WIKIPEDIA HADOOP

STEVEN STULGA
SPRING 2016
CS 4624 MULTIMEDIA/HYPERTEXT/INFORMATION ACCESS
PROFESSOR EDWARD FOX
CLIENT: SHIVAM MAHARSHI
Goal: Import the entire English Wikipedia with full revision history into Apache Hadoop and configure it to be searchable from Apache Solr

A prototype outlining the process
- importing the data
- converting the data
- configuring the system
IMPLEMENTATION

- Creating a prototype
  - Downloading small segment of dataset
  - Uncompressing, extracting, converting this data
  - Import it into local HDFS
  - Configure Solr to index data
- Follow same steps for full data on cluster
  - MUCH larger set of data
  - Working on true cluster
<table>
<thead>
<tr>
<th></th>
<th>Prototype Test Data</th>
<th>English Wikipedia Current Version</th>
<th>English Wikipedia with Full Revision History</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size Compressed</strong></td>
<td>365 KB</td>
<td>23.2 GB</td>
<td>107 GB</td>
</tr>
<tr>
<td><strong>Size Decompressed</strong></td>
<td>46 MB</td>
<td>51 GB</td>
<td>10 TB</td>
</tr>
<tr>
<td><strong>Number of Articles</strong></td>
<td>21</td>
<td>5,137,000</td>
<td>5,137,000</td>
</tr>
</tbody>
</table>
Succesess

107GB of compressed XML data

This dataset is ~10TB decompressed
SUCCESSES

Query

Response with list of 21 Documents

Document title
LESSONS LEARNED

- Work with data in small pieces
- There isn’t always the “how to” guide you need
- Implement full prototype and outline all steps
- Effective learning requires more time and research than you expect
- Communication can be slow and hinder progress
Shivam Maharshi, the client, for presenting the project and giving guidance about how to start and tackle challenges

Sunshin Lee, who provided access to the remote machine and answered questions about execution on the machine

Dr. Fox, for great feedback and help with the presentation and documentation
THANK YOU!

Questions?