TPCx-HS Experiments

Todor Ivanov  (todor@dbis.cs.uni-frankfurt.de)
Sead Izberovic

Frankfurt Big Data Lab
-understanding and applying technologies for Big Data-

Goethe University Frankfurt am Main, Germany
http://www.bigdata.uni-frankfurt.de/
New Cluster Setup

- Operating System: Ubuntu Server 14.04.1. LTS
- Cloudera’s Hadoop Distribution - CDH 5.2
- Replication Factor of 2 (only 3 worker nodes)

Goal → Run end-to-end, analytical Big Data benchmark (BigBench) to evaluate the platform!

Initial results → BigBench performance was very slow! → Shared 1Gbit Network

Solution → Upgrade to Dedicated 1Gbit Switch (around 30 €)
Questions

• What about the performance differences between the two setups (shared vs. dedicated networks)?

• How can we measure the network improvement in the new setup (dedicated network)?

→ Use network intensive workload/benchmark like TPCx-HS.
**TPCx-HS: TPC Express for Hadoop Systems**

- **X:** Express, **H:** Hadoop, **S:** Sort
- TPCx-HS [1],[2] is the **first industry standard Big Data Benchmark** released in July 2014
- Based on **TeraSort** and consists of 4 modules: HSGen, HSDataCheck, HSSort & HValidate
- Scale Factors following stepped size model: 100GB, 300GB, 1TB, 3TB, 10TB ....
- The TPCx-HS specification defines three major metrics:
  - Performance metric (HSph@SF)
  - Price-performance metric ($/HSph@SF)
  - Power per performance metric (Watts/HSph@SF)
- Use **TPCx-HS Kit 1.1.2**

---

Performance (Shared vs. Dedicated)

- The presented results are obtained by executing the TPCx-HS kit provided on the official TPC website (www.tpc.org). However, the reported times and metrics are experimental, not audited by any authorized organization and therefore not directly comparable with other officially published full disclosure reports.

- Tested with 3 scale factors: 100GB, 300GB, 1TB

- The shared setup is 5 times slower compared to the dedicated one.
Results

- **Data Δ** column represents the difference in percent of the Data Size to the data baseline in our case 100GB.
- **Time (Sec)** shows the average time in seconds of two complete TPCx-HS runs.
- **Time Stdv (%)** shows the standard deviation of Time (Sec) in percent between the two runs.
- **Time Δ (%)** represents the difference in percent of Time (Sec) to the time baseline in our case scale factor 0.1.

<table>
<thead>
<tr>
<th>Scale Factor</th>
<th>Data Size</th>
<th>Data Δ (%)</th>
<th>Network</th>
<th>Metric (HSph@SF)</th>
<th>Time (Sec)</th>
<th>Time Stdv (%)</th>
<th>Time Δ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>100 GB</td>
<td>baseline</td>
<td>shared</td>
<td>0.03</td>
<td>10721.75</td>
<td>0.59</td>
<td>baseline</td>
</tr>
<tr>
<td>0.3</td>
<td>300 GB</td>
<td>+200</td>
<td>shared</td>
<td>0.03</td>
<td>32142.75</td>
<td>0.50</td>
<td>+199.79</td>
</tr>
<tr>
<td>1</td>
<td>1 TB</td>
<td>+900</td>
<td>shared</td>
<td>0.03</td>
<td>105483.00</td>
<td>0.41</td>
<td>+883.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>100 GB</td>
<td>baseline</td>
<td>dedicated</td>
<td>0.16</td>
<td>2234.75</td>
<td>0.72</td>
<td>baseline</td>
</tr>
<tr>
<td>0.3</td>
<td>300 GB</td>
<td>+200</td>
<td>dedicated</td>
<td>0.18</td>
<td>6001.00</td>
<td>0.89</td>
<td>+168.53</td>
</tr>
<tr>
<td>1</td>
<td>1 TB</td>
<td>+900</td>
<td>dedicated</td>
<td>0.17</td>
<td>21047.75</td>
<td>1.53</td>
<td>+841.84</td>
</tr>
</tbody>
</table>
Data Scaling Behavior

- TPCx-HS Scaling Behavior (0 on the X and Y-axis is equal to the baseline of SF 0.1/100GB)
**CPU (Shared vs. Dedicated)**

- Performance Analysis Tool (PAT) ([https://github.com/intel-hadoop/PAT](https://github.com/intel-hadoop/PAT))
- *Worker Node* average statistics - measured for 100GB scale factor.

- The *shared* setup **performs 5-6 times slower** than the *dedicated* setup.
Disk Bandwidth (Shared vs. Dedicated)

- **Dedicated** setup: on average *read throughput* is around **6.4 MB** per second and *write throughput* is around **18.6 MB** per second.

- **Shared** setup: on average *read throughput* is around **1.4 MB** per second and *write throughput* is around **4 MB** per second.
Network (Shared vs. Dedicated)

- **Dedicated** setup: on average received **32.8MB per second** and transmitted **30.6MB per second**.

- **Shared** setup: on average are received **7.1MB per second** and transmitted **6.4MB per second**

- The dedicated setup achieves **almost 5 times better network utilization**.
# Summary of Resource Utilization

<table>
<thead>
<tr>
<th>Worker Nodes</th>
<th>Dedicated 1Gbit</th>
<th>Shared 1Gbit</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Network Type:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Scale Factor:</strong></td>
<td>100GB</td>
<td>100GB</td>
</tr>
<tr>
<td><strong>Avg. CPU Utilization - User %</strong></td>
<td>56.04</td>
<td>12.44</td>
</tr>
<tr>
<td><strong>Avg. CPU Utilization - System %</strong></td>
<td>9.52</td>
<td>3.71</td>
</tr>
<tr>
<td><strong>Avg. CPU Utilization - IOwait %</strong></td>
<td>3.61</td>
<td>1.28</td>
</tr>
<tr>
<td><strong>Avg. Context Switches per Second</strong></td>
<td>20788.16</td>
<td>14233.08</td>
</tr>
<tr>
<td><strong>Memory Utilization %</strong></td>
<td>92.31</td>
<td>92.93</td>
</tr>
<tr>
<td><strong>Avg. Kbytes Transmitted per Second</strong></td>
<td>31363.57</td>
<td>6548.16</td>
</tr>
<tr>
<td><strong>Avg. Kbytes Received per Second</strong></td>
<td>33636.93</td>
<td>7297.27</td>
</tr>
<tr>
<td><strong>Avg. Kbytes Read per Second</strong></td>
<td>6532.24</td>
<td>1438.23</td>
</tr>
<tr>
<td><strong>Avg. Kbytes Written per Second</strong></td>
<td>19010.01</td>
<td>4087.33</td>
</tr>
<tr>
<td><strong>Avg. Read Requests per Second</strong></td>
<td>111.75</td>
<td>24.70</td>
</tr>
<tr>
<td><strong>Avg. Write Requests per Second</strong></td>
<td>39.50</td>
<td>9.71</td>
</tr>
<tr>
<td><strong>Avg. I/O Latencies in Milliseconds</strong></td>
<td>136.87</td>
<td>69.83</td>
</tr>
</tbody>
</table>

**CPU**
**Memory**
**Network**
**Disk**
Lessons Learned

• The dedicated network setup is around **5 times faster** than the shared network setup in terms of:
  – Execution time
  – HSp@SF metric
  – Average read and write throughput per second
  – Network utilization

• The TPCx-HS benchmark is a good choice for testing/comparing the network performance of clusters!
Evaluating Hadoop Clusters with TPCx-HS
http://arxiv.org/abs/1509.03486

Todor Ivanov and Sead Izberovic
Contact

Todor Ivanov
todor@dbis.cs.uni-frankfurt.de

Frankfurt Big Data Lab
-understanding and applying technologies for Big Data-

Goethe University Frankfurt am Main, Germany
http://www.bigdata.uni-frankfurt.de/