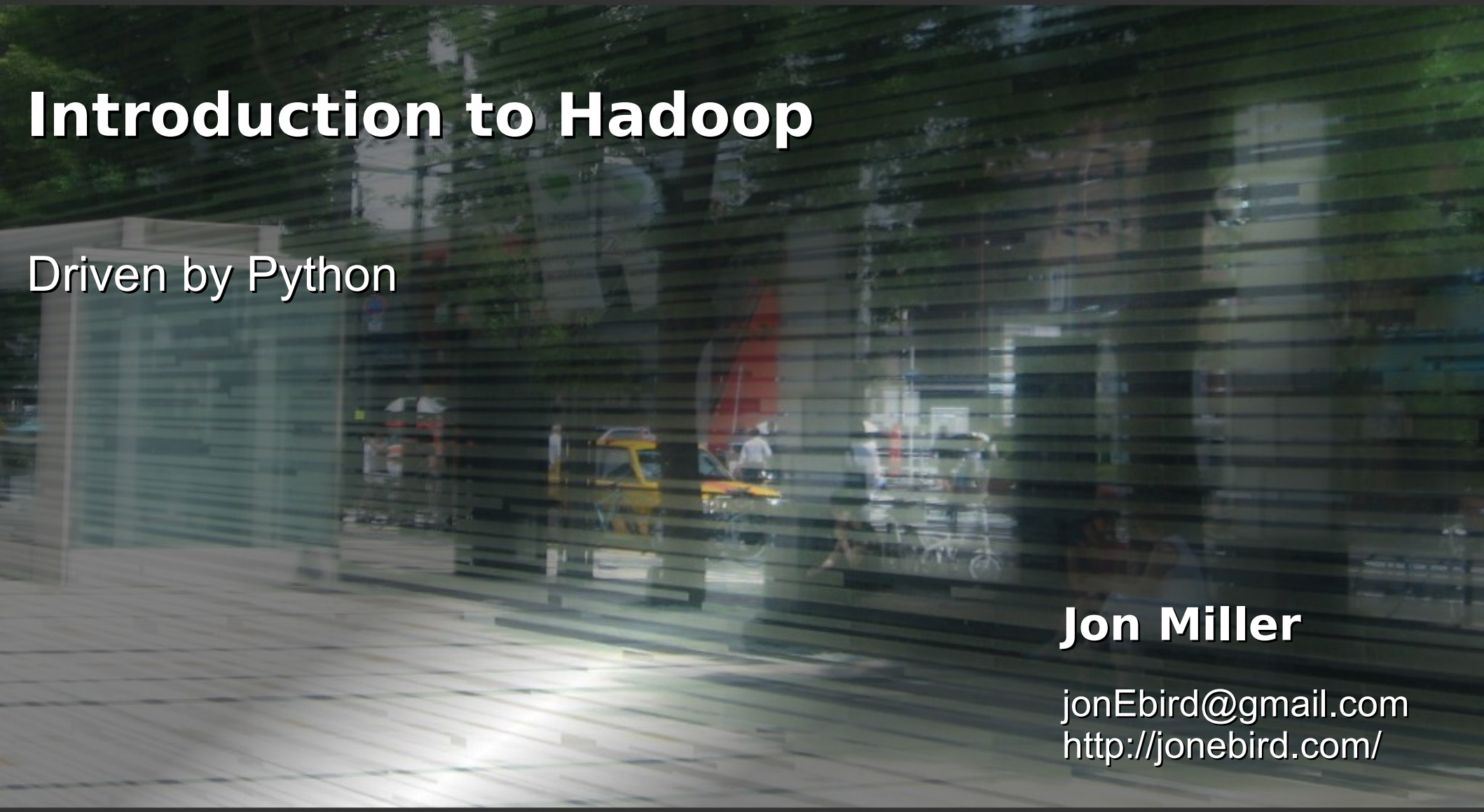


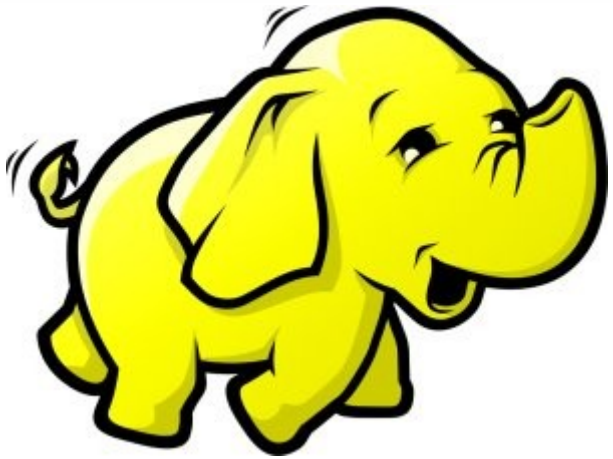
Introduction to Hadoop

Driven by Python

Jon Miller

jonEbird@gmail.com
http://jonebird.com/


What is Hadoop?



What is Hadoop?

- Doug Cutting's daughter's stuffed toy elephant
- Distributed MapReduce System
- Apache Project with multiple sub-projects
  Core, HDFS then HBase, Hive, Pig, ZooKeeper


Where is the Python?

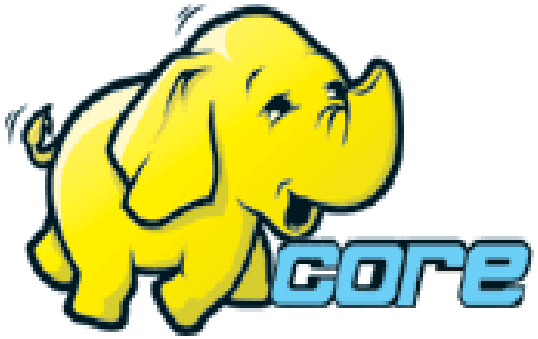
Where is the Python?

- Hadoop Streaming
- Automatically copies your python script to nodes
- Uses STDIN / STDOUT to communicate


Hadoop Architecture

Hadoop Architecture

- Expect hardware failures
- Take the computing to the data,
  NOT pull data to compute
- Datanodes, Tasktrackers & Jobtracker



Web Analytics Example

Mapper

```
#!/usr/bin/env python

import sys

IGNORE_SITES = [ 'http://jonebird.com/', 'http://www.jonebird.com/' ]

for line in sys.stdin:
    if line.count('"') == 6:
        # some entries I do not care about:
        # 1. Discard if referer is myself
        # 2. Discard if there is _no_ referer. i.e. "-"
        referer = line.split('"')[3]
        can_ignore = any( referer.startswith(site) for site in IGNORE_SITES )
        if referer != '-' and not can_ignore:
            print '%s\t%d' % (referer, 1)
```

Reducer

```
#!/usr/bin/env python

import sys

referer_count = {}

# parse input from the mapping process
for line in sys.stdin:
    try:
        referer, count = line.strip().split('\t', 1)
        count = int(count)
        referer_count[referer] = referer_count.get(referer, 0) + count
    except ValueError:
        # ignoring odd failures
        pass

# Report our results
for referer, count in referer_count.iteritems():
    print '%s\t%s' % (referer, count)
```

Invocation

```
# With $HADOOP_HOME
PATH=$PATH:${HADOOP_HOME}/bin

hadoop dfs -copyFromLocal /var/log/httpd/ apache_logs

export HSTREAM="${HADOOP_HOME}/bin/hadoop jar \
  ${HADOOP_HOME}/contrib/streaming/hadoop-${HADOOP_VERSION}-streaming.jar"

# Now run the following command to get a quick
# usage statement about using the streamer
$HSTREAM -info

$HSTREAM -D mapred.job.name='Apache Referer' \
  -input apache_logs/access_log* \
  -output apache_referer \
  -mapper $(pwd)/mapper.py \
  -reducer $(pwd)/reducer.py
```

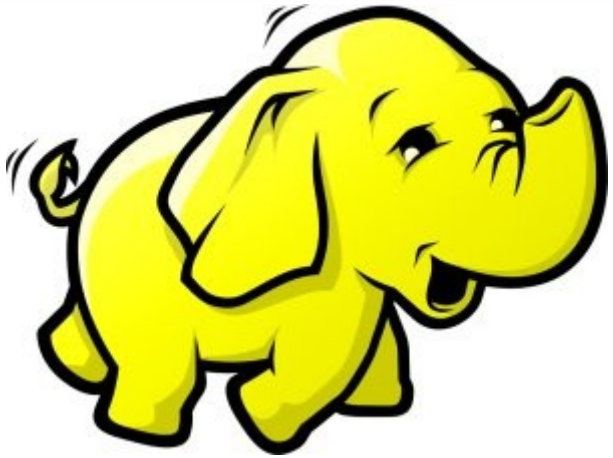
Results

```
# With $HADOOP_HOME
PATH=$PATH:${HADOOP_HOME}/bin

# View the resultant data sets in the HDFS
hadoop dfs -ls apache_referer

hadoop dfs -cat apache_referer/part*
```


Why Should I Care?



Google

IBM

YAHOO!

facebook

amazon.com

LinkedIn

last.fm

cloudera

Interwebs

http://hadoop.apache.org/
http://cloudera.com/
http://developer.yahoo.com/hadoop/tutorial/

Books

Hadoop: The Definitive Guide by Tom White
Pro Hadoop by Jason Venner

Videos

Google MapReduce Lectures
http://www.youtube.com/watch?v=yjPBkvYh-ss