Datameer

What's Really the Buzz?

Stefan Groschupf
sg@datameer.com

Street Cred

- Long time open source contributor

nutch

hadoop

katta

Bixo

http://github.com/sgroschupf/zkclient

http://github.com/sgroschupf/aws-tasks

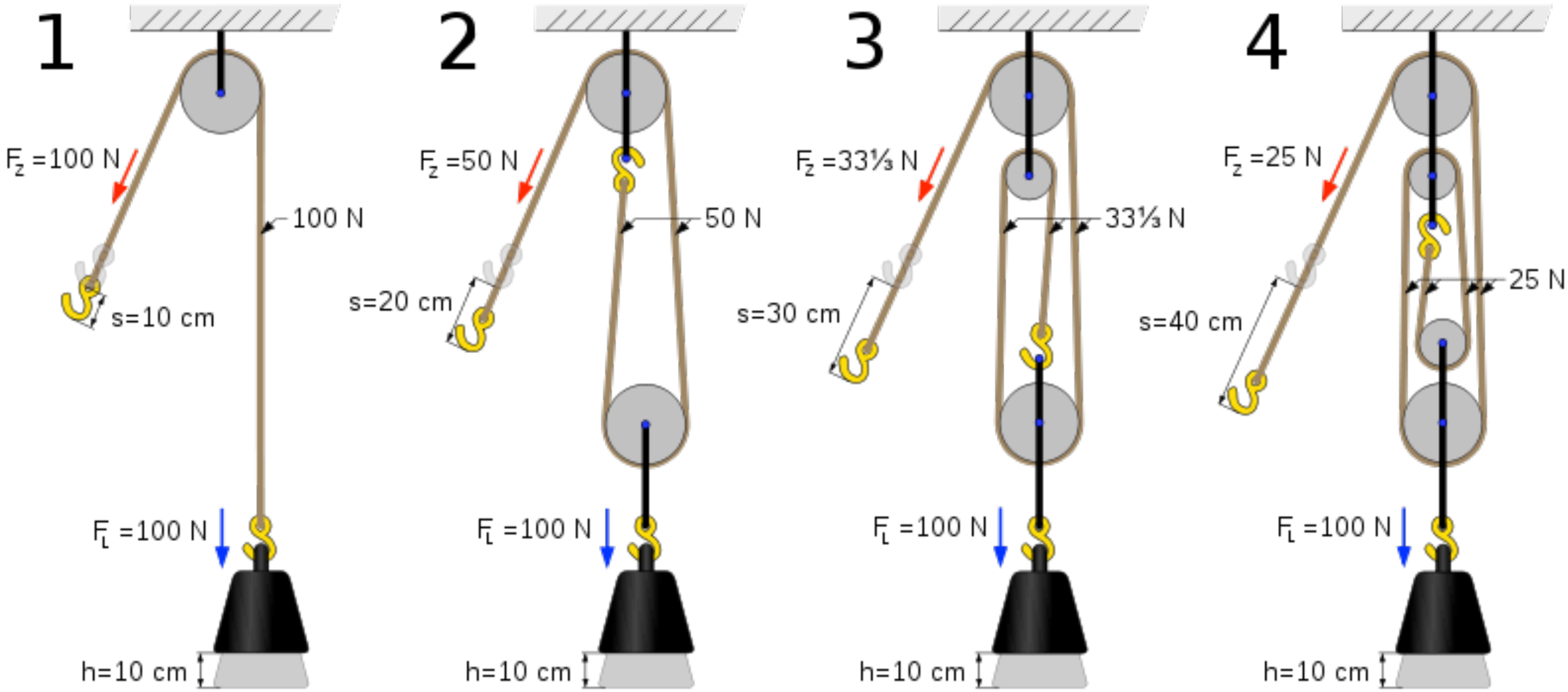
Cubicle Cred

- Cloud Computing Architect EMI
- Hadoop consultant at e.g.

NOKIA
Connecting People

- Co-Founder/Instructor Scale Unlimited
- **Co-Founder / CTO Datameer Inc.**

Hadoop vs. DB



Hadoop(mergesort)

DB(b-tree/index)

Tweets per Day

Source: twitter

Twitter API

- **Streaming API**

  `curl http://stream.twitter.com/1/statuses/sample.json \`
  `-ustefan:somepwd`

  streams, once a while interruped
  5% of public statuses by default

- **Search API**

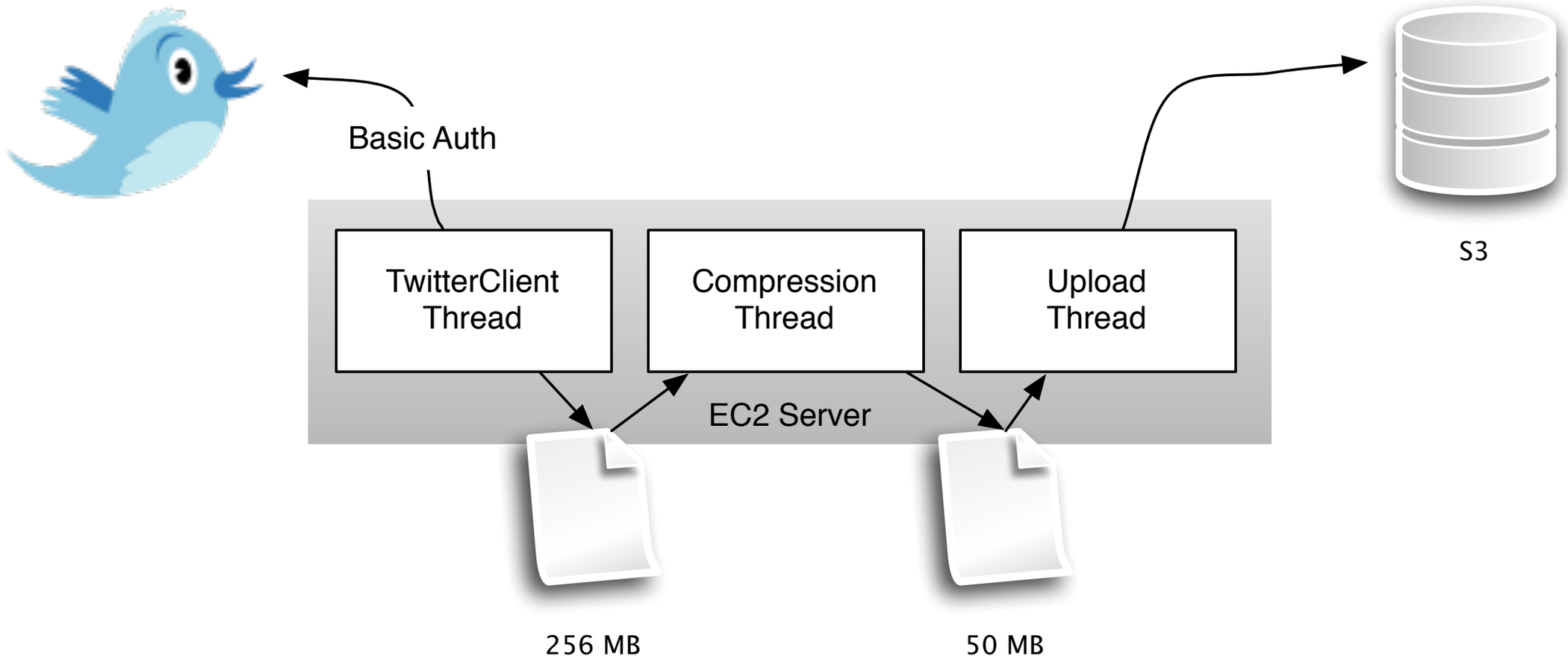
  `curl http://search.twitter.com/search.json?q=datameer \`
  `-ustefan:somepwd`

  150 requests per hour

Some Twitter Fields

- user_screen_name
- user_followers_count
- user_friends_count
- user_statuses_count
- user_location
- created_at
- text
- source
- geo_type
- geo_longitude
- geo_latitude

Twitter Client



Analytics

Datameer

Amazon Elastic MapReduce

1.) Trending Topic

- Tokenizing,lower
- Filter only hash tags
- Group hash tags
- Count group size
- Sort top groups

2.) Topic Reach

- Tokenizing, Lower , Copy followers
- Filter only hash tags
- Group hash tags
- Count group size, Sum followers
- Sort top by Followers

3.) Tweet spreading timing

- Extract pure tweet msg, copy timestamp
- Group by msg, group count, group min, group max, max-min
- Sort by min-max column

Gephi / gexf

- Gephi
  - cross platform
  - GPL
  - Graph exploration
  - 0.7 Alpha
- Graph Exchange XML Format
  - support dynamics

Twitter > S3 > Hadoop > DAS > CSV > gexf > Gephi

Gephi

- Conversation groups
  - YifanHu Multilevel Layout Algorithm
  - Pagerank
  - Color by Pagerank
  - Find highest score node

Gephi

- **Spreading**
  -

So much more.

- **Reach**
  - The degree any member of a network can reach other members of the network.
- **Prestige**
  - In a directed graph prestige is the term used to describe a node's centrality. "Degree Prestige", "Proximity Prestige", and "Status Prestige" are all measures of Prestige. See also degree (graph theory).
- **Path Length**
  - The distances between pairs of nodes in the network. Average path-length is the average of these distances between all pairs of nodes.

http://en.wikipedia.org/wiki/Social_network

Resources…

- www.datameer.com
- twitter.com/datameer
- http://xrime.sourceforge.net/
  - SNA Metrics and Structures
- http://github.com/sgroschupf
- sg@datameer.com