Hadoop tutorial

5 - The Pig data processing language

A. Hammad, A. García | September 7, 2011
The background

<table>
<thead>
<tr>
<th>Hadoop MapReduce</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ data intensive computing on commodity hardware</td>
</tr>
<tr>
<td>+ great scalability</td>
</tr>
<tr>
<td>- too low level for some analysis tasks</td>
</tr>
<tr>
<td>- gets messy if need to combine several MR steps</td>
</tr>
</tbody>
</table>

⇒ need to wrap complexity in higher level framework
The Pig data processing language

- higher level language
- data flow language
- adapted for processing big amounts of moderately structured data
- more structured data than MapReduce, less structured than RDBMS
- powerful transformation functions and data model
  “Pigs eat anything” :-)

- Pig Latin language
- execution environment/interpreter
  ⇒ converts Pig instructions into (series of) MapReduce programs
  transparently
The Pig data processing language

- Pig is a scripting language for exploring large data sets
  ⇒ 40% of Yahoo! Hadoop usage
- has powerful debugging concepts: ILLUSTRATE to operate on representative sample subset
- extensible: **User Defined Functions** for loading, evaluating, storing
- language structure dictated by MapReduce paradigm: ideal for batch processing large dataset, touching most of the data (cf, databases optimized for tiny operations)
Hands on!

Questions?
Let’s try it!