10-405 Big ML
Hadoop
Large-vocabulary Naïve Bayes with stream-and-sort

- For each example \( id, y, x_1, \ldots, x_d \) in train:
  - Print “Y=ANY += 1”
  - Print “Y=y += 1”
  - For \( j \) in 1..d:
    - Print “Y=y ^ X=x_j += 1”
- Sort the event-counter update “messages”
- Scan and add the sorted messages and output the final counter values

\[
\begin{align*}
Y=\text{business} ^ X=\text{zynga} & += 1 \\
Y=\text{sports} ^ X=\text{hat} & += 1 \\
Y=\text{sports} ^ X=\text{hockey} & += 1 \\
Y=\text{sports} ^ X=\text{hockey} & += 1 \\
Y=\text{sports} ^ X=\text{hockey} & += 1 \\
\end{align*}
\]

\[
\text{python MyTrainer.py train | sort | python MyCountAdder.py >model}
\]
MORE STREAM-AND-SORT EXAMPLES
Some other stream and sort tasks

• A task: classify Wikipedia pages
  – Features:
    • words on page: \( \text{src y } w_1 w_2 \ldots \)
    • or, outlinks from page: \( \text{src y } \text{dst}_1 \text{dst}_2 \ldots \)
    • how about \textbf{inlinks} to the page?
Some other stream and sort tasks

- outlinks from page: src dst₁ dst₂ ...

  - Algorithm:
    - For each input line src dst₁ dst₂ ... dstₙ print out
      - dst₁ inlinks.= src
      - dst₂ inlinks.= src
      - ...
      - dstₙ inlinks.= src
    - Sort this output
    - Collect the messages and group to get
      - dst src₁ src₂ ... srcₙ
Some other stream and sort tasks

\[ \text{\textbf{prevKey} = \text{Null}} \]
\[ \text{\textbf{sumForPrevKey} = 0} \]
\[ \text{\textbf{For each \textit{(event, delta)} in input:}} \]
  \[ \text{\textbf{If \textit{event} == \textbf{prevKey}}}
    \text{\quad \textbf{sumForPrevKey} += \text{\textit{delta}}} \]
  \[ \text{\textbf{Else}}
    \text{\quad \textbf{OutputPrevKey}()}
    \text{\quad \textbf{prevKey} = \textit{event}}
    \text{\quad \textbf{sumForPrevKey} = \text{\textit{delta}}} \]
\[ \text{\textbf{OutputPrevKey}()} \]

\text{define \textbf{OutputPrevKey}():}
\[ \text{\textbf{If \textbf{PrevKey} != \text{Null}}}
  \text{\quad \textbf{print \textbf{PrevKey}, \textbf{sumForPrevKey}}} \]

\[ \text{\textbf{dst}_1 \text{ inlinks.} = \text{src}} \]
\[ \text{\textbf{dst}_2 \text{ inlinks.} = \text{src}} \]
\[ \text{...} \]
\[ \text{\textbf{dst}_n \text{ inlinks.} = \text{src}} \]

\[ \text{\textbf{prevKey} = \text{Null}} \]
\[ \text{\textbf{docsForPrevKey} = [ ]} \]
\[ \text{\textbf{For each \textit{(dst, src)} in input:}} \]
  \[ \text{\textbf{If \textit{dst} == \textbf{prevKey}}}
    \text{\quad \textbf{docsForPrevKey}.append(\textit{src})} \]
  \[ \text{\textbf{Else}}
    \text{\quad \textbf{OutputPrevKey}()}
    \text{\quad \textbf{prevKey} = \textit{dst}}
    \text{\quad \textbf{docsForPrevKey} = [\textit{src}]} \]
\[ \text{\textbf{OutputPrevKey}()} \]

\text{define \textbf{OutputPrevKey}():}
\[ \text{\textbf{If \textbf{PrevKey} != \text{Null}}}
  \text{\quad \textbf{print \textbf{PrevKey}, \textbf{docsForPrevKey}}} \]
Some other stream and sort tasks

- What if we run this same program on the words on a page?
  - Features:
    - words on page: $src \ w_1 \ w_2 \ldots$
    - outlinks from page: $src \ dst_1 \ dst_2 \ldots$

an inverted index for the documents
Some other stream and sort tasks

• Finding unambiguous geographical names
• GeoNames.org: for each place in its database, stores
  – Several alternative names
  – Latitude/Longitude
  – …
• Problem: you need to soft-match names, and many names are ambiguous, if you allow an approximate match
Point Park College
Point Park University

Carnegie Mellon
Carnegie Mellon University
Carnegie Mellon School
Some other stream and sort tasks

• Finding almost unambiguous geographical names
• GeoNames.org: for each place in the database
  – print all plausible soft-match substrings in each alternative name, paired with the lat/long, e.g.
    • Carnegie Mellon University at lat1,lon1
    • Carnegie Mellon at lat1,lon1
    • Mellon University at lat1,lon1
    • Carnegie Mellon School at lat2,lon2
    • Carnegie Mellon at lat2,lon2
    • Mellon School at lat2,lon2
    • ...
  – Sort and collect… and filter
Some other stream and sort tasks

- prevKey = Null
- sumForPrevKey = 0
- For each \((event += delta)\) in input:
  - If \(event==prevKey\)
    - sumForPrevKey += delta
  - Else
    - OutputPrevKey()
    - prevKey = event
    - sumForPrevKey = delta

- OutputPrevKey()

define OutputPrevKey():
- If PrevKey!=Null
  - print PrevKey,sumForPrevKey

• prevKey = Null
• locOfPrevKey = Gaussian()
• For each \((place at lat,lon)\) in input:
  • If \(dst==prevKey\)
    • locOfPrevKey.observe(lat, lon)
  • Else
    • OutputPrevKey()
    • prevKey = place
    • locOfPrevKey = Gaussian()
    • locOfPrevKey.observe(lat, lon)

• OutputPrevKey()

define OutputPrevKey():
- If PrevKey!=Null and locOfPrevKey.stdDev() < 1 mile
  - print PrevKey, locOfPrevKey.avg()
Some other stream and sort tasks

Can incrementally maintain estimates for mean + variance as you add points to a Gaussian distribution

- prevKey = Null
- locOfPrevKey = Gaussian()
- For each (place at lat,lon) in input:
  - If dst==prevKey
    - locOfPrevKey.observe(lat, lon)
  - Else
    - OutputPrevKey()
    - prevKey = place
    - locOfPrevKey = Gaussian()
    - locOfPrevKey.observe(lat, lon)
  - OutputPrevKey()

**define OutputPrevKey():**
- If PrevKey!=Null and locOfPrevKey.stdDev() < 1 mile
  - print PrevKey, locOfPrevKey.avg()
PARALLELIZING STREAM AND SORT
Stream and Sort Counting → Distributed Counting

- example 1
- example 2
- example 3
- ....

C[x] += D

Logic to combine counter updates

process A

process B

process C

C[x1] += D1
C[x1] += D2
....
Stream and Sort Counting $\rightarrow$ Distributed Counting

- Counting logic
  - example 1
  - example 2
  - example 3
  - ...

- Standardized message routing logic
  - "C[x] += D"

- Logic to combine counter updates
  - C[x1] += D1
  - C[x1] += D2
  - ...

- Machines A1,…,
- Machines B1,…,
- Machines C1,…,

Can parallelize! Can parallelize! Can parallelize!
Stream and Sort Counting → Distributed Counting

- example 1
- example 2
- example 3
- ....

```
"C[x] += D"
```

Machines A1, ...

```
• C[x1] += D1
• C[x1] += D2
• ....
```

Logic to combine counter updates

Machines C1, ..,

Machines B1, ...

Counting logic

Sort

Machines A1, ...

Merges Spill Files

Spill 1

Spill 2

Spill 3

....

Stream and Sort Counting → Distributed Counting

- example 1
- example 2
- example 3
- ...

Counting logic

“C[x] += D”

Counter Machine

Sort

Sort key

- C[x1] += D1
- C[x1] += D2
- ...

Logic to combine counter updates

Reducer Machine

Combiner Machine

Observation: you can “reduce” in parallel (correctly) as no sort key is split across multiple machines
Observation: you can “reduce” in parallel (correctly) as no sort key is split across multiple machines
Stream and Sort Counting → Distributed Counting

Observation: you can “reduce” in parallel (correctly) as no sort key is split across multiple machines
Same holds for counter machines: you can count in parallel as no **sort key** is split across multiple reducer machines.
Stream and Sort Counting → Distributed Counting

Counter Machine 1
- example 1
- example 2
- example 3
- ...

Counting logic

Partition/Sort

Counter Machine 2
- example 1
- example 2
- example 3
- ...

Counting logic

Partition/Sort

Reducer Machine 1
- \(C[x1] += D1\)
- \(C[x1] += D2\)
- ...

Logic to combine counter updates

Reducer Machine 2
- \(C[x2] += D1\)
- \(C[x2] += D2\)
- ...

Combine counter updates

Merge Spill Files

Spill 1

Spill 2

Spill 3

Spill n
Mapper/counter machines run the “Map phase” of map-reduce

- Input different subsets of the total inputs
- Output (sort key,value) pairs
- The (key,value) pairs are partitioned based on the key
- Pairs from each partition will be sent to a different reducer machine.
Stream and Sort Counting $\rightarrow$ Distributed Counting

The shuffle/sort phrase:

- (key, value) pairs from each **partition** are sent to the right reducer
- The reducer will **sort** the pairs together to get **all the pairs with the same key** together.

The reduce phase:
- Each reducer will scan through the sorted key-value pairs.
Stream and Sort Counting → Distributed Counting

Counter Machine 1

- example 1
- example 2
- example 3
- ...

Counting logic

Partition/Sort

Spill 1
Spill 2
Spill 3
Spill n

Reducer Machine 1

- C[x1] += D1
- C[x1] += D2
- ...

Logic to combine counter updates

Merge Spill Files

Reducer Machine 2

- C[x2] += D1
- C[x2] += D2
- ...

combine counter updates

Merge Spill Files

Counter Machine 2

- example 1
- example 2
- example 3
- ...

Counting logic

Partition/Sort

Spill 1
Spill 2
Spill 3
Spill n
Distributed Stream-and-Sort: Map, Shuffle-Sort, Reduce

Map Process 1
- example 1
- example 2
- example 3
- ....

Counting logic

Partition/Sort

Distributed Shuffle-Sort

Spill 1

Spill 2

Spill 3

Spill n

Merge Spill Files

Reducer 1
- $C[x1] += D1$
- $C[x1] += D2$
- ....

Logic to combine counter updates

Reducer 2
- $C[x2] += D1$
- $C[x2] += D2$
- ....

Combine counter updates

Map Process 2
- example 1
- example 2
- example 3
- ....

Counting logic

Partition/Sort
HADOOP AS PARALLEL STREAM-AND-SORT
Hadoop: Distributed Stream-and-Sort

Local implementation:

cat input.txt | MAP | sort | REDUCE > output.txt

How could you parallelize this?
Local implementation:

cat input.txt | MAP | sort | REDUCE > output.txt

In parallel

1. Run N mappers
   mout.txt, mout2.txt, ...
2. Partition mouti.txt for the M mapper machines:
   part1.1.txt, part1.2..txt, ...
   partN.M.txt
3. Send each partition to the right reducer
**Hadoop: Distributed Stream-and-Sort**

Local implementation:

```
cat input.txt | MAP | sort | REDUCE > output.txt
```

In parallel:

1. Accept N partition files, part1.j.txt, ... partN.j.txt
2. Sort/merge them together to rinj.txt
3. Reduce to get the final result (reduce output for each key in partition j)

If necessary concatenate the reducer outputs to a single file.
Hadoop: Distributed Stream-and-Sort

Local implementation:

```
cat input.txt | MAP | sort | REDUCE > output.txt
```

In parallel:
1. Map
2. Partition
3. Send

In parallel:
1. Accept
2. Merge
3. Reduce

You could do this as a class project ...

but for really big data …
Hadoop: Distributed Stream-and-Sort

**Robustness:** with 10,000 machines, machines and disk drives fail *all the time.* How do you detect and recover?

**Efficiency:** with many jobs and programmers, how do you balance the loads? How do you limit network traffic and file i/o?

**Usability:** can programmers keep track of their data and jobs?

---

<table>
<thead>
<tr>
<th>In parallel</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Map</td>
<td></td>
</tr>
<tr>
<td>2. Partition</td>
<td></td>
</tr>
<tr>
<td>3. Send</td>
<td></td>
</tr>
</tbody>
</table>

In parallel:
1. Accept
2. Merge
3. Reduce
Hadoop: Intro

• pilfered from: Alona Fyshe, Jimmy Lin, Google, Cloudera

  • http://code.google.com/edu/submissions/mapreduce-minilecture/listing.html
  • http://code.google.com/edu/submissions/mapreduce/listing.html
Surprise, you mapreduced!

- Mapreduce has three main phases
  - Map (send each input record to a key)
  - Sort (put all of one key in the same place)
    - handled behind the scenes
  - Reduce (operate on each key and its set of values)
    - Terms come from functional programming:
      - `map(lambda x:x.upper(),["william","w","cohen"])->['WILLIAM', 'W', 'COHEN']`
      - `reduce(lambda x,y:x+-"-"+y,["william","w","cohen"])->"william-w-cohen"`
Mapreduce overview
Mapreduce: slow motion

- The canonical mapreduce example is word count
- Example corpus:
  
  Joe likes toast
  
  Jane likes toast with jam
  
  Joe burnt the toast
MR: slow motion: Map

Input

Joe likes toast  Map 1
Jane likes toast with jam  Map 2
Joe burnt the toast  Map 3

Output

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe</td>
<td>1</td>
</tr>
<tr>
<td>likes</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
<tr>
<td>Jane</td>
<td>1</td>
</tr>
<tr>
<td>likes</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
<tr>
<td>with</td>
<td>1</td>
</tr>
<tr>
<td>jam</td>
<td>1</td>
</tr>
<tr>
<td>Joe</td>
<td>1</td>
</tr>
<tr>
<td>burnt</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
</tbody>
</table>
MR: slow motion: Sort

**Input**

<table>
<thead>
<tr>
<th>Joe</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>likes</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
<tr>
<td>Jane</td>
<td>1</td>
</tr>
<tr>
<td>likes</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
<tr>
<td>with</td>
<td>1</td>
</tr>
<tr>
<td>jam</td>
<td>1</td>
</tr>
<tr>
<td>Joe</td>
<td>1</td>
</tr>
<tr>
<td>burnt</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
</tbody>
</table>

**Output**

<table>
<thead>
<tr>
<th>Joe</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jane</td>
<td>1</td>
</tr>
<tr>
<td>likes</td>
<td>1</td>
</tr>
<tr>
<td>likes</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
<tr>
<td>with</td>
<td>1</td>
</tr>
<tr>
<td>jam</td>
<td>1</td>
</tr>
<tr>
<td>burnt</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
</tr>
</tbody>
</table>
### MR: slow mo: Reduce

**Input**

<table>
<thead>
<tr>
<th>Input</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe</td>
<td>1</td>
</tr>
<tr>
<td>Joe</td>
<td>1</td>
</tr>
<tr>
<td>Jane</td>
<td>1</td>
</tr>
<tr>
<td>likes</td>
<td>1</td>
</tr>
<tr>
<td>likes</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
<tr>
<td>toast</td>
<td>1</td>
</tr>
<tr>
<td>with</td>
<td>1</td>
</tr>
<tr>
<td>jam</td>
<td>1</td>
</tr>
<tr>
<td>burnt</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
</tr>
</tbody>
</table>

**Reduce**

<table>
<thead>
<tr>
<th>Reduce</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduce 1</td>
<td>1</td>
</tr>
<tr>
<td>Reduce 2</td>
<td>1</td>
</tr>
<tr>
<td>Reduce 3</td>
<td>1</td>
</tr>
<tr>
<td>Reduce 4</td>
<td>1</td>
</tr>
<tr>
<td>Reduce 5</td>
<td>1</td>
</tr>
<tr>
<td>Reduce 6</td>
<td>1</td>
</tr>
<tr>
<td>Reduce 7</td>
<td>1</td>
</tr>
<tr>
<td>Reduce 8</td>
<td>1</td>
</tr>
</tbody>
</table>

**Output**

<table>
<thead>
<tr>
<th>Output</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joe</td>
<td>2</td>
</tr>
<tr>
<td>Jane</td>
<td>1</td>
</tr>
<tr>
<td>likes</td>
<td>2</td>
</tr>
<tr>
<td>toast</td>
<td>3</td>
</tr>
<tr>
<td>with</td>
<td>1</td>
</tr>
<tr>
<td>jam</td>
<td>1</td>
</tr>
<tr>
<td>burnt</td>
<td>1</td>
</tr>
<tr>
<td>the</td>
<td>1</td>
</tr>
</tbody>
</table>
Issue: reliability

• Questions:
  – How will you know when each machine is done?
    • Communication overhead
  – How will you know if a machine is dead?
    • Remember: we can to use a huge pile of cheap machines, so failures will be common!
  – Is it dead or just really slow?
Issue: reliability

- What’s the difference between slow and dead?
  - Who cares? Start a backup process.
    - If the process is slow because of machine issues, the backup may finish first
    - If it’s slow because you poorly partitioned your data... waiting is your punishment
Issue: reliability

• If a disk fails you can lose some intermediate output
  • Ignoring the missing data could give you wrong answers

• Who cares? if I’m going to run backup processes I might as well have backup copies of the intermediate data also
HDFS: The Hadoop File System

- Distributes data across the cluster
  - distributed file *looks like* a directory with shards as files inside it
  - makes an effort to run processes *locally* with the data
- Replicates data
  - default 3 copies of each file
- Optimized for streaming
  - really really big “blocks”
weak as the arrival of arbitraged cargoes from the West has put the local market under pressure… M14,M143,MCAT The Brent crude market on the Singapore International …
Hadoop job_201301231150_0778 on hadoopjt

User: wcohen
Job Name: streamjob6055532903853567038.jar
Job Setup: Successful
Status: Failed
Started at: Wed Jan 30 11:46:47 EST 2013
Failed in: 41sec
Job Cleanup: Successful
Black-listed TaskTrackers: 2
Job Scheduling information: 5 running map tasks using 5 map slots, 0 running reduce tasks using 0 reduce slots.

<table>
<thead>
<tr>
<th>Kind</th>
<th>% Complete</th>
<th>Num Tasks</th>
<th>Pending</th>
<th>Running</th>
<th>Complete</th>
<th>Killed</th>
</tr>
</thead>
<tbody>
<tr>
<td>map</td>
<td>100.00%</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reduce</td>
<td></td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
</tbody>
</table>

Failed/Killed Task Attempts:
- map: 35 / 5
- reduce: 0 / 0

Job Counters

<table>
<thead>
<tr>
<th>Counter</th>
<th>Map</th>
<th>Reduce</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rack-local map tasks</td>
<td>0</td>
<td>0</td>
<td>38</td>
</tr>
<tr>
<td>Launched map tasks</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>Data-local map tasks</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Failed map tasks</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Map Completion Graph - close
<table>
<thead>
<tr>
<th>Task</th>
<th>Complete</th>
<th>Status</th>
<th>Start Time</th>
<th>Finish Time</th>
<th>Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>task_201301231150_0778_m_000000</td>
<td>0.00%</td>
<td></td>
<td>30-Jan-2013 11:47:01</td>
<td>30-Jan-2013 11:47:25 (24sec)</td>
<td>java.lang.RuntimeException: PipeMa</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>at org.apache.hadoop.stream</td>
</tr>
<tr>
<td>Task Attempts</td>
<td>Machine</td>
<td>Status</td>
<td>Progress</td>
<td>Start Time</td>
<td>Finish Time</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------------</td>
<td>--------</td>
<td>----------</td>
<td>---------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>attempt_201301231150_0778_m_000000_0</td>
<td>/default-rack/cloud3u12.opencloud</td>
<td>FAILED</td>
<td>0.00%</td>
<td>30-Jan-2013 11:47:01</td>
<td>30-Jan-2013 11:47:06 (4sec)</td>
</tr>
<tr>
<td>attempt_201301231150_0778_m_000000_1</td>
<td>/default-rack/cloud2u28.opencloud</td>
<td>FAILED</td>
<td>0.00%</td>
<td>30-Jan-2013 11:47:07</td>
<td>30-Jan-2013 11:47:11 (4sec)</td>
</tr>
<tr>
<td>Name</td>
<td>Errors</td>
<td>Task Logs</td>
<td>Counters</td>
<td>Actions</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------------------------------------------</td>
<td>-----------</td>
<td>----------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1</td>
<td>Last 4KB</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:311)</td>
<td>Last 8KB</td>
<td></td>
<td>All</td>
<td></td>
</tr>
<tr>
<td></td>
<td>at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:540)</td>
<td>All</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:132)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:57)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:36)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:358)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>at org.apache.hadoop.mapred.MapTask.run(MapTask.java:307)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>at org.apache.hadoop.mapred.Child.main(Child.java:170)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2013 | java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1 | Last 4KB  | 1        |         |
|      | at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:311) | Last 8KB  |          | All     |
|      | at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:540)     | All       |          |         |
|      | at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:132)             |           |          |         |
|      | at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:57)                    |           |          |         |
|      | at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:36)         |           |          |         |
|      | at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:358)             |           |          |         |
|      | at org.apache.hadoop.mapred.MapTask.run(MapTask.java:307)                      |           |          |         |
|      | at org.apache.hadoop.mapred.Child.main(Child.java:170)                        |           |          |         |

2013 | java.lang.RuntimeException: PipeMapRed.waitOutputThreads(): subprocess failed with code 1 | Last 4KB  | 1        |         |
|      | at org.apache.hadoop.streaming.PipeMapRed.waitOutputThreads(PipeMapRed.java:311) | Last 8KB  |          | All     |
|      | at org.apache.hadoop.streaming.PipeMapRed.mapRedFinished(PipeMapRed.java:540)     | All       |          |         |
|      | at org.apache.hadoop.streaming.PipeMapper.close(PipeMapper.java:132)             |           |          |         |
|      | at org.apache.hadoop.mapred.MapRunner.run(MapRunner.java:57)                    |           |          |         |
|      | at org.apache.hadoop.streaming.PipeMapRunner.run(PipeMapRunner.java:36)         |           |          |         |
|      | at org.apache.hadoop.mapred.MapTask.runOldMapper(MapTask.java:358)             |           |          |         |
|      | at org.apache.hadoop.mapred.MapTask.run(MapTask.java:307)                      |           |          |         |
|      | at org.apache.hadoop.mapred.Child.main(Child.java:170)                        |           |          |         |
Task Logs: 'attempt_201301231150_0778_m_000000_0'

stdout logs

stderr logs

Exception in thread "main" java.lang.NoSuchMethodError: com/wcohen/StreamNB
Caused by: java.lang.ClassNotFoundException: com.wcohen.StreamNB
  at java.net.URLClassLoader$1.run(URLClassLoader.java:202)
  at java.security.AccessController.doPrivileged(Native Method)
  at java.net.URLClassLoader.findClass(URLClassLoader.java:190)
  at java.lang.ClassLoader.loadClass(ClassLoader.java:306)
  at sun.misc.Launcher$AppClassLoader.loadClass(Launcher.java:301)
  at java.lang.ClassLoader.loadClass(ClassLoader.java:247)
Could not find the main class: com.wcohen.StreamNB. Program will exit.
1. This would be pretty systems-y (remote copy files, waiting for remote processes, …)

2. It would take work to make run for 500 jobs

- **Reliability**: Replication, restarts, monitoring jobs,…
- **Efficiency**: load-balancing, reducing file/network i/o, optimizing file/network i/o,…
- **Useability**: stream defined datatypes, simple reduce functions, …
Map reduce with Hadoop streaming
Breaking this down…

• Our imaginary assignment uses key-value pairs. What’s the data structure for that? How do you interface with Hadoop?

• One very simple way: Hadoop’s streaming interface.
  – Mapper outputs key-value pairs as:
    • One pair per line, key and value tab-separated
  – Reducer reads in data in the same format
    • Lines are sorted so lines with the same key are adjacent.
An example (in Java, sorry):

- SmallStreamNB.java and StreamSumReducer.java:
To run locally:

test-small: small-events.txt nb.jar
time java -cp nb.jar com.wcohen.SmallStreamNB \ 
RCV1.small_test.txt MCAT,CCAT,GCAT,ECAT 2000 < small-events.txt \ 
| cut -f3 | sort | uniq -c

small-events.txt: nb.jar
time java -cp nb.jar com.wcohen.SmallStreamNB \ 
< RCV1.small_train.txt | sort -k1,1 \ 
| java -cp nb.jar com.wcohen.StreamSumReducer> small-events.txt
To run locally:

```
test-small: small-events.txt nb.jar
    time java -cp nb.jar com.wcohen.SmallStreamNB \
    RCV1.small_test.txt MCAT,CCAT,GCAT,ECAT 2000 < small-events.txt \\
    | cut -f3 | sort | uniq -c

small-events.txt: nb.jar
    time java -cp nb.jar com.wcohen.SmallStreamNB \
    < RCV1.small_train.txt | sort -k1,1 \\
    | java -cp nb.jar com.wcohen.StreamSumReducer > small-events.txt
```

python streamNB.py

python streamSumReducer.py
To train with streaming Hadoop you do this:

STRJAR = /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar

small-events-hs:
  hadoop fs -rmr rcv1/small/events
  hadoop jar $(STRJAR) \
    -input rcv1/small/sharded -output rcv1/small/events \
    -mapper 'java -Xmx512m -cp nb.jar com.wcohen.StreamNB' \
    -reducer 'java -Xmx512m -cp nb.jar com.wcohen_STREAMSumReducer' \
    -file nb.jar -numReduceTasks 10

But first you need to get your code and data to the “Hadoop file system”
hadoop fs -rmr rcv1/small/events
Moved to trash: hdfs://t1disc-pnn:8020/user/wcohen/rcv1/small/events
time hadoop jar /usr/lib/hadoop/contrib/streaming/hadoop-streaming-1.2.0.1.3.0.0-107.jar \  
   -input rcv1/small/sharded -output rcv1/small/events \  
   -mapper 'java -Xmx512m -cp ./lib/nb.jar com.wcohen.StreamNB' \  
   -reducer 'java -Xmx512m -cp ./lib/nb.jar com.wcohen.StreamSumReducer' \  
   -file nb.jar -numReduceTasks 10
packageJobJar: [nb.jar, /tmp/hadoop-wcohen/hadoop-unjar8935719391413249732/] [] /tmp/streamjob

14/02/05 11:24:56 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library
14/02/05 11:24:56 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo library [ha08101c2729dc0c9ff3]
14/02/05 11:24:56 WARN snappy.LoadSnappy: Snappy native library is available
14/02/05 11:24:56 INFO util.NativeCodeLoader: Loaded the native-hadoop library
14/02/05 11:24:56 INFO snappy.LoadSnappy: Snappy native library loaded
14/02/05 11:24:56 INFO mapred.FileInputFormat: Total input paths to process : 10
14/02/05 11:24:57 INFO streaming.StreamJob: getLocalDirs(): [/l/a/hadoop/mapred, /l/b/hadoop/tdoop/mapred]
14/02/05 11:24:57 INFO streaming.StreamJob: Running job: job_201312100900.1189
14/02/05 11:24:57 INFO streaming.StreamJob: To kill this job, run:
14/02/05 11:24:57 INFO streaming.StreamJob: /usr/lib/hadoop/libexec/./bin/hadoop job -Dmapreduce.local:50300 -kill job_201312100900.1189
14/02/05 11:24:57 INFO streaming.StreamJob: Tracking URL: http://t1disc-jt.disc.pdl.cmu.local:312100900.1189
14/02/05 11:24:58 INFO streaming.StreamJob: map 0% reduce 0%
14/02/05 11:25:09 INFO streaming.StreamJob: map 20% reduce 0%
14/02/05 11:25:14 INFO streaming.StreamJob: map 57% reduce 0%
14/02/05 11:25:15 INFO streaming.StreamJob: map 77% reduce 0%
14/02/05 11:25:16 INFO streaming.StreamJob: map 79% reduce 0%
14/02/05 11:25:17 INFO streaming.StreamJob: map 79% reduce 4%
14/02/05 11:25:18 INFO streaming.StreamJob: map 80% reduce 6%
14/02/05 11:25:19 INFO streaming.StreamJob: map 70% reduce 7%
14/02/05 11:25:24 INFO streaming.StreamJob: map 80% reduce 7%
14/02/05 11:25:25 INFO streaming.StreamJob: map 80% reduce 10%
14/02/05 11:25:26 INFO streaming.StreamJob: map 90% reduce 17%
14/02/05 11:25:27 INFO streaming.StreamJob: map 100% reduce 20%
14/02/05 11:25:28 INFO streaming.StreamJob: map 100% reduce 22%
14/02/05 11:25:31 INFO streaming.StreamJob: map 100% reduce 34%
14/02/05 11:25:32 INFO streaming.StreamJob: map 100% reduce 38%
14/02/05 11:25:33 INFO streaming.StreamJob: map 100% reduce 40%
14/02/05 11:25:34 INFO streaming.StreamJob: map 100% reduce 60%
14/02/05 11:25:35 INFO streaming.StreamJob: map 100% reduce 87%
14/02/05 11:25:36 INFO streaming.StreamJob: map 100% reduce 100%
14/02/05 11:25:37 INFO streaming.StreamJob: Job complete: job_201312100900.1189
14/02/05 11:25:37 INFO streaming.StreamJob: Output: rcv1/small/events
2.59User 0.13system 0:42.14elapsed 6%CPU (0avgtext+0avgdata 314512maxresident)k
0inputs+1080outputs (0major+2550minor)pagefaults 0swaps
To train with streaming Hadoop:

• First, you need to prepare the corpus by splitting it into *shards*
• … and distributing the shards to different machines:

```bash
wcohen@shell2:~/naive-bayes-demo$ hadoop fs --help
-help: Unknown command
Usage: java FsShell
    [-ls <path>]
    [-lsr <path>]
    [-du <path>]
    [-dus <path>]
    [-count[-q] <path>]
    [-mv <src> <dst>]
    [-cp <src> <dst>]
    [-rm [-skipTrash] <path>]
    [-rmdir [-skipTrash] <path>]
    [-expunge]
    [-put <localsrc> ... <dst>]
    [-copyFromLocal <localsrc> ... <dst>]
    [-moveFromLocal <localsrc> ... <dst>]
    [-get [-ignoreCrc] [-crc] <src> <localdst>]
    [-getmerge <src> <localdst> [addnl]]
    [-cat <src>]
```
weak as the arrival of arbitrage cargoes from the West has put the local market under pressure. In Singapore, May swaps fell to $21.30/$21.50 per barrel in late trade on Thursday from $21.70/$21.90 on Wednesday. While in Tokyo, first-half June open-spec naphtha was assessed at $207.00/$208.00 per tonne, compared with late-Wednesday's $213.00/$214.00. Added to these factors, traders said that a few petrochemicals were due to go into turnaround in the next few weeks which would further dampen demand. -- Melanie Goodfellow, London Newsroom, +44 171 542 7714.

M14,M143,MCAT The Brent crude market on the Singapore International Monetary Exchange (SIMEX) will be closed on Friday and Monday, SIMEX officials said on Tuesday. The closure will mark the corresponding close on Friday and Monday of the Brent market on the International Petroleum Exchange (IPE) in London. SIMEX Brent operates a mutual offset system with the IPE in London, so SIMEX tends to close in line with the U.K.. -- Singapore Newsroom (+65 870 3081)
To train with streaming Hadoop:

• One way to shard text:
  – `hadoop fs -put LocalFileName HDFSName`
  – then run a streaming job with ‘cat’ as mapper and reducer
  – and specify the number of shards you want with option
    `-numReduceTasks`
To train with streaming Hadoop:

- Next, prepare your code for upload and distribution to the machines cluster

```bash
nb.jar: StreamSumReducer.java StreamNB.java SmallStreamNB.java
javac -d classes StreamSumReducer.java StreamNB.java SmallStreamNB.java
jar -cvf nb.jar -C classes .
```
Now you can run streaming Hadoop:

```bash
small-events-hs:
  hadoop fs -rmr rcv1/small/events
  hadoop jar $(STRJAR) \
    -input rcv1/small/sharded -output rcv1/small/events \
    -mapper 'java -Xmx512m -cp nb.jar com.wcohen.StreamNB' \
    -reducer 'java -Xmx512m -cp nb.jar com.wcohen.StreamSumReducer' \
    -file nb.jar -numReduceTasks 10
```

```bash
small-events-hs:
  hadoop fs -rmr rcv1/small/events
  hadoop jar $(STRJAR) \
    -input rcv1/small/sharded -output rcv1/small/events \
    -mapper 'python streamNB.py' \
    -reducer 'python streamSumReducer.py' \
    -file streamNB.py -file streamSumReducer.py
```
Map reduce without Hadoop streaming
“Real” Hadoop

- Streaming is simple but
  - There’s no typechecking of inputs/outputs
  - You need to parse strings a lot
  - You can’t use compact binary encodings
  - ...
  - basically you have limited control over the messages you’re sending

- i/o costs = O(message size) often dominates
public static void main(String[] args) throws Exception {

    Configuration conf = new Configuration();
    Job job = new Job(conf, "wordcount");

    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);

    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);

    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);

    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));

    job.waitForCompletion(true);
}
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

private final static IntWritable one = new IntWritable(1);
private Text word = new Text();

public void map(LongWritable key, Text value, Context context) throws <stuff> {
    String line = value.toString();
    StringTokenizer tokenizer = new StringTokenizer(line);
    while (tokenizer.hasMoreTokens()) {
        word.set(tokenizer.nextToken());
        context.write(word, one);
    }
}
}
public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

    public void reduce(Text key, Iterable<IntWritable> values, Context context)
    throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        context.write(key, new IntWritable(sum));
    }
}
“Real” Hadoop vs Streaming Hadoop

• Tradeoff: simplicity vs control
• In general:
  – If you want to really optimize you need to get down to the actual Hadoop layers
  – Often it’s better to work with abstractions “higher up”
Debugging Map-Reduce
Some common pitfalls

• You have no control over the order in which reduces are performed
• You have no* control over the order in which you encounter reduce values
  • *by default anyway
• The only ordering you should assume is that Reducers always start after Mappers
Some common pitfalls

• You should assume your Maps and Reduces will be taking place on different machines with different memory spaces
• Don’t make a static variable and assume that other processes can read it
  – They can’t.
  – It appear that they can when run locally, but they can’t
  – No really, don’t do this.
Some common pitfalls

• Do not communicate between mappers or between reducers
  • overhead is high
  • you don’t know which mappers/reducers are actually running at any given point
  • there’s no easy way to find out what machine they’re running on
    – because you shouldn’t be looking for them anyway
When mapreduce doesn’t fit

• The beauty of mapreduce is its separability and independence

• If you find yourself trying to communicate between processes
  • you’re doing it wrong
    » or
  • what you’re doing is not a mapreduce
When mapreduce doesn’t fit

• Not everything is a mapreduce
• Sometimes you need more communication
  – We’ll talk about other programming paradigms later
What’s so tricky about MapReduce?

• Really, nothing. It’s easy.
• What’s often tricky is figuring out how to write an algorithm as a series of map-reduce substeps.

  – How and when do you parallelize?
  – When should you even try to do this?

when should you use a different model?
Performance

• IMPORTANT
  – You may not have room for all reduce values in memory
    • In fact you should PLAN not to have memory for all values
    • Remember, small machines are much cheaper
      – you have a limited budget
Combiners in Hadoop
Some of this is wasteful

• Remember - moving data around and writing to/reading from disk are very expensive operations

• No reducer can start until:
  • all mappers are done
  • data in its partition has been sorted
How much does buffering help?

<table>
<thead>
<tr>
<th>BUFFER_SIZE</th>
<th>Time</th>
<th>Message Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td></td>
<td>1.7M words</td>
</tr>
<tr>
<td>100</td>
<td>47s</td>
<td>1.2M</td>
</tr>
<tr>
<td>1,000</td>
<td>42s</td>
<td>1.0M</td>
</tr>
<tr>
<td>10,000</td>
<td>30s</td>
<td>0.7M</td>
</tr>
<tr>
<td>100,000</td>
<td>16s</td>
<td>0.24M</td>
</tr>
<tr>
<td>1,000,000</td>
<td>13s</td>
<td>0.16M</td>
</tr>
<tr>
<td>limit</td>
<td></td>
<td>0.05M</td>
</tr>
</tbody>
</table>

Recall idea here: in stream-and-sort, use a buffer to *accumulate* counts in messages for common words *before* the sort so *sort input was smaller*
Combiners

• Sits between the map and the shuffle
  – Do some of the reducing while you’re waiting for other stuff to happen
  – Avoid moving all of that data over the network
  – Eg, for wordcount: instead of sending \((\text{word,1})\) send \((\text{word,n})\) where \(n\) is a partial count (over data seen by that mapper)
    • Reducer still just sums the counts

• Only applicable when
  – order of reduce values doesn’t matter (since order is undetermined)
  – effect is cumulative
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

private final static IntWritable one = new IntWritable(1);
private Text word = new Text();

public void map(LongWritable key, Text value, Context context) throws <stuff> {
    String line = value.toString();
    Stringtokenizer tokenizer = new Stringtokenizer(line);
    while (tokenizer.hasMoreTokens()) {
        word.set(tokenizer.nextToken());
        context.write(word, one);
    }
}

public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {

public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
    int sum = 0;
    for (IntWritable val : values) {
        sum += val.get();
    }
    context.write(key, new IntWritable(sum));
}

public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    Job job = new Job(conf, "wordcount");
    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);
    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    job.waitForCompletion(true);
}
public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();

    public void map(LongWritable key, Text value, Context context) throws <stuff> {
        String line = value.toString();
        StringTokenizer tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()) {
            word.set(tokenizer.nextToken());
            context.write(word, one);
        }
    }
}

public static void main(String[] args) throws Exception {

    Configuration conf = new Configuration();
    Job job = new Job(conf, "wordcount");
    job.setMapperClass(Map.class);
    job.setReducerClass(Reduce.class);
    job.setCombinerClass(Reduce.class);
    job.setInputFormatClass(TextInputFormat.class);
    job.setOutputFormatClass(TextOutputFormat.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    job.waitForCompletion(true);
}
Combiners in streaming Hadoop

```
STRJAR = /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar

small-events-hs:
    hadoop fs -r mr rcv1/small/events
    hadoop jar $(STRJAR) \
        -input rcv1/small/sharded -output rcv1/small/events \
        -mapper 'java -Xmx512m -cp nb.jar com.wcohen.StreamNB' \
        -reducer 'java -Xmx512m -cp nb.jar com.wcohen.StreamSumReducer' \
        -file nb.jar -num Reduce Tasks 10
```
Deja vu: Combiner = Reducer

• Often the combiner is the reducer.
  – like for word count
  – but not always

  – remember you have no control over when/whether the combiner is applied