Towards Optimizing Hadoop Provisioning in the Cloud

Karthik Kambatla, Purdue University
Abhinav Pathak, Purdue University
Himabindu Pucha, IBM Research Almaden
Data analytics is important/prevalent
  - MapReduce – highly scalable solution

Performing Hadoop–like data analytics in the cloud is particularly synergistic
  - Utility model
    - Request/Relinquish resources on demand
    - Billed by machine hours
      - Not limited by number of machines
Challenge: Hadoop Provisioning

- Provisioning
  - Allocate resources
  - Configure for best utilization

- Current tools
  - Hadoop on Demand, Cloudera, etc.
  - Automate deployment, Do Not Optimize Resources!

- Our Contribution: Optimized provisioning
  - Minimize cost, Maximize Performance
Our Proposal

Hadoop Application

Input Data

RS Maximizer

<Conf, Cluster>

<table>
<thead>
<tr>
<th>Config</th>
<th># node</th>
<th>Cluster</th>
<th>Est. Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>N1</td>
<td>Cl x</td>
<td>T1</td>
</tr>
<tr>
<td>C2</td>
<td>N2</td>
<td>Cl y</td>
<td>T2</td>
</tr>
<tr>
<td>C3</td>
<td>N3</td>
<td>Cl z</td>
<td>T3</td>
</tr>
</tbody>
</table>

RS Sizer
Number of Reduces doesn’t affect performance

Significant Performance Difference

Optimal: 8 maps
Grep

Too low doesn’t work! Too high doesn’t work either!

Grep - 80GB

Time (Seconds)

Number of Reduces

1 map
4 maps
8 maps
16 maps
24 maps

Karthik Kambatla - HotCloud 6/19/2009
Sort

Same configuration would not work across applications

Number of Reduces also affects performance

So does number of maps

Best performance at (8, 8)
RS Maximizer

1. INPUT DATA
2. Generate Small Chunk
3. Run Application On Hadoop
4. Generate Signature for Small Data Run
5. Generate New Config. File
6. RUN App. On Entire Data With New Configuration

SIG. Database
App Sig
Optimum Config.
Preliminary Results

- Matrix addition, multifile–wordcount
  - Signature similar to wordcount
  - Optimal configuration is the same
Future Work

- Add a feedback phase
  - Check if predicted values are optimal
  - Else predict new optimal configuration

- RS Sizer
Thank You

Questions?