Hadoop tutorial

1 - Introduction to Hadoop

A. Hammad, A. García | September 7, 2011
The Hadoop Framework

Data intensive computing on commodity hardware

- Yahoo’s (re)implementation of Google’s Map-Reduce
  - simple-process huge amounts of data in efficient way
- highly scalable filesystem, computing coupled to storage
The Map-Reduce paradigm

- (Split step)
- Map step
- (Shuffle step)
- Reduce step
Scalability

- Scalability achieved through data locality
- computing goes to data, not otherwise
Scalability

Daily production usage in Yahoo!, Facebook

- clusters with thousands of nodes
- 30 PB of data and growing
  - whole dataset processed daily!
  - sorting benchmarks winners, e.g. 1 TB data sorted in 1 minute by 3800 nodes (2009)
Applications

- Programming language Java
- Hadoop Pipes API for C++
- Streaming for any executables (e.g. shell utilities) as mapper or reducer

Example

```
hadoop jar $HADOOP_LIB/hadoop-streaming.jar
    -input /dfsInputDir/myInputData -mapper "shellMapper.sh"
    -reducer "shellReducer.sh" -output /dfsOutputDir/myResults
```
Use cases

- Data intensive computing (high IO)
- High parallelization

Complementary to

- message passing (MPI, ...)
- RDBMS, traditional databases

<table>
<thead>
<tr>
<th></th>
<th>Traditional RDBMS</th>
<th>MapReduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data size</td>
<td>Gigabytes</td>
<td>Petabytes</td>
</tr>
<tr>
<td>Access</td>
<td>Interactive and batch</td>
<td>Batch</td>
</tr>
<tr>
<td>Updates</td>
<td>Read and write many times</td>
<td>Write once, read many times</td>
</tr>
<tr>
<td>Structure</td>
<td>Static schema</td>
<td>Dynamic schema</td>
</tr>
<tr>
<td>Integrity</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Scaling</td>
<td>Nonlinear</td>
<td>Linear</td>
</tr>
</tbody>
</table>
The Hadoop ecosystem

- Apache project, **open source**
- many subprojects
  - common, HDFS, MapReduce
  - **Pig**: data flow language
  - **Hive**: a distributed data warehouse, SQL-based language inspired by Google’s Bigtables; billions rows, million columns
  - **HBase**: a distributed, column-oriented database
  - **Zookeeper**: a distributed, highly available coordination service
  - **Oozie**: a MapReduce workflow service
  - ...

- **backed by big web players** (Yahoo!, Facebook, Amazon, Twitter, ...)
  - available as a Service: Amazon’s Elastic MapReduce
http://hadoop.apache.org/common/docs/current/mapred_tutorial.html
Thanks for listening!

Questions?