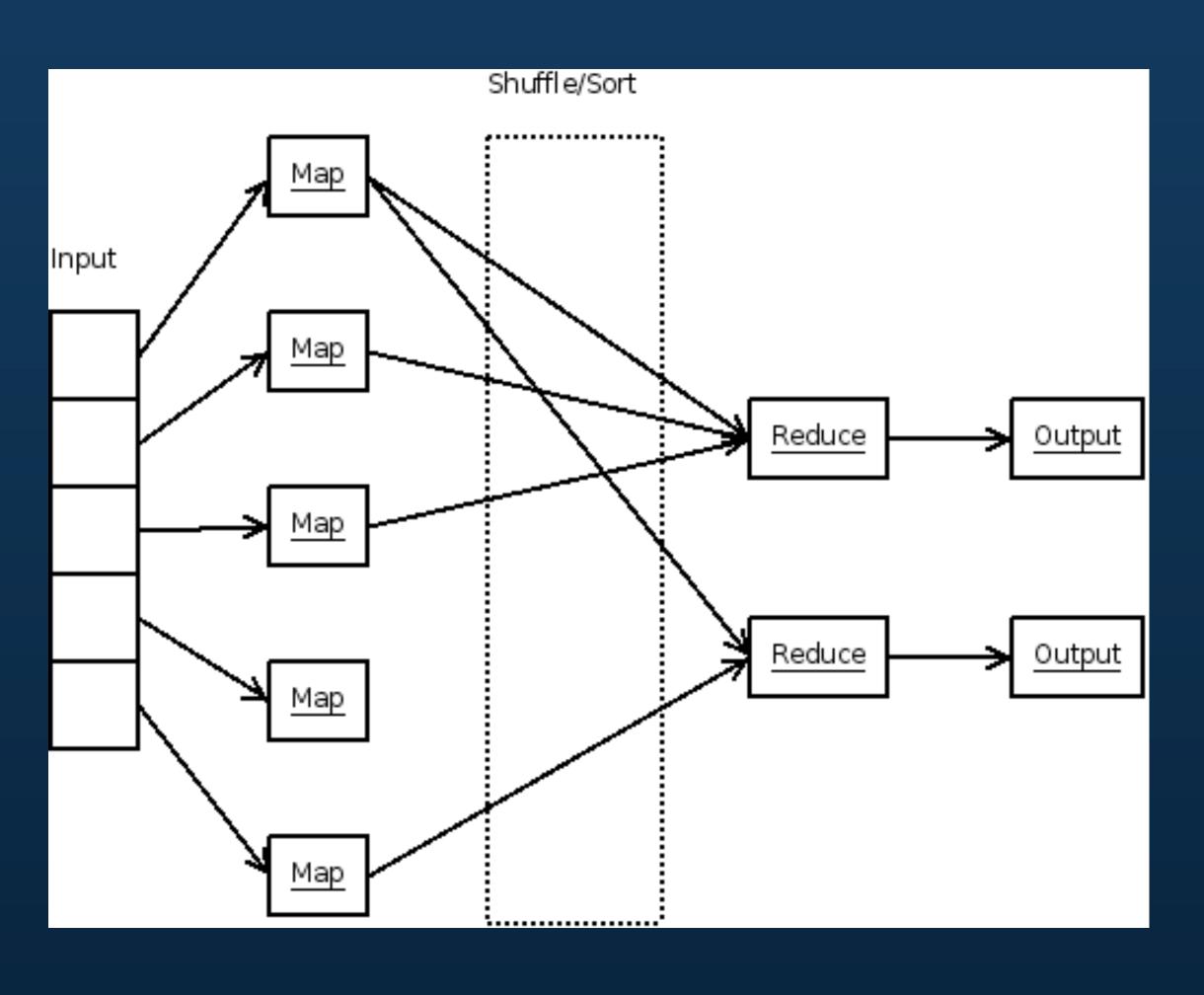
- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user\_id, value=1
- Shuffle: sort by user\_id
- Reduce: for each user\_id, sum
- Output: user\_id, tweet count
- With 2x machines, runs close to 2x faster.



- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user\_id, value=1
- Shuffle: sort by user\_id
- Reduce: for each user\_id, sum
- Output: user\_id, tweet count
- With 2x machines, runs close to 2x faster.



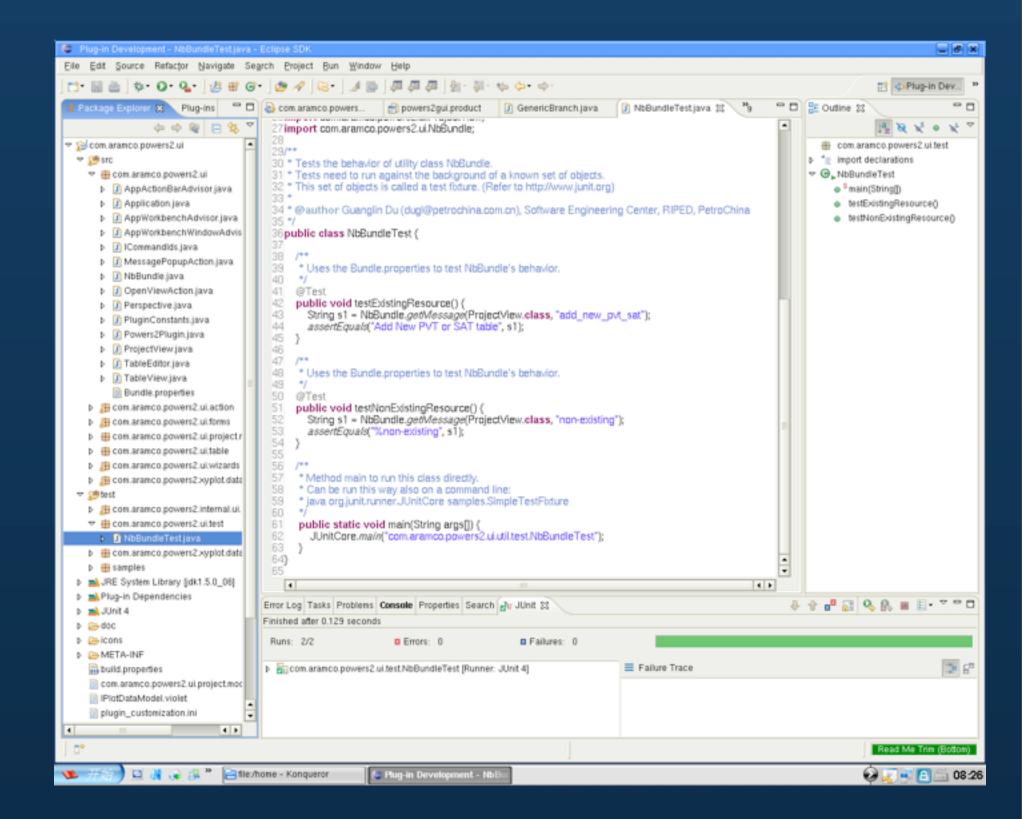
- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user\_id, value=1
- Shuffle: sort by user\_id
- Reduce: for each user\_id, sum
- Output: user\_id, tweet count
- With 2x machines, runs close to 2x faster.



- Challenge: how many tweets per user, given tweets table?
- Input: key=row, value=tweet info
- Map: output key=user\_id, value=1
- Shuffle: sort by user\_id
- Reduce: for each user\_id, sum
- Output: user\_id, tweet count
- With 2x machines, runs close to 2x faster.

#### But...

- Analysis typically done in Java
- Single-input, two-stage data flow is rigid
- Projections, filters: custom code
- Joins: lengthy, error-prone
- n-stage jobs: Hard to manage
- Prototyping/exploration requires compilation



analytics in Eclipse? ur doin it wrong...

## Enter Pig



- High level language
- Transformations on sets of records
- Process data one step at a time
- Easier than SQL?

# Why Pig?

Because I bet you can read the following script.

## A Real Pig Script

```
\Theta
                                                                top_5.pig
 users = load 'users.csv' as (username: chararray, age: int);-
  users_1825 = filter users by age >= 18 and age <= 25;-
  pages = load 'pages.csv' as (username: chararray, url: chararray); -
  joined = join users_1825 by username, pages by username;
  grouped = group joined by url;-
  summed = foreach grouped generate group as url, COUNT(joined) AS views; -
  sorted = order summed by views desc; -
  top 5 = limit sorted 5;
  store top_5 into 'top_5_sites.csv';
```

Now, just for fun... the same calculation in vanilla Hadoop MapReduce.

#### No, seriously.

```
lp.setOutputEmpClass(Test.olass);
                                                                                                                reporter_setStatus("OK");
                                                                                                                                                                                                   1p.setOutputValueClass(Text, class);
        lava.util.Arrayblat;
                                                                                                                                                                                                   ip.estMapperClass(LoadFages, class);
 mport laws.will.literatory
 mport lave, while, highly
                                                                                                           // Do the cross product and collect the values
                                                                                                                                                                                                  FileInputFormat.eddInputFath(1p, new
                                                                                                                                                                                        Fath("/user/quies/pages"));
FileOutputFormet.seqOutputFath(lp.
                                                                                                           for couring at a first) c
                                                                                                               for cotring at a second; (
String outval - key - "," + at - "," + at;
oc.cottect(matt, new Yestcoutval;;;
 mport org.apache.hadoop.fs.Path;
 mport org.apache.hadoop.to.LongWritable;
                                                                                                                                                                                                       new Fath; "/user/gates/tmp/indexed_pages";;;
  sport org.apache.hadoop.io.Text.
                                                                                                                                                                                                   lp-setMonReduceTasks(0);
                                                                                                                                                                                                  nob loadFages - new nobclipts
  sport org.spacks.badoop.to.Mritable;
                                                                                                                     reporter;setStatum;"DE");
  sport org.apache.badoop.lo.WritableComparable;
 myorn organization to doop a neighbor to be imported as
                                                                                                                                                                                                  Ashriand life " new Ashriand ("Minnample-slane);
  moort org.apache.hadoop.mapred.FileDutputFormat;
                                                                                                                                                                                                  bfo.setJobHame; "Load and Filter Corrs");
  aport org.apache.hadoop.mapred.JubConf;
                                                                                                                                                                                                   lfu.metInputFormat(TestInputFormat.class);
  moors org.spacks.badoop.magoed.EspValueTeatInputFormat;
                                                                                                 public static class Load/blood satends Rapheducebase
                                                                                                                                                                                                   lfu.metCoiputEeyClass(Test.class);
  aport org.apache.hadoop.mapred.Happery
                                                                                                      implements Happer-Text, Text, Text, LongWritable> (
                                                                                                                                                                                                   lfu.metGutputTalueClade;Test.olade);
  mport org. apacks. hadrop. majored. Hapkedordhase;
                                                                                                                                                                                                   1fs.setMapperClass(LoadhodFiltsrUmers,class);
  mport org.apacke.hadoop.mapred.CukputCollector;
                                                                                                      public veid map:
                                                                                                                                                                                                  FileImputFormat.addImputPathilfs, new
                                                                                                                                                                                       Fath: "/waer/qakee/teece"));
FiledutpotFormat: astOutputFath: 1fo,
new Fath: "/weer/qatee/tep/filtered_waers"));
lfo.mathumbedureTasks: 0);
  mport org. mpache. hadrop. majored. Recordheader;
  mport org: apache: hadoop: majred: Redocer:
                                                                                                                Teat eat.
                                                                                                                OutputCallector(Text, LongWeltable: or, Reporter reporter) throws Differentian (
  aport org.apache.badoop.mapred.Emporter;
  mport org. mpache.tedoop.mapred.Empleocefile(sputformat)
                                                                                                           // Find the url
                                                                                                                                                                                                  Ash loadWeers - new Ashilfut;
  miort org. miache-hadoop majored. SeguenceFileSutputFormat:
                                                                                                          Orring time - Val. toString():

Lot FirstComma - line. Lodestof(',');
Lot secondComma - line. Lodestof(',', firstComma);
String bey - line.substring(firstComma, secondComma);
// drop the rest of the record, I don't need it anymore,
// just pass a 1 for the combiner/reducer to sum listeed.
 sport org.apacks.kadoop.mapred.TestInputFormat;
 aport org.apache.tadoop.mapred.loboontrol.Fob;
aport org.apache.tadoop.mapred.loboontrol.fobControl;
                                                                                                                                                                                                  John setJohnsen | John Users and Pages' | |
| 2010.setJohnsen | John Users and Pages' | |
| 2010.setJohnsen | RepvalueTextInputPormat_class) |
 sport org.apache.badcop.mapred.lib.Edentitymapper;
                                                                                                                                                                                                   oun-setOutputReyClassCTest.class);
                                                                                                                                                                                                   min.matOutputValueClass(Text.class);
public class MMExample (
    public static class LoadPapes extends MapReduceRase
                                                                                                           Test outley - new Test(key);
                                                                                                                                                                                                   oin-setMapperClass; IdentityMapper.class;
           implements Happer-LongWritable, Yeat, Yest, Test: {
                                                                                                           or collectioutRey, new LongWritable(16));
                                                                                                                                                                                                   join-setHedisser(Less; Join-class);
                                                                                                                                                                                                  FileisputFormat-addingutFath; join, new
          public void map; LongWritable h. Text val.
                                                                                                                                                                                        Path("/user/gates/tep/indexed_pages"));
                                                                                                 public static class ReduceOrls extends MagReduceRape
                    ComputCallector/Text, Text> oc.
                                                                                                                                                                                                 FileImputFormat.addImputPath()ois, new
                    Reporter reporter) throws lottereption (
                                                                                                     implements Reducer-Text, LongWritzble, WritzbleComparable,
                                                                                                                                                                                        Path; "/uses/gates/tmp/filtered_users"));
                                                                                                                                                                                                  FileDutputFormst.setOutputFath:join, new
               // Full the key out
                                                                                                                                                                                        Path; "/uses/gates/tmp/joined");;
               String line - val.toStringill
               int firetcomms = line.indexOf(", ');
                                                                                                      public void reduce;
                                                                                                                                                                                                   loin.setMunReduceTeaks(50);
               String key - line, substring(), firstComma;
                                                                                                                                                                                                  Job Joinfob - new Jobijechii
               String value - line.substring(firetcomma - 1);
Text outbey - new Text(Rey);
                                                                                                                Discussion Stables itself.
                                                                                                                                                                                                   joinJob:addDependingFobijoadPages);
                                                                                                                OutputCollector WritableComparable, Writable oc.
                                                                                                                                                                                                   ]ointob:addrependingtob:boadteirs;
               // Frepend an index to the value so we know which file
                                                                                                                Reporter reporter; throws DEException (
                                                                                                           // Add up all the values we see
               // 15 came from.
                                                                                                                                                                                                  JobConf group - new JobConfipRExample.class;;
               Test outyal " new Test("1" + value);
                                                                                                                                                                                                  dronb'settoppase(,cooh nare,);
               oc.collect/outHey, outWall;
                                                                                                                                                                                                  group.setInputFormat(ReyValueTextInputFormat-class);
                                                                                                           while (iter-headest()) (
                                                                                                                                                                                                   group.setOutputKeyClass(Text.class);
                                                                                                                sum or iter-next()-pet():
                                                                                                                                                                                                   group.setOutputValueClass(LongWritsble.class);
     public static class LoadhodFilterOpers extends HapReduceBase
                                                                                                                reporter-setStatus("OK");
                                                                                                                                                                                                   group.setOutputFormat(SequenceFileOutputFormat.class);
          implements Mapper-LongMritable, Text, Text: {
                                                                                                                                                                                                   group.setMapperClass(LoadJoined.class);
                                                                                                                                                                                                   group.setCoshinerClass(ReduceOrls_class);
          public void map; LongWritable h, Test val,
                                                                                                           or.collert(key, new LongWritable(sum));
                                                                                                                                                                                                   group.setReducerClass(ReduceOrls.class);
                                                                                                                                                                                                  FileInputFormat.addInputPath;group, new
                    Computabilization (Text. Text. oc.
                                                                                                                                                                                       Path! "/user/gakes/tmp/joined";;;
FileDutgutFormat.setOutgutPath;group, new
                    Reporter reporter) throws DOEsception (
                 Full the key out
                                                                                                 public static class LoadClicks extends MapReduceRase
               String line - Val. toString();
                                                                                                      implements Mapper Writable Comparable, Writable, LongWritable,
                                                                                                                                                                                        Faith!"/com/gates/imp/grouped"[]:
              int firstcomms = line.todexoft; ');

thring value = line.substring(firstcomms = l);

int age = ittoper.parssint(value);

if case = it | age = 25; return;

thring key = line.substring(f, firstcomms);
                                                                                                                                                                                                  group, setMonReduceTeaks; 50);
                                                                                                                                                                                                  Job group/ob = new Job/group()
                                                                                                                                                                                                  group/ob:addrepending/ob; join/ob;;
                                                                                                      public void map:
                                                                                                                MritableComparable key,
                                                                                                                                                                                                  Asbelond topics - new Jobelond (MREsample class);
                                                                                                                mrinante val-
                                                                                                                                                                                                  topico.set/onmame; fop 100 sites');
               Text outley " new Text(bey);
                                                                                                                Outputchilector/LongMritable, Text/ or,
                                                                                                                                                                                                  top:000:set:imputPormat::SequenceFile:imputPormat:-01eee;;
top:000:set:imputPormat::SequenceFile::state:::
top:000:set:imputPormat::set:Text.-01eee;;
                                                                                                          Reporter reporter; throws commomption of
mr.collect; thoughoutentaire; cresting;;
               // Frepend an index to the value so we know which file
               Yest matvat " new Test; 3" a value);
                                                                                                                                                                                                  top:00.setOutputFormat:Emplement() #004gotFormat.class);
               no.collectiontEsp. suttall:
                                                                                                 public static class LimitClicks saturds MapRedyceRess
                                                                                                                                                                                                  top:00.matHapperClass(LoadClicks.slass);
                                                                                                      implements Reduces LongHyltskie, Text, LongWyltskie, Text: [
                                                                                                                                                                                                  topido anticidamentiadaciamitii intercitada ;
     public static class Join extends HapkeduceBase
                                                                                                                                                                                                   top:000.astReducer01esscLin44014chs-class;;
                                                                                                                                                                                                  FileInputFormat.addInputPath(topid), new
          implements Reducer-Text, Text, Text, Text: (
                                                                                                      public youd reduces
                                                                                                                                                                                        Parth ("/oser/gahes/tep/grouped")))
                                                                                                                                                                                                  FileOutputFormet.setOutputPath(top100, new
          public wold reduce; Test key,
                                                                                                           Longhritable Key,
                                                                                                                                                                                        Path; "/oser/gates/togl@baltesforusersl#tolb"));
                    Iterator-Teath Ster,
                                                                                                           Distance-Text- list.
                  OutputCollector Text. Text> oc.
Reporter reporter; throws DoException (
For each value, figure out which file it's from and
                                                                                                           OutputCollector-LongWritable, Text> co. Reporter reporter) throws 10Emception (
                                                                                                                                                                                                  top: 00 , setHumbeduceTanks ( ) )
                                                                                                                                                                                                  Job limit - new Jobitopissi;
store it
                                                                                                           // Only output the first 100 records
                                                                                                           while coount = 100 as iter-hashwat()) (
                                                                                                                                                                                                  PobCuntrol je = new PobCuntrol("Find top 100 mites for users
               List-String- first = new ArrayList-String-();
                                                                                                                oc.collectivey, iter.next());
                                                                                                                                                                                        FR 20 52,11
                                                                                                                                                                                                   (ir-add/ob) loadPages);
               List-String- second - new ArrayList-String-[];
                                                                                                                CONTRACTOR
                                                                                                                                                                                                   te-add/objicad@acrets
               while (ther.hashest()) (
                                                                                                                                                                                                   per-makel Joseph Service Joseph La
                    Test t - Ster. nest();
                                                                                                                                                                                                   5e-mid-hole (persup-hole) ;
                                                                                                 public static void main(String() args) throws Differentian (
Josefond by - new Josefond(Stringers-views);
                    String value = t.toString();
                                                                                                                                                                                                   ter-matching Limits;
                   if (value.charated) -- 'l';
                                                                                                                                                                                                   DE-PUBLIC
firet.addivelue.eubetring(l));
                                                                                                       hp.met.Pobstane; "Load Pages");
                                                                                                       p.setInputFormat(TextInputFormat-class);
```

# Pig Democratizes Large-scale Data Analysis

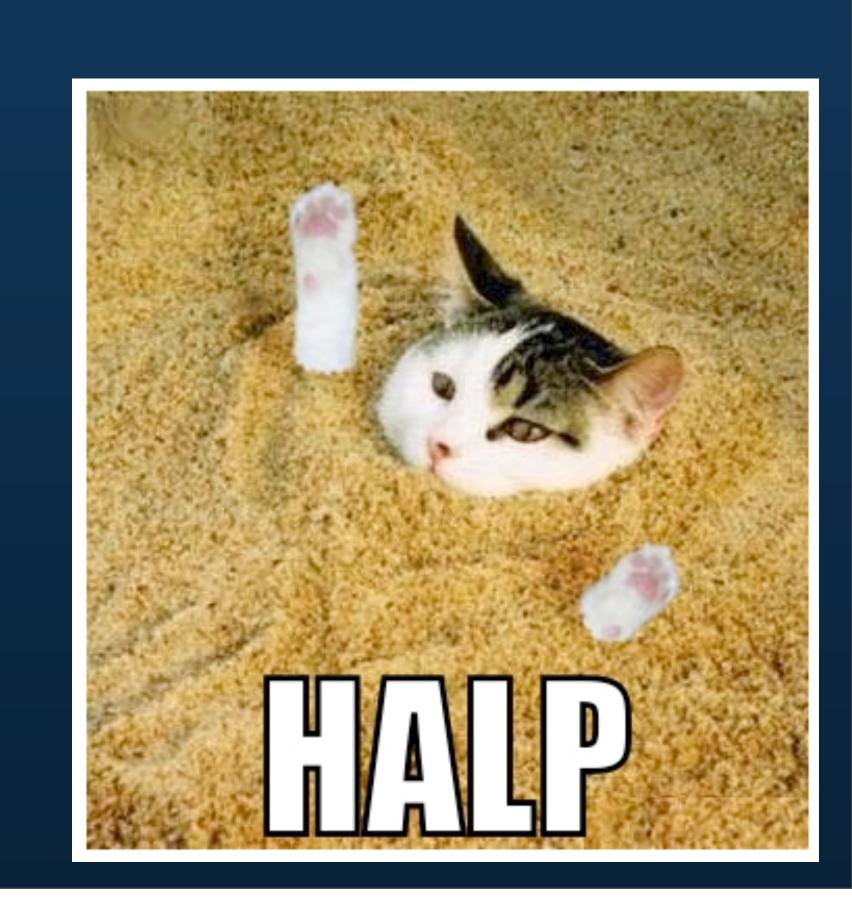
- The Pig version is:
- 5% of the code
- 5% of the time
- Within 50% of the execution time.
- Innovation increasingly driven from large-scale data analysis
- Need fast iteration to understand the *right questions*
- More minds contributing = more value from your data

#### Introduction

- Hadoop Overview
- Why Pig?
- **Evolution of Data Processing at Twitter**
- Pig for Counting
- Pig for Correlating
- Pig for Research and Data Mining
- Conclusions and Next Steps

## MySQL, MySQL, MySQL

- We all start there.
- But MySQL is not built for analysis.
- select count(\*) from users? Maybe.
- select count(\*) from tweets? Uh...
- Imagine joining them.
- And grouping.
- Then sorting.



#### The Hadoop Ecosystem at Twitter

- Cloudera's free distribution, running Hadoop 0.20.1
- Heavily modified Facebook Scribe for log collection -> HDFS\*
- Heavily modified LZO code for fast, splittable data compression\*\*
- Data stored either as LZO-compressed flat files (logs, etc) or serialized, LZO-compressed protocol buffers (structured data).
- Custom InputFormats, Pig LoadFuncs for the above\*
- Some Java-based MapReduce, some Hadoop Streaming
- Most analysis, and most interesting analyses, done in Pig.
- \* Open sourced, or on the way. Please come talk afterwards if you're interested.
- \*\* http://www.github.com/kevinweil/hadoop-lzo

#### Data?



- Semi-structured: apache logs, search logs, RoR logs, mysql query logs, rate limiter logs, per-application logs
- Structured: tweets, users, block notifications, phones, favorites, saved searches, retweets, authentications, sms usage, third party clients, followings
- Entangled: the social graph

#### Introduction

- Hadoop Overview
- Why Pig?
- Evolution of Data Processing at Twitter
- Pig for Counting
- Pig for Correlating
- Pig for Research and Data Mining
- Conclusions and Next Steps

standard counts, min, max, std dev

How many requests do we serve in a day?



- How many requests do we serve in a day?
- What is the average latency? 95% latency?



- How many requests do we serve in a day?
- What is the average latency? 95% latency?
- Group by response code. What is the hourly distribution?



- How many requests do we serve in a day?
- What is the average latency? 95% latency?
- Group by response code. What is the hourly distribution?
- How many searches happen each day on Twitter?



- How many requests do we serve in a day?
- What is the average latency? 95% latency?
- Group by response code. What is the hourly distribution?
- How many searches happen each day on Twitter?
- How many unique queries, how many unique users?



- How many requests do we serve in a day?
- What is the average latency? 95% latency?
- Group by response code. What is the hourly distribution?
- How many searches happen each day on Twitter?
- How many unique queries, how many unique users?
- What is their geographic distribution?



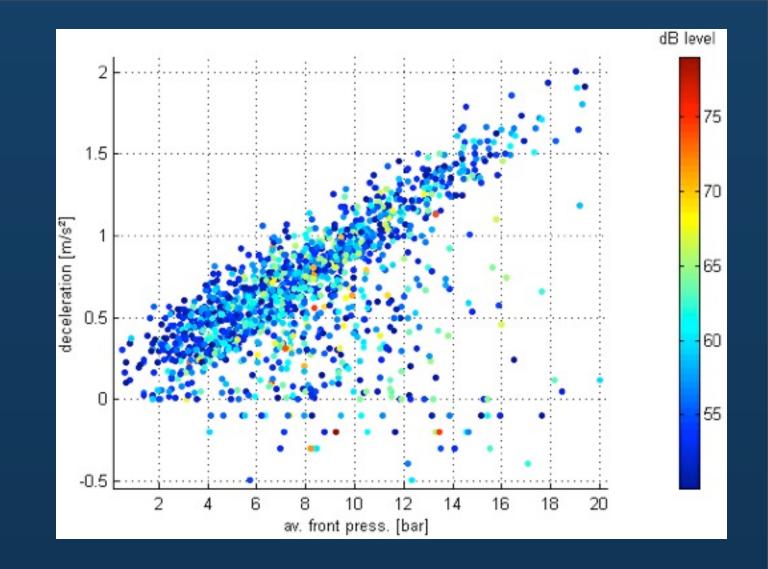
Where are users querying from? The API, the front page, their profile page, etc?



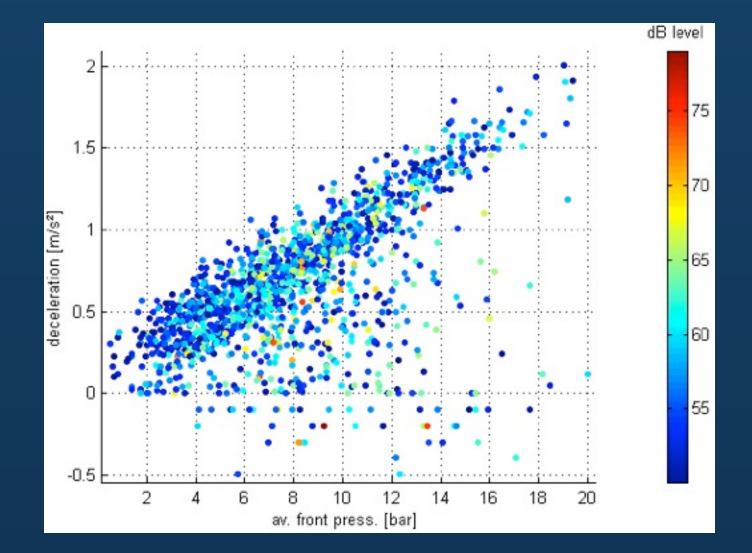
```
search_origin.pig
raw_data = load '$INPUT_FILES' using com.twitter.twadoop.pig.storage.LzoTwitterApacheLogLoader() as -
    (..., virtual_host: chararray, apache_time: chararray, request_method: chararray, request_url: chararray,
     request_protocol: chararray, response_code: chararray, response_size: int, referrer: chararray,
     user_agent: chararray, response_microseconds: int, ...); -
searches_only = filter raw_data by com.twitter.twadoop.pig.piggybank.IsSearchUrl(request_url);-
searches_with_type = foreach searches_only generate ¬
    com.twitter.twadoop.pig.piggybank.ExtractSearchOrigin(virtual_host, request_url) as origin; -
grouped = group searches_with_type by origin parallel $PARALLEL; -
counted = foreach grouped generate group, COUNT(searches_with_type);¬
store counted into 'searches_by_type.tsv' using PigStorage('\t');
```

probabilities, covariance, influence

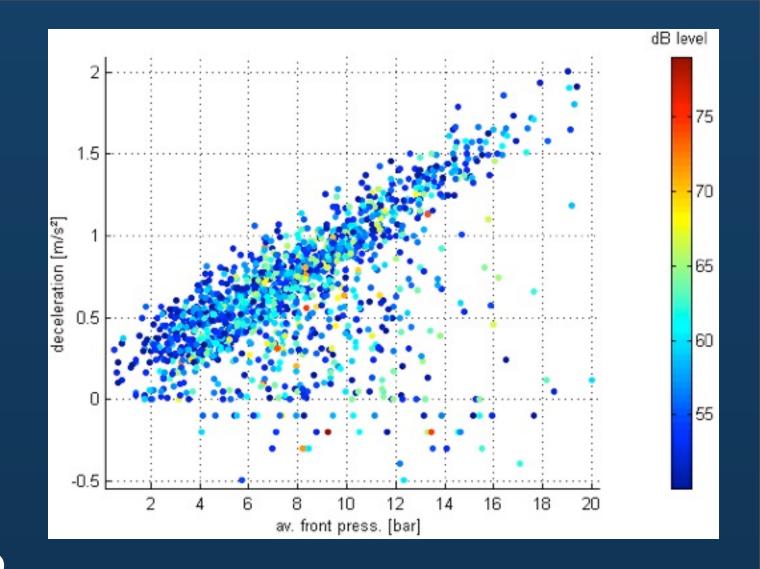
How does usage differ for mobile users?



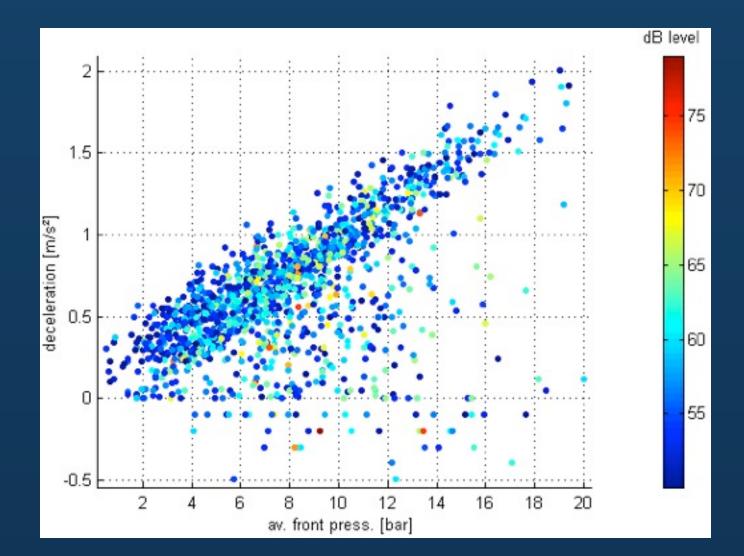
- How does usage differ for mobile users?
- How about for users with 3rd party desktop clients?



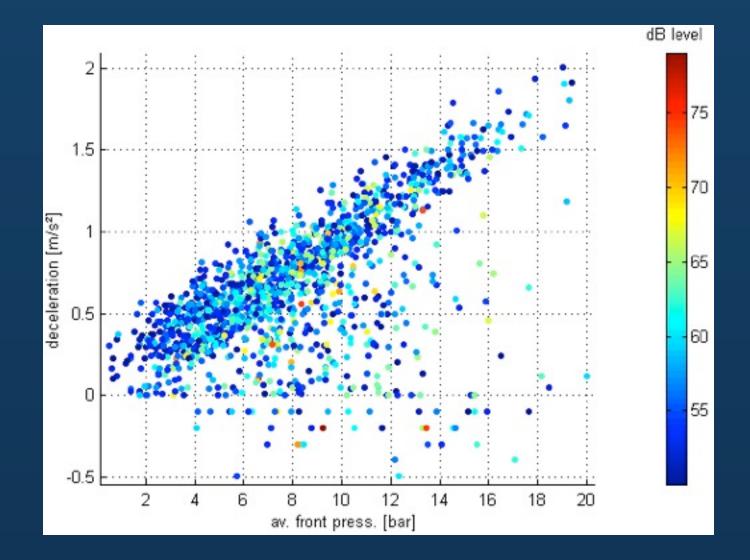
- How does usage differ for mobile users?
- How about for users with 3rd party desktop clients?
- Cohort analyses



- How does usage differ for mobile users?
- How about for users with 3rd party desktop clients?
- Cohort analyses
- Site problems: what goes wrong at the same time?

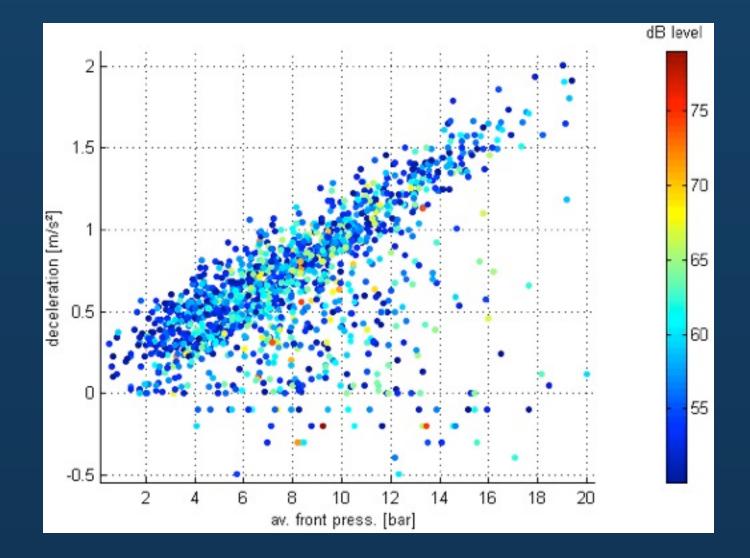


- How does usage differ for mobile users?
- How about for users with 3rd party desktop clients?
- Cohort analyses
- Site problems: what goes wrong at the same time?
- Which features get users hooked?



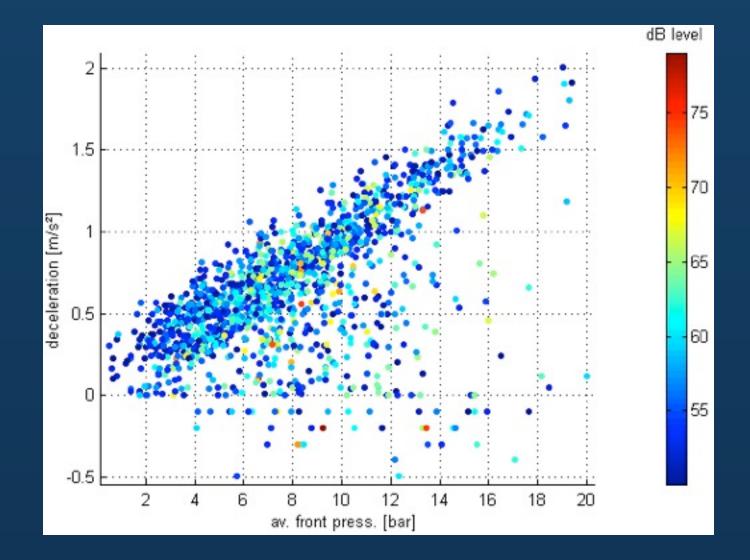
#### probabilities, covariance, influence

- How does usage differ for mobile users?
- How about for users with 3rd party desktop clients?
- Cohort analyses
- Site problems: what goes wrong at the same time?
- Which features get users hooked?
- Which features do successful users use often?



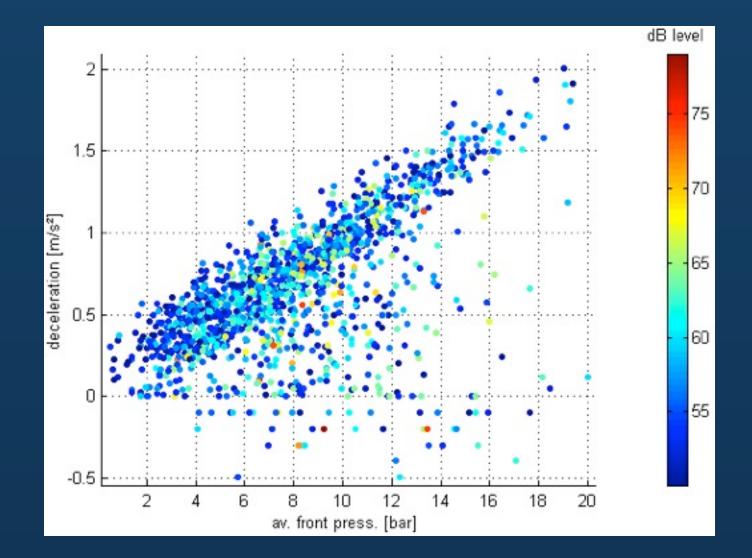
#### probabilities, covariance, influence

- How does usage differ for mobile users?
- How about for users with 3rd party desktop clients?
- Cohort analyses
- Site problems: what goes wrong at the same time?
- Which features get users hooked?
- Which features do successful users use often?
- Search corrections, search suggestions



#### probabilities, covariance, influence

- How does usage differ for mobile users?
- How about for users with 3rd party desktop clients?
- Cohort analyses
- Site problems: what goes wrong at the same time?
- Which features get users hooked?
- Which features do successful users use often?
- Search corrections, search suggestions
- A/B testing

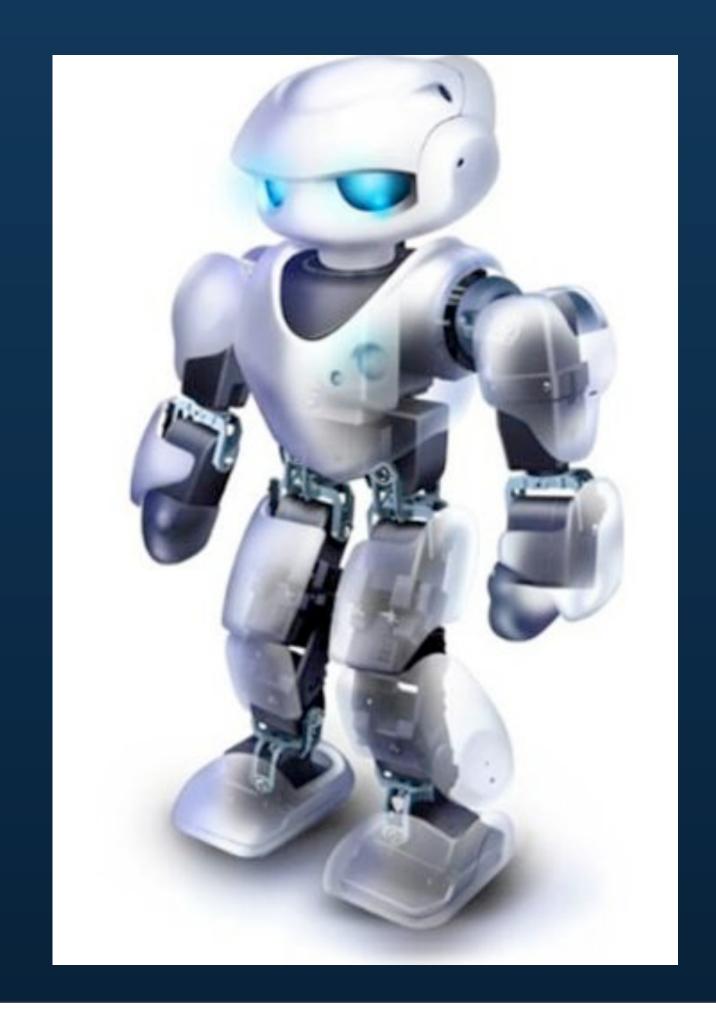


What is the correlation between users with registered phones and users that tweet?

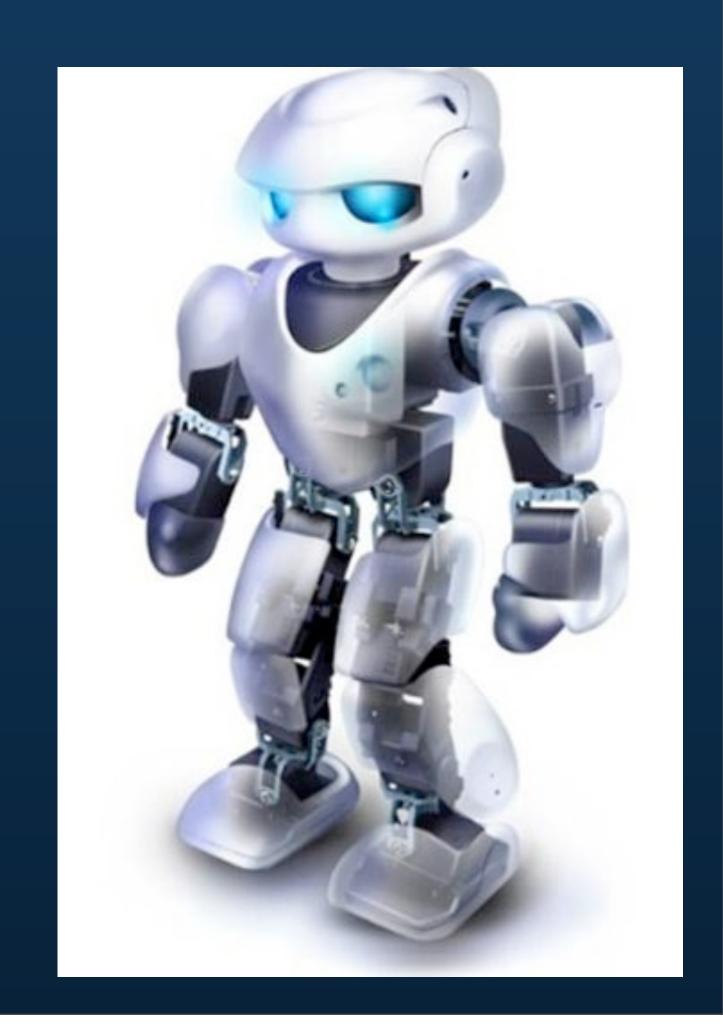
```
correlation.pig
  register piggybank.jar;-
  devices = load '$DEVICES' using com.twitter.twadoop.pig.storage.LzoDevicesLoader() as -
    (device_id: long, user_id: long, ...);-
  tweets = load '$TWEETS' using com.twitter.twadoop.pig.storage.LzoStatusLoader() as-
     (tweet_id: long, user_id: long, text: chararray, ...);
  tweet_user_ids = foreach tweets generate user_id; -
  tweets_grouped = group tweet_user_ids by user_id;-
  tweets_by_user = foreach tweets_grouped generate group as user_id, COUNT(tweet_user_ids) as count;-
  combined = cogroup devices by user_id, tweets_by_user by user_id;-
  summed = foreach combined generate user_id, SIZE(devices) as has_device, ¬
      (SIZE(tweets_by_user) == 0 ? 0 : SUM(tweets_by_user.count)) as tweet_count; -
  summed_grouped = group summed all; -
  covariance = foreach summed_grouped generate group, -
      COR(summed_grouped.has_device, summed_grouped.tweet_count); -
```

prediction, graph analysis, natural language

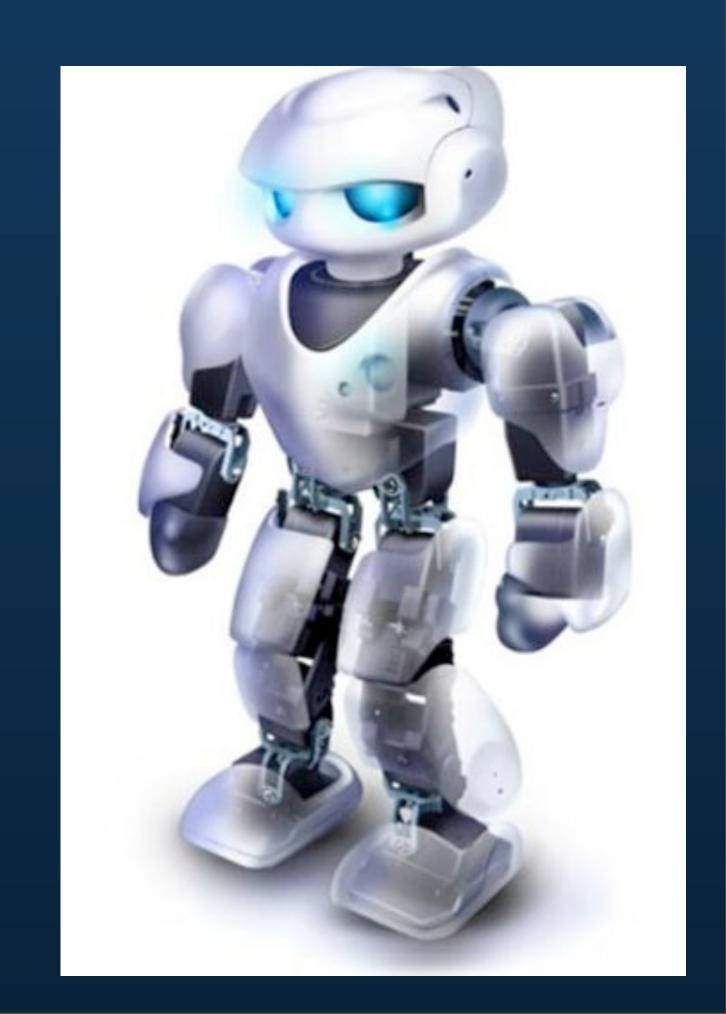
What can we tell about a user from their tweets?



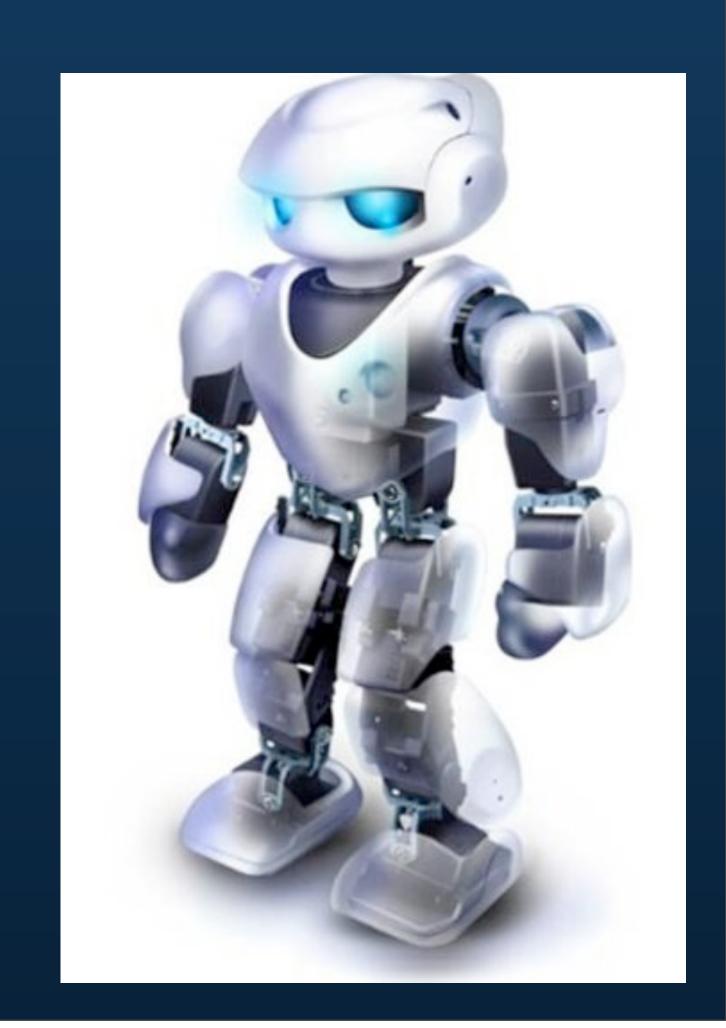
- What can we tell about a user from their tweets?
- From the tweets of those they follow?



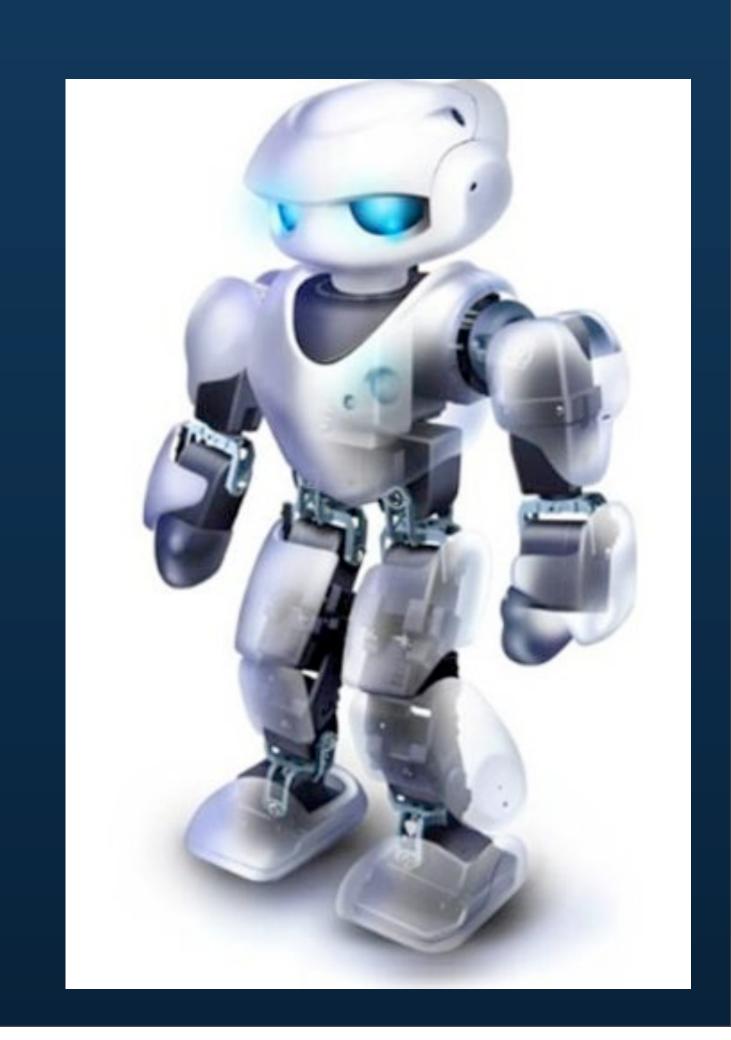
- What can we tell about a user from their tweets?
- From the tweets of those they follow?
- From the tweets of their followers?



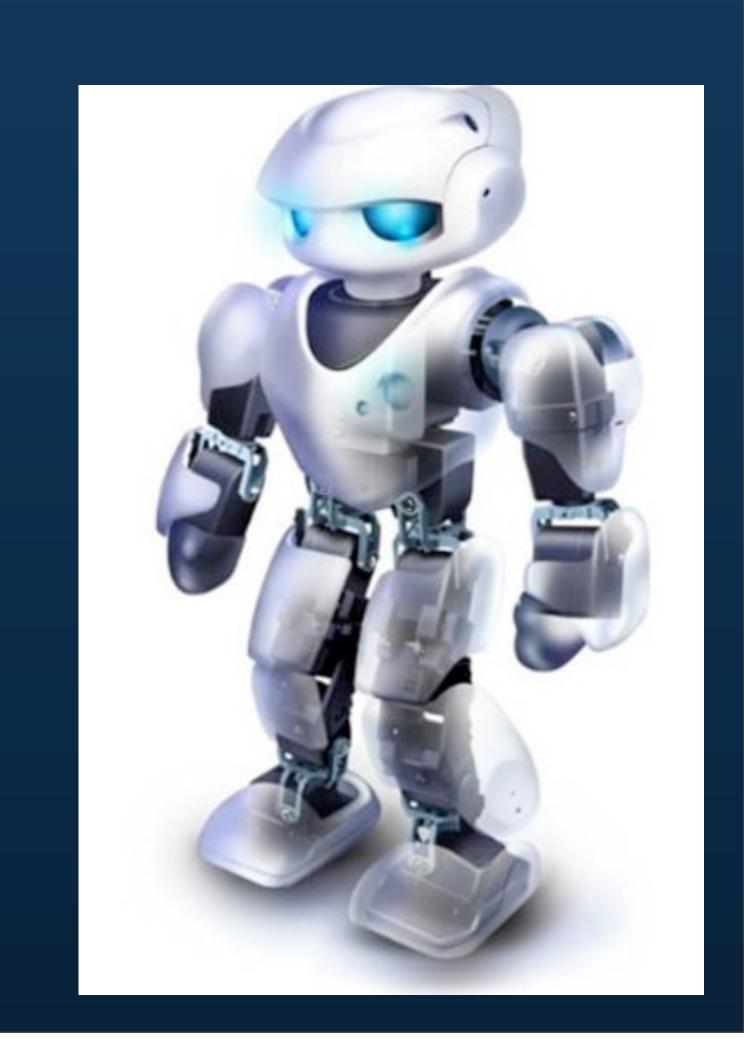
- What can we tell about a user from their tweets?
- From the tweets of those they follow?
- From the tweets of their followers?
- From the ratio of followers/following?



- What can we tell about a user from their tweets?
- From the tweets of those they follow?
- From the tweets of their followers?
- From the ratio of followers/following?
- What graph structures lead to successful networks?

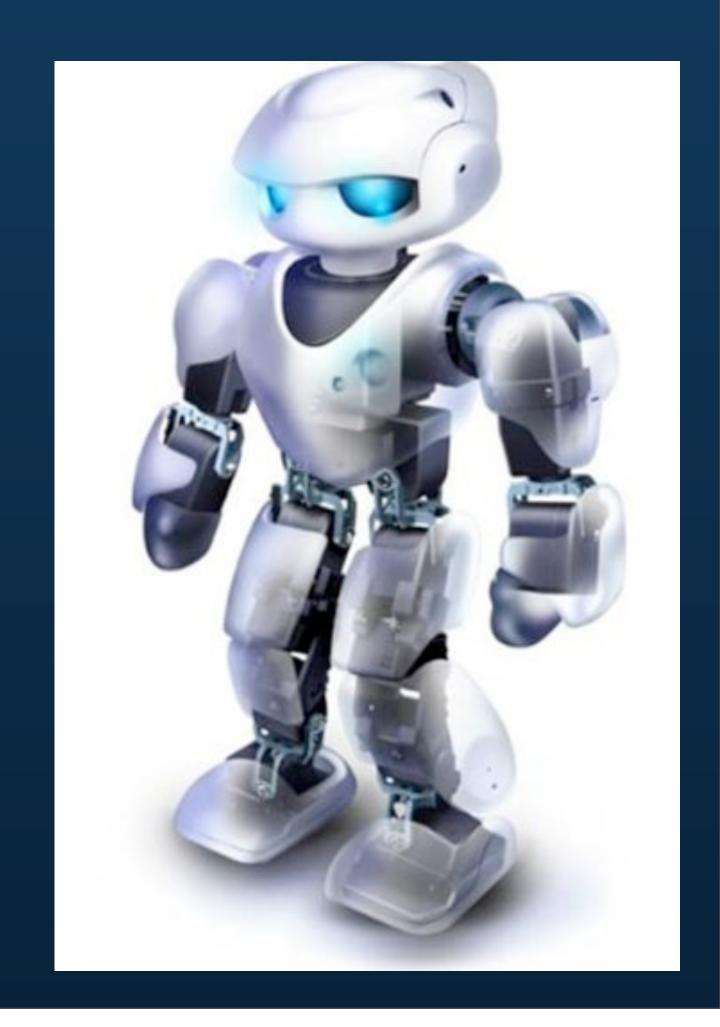


- What can we tell about a user from their tweets?
- From the tweets of those they follow?
- From the tweets of their followers?
- From the ratio of followers/following?
- What graph structures lead to successful networks?
- User reputation

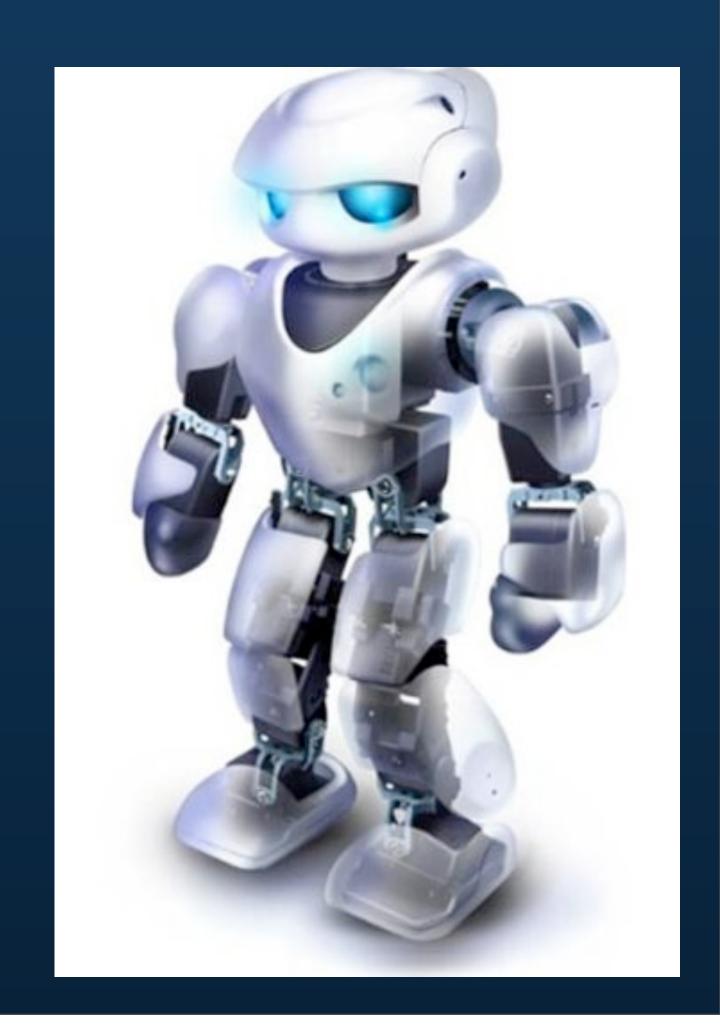


prediction, graph analysis, natural language

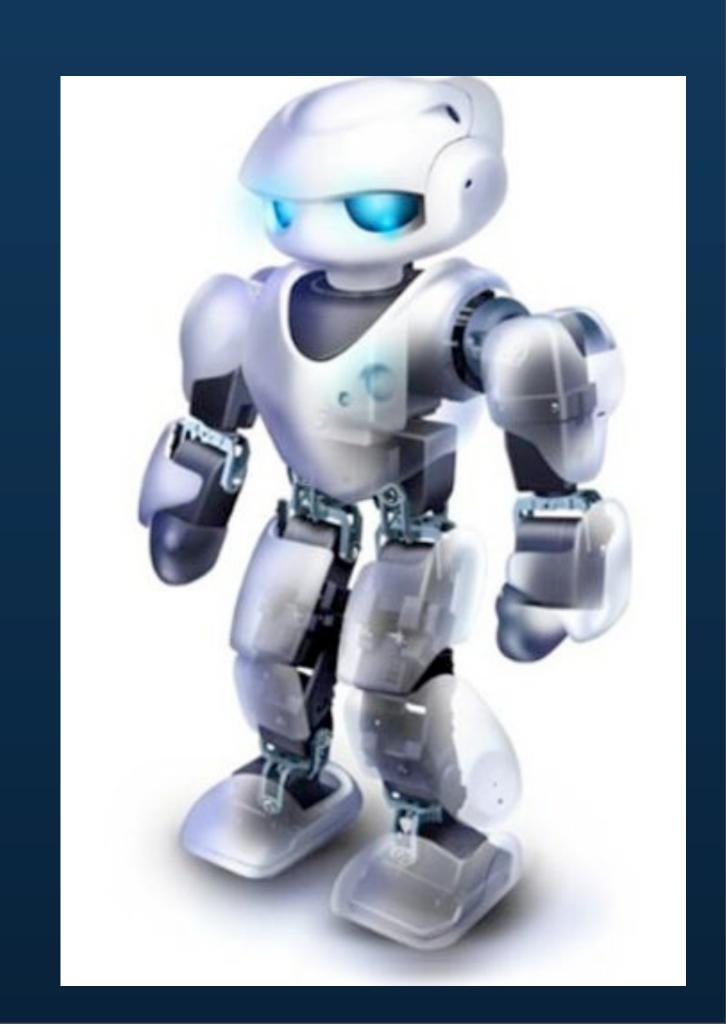
Sentiment analysis



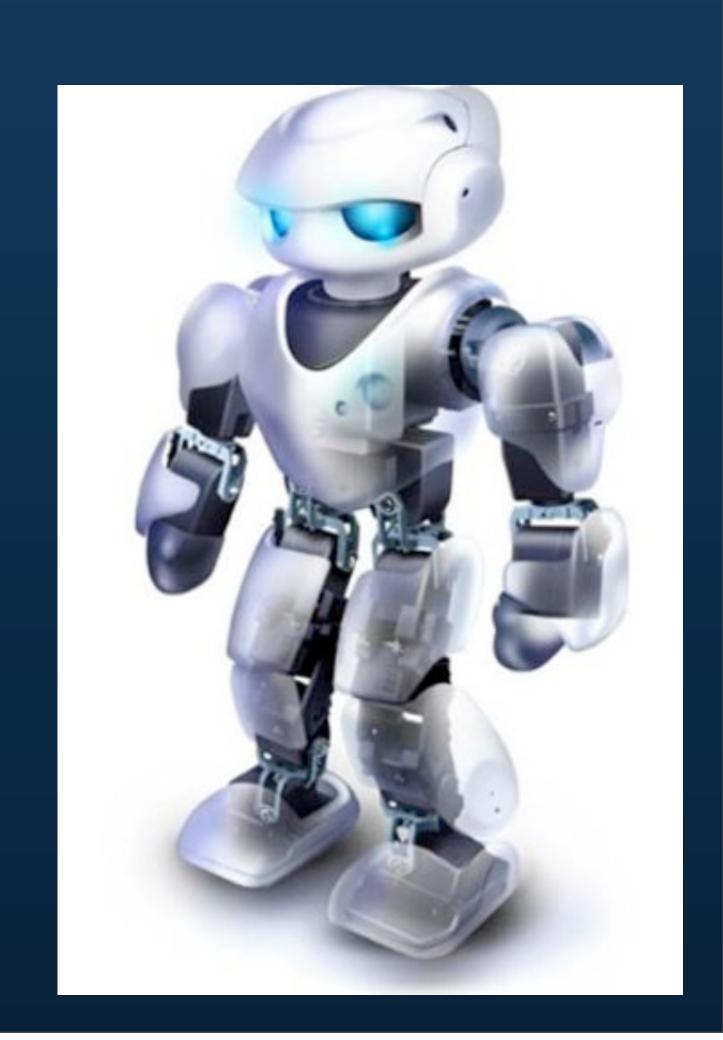
- Sentiment analysis
- What features get a tweet retweeted?



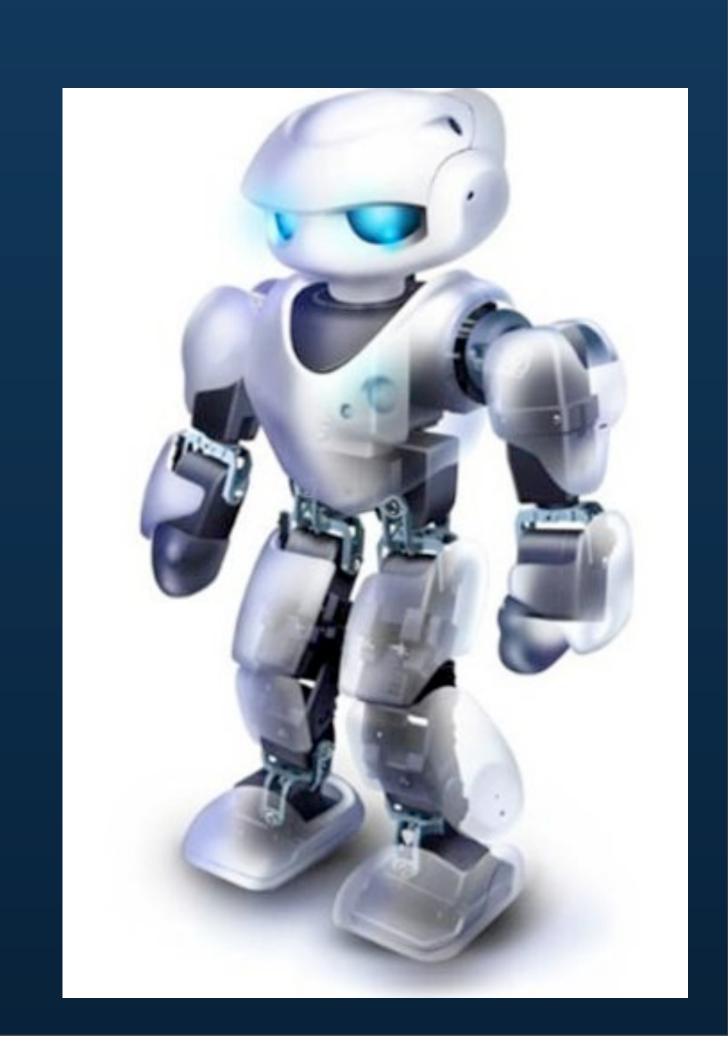
- Sentiment analysis
- What features get a tweet retweeted?
- How deep is the corresponding retweet tree?



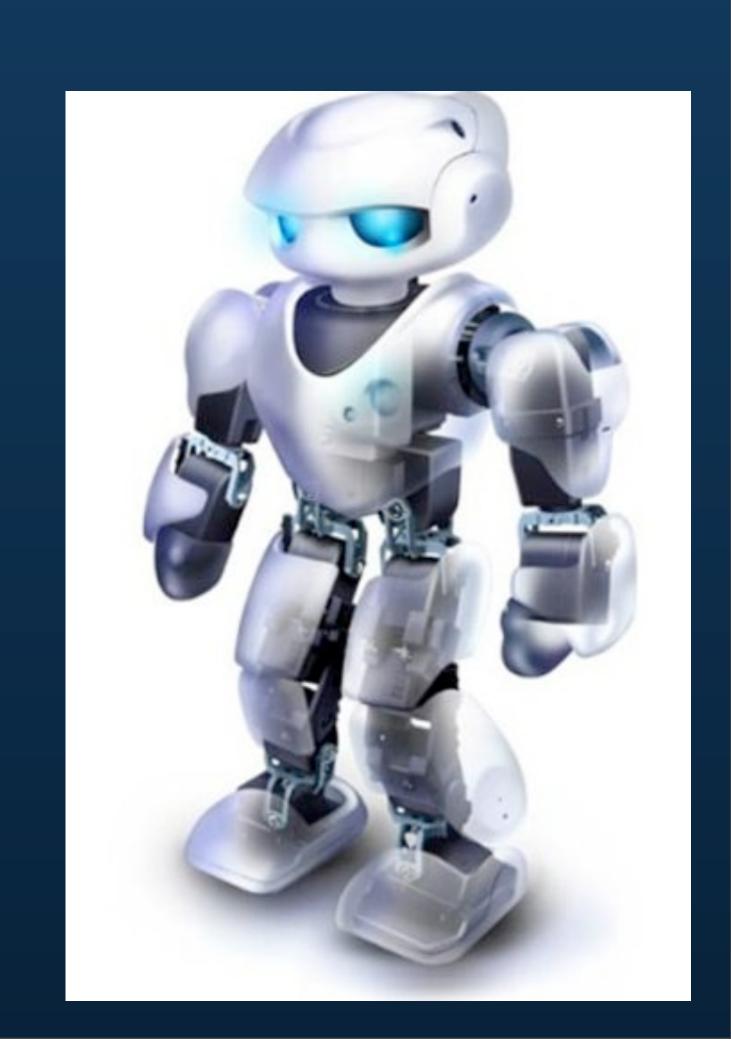
- Sentiment analysis
- What features get a tweet retweeted?
- How deep is the corresponding retweet tree?
- Long-term duplicate detection



- Sentiment analysis
- What features get a tweet retweeted?
- How deep is the corresponding retweet tree?
- Long-term duplicate detection
- Machine learning



- Sentiment analysis
- What features get a tweet retweeted?
- How deep is the corresponding retweet tree?
- Long-term duplicate detection
- Machine learning
- Language detection



- Sentiment analysis
- What features get a tweet retweeted?
- How deep is the corresponding retweet tree?
- Long-term duplicate detection
- Machine learning
- Language detection
- ... the list goes on.



How well can we detect bots and other non-human tweeters?

```
bot_probability.pig
raw_data = load '$INPUT_FILES' using com.twitter.twadoop.pig.storage.LzoStatusLoader() as-
    (id: long, -
     user_id: long, -
     text: chararray, -
     created_at: chararray, -
     ...);¬
projected = foreach raw_data generate user_id, text; -
grouped_by_user = group projected by user_id parallel $PARALLEL; -
bot_probabilities = foreach grouped_by_user generate group as user, ¬
  com.twitter.twadoop.pig.piggybank.UserBotProbability(projected.text) as bot_probability;-
store bot_probabilities into 'bot_test.tsv' using PigStorage('\t');
```

#### Introduction

- Hadoop Overview
- Why Pig?
- Evolution of Data Processing at Twitter
- Pig for Counting
- Pig for Correlating
- Pig for Research and Data Mining
- Conclusions and Next Steps

# Why Hadoop?

- Data is growing rapidly; need horizontally scalable computation
- Fault tolerant; gracefully handles machine failure

... but writing MapReduce jobs in Java is harder than it should be.

# Why Pig?

- Pig makes Hadoop accessible
- Pig chains together complex job flows
- User-defined functions are first class citizens
- Vibrant OS community, dedicated team at Y! improving it daily
- At Twitter, Pig helps us understand our business faster.

## Questions?

Follow me at twitter.com/kevinweil

If this sounded interesting to you -- that's because it is. And we're hiring.

