HLoader

CMS Data-Services Ingestion into CERN’s Hadoop Big Data Analytics Infrastructure
The Challenge

- Schedule, control and monitor bulk data transfer

- Frequent requests
  - CMS data popularity
  - ATLAS and CMS Job monitoring
  - ACCLOG

- Manually executed jobs
  - could be partially automated
  - incrementally
  - communicate with end user

Security
### EVENT_ID

<table>
<thead>
<tr>
<th>EVENT_ID (PK)</th>
<th>MUONS</th>
<th>ALPHAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>251</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>501</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>750</td>
<td></td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Number of Mappers: 4

MAP 1

MAP 2

MAP 3

MAP 4
Solution

Oracle Databases

REST API

Agent

Meta DB

Hadoop Clusters

Secure

- CERN SSO authentication
- Authorization
- Secure SSH tunneling
- Secure password input

Modular
Flexible, independent and interchangeable components

Architecture
Summary

– Easily expandable framework and service for transferring data from Oracle to Hadoop

– Designed with automation in mind
  minimal administrator intervention needed

– Service built for easy usage
  easy to use for the routine jobs
### Current jobs

<table>
<thead>
<tr>
<th>Edit</th>
<th>Owner</th>
<th>Started</th>
<th>Source database</th>
<th>Schema</th>
<th>Object</th>
<th>Destination path</th>
<th>Cluster id</th>
<th>Incremental</th>
</tr>
</thead>
<tbody>
<tr>
<td>edit</td>
<td>kzbledzi</td>
<td>Aug 10, 2015 7:14:17 PM</td>
<td>d3r.cern.ch</td>
<td>gp.user</td>
<td>DATA_NUMERIC</td>
<td>transfers</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>edit</td>
<td>kzbledzi</td>
<td>Aug 10, 2015 7:14:27 PM</td>
<td>d3r.cern.ch</td>
<td>gp.user</td>
<td>DATA_NUMERIC</td>
<td>transfers</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>edit</td>
<td>kzbledzi</td>
<td>Aug 11, 2015 8:15:23 PM</td>
<td>d3r.cern.ch</td>
<td>gp.user</td>
<td>DATA_NUMERIC</td>
<td>transfers</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### Create new job

Select data source

<table>
<thead>
<tr>
<th>Database name</th>
<th>Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>devdb11</td>
<td>kzbledzi -</td>
</tr>
<tr>
<td>acccon</td>
<td>kzbledzi -</td>
</tr>
</tbody>
</table>

Warning! Schemas below belong to you but don’t belong to Hadoop Setup. Contact hadoopAdmins.

cdr
accclog

Select cluster

<table>
<thead>
<tr>
<th>Name</th>
<th>Version</th>
</tr>
</thead>
<tbody>
<tr>
<td>kib0955</td>
<td></td>
</tr>
</tbody>
</table>

Give relative path

Your path is: http://users/roader/kzbledzi/acccon/kzbledzi/test/object/biabol...
Contributors

- Anirudha Bose
- Dániel Stein
- Antonio Romero Marin
- Domenico Giordano
- Kacper Surdy
- Katarzyna Maria Dziedziniewicz-Wójcik
- Manuel Martín Márquez
- Zbigniew Baranowski