Lecture 16:
- Quiz 3
- More on MapReduce
- Data Processing Architectures (intro)

Announcements:
- HW-2 due Mon
- Exam 1 next Friday

MapReduce: Fault Tolerance (Basics)

The Master worker finds failures:
- Master periodically pings workers
- Machines (workers) that don’t respond are considered “failed”

In-progress tasks on failed workers:
- Map and Reduce tasks on failed workers are rescheduled
- Master failure requires entire job to be restarted

Completed tasks on failed workers:
- Reducer tasks are fine – output stored in GFS
- Mapper tasks rescheduled since no access to output files

Reduce tasks are notified of failed workers (to read from restarted Mapper)
Map Reduce: Data-Mining/Machine-Learning Examples

Many “standard” ML algorithms can be converted to MapReduce
• e.g., many follow the “statistical query” pattern (*)
• expressed as sums applying (statistical) functions over data points
• includes regressions, naive bayes, $k$-means, neural nets, PCA, SVM, etc.
• shown to scale linearly on MapReduce (with number of cores)

We’ll go through some “sketches” of how these can be modeled
• specifically for $k$-means and naive bayes

(*) Chu et al., “Map-Reduce for Machine Learning on Multicore” (NIPS’06)

Map Reduce: $k$-means review

Goal of $k$-means is to divide data (rows) into $k$ clusters (partitions)
• A centroid is the center “point” (row) of a cluster (not nec. a row)
• Using Euclidean distance, cluster’s centroid is average of each feature value

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

What is the centroid? ... (4, 6)

$k$-means computes the fixed-point of cluster centroids:
1. pick $k$ rows to be centroids
2. find rows closest to each centroid (creates $k$ initial clusters)
3. recompute the centroids
4. repeat 2 and 3 until centroids don’t change
Map Reduce: $k$-means sketch

Sketch of modeling $k$-means centroid computation in MapReduce ...

Centroids

Split 0
Split 1
...
Split M-1

Data rows
(feature vectors)

Map 1
(centroid, row)

...(centroid, cluster rows)

Reduce 1
new centroid

© S. Bowers
CPSC 324, Spring 2024