Federated data usage statistics in the Earth System Grid federation

A. Nuzzo, M. Mirto, P. Nassisi, S. Fiore, G. Aloisio
CMCC Foundation
(Euro Mediterranean Center on Climate Change)
Outline

- Goals and main tasks
- Architecture in the large – single node level
- Architecture in the large – federation level
- Federation protocol
- New Dashboard-UI module
Goals and main tasks

The main goal of the DWT was to provide a distributed and scalable monitoring framework responsible for:

- capturing usage metrics, system status and aggregated information at the single site level and at the federated level
- providing the user with a user friendly interface including widget showing aggregated statistics and monitoring information.

The Dashboard system faces this important challenge through two main components:

ESGF Dashboard (back-end engine)  ESGF Dashboard-UI (front-end layer)
Architecture in the large – Single node level

**ESGF DATA NODE**

<table>
<thead>
<tr>
<th>ID</th>
<th>url_path</th>
<th>duration</th>
<th>size</th>
<th>timestamp</th>
<th>success</th>
<th>processed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DASHBOARD_QUEUE**

- Multi Tier Database
  - 1 Datawarehouse +
  - A set of data marts

**Features:**
- Extended set of statistics
- Fine grain level
- Project specific views
- More scalable design

**ESGF DASHBOARD-UI**

- ETL
- METADATA

**ESGF dashboard**

- getMetadata(url_path)
- METADATA

**SOLR**

- data download
The federated protocol is based on a hierarchical view of the system.

Two kinds of nodes:
- Collector node
- Leaf node

Till now, the federated statistics have been collected by manually executing a set of different queries on the various data nodes and importing the results into a single database.
Leaf node

- Dashboard back-end engine
  - A data warehouse storing all the data related to the downloads
  - A set of data marts containing specific statistics information
- A set of RESTful API providing the collector node the possibility to access data marts and getting the statistics.
Collector node

The collector node has a more complex structure because, in addition to making its information available to the collector through the RESTful API, is in charge to query its leaf nodes.

The collector node is composed by:
- data warehouse and data marts
- RESTful API
- xml configuration file
- federation component

A first prototype of such protocol has been successfully installed and tested on four sites: CMCC, DKRZ, NASA/JPL, PCMDI
New Dashboard-UI – Statistics Overview

Statistics Overview

- **ESGF total number of datasets**: 672,628
- **ESGF total data volume**: 4,632.704 TB
- **CMIP5 total number of datasets**: 150,824
- **CMIP5 total data volume**: 4,261.921 TB
- **Obs4MIPs total number of datasets**: 365
- **Obs4MIPs total data volume**: 0.285 TB
- **CORDEX total number of datasets**: 68,709
- **CORDEX total data volume**: 59.813 TB

Registered Users per Project:

- **CMIP5**: 20,879
New Dashboard-UI – Statistics Overview

Registered Users and Number of Downloads per Project

- **CMIP5**
  - Registered Users: 20,879
  - Number of Downloads: 3,852,332
New Dashboard-UI

Number of Downloads by Continent and Countries

<table>
<thead>
<tr>
<th>Continent</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>625805</td>
</tr>
<tr>
<td>North America</td>
<td>1306321</td>
</tr>
<tr>
<td>South America</td>
<td>1070949</td>
</tr>
<tr>
<td>Asia</td>
<td>16328357</td>
</tr>
<tr>
<td>Oceania</td>
<td>4218743</td>
</tr>
</tbody>
</table>

Europe
- Germany: 3014525
- Spain: 1044863
- Switzerland: 685415
- United Kingdom: 500526
- France: 365729
- Netherlands: 105944
- Norway: 96965
- Italy: 673988
- Sweden: 59546
- Portugal: 31321
- Greece: 20520

Oceania
- Australia: 2157734
- New Zealand: 216397
- New Caledonia: 7
New Dashboard-UI

Number of Users by Continent and Countries

Europe
- Germany: 1065
- France: 356
- Italy: 121
- Spain: 139
- Netherlands: 83
- Russian Federation: 79
- Norway: 75
- Sweden: 66
- Switzerland: 67
- Belgium: 57
- Denmark: 36

North America
- North America: 4032

South America
- Brazil: 104
- Colombia: 71
- Chile: 56
- Argentina: 39
- Peru: 20
- Bolivia: 6
- Suriname: 3
- Venezuela: 2
- Ecuador: 2
- Paraguay: 1
- Uruguay: 1

Asia
- China: 1094
- India: 466
- Japan: 217
- Iran: 171
- Korea, South: 135
- Thailand: 97
- Taiwan: 46
- Indonesia: 40
- Pakistan: 37
- Malaysia: 25
- Israel: 21

Oceania
- Australia: 194
- New Zealand: 19
- Norfolk Island: 1
New Dashboard-UI – Data Usage Statistics section

Number of download over time

Number of download per host
New Dashboard-UI – Project specific section

Obs4MIP4 project

- Top ten datasets
- Top ten sources
- Top ten variables
- Number of downloads by variable (Pie Chart)
- Number of downloads by main
Number of users who made a download
New Dashboard-UI – Federated Data Archive section

Federated data archive

Select a Data Node:

Filter by data node
New Dashboard-UI – Federated Data Archive section

Total number of datasets and related size for each Model and Modeling Institute for CMIP5 project (data obtained by SOLR module).

<table>
<thead>
<tr>
<th>#</th>
<th>Model</th>
<th># of datasets</th>
<th>Size (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ACCESS1.0</td>
<td>330</td>
<td>35.476 TB</td>
</tr>
<tr>
<td>2</td>
<td>ACCESS1.3</td>
<td>316</td>
<td>33.825 TB</td>
</tr>
<tr>
<td>3</td>
<td>BCC-CSM1.1</td>
<td>5,101</td>
<td>41.844 TB</td>
</tr>
<tr>
<td>4</td>
<td>BCC-CSM1.1(m)</td>
<td>400</td>
<td>3,610.014 TB</td>
</tr>
<tr>
<td>5</td>
<td>BNU-ESM</td>
<td>513</td>
<td>19.96 TB</td>
</tr>
<tr>
<td>6</td>
<td>CCSR/MESS</td>
<td>4,986</td>
<td>166.347 TB</td>
</tr>
<tr>
<td>7</td>
<td>CESM1(BGC)</td>
<td>300</td>
<td>22.716 TB</td>
</tr>
<tr>
<td>8</td>
<td>CESM1(CAM5)</td>
<td>370</td>
<td>31.279 TB</td>
</tr>
<tr>
<td>9</td>
<td>CESM1(CAM5,1-FV2)</td>
<td>51</td>
<td>2.629 TB</td>
</tr>
<tr>
<td>10</td>
<td>CESM1(FASTCHEM)</td>
<td>51</td>
<td>3.269 TB</td>
</tr>
<tr>
<td>11</td>
<td>CESM1(WACCM)</td>
<td>155</td>
<td>3.873 TB</td>
</tr>
<tr>
<td>12</td>
<td>CFV2-2011</td>
<td>2,644</td>
<td>34.739 TB</td>
</tr>
<tr>
<td>13</td>
<td>CMCC-CESM</td>
<td>87</td>
<td>1.284 TB</td>
</tr>
<tr>
<td>14</td>
<td>CMCC-CM</td>
<td>992</td>
<td>144.024 TB</td>
</tr>
<tr>
<td>15</td>
<td>CMCC-GMS</td>
<td>110</td>
<td>4.227 TB</td>
</tr>
<tr>
<td>16</td>
<td>CNRM-CM5</td>
<td>3,390</td>
<td>135.5 TB</td>
</tr>
<tr>
<td>17</td>
<td>CNRM-CM5-2</td>
<td>263</td>
<td>11.909 TB</td>
</tr>
<tr>
<td>18</td>
<td>CSIRO-Mk3.6.0</td>
<td>3,120</td>
<td>57.503 TB</td>
</tr>
<tr>
<td>19</td>
<td>CSIRO-Mk3.6.1-2</td>
<td>26</td>
<td>0.03 TB</td>
</tr>
<tr>
<td>20</td>
<td>CanAM4</td>
<td>184</td>
<td>8.027 TB</td>
</tr>
<tr>
<td>21</td>
<td>CanCM4</td>
<td>19,118</td>
<td>23.317 TB</td>
</tr>
<tr>
<td>22</td>
<td>CanESM2</td>
<td>2,577</td>
<td>39.844 TB</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>#</th>
<th>Modeling institute</th>
<th># of datasets</th>
<th>Size (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BCC</td>
<td>5,504</td>
<td>97.992 TB</td>
</tr>
<tr>
<td>2</td>
<td>BNU</td>
<td>513</td>
<td>19.93 TB</td>
</tr>
<tr>
<td>3</td>
<td>CCCMA</td>
<td>21,879</td>
<td>71.187 TB</td>
</tr>
<tr>
<td>4</td>
<td>CMCC</td>
<td>1,199</td>
<td>149.504 TB</td>
</tr>
<tr>
<td>5</td>
<td>CNRM-CERFACS</td>
<td>3,620</td>
<td>147.409 TB</td>
</tr>
<tr>
<td>6</td>
<td>COLA-CFS</td>
<td>1,199</td>
<td>7.953 TB</td>
</tr>
<tr>
<td>7</td>
<td>CSIRO-BOM</td>
<td>616</td>
<td>69.301 TB</td>
</tr>
<tr>
<td>8</td>
<td>CSIRO-OCCCE</td>
<td>2,120</td>
<td>57.593 TB</td>
</tr>
<tr>
<td>9</td>
<td>FRO</td>
<td>230</td>
<td>6.834 TB</td>
</tr>
<tr>
<td>10</td>
<td>ICHEC</td>
<td>2,620</td>
<td>134.946 TB</td>
</tr>
<tr>
<td>11</td>
<td>INM</td>
<td>496</td>
<td>21.492 TB</td>
</tr>
<tr>
<td>12</td>
<td>INPE</td>
<td>24</td>
<td>7.957 TB</td>
</tr>
<tr>
<td>13</td>
<td>IPSL</td>
<td>10,757</td>
<td>699.236 TB</td>
</tr>
<tr>
<td>14</td>
<td>LASSCESS</td>
<td>1,553</td>
<td>40.014 TB</td>
</tr>
<tr>
<td>15</td>
<td>LASSIMI</td>
<td>418</td>
<td>7.597 TB</td>
</tr>
<tr>
<td>16</td>
<td>LMGOC</td>
<td>16,701</td>
<td>823.919 TB</td>
</tr>
<tr>
<td>17</td>
<td>MOHC</td>
<td>24,720</td>
<td>148.004 TB</td>
</tr>
<tr>
<td>18</td>
<td>MPI-MI</td>
<td>11,825</td>
<td>195.248 TB</td>
</tr>
<tr>
<td>19</td>
<td>MRI</td>
<td>7,271</td>
<td>404.758 TB</td>
</tr>
<tr>
<td>20</td>
<td>NASA-GISS</td>
<td>7,681</td>
<td>227.295 TB</td>
</tr>
<tr>
<td>21</td>
<td>NASA-GMAO</td>
<td>2,520</td>
<td>8.531 TB</td>
</tr>
<tr>
<td>22</td>
<td>NCAR</td>
<td>4,906</td>
<td>156.347 TB</td>
</tr>
</tbody>
</table>
New Dashboard-UI – Service status section

Deployment distribution
Thank you