# Distributed Machine Learning

# Implement Your MapReduce

Yi Wang

# Goal

- People think Hadoop MapReduce and Google MapReduce as unsurpassable.
- The truth is that MapReduce implementations could be extremely easy.

# One Liner

```
text=$(cat <<EOF
This is my cup
It is not your cup
My cup is white
Your cup is blue
EOF
)

echo $text \
| awk '{for (i=1; i<=NF; i++) print $i, 1;}' \
| sort \
| awk '{if ($1 != prev) {print prev, c; c=0; prev=$1;} c+=$2;}'
```

# One Liner

- The first awk command works on mapping.
- The sort command works on shuffling.
- The second awk command works on reducing.

# More Computers

- How if we want to make use remote computers?
- We can start worker processes remotely using ssh.

```
cat “hello world” | ssh yiwang-ld1 ‘cat’
```

- What is the role of ssh/sshd in this case?

# More Computers

```
text=$(cat <<EOF
This is my cup
It is not your cup
My cup is white
Your cup is blue
EOF
)

echo $text \
| ssh 192.168.1.2 ‘awk '{for (i=1; i<=NF; i++) print $i, 1;}'' \
| sort \
| awk '{if ($1 != prev) {print prev, c; c=0; prev=$1;} c+=$2;}'
```

# Distributing

- Big data are out of storage of any single computer.
- Distributing both computation and data.

# Distributing

```
Map='{for (i=1; i<=NF; i++) print $i, 1;}'

ssh worker1 awk $Map /input*.txt > /tmp/o1 &
ssh worker2 awk $Map /input*.txt > /tmp/o2 &
ssh worker3 awk $Map /input*.txt > /tmp/o3 &

cat /tmp/o*
| sort \
| awk '{if ($1 != prev) {print prev, c; c=0; prev=$1;} c+=$2;}'
```

# Distributed Shuffling

- Only one requirement: the same "word" goes to the same reduce worker.
- Each map worker generates NxM map output files, where N/M count map/reduce workers.
- Each file contains sorted map outputs, so merge sort works at combining every N outputs for each reduce worker.

# Bashreduce

- A MapReduce implementation in Bash:

  https://github.com/erikfrey/bashreduce

# Hadoop Streaming

- A MapReduce implementation in Java:

  http://princetonits.com/technology/hadoop-mapreduce-streaming-using-bash-script/

# Hadoop Streaming

- Discussion:
  - Hadoop MapReduce Java API allows customization of input/output file format. Is this a good design?
  - Hadoop MapReduce Java API supports partitioners. Is this a good design?

# MapReduce Lite

- A MapReduce implementation in C++:

  http://princetonits.com/technology/hadoop-mapreduce-streaming-using-bash-script/

# Code It!

- https://github.com/wangkuiyi/mapreduce-lite/blob/master/src/mapreduce_lite/design.txt