The Cloud-Based Sensor Data Warehouse

Wen-Yuan Ku, 卞文元
GIS Center, Feng Chia University, Taiwan
2011/3/25
Outline

- Introduction
- Cloud-Based Database – HBase
- Design of Sensor Data Structure
- Experiment
- Conclusion
Introduction

- The sensors have been widely used in human observation, environment monitoring or biological activities.
- Sensor data with time characteristic.
- Historic sensor data will require a large amount of data storage.
Debris Flow Observation Station

Satellite Monitoring

Data receiving

Debris Flow Observation Station Data Center

Monitoring sensor

10GB per day
Introduction (cont’d)

- The RDBMS is the most widely used database
- Advantage: easy to manage data (like SQL supported, join…)
- Disadvantage: extremely expensive if huge amount of data storage (PB above) is needed
Cloud-Based Database-HBase

**What is HBase**
- is the Hadoop database
- is a distributed column-oriented
- is a distributed data store that can scale to 1,000s of commodity servers
- integrated into the Hadoop MapReduce framework

**Benefits**
- High scalability
- High availability
- High performance
## Design of Sensor Data Structure

### Rowkey design
- **HBase** is a Key-Value database.
- HBase only has an indexed key: **RowKey**.
- **RowKey format:** 
  - `<SensorID>_<YYYYMMDDHHmmss>`

Ex. SensorID: Geophone001
Time: 2011/03/01 08:00:00

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>“Geophone001_20110301080000”</td>
<td>124.4</td>
<td>102.6</td>
<td>95.1</td>
</tr>
</tbody>
</table>
Experiment (1/5)

Experimental Environment

- AMD Phenom 2.3G X 4, 4G RAM, 3 machines
- Ubuntu 9.10
- Hadoop 0.20.1
- HBase 0.20.3
Experiment(2/5)

- Imported 100 million records by MapReduce
Experiment (3/5)

Source data: 13.6GB
Spent time: 6hr 3mins 5sec
Written 5000 records/sec
Experiment (4/5)

**Reading performance**

- **simulate** 50, 100, 200, 300, 400, 500 client
- Read data randomly for 2 minutes
- Take average response time
  - 50 clients: 1.9ms
  - 500 clients: 5.6ms
## Experiment (5/5)

**Writing performance**

Writing 1000 records into HBase

**Number of Columns**

1, 10, 100, 500

<table>
<thead>
<tr>
<th></th>
<th># of columns</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment</strong></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>500</td>
</tr>
<tr>
<td><strong>Write Columns/Sec</strong></td>
<td>944</td>
</tr>
<tr>
<td></td>
<td>949</td>
</tr>
<tr>
<td></td>
<td>940</td>
</tr>
<tr>
<td></td>
<td>951</td>
</tr>
<tr>
<td><strong>Write Rows/Sec</strong></td>
<td>944</td>
</tr>
<tr>
<td></td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>1.92</td>
</tr>
</tbody>
</table>
Conclusion

- Sensor-produced data is calculated in GBs.
- If using distributed column-oriented database, e.g. HBase, data will be stored on separated machines for more efficient I/O.
- From our experimental test results, the number of columns in the table will affect performance of data accessing. More columns a data row has, more data access time it will increase.
Conclusion (cont’d)

- It will increase the efficiency of database I/O if data can be converted to XML format and save XML data to **single column**.
- We imported **100 million** records into Hbase and simulated **50 to 500 clients** for accessing the HBase at the same time. The average response time is less than **6ms**.
- It proves that HBase is very suitable for sensor data warehouse.
Thanks for your listening.