EUROPEAN COMMISSION

Community research

is-enes

INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

# IS-ENES2 project

**(Contract Number: 312979)**

# *Report on service monitoring and dashboard*

## DELIVERABLE (N°: D113)


Author(s): ***Alessandra Nuzzo, Paola Nassisi, Maria Mirto, Sandro Fiore, Giovanni Aloisio***

Reviewer (s): ***Michael Lautenschlager (DKRZ), Sébastien Denvil (IPSL)***


Reporting period: 01/10/2014 – 31/03/2016


Date for review: 01/02/2017

Final date of issue: 22/02/2017

Start date of project: **01/04/2013** Duration: **48 Months**




CAPACITIES

| Revision table | | | |
|---|---|---|---|
| **Version** | **Date** | **Reviewer** | **Comments** |
| 0.1 | 08/01/2017 | Giovanni Aloisio | First formal release |
| 0.2 | 30/01/2017 | Michael Lautenschlager | Glossary section added. |
| 0.3 | 01/02/2017 | Sébastien Denvil | Data dictionary moved into Appendix section and comments addressed. |
| 1.0 | 21/02/2017 | Giovanni Aloisio | Final version – reviewers' comments integrated |



| Project co-funded by the European Commission's Seventh Framework Programme (FP7; 2007-2013) under the grant agreement n°312979 | | |
|---|---|---|
| **Dissemination Level** | | |
| **PU** | Public | X |
| **PP** | Restricted to other programme participants including the Commission Services | |
| **RE** | Restricted to a group specified by the partners of the **IS-ENES2** project | |
| **CO** | Confidential, only for partners of the **IS-ENES2** project | |

# Abstract

***This document is part of WP11/JRA3 "Developing software infrastructure for data archive services" and its purpose is to provide a detailed view of the Federated Archive System Monitoring (FASM) package which provides statistics about the usage and demographics not only for the European nodes but also for the global ESGF federation.***

***With respect to the "Monitoring system and dashboard design" document (MILESTONE n° MS111), in this report a description of the implementation of the FASM-N and the FASM-D modules is provided. In particular, the former is responsible for collecting and storing a high volume of heterogeneous metrics while the latter provides the user with a dashboard including local (node-level) and global (federation-level) views, aggregated statistics and monitoring information.***

***A review of the design phase is presented, through the description of the architecture in the large and the design of the new data warehouse system. Moreover, the architecture in the small of FASM-N and FASM-D is presented. For FASM-N, the implementation of the single node and the collector modules is reported whereas, for FASM-D, the main views are presented.***

# Table of contents

# Table of figures

# Executive Summary

This report provides a complete overview about the development of the Federated Archive System Monitoring (FASM) in the IS-ENES2 project.
The report describes the two main monitoring components, starting from the analysis of both the functional and non-functional requirements gather over time, such as the scalability of the entire system as well as the plurality of the collected statistics.
From a software engineering point of view, the document provides an overview of the framework in terms of design of the data warehouse and architecture in the large and in the small of the notification (FASM-N) and dashboard (FASM-D) modules.
More specifically, an extension of the *esgcet* database is described; such extension has foreseen a detailed design of distinct data warehouse systems with the aim to collect an extended set of data usage statistics related to cross-project and project-specific information.
Going through the document, the main two modules of the framework are explained: the notification module is responsible for managing usage statistics at each site, ENES level and federation level; the dashboard module represents a dashboard providing a flexible, transparent and easy-to-access user interface. Both are discussed in detail as two separate (loosely coupled) modules of the same architecture.
The FASM-N section describes how the notification component retrieves the information related to the downloaded files both from logging data and search service provided by the Earth System Grid Federation (ESGF) index node. Also, to collect statistics at ENES-level and at global federation-level, the difference between two types of configuration is introduced: the leaf and the collector data nodes. In particular, a leaf node is responsible to collect local statistics only, while a collector node is in charge of gather the complete set of statistics from its leaf nodes through the use of a REpresentational State Transfer (REST) service.
Finally, an overview of the system front-end is presented in the FASM-D section. The module is described in terms of structure of the main components and views provided to the final user.
The statistics presented include information at cross-project and project-specific level for CMIP5 (Coupled Model Intercomparison Project Phase 5), Obs4MIPs (Observations for Model Intercomparisons Project) and CORDEX (COordinated Regional Climate Downscaling Experiment); the distribution of the clients is also shown in a different section as well as information about the federated data archive and the ESGF nodes service status.

## 1. Introduction and goals

The Federated Archive System Monitoring (FASM) is a key component for the IS-ENES2 project. The FASM faces this important challenge through two main modules: the FASM-N for managing coarse and fine grain statistics at each site and the FASM-D, a dashboard module providing a flexible user interface.

The main goal of the FASM system is to provide a distributed and scalable monitoring framework responsible for capturing usage metrics at the single site level, at the ENES archive level and at the global ESGF level.

The FASM has to (i) collect and store a high volume of heterogeneous metrics, covering coarse and fine grain measures such as downloads and clients statistics, aggregated cross and project-specific download statistics as well as to (ii) provide a rich set of charts and reports through a web interface, allowing users and system managers to visualize the status of the IS-ENES infrastructure through a set of smart and attractive web gadgets.

Relevant goals of the FASM-N architecture are the management of local and federated metrics. On the other hand, key goals for the FASM-D module are an effective, transparent, robust and easy access to all the metrics, reports, charts and statistics provided by the system.

FASM offers an analytical dashboard focused on gaining insights from a volume of data collected over time and using them to understand what happened, why, and what changes should be made in the future.

# 2. Requirements analysis

With respect to the MS111 document, in this section the requirements of the FASM have been reviewed, taking into account specific needs and requests of the community, gathered over time during project meetings and web conferences.

## 2.1 Non-functional requirements

The most relevant non-functional requirements related to the FASM system are presented in the following:

- **[NFR1] transparency**: the system has to provide the metrics and statistics in a transparent way, which means hiding the back-end complexity as well as low level technical details;
- **[NFR2] robustness**: the system has to be robust, which means stable with regard to (i) operate under stress or (ii) tolerate unpredictable or invalid input;
- **[NFR3] scalability**: the system has to scale well both with regard to the set of metrics stored in the back-end and to the statistics displayed in the web-gadgets; in the same way the system has to support an increasing number of data nodes;
- **[NFR4] efficiency**: the system has to be efficient with regard to the large set of metrics to be managed and stored in the back-end of the system, as well as the set of sensors it will run;
- **[NFR5] security**: the system has to provide a security layer able to address both authentication (in strong synergy with the authentication mechanisms currently available in the IS-ENES/ESGF federation) and authorization (in terms of classes of users previously defined);
- **[NFR6] reusability**: the system has to address software re-usability. This means that new interfaces or metrics should be straightforwardly added to the system with few code changes/additions;
- **[NFR7] extensibility**: the system has to be easily extensible in terms of new metrics and interfaces. Such a requirement is very important as the set of metrics and statistics could be extended in the future, based on new user/system requirements;
- **[NFR8] configurability/flexibility**: the system (and the sensors) has to be highly configurable and flexible to allow site administrators carrying out site-specific configurations. Such a requirement address local autonomy and preserve federation-level needs;
- **[NFR9] usability**: the dashboard user interface must be usable for the target user/audience, which includes computational scientists as well as domain-based experts (climate change scientists, etc.) ;

- **[NFR10] look and feel**: the dashboard user interface must provide a strong look and feel to address a very large adoption of all the provided functionalities;
- **[NFR11] programmability**: the dashboard back-end must provide a complete set of REST APIs to give the users the opportunity to programmatically get access to the dashboard metrics.
- **[NFR12] zero-conf**: the dashboard has to be „zero-conf", w.r.t. the list of nodes to be monitored, geo-location information, list of available services, deployment information, etc. This will allow the node administrator to test and use the dashboard right after the installation without any intermediate setup/configuration steps.

## 2.2 Functional requirements

The most relevant functional requirements related to the FASM system are presented and discussed in the following.
They are related to an easy and transparent access to:

- **[FR1]** data download statistics (both per-node, per-project, institution-based and federated-level view). In this regard the download statistics should be provided according to different views based on time, models, hosts, institutions, variables, datasets, experiments or a combination of two or three of them. In general a support for several dimensions like the facets provided by ESGF or CORDEX should be provided too;
- **[FR2]** data client usage statistics (a per-node, per-project, institution-based or federation-level view related to all of the clients that carried out at least one download from the selected data nodes). In this regard the clients statistics should be grouped by country and/or continent and they should be aggregated also over time. Moreover top-ten lists of variables, datasets, models, experiments and countries should be provided as well;
- **[FR3]** ENES-level and federation-level registered users information;
- **[FR4]** ENES-level and federation-level deployment nodes information;
- **[FR5]** current status of the federation in terms of data volume and number of published dataset at project and global level.

Moreover the system has to be able:

- **[FR6]** to offer a configurable way to instanciate and configure new metrics, based on a well-known set of pre-defined metrics classes provided by the system.
- **[FR7]** to provide an Event-Condition-Action (ECA) based system to associate events, conditions and actions regarding the monitored metrics.
- **[FR8]** to support a *single-node running* mode for separate portal instances like the climate impact portal.

# 3. Design and implementation

In the following chapter there is an overview of the FASM design and implementation, both for back-end and front-end side, starting from a general view on the architecture and going down into the details of the database extension, the collection of local and federated statistics and the description of the front-end user interface.

## 3.1 Architecture in the large

Through the old version of the monitoring framework, only a reduced set of statistics was provided, those strictly related to the logging information gathered by the node manager filter. As shown in the Figure 1, disregarding all the elements unrelated to the statistics, the ESGF data node exploits the ESGF node manager filter to record the downloads performed by the users. The information retrieved were saved into the *access_logging_table* and, after an ETL (Extraction – Transformation – Loading) process, stored into the *esgf_dashboard* namespace to be aggregated and visualized through the ESGF Desktop.

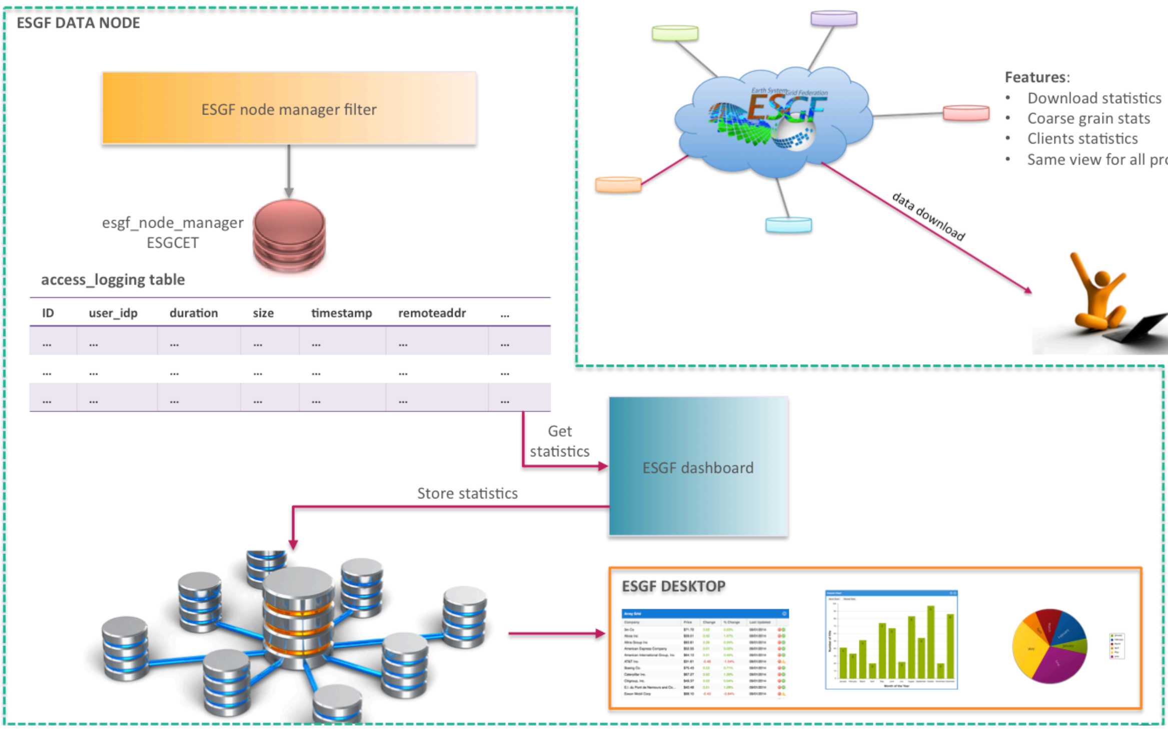



*Fig. 1: Previous FASM architecture and workflow*

To fulfill the new requirements, in particular **[FR1]**, a new architecture has been defined to collect, besides the metrics contained in the logging information, new kinds of data usage

statistics about the models, the variables, the datasets and so on. For this reason, alongside the dashboard and the *access_logging* table, a new component of the ESGF distributed architecture was involved in the process: the SOLR search engine.
In the following, the Figure 2 represents the current configuration and workflow of the monitoring framework.

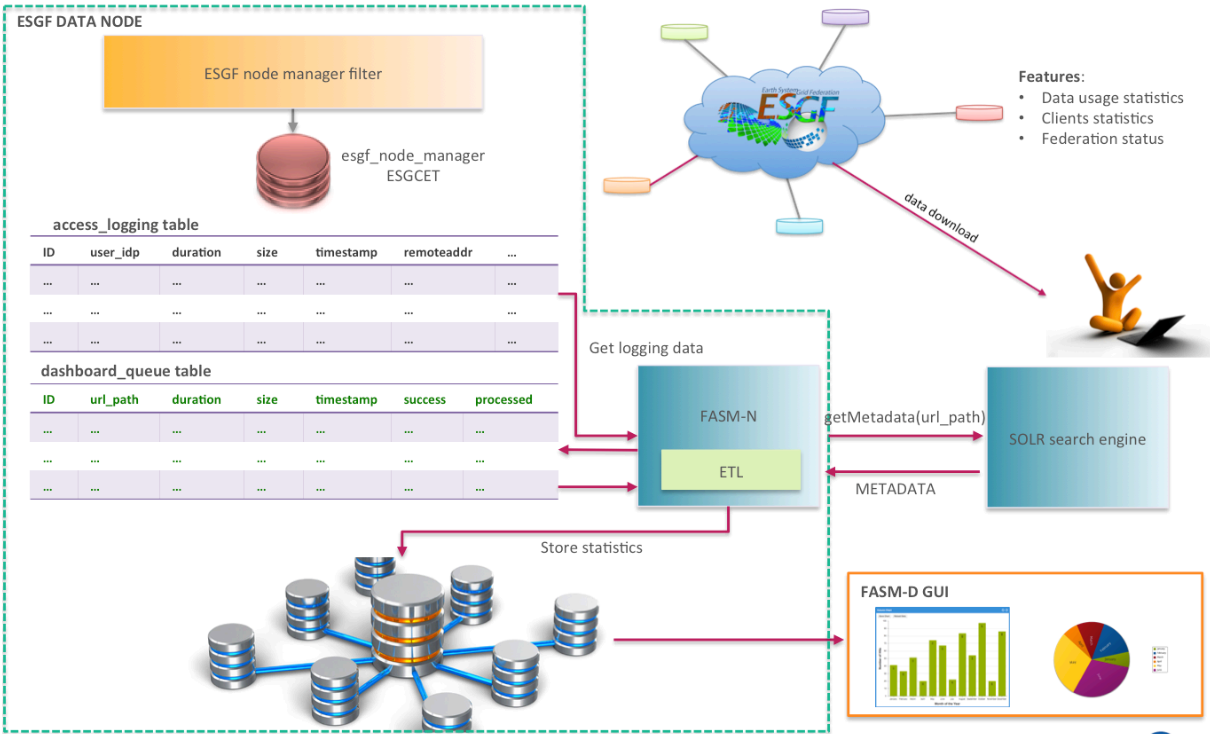



*Fig. 2: Current FASM architecture and workflow*

As the previous one, the current architecture also relies on the node manager filter to gather the basic logging information from the *access_logging* table. However, such data are collected by an ETL subroutine, normalized and stored into the *esgf_dashboard* namespace. Other information will be collected by querying the index node SOLR search engine, which is able to provide a wide collection of metadata coupled with the downloaded datasets.
Another important difference relies on the extension of the *esgcet* database, completely remodeled to contain the new statistics and to make them easily accessible.
It is important to notice that, since the node manager filter is enabled to trace THREDDS downloads, at the moment the statistics collected will refer only to this kind of service. In the future, the framework will be extended to support other kind of download services (i.e. Globus, wget, etc.) in order to collect a more extended set of data usage statistics.
Regarding the FASM-D GUI, as web application, could be the target of possible attacks to the federation. For this reason, the deployment has foreseen a centralized approach with the web interface which is no longer installed on the data nodes but hosted on a separated site.
All of these changes will be widely explained in the next sections.

### 3.2 *esgcet* extension: a log table and the data warehouse system

A solid and well-designed data warehouse system is crucial to store a huge collection of data in any field of application, from health care to financial services and, above all, to extract relevant and reliable information to be used as solid support for decision-makers.
The extension of the *esgcet* database was required to enable the collection of an extended set of statistics about not only logging information but also project-specific metadata, geolocation of the clients, status of the published datasets.
To this purpose, the following groups of tables were added to the original *esgf_dashboard* namespace:

1. a table called *dashboard_queue* where the logging information are queued, waiting to be further processed by the ETL module of the information provider and to be stored in the data warehouse system;
2. a data warehouse system containing general cross-project information;
3. a data warehouse system for each project in ESGF (e.g. CMIP5, Obs4MIPs, CORDEX, etc.),
4. a set of data marts as specific aggregated views on the data to support the graphical user interface.

Each ESGF data node will expose the same database schema and the FASM-N will take care of properly feed the system with data coming from local sources and additional metadata retrieved from the index ESGF node.

#### a) Dashboard_queue table

The *dashboard_queue* table was designed to be a rolling table quite similar to the *access_logging* table of the *esgf_node_manager* schema. It contains information about the download operations performed by the users through the distributed ESGF infrastructure. Each node records the files downloaded, the time instant, the duration and the outcome of the operation and some sensible user information.
The need to have an additional table was related to the possibility of marking the entries processed by the FASM-N module without modifying the existing *esgf_node_manager* schema. Moreover, this choice doesn't add redundancy to the database since the entries, once processed, after a predefined time, will be deleted.
In the Figure 3, below, a complete description of the table:

```sql
CREATE TABLE dashboard_queue (
    id integer NOT NULL,                              -- unique id
    url_path character varying NOT NULL,              -- path of the downloaded file
    remote_addr character varying NOT NULL,           -- user ip address
    user_id_hash character varying,                   -- hash code of the user id
    user_idp character varying,                       -- user identity provider
    service_type character varying,                   -- download service type
    success boolean,                                  -- outcome of the download operation
    duration double precision,                        -- duration of the download operation
    size bigint DEFAULT (-1),                         -- file dimensions
    "timestamp" double precision NOT NULL,            -- download time instant
    processed smallint DEFAULT 0 NOT NULL             -- dashboard flag
);
```

*Fig. 3: dashboard_queue schema*

Some relevant logging information can be retrieved directly from the *dashboard_queue* table. As highlighted in the figure above, the fields could be grouped in three main categories:

- user information (remote address, user identification, IdP node where the user is registered);
- file information (path to the file and its size);
- download information (outcome, duration, time instant, type of download service, as described in Section 3.1).

The last field '*processed*' is a flag used by the back-end to mark the entries that have been analysed, in order to be deleted as soon as possible.

The table is the starting point for the ETL process undertaken by the back-end: some of the information will be normalized; some other will be used to contact the index node and collect the related metadata. At the end, all these data will be ingested in the data warehouse system.

### b) Cross-project Data Warehouse

Cross-project data warehouse contains general information shared between different ESGF projects (Figure 4). The goal is to keep track of the user downloads providing a cross-sectional view; in other words, it allows node administrators and the overall IS-ENES and ESGF communities to access these data analysing them from different point of views without taking into account project-specific attributes.
The different colours on the diagram below, representing the Dimensional Fact Model (DFM) of the system, refer to the different data sources from which the information are retrieved. Most of the data collected here comes from the *dashboard_queue* table.

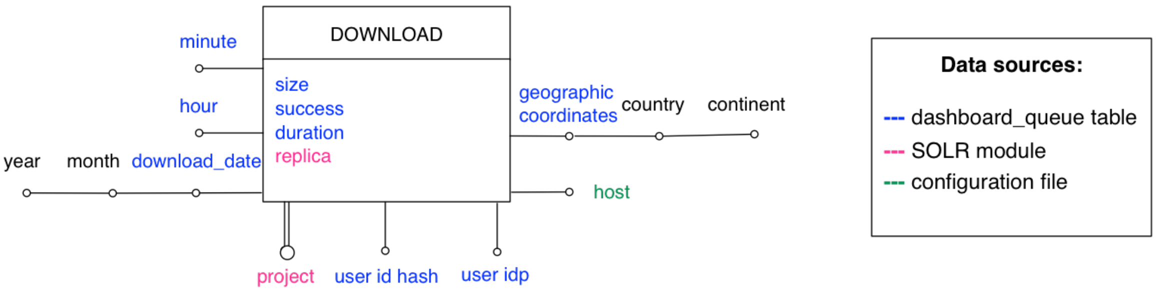



*Fig. 4: Cross-project DFM*

The statistics related to the usage of data provided by the ESGF nodes are the result of a targeted mining operation through a large amount of information related to an elementary fact, the download; the DFM of the cross-project data warehouse system highlights how this fact is characterized by a set of relevant measures and dimensions of analysis. More specifically, the quantities to be measured are:

- the size of the download;
- the eventual success of the operation;
- the duration;
- if the downloaded file is a replica of the original one or not.

These quantities could be considered from different point of views, over time, by project, by user, by identity provider, by host or spread out on a global map, in order to have a more accurate overview of what these information mean.

The Figure 5 is related to the snowflake schema of the previous DFM, where fact, measures and dimension are translated in tables connected by foreign keys.

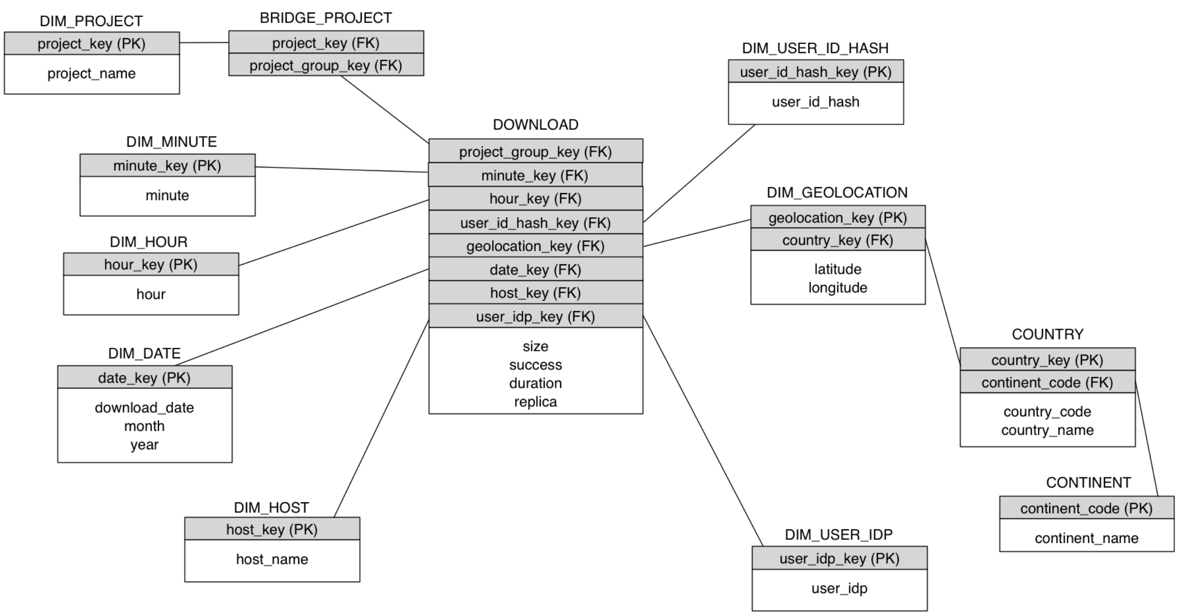



*Fig. 5: Cross-project snowflake schema*

Data marts are aggregated views of the entire system focusing on some of the dimensions available and providing an easy access to frequently needed data in order to improve the response time of the web front-end.
In the Figures 6 and 7 are presented two data marts of the cross-project data warehouse.

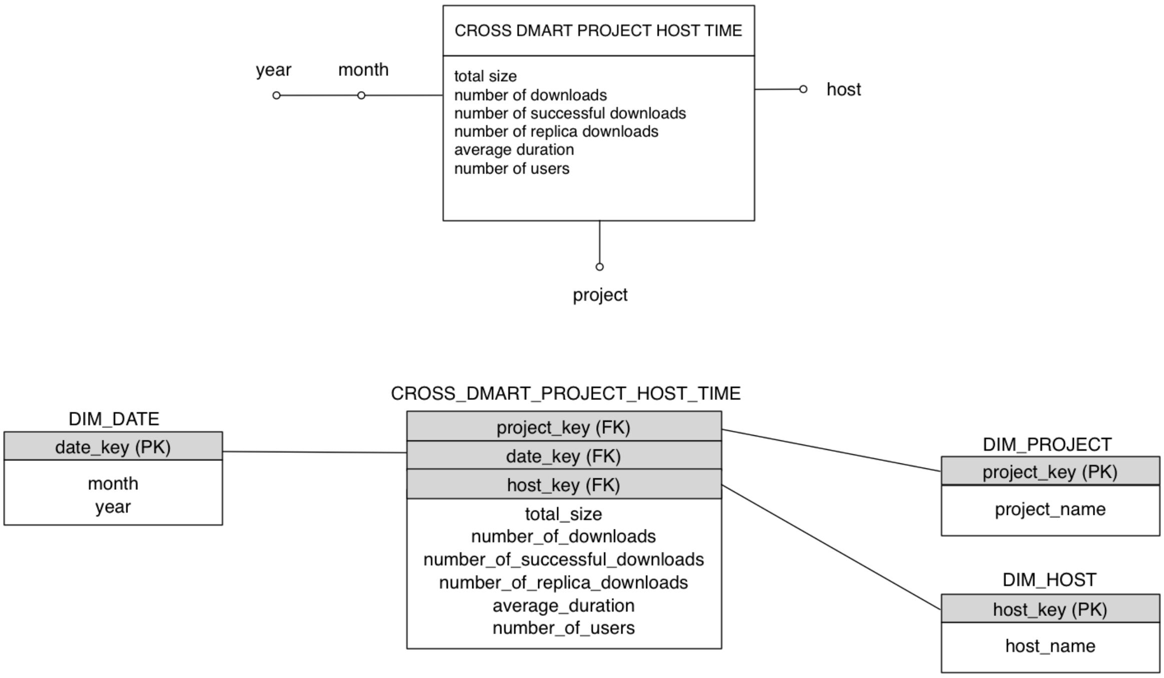



*Fig. 6: Cross-project data mart (correlation between time, host and project dimensions)*

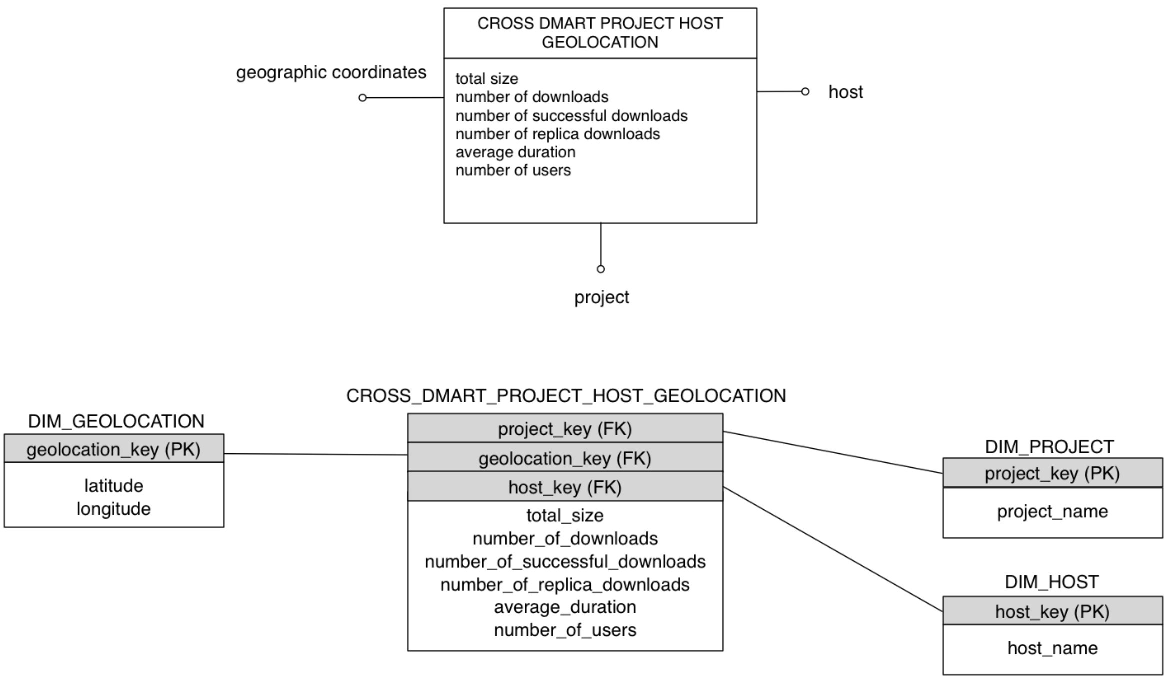



*Fig. 7: Cross-project data mart (correlation between geolocation, host and project dimensions)*

In both cases, the metrics resulting from the aggregation process are:

- total volume of the downloads;
- total number of the downloads;
- total number of successful downloads;
- total number of downloads related to a replicated file;
- average duration of the download time;
- total number of users that performed a download.

The two data marts represent respectively a geographical and a temporal point of view on the metrics analysed by project and host. To be practical, they represent for instance a support for the number of users distributed on a map through proportional markers or a multiple time series about the number of downloads per project.

In the future, other data marts could be produced, on the basis of the user requirements, since the operation would simply require the creation of a new table in the database and the code to aggregate data in the FASM-N module.

### c) Project-specific Data Warehouse

A project-specific data warehouse aims to support the extraction of project-related data usage statistics. For this reason, with the collaboration of other ESGF partners, a number of relevant features have been defined for the projects CMIP5 and Obs4MIPs and they have been translated into dimensions of analysis.

The following diagrams refer to CMIP5 while next there will be a quick overview of the Obs4MIPs data warehouse schemes.

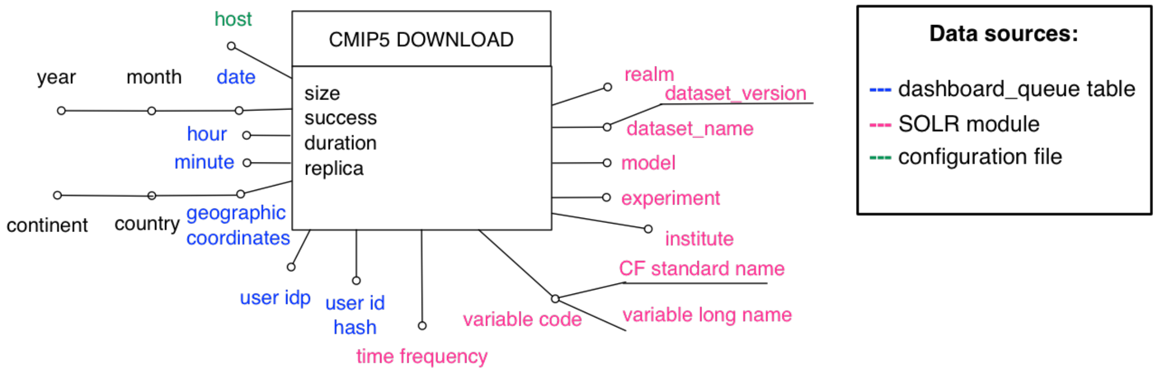



*Fig. 8: CMIP5-specific DFM*

As for the cross-project data warehouse, the metrics and the dimensions in blue and green in the figure above are the same. The main difference is in the information gathered through the SOLR interrogation (pink hierarchies), which allows the information provider to retrieve peculiar metadata related to the file downloaded by the user as, in this case for CMIP5:

- realm;
- dataset name and version;
- model;
- experiment;
- institute;
- variable code, long name and Climate and Forecast (CF) standard name;
- time frequency.

These facets enrich the possibilities of data analysis providing the community with a helpful tool for better understanding the most downloaded and, consequently, interesting features.
In the Figure 9 is represented the snowflake schema related to the CMIP5 data warehouse.

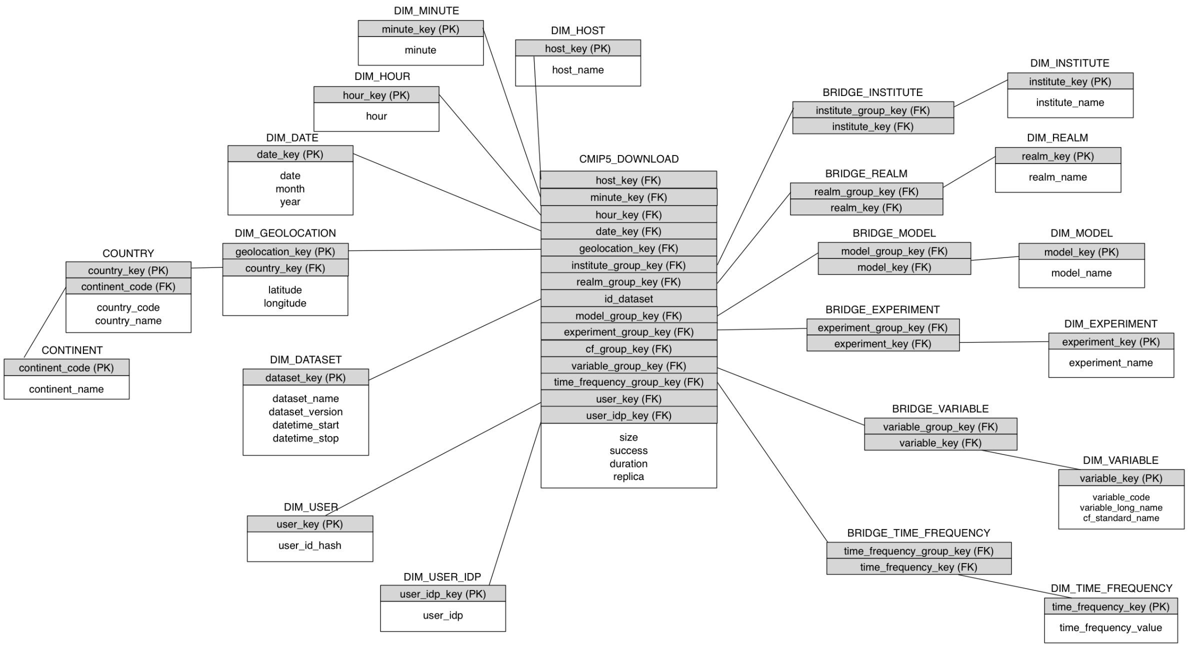



*Fig. 9: CMIP5-specific snowflake schema*

The data marts developed to support the CMIP5 specific views take into consideration the same metrics of the cross-project ones. In particular, one of them explicits the measures according to the geographic coordinates and the time (Figure 10), while the others focus on the time and one specific CMIP5 facet (Figures 11, 12, 13): model, experiment and variable. Moreover, the host dimension, present in all the data marts, besides providing a local view on data, supports the FASM-N module in the collection of the federated statistics allowing the traceability of the gathered information.

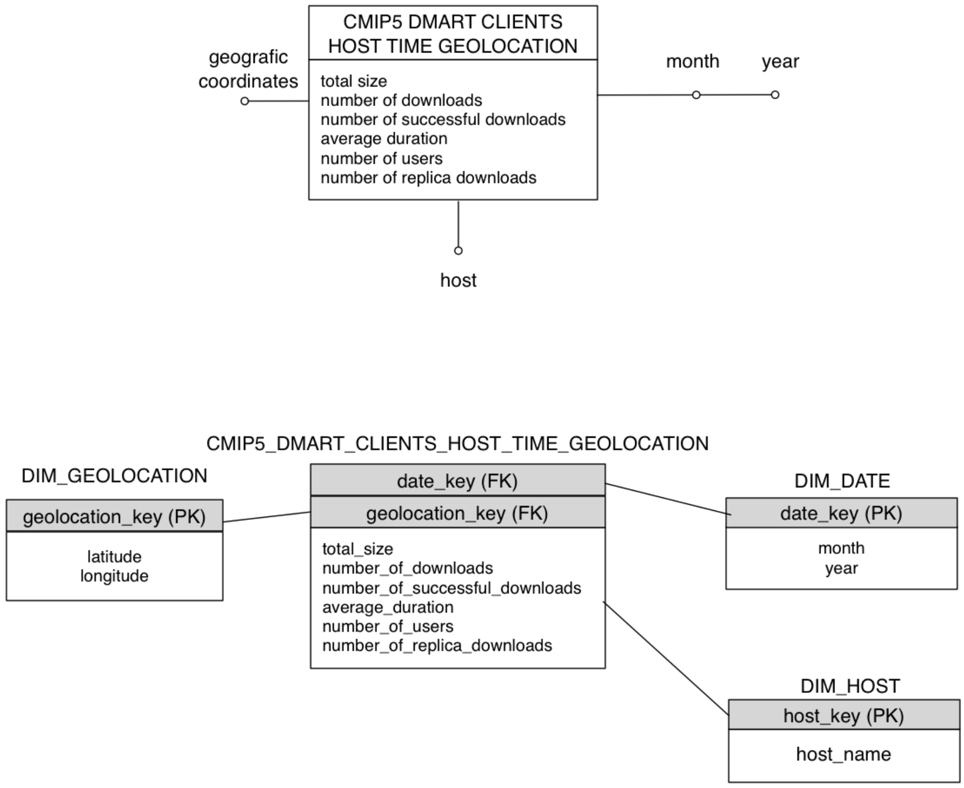



*Fig. 10: CMIP5-specific data mart (correlation between geolocation, host and time dimensions)*

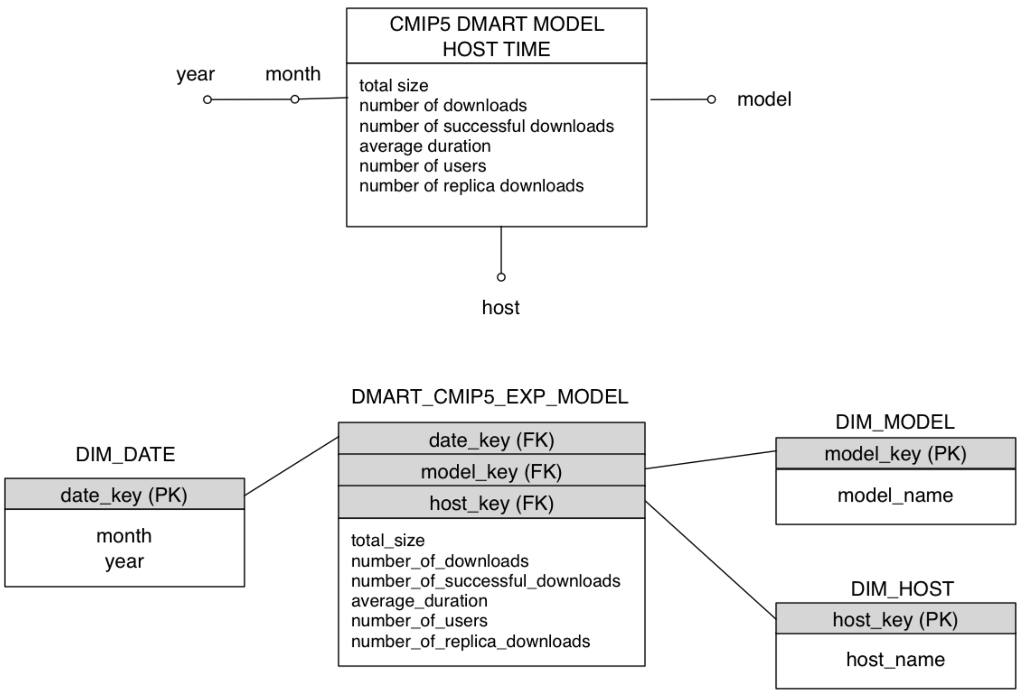
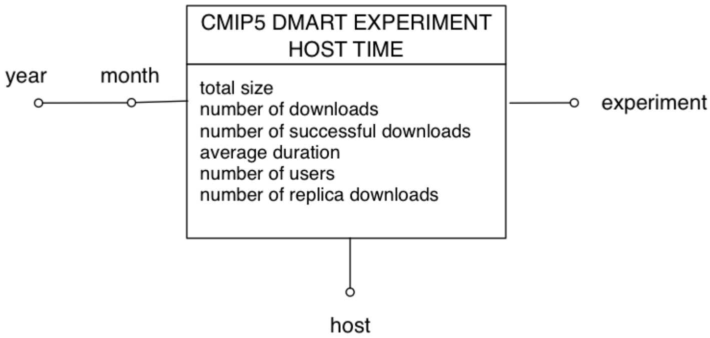



*Fig. 11: CMIP5-specific data mart (correlation between model, host and time dimensions)*



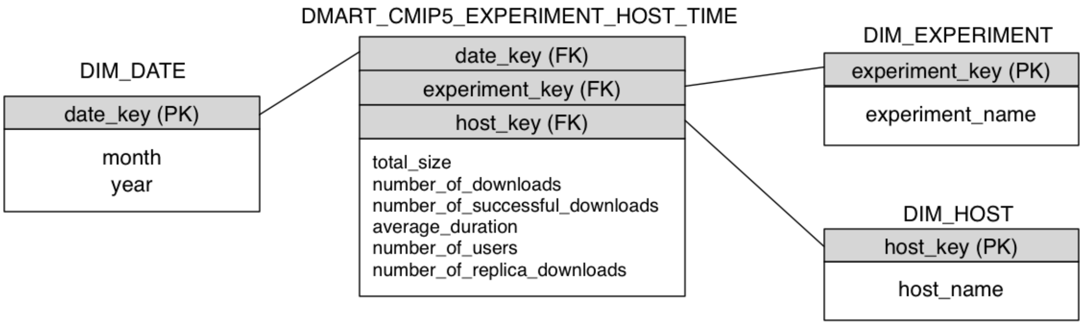
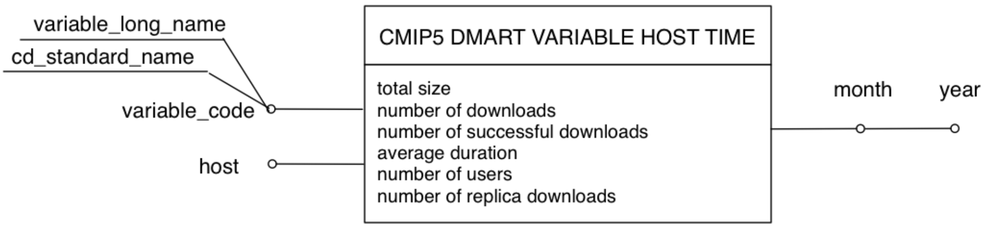




*Fig. 12: CMIP5-specific data mart (correlation between experiment, host and time dimensions)*



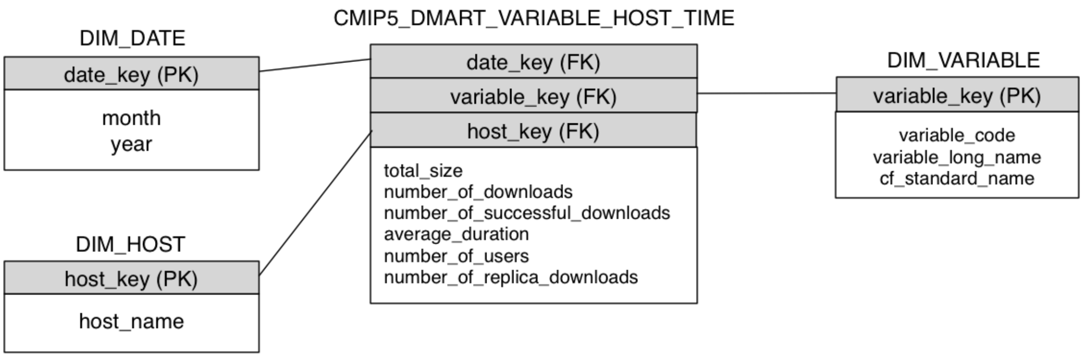




*Fig. 13: CMIP5-specific data mart (correlation between variable, host and time dimensions)*

In the Figures 14 and 15, the DFM and the snowflake schema related to the Obs4MIPs project. There are some dimensions in common with the CMIP5 project: obviously the data retrieved from the *dashboard_queue* data, the host name and some of the dimensions (realm, time frequency, variable-related attributes, etc.). Some other attributes are specific of Obs4MIPs as the source of the data and the processing level.

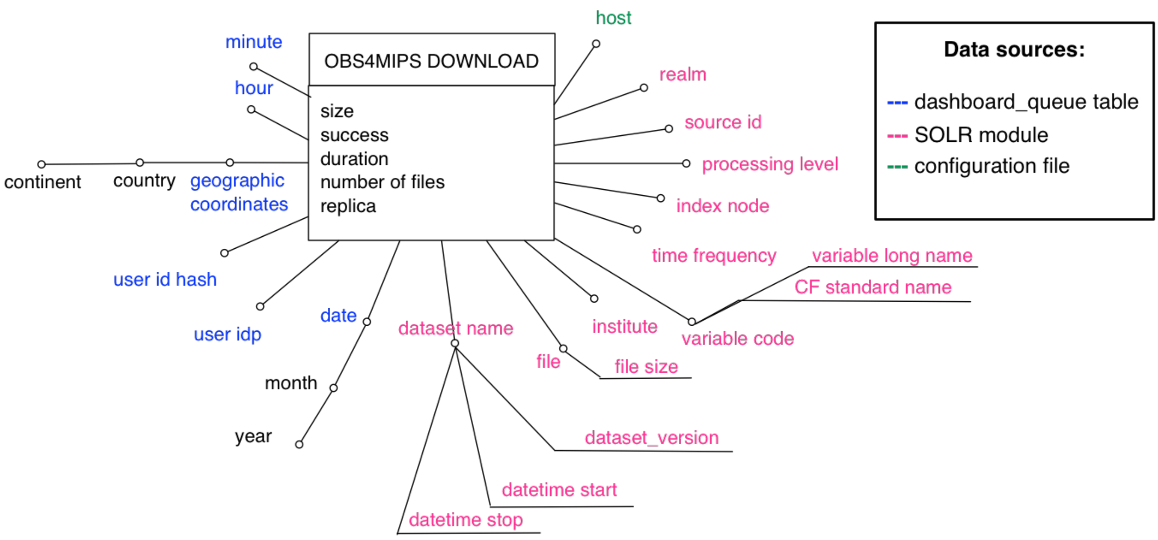



*Fig. 14: Obs4MIPs-specific DFM*

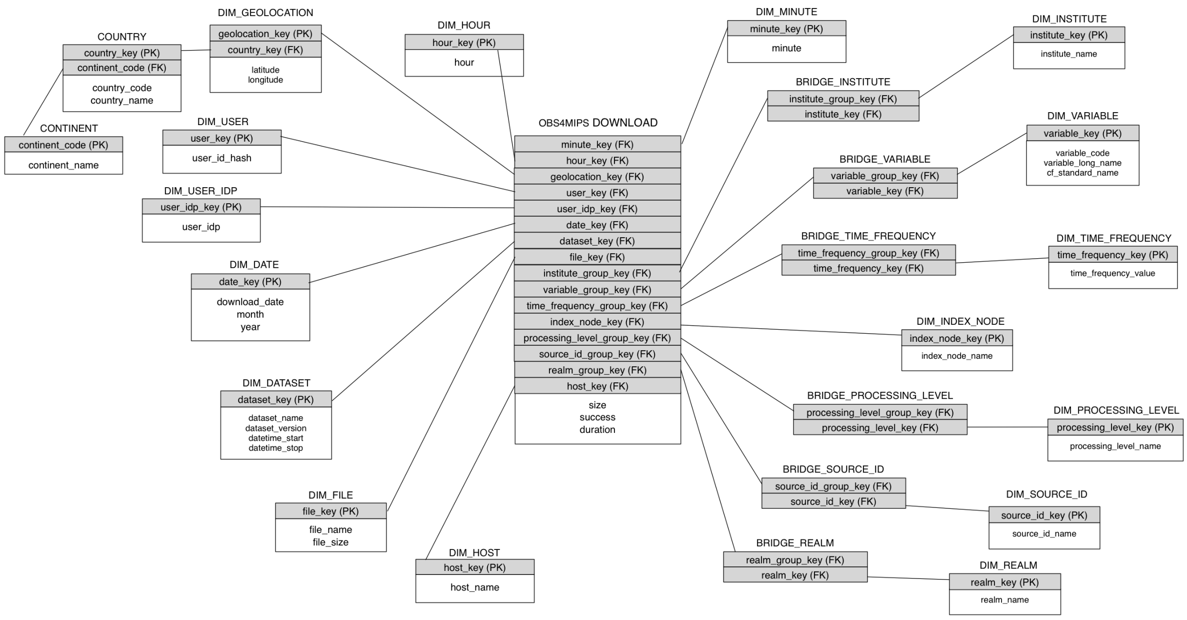



*Fig. 15: Obs4MIPs-specific snowflake schema*

Similar to the CMIP5 use case, also for Obs4MIPs one data mart highlights the correlation between the spatial and the temporal dimension (Figure 16) while the others combine the time with respectively variable, source, realm and dataset (Figures 17, 18, 19, 20). To be practical, the first one could support a set of maps showing the geographic distribution of one or more metrics over time, while the seconds would provide, for example, time series of the downloads or of the number of users performed at list a download operation from an IS-ENES data node.

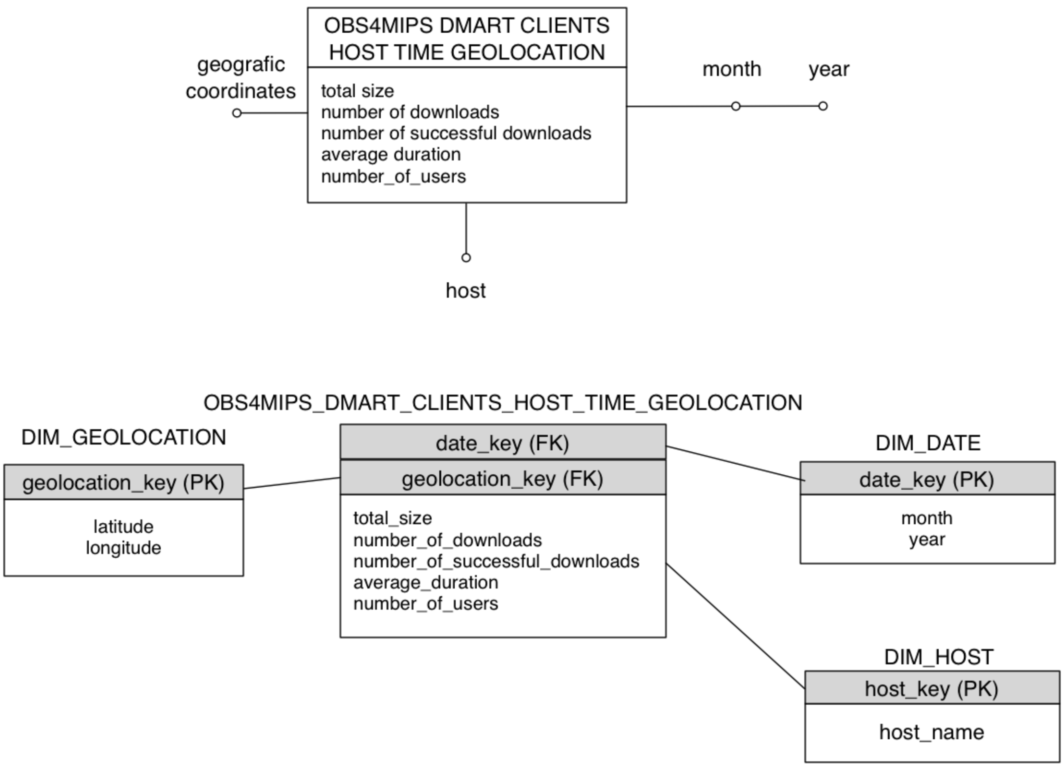



*Fig. 16: Obs4MIPs-specific data mart (correlation between geolocation, host and time dimensions)*

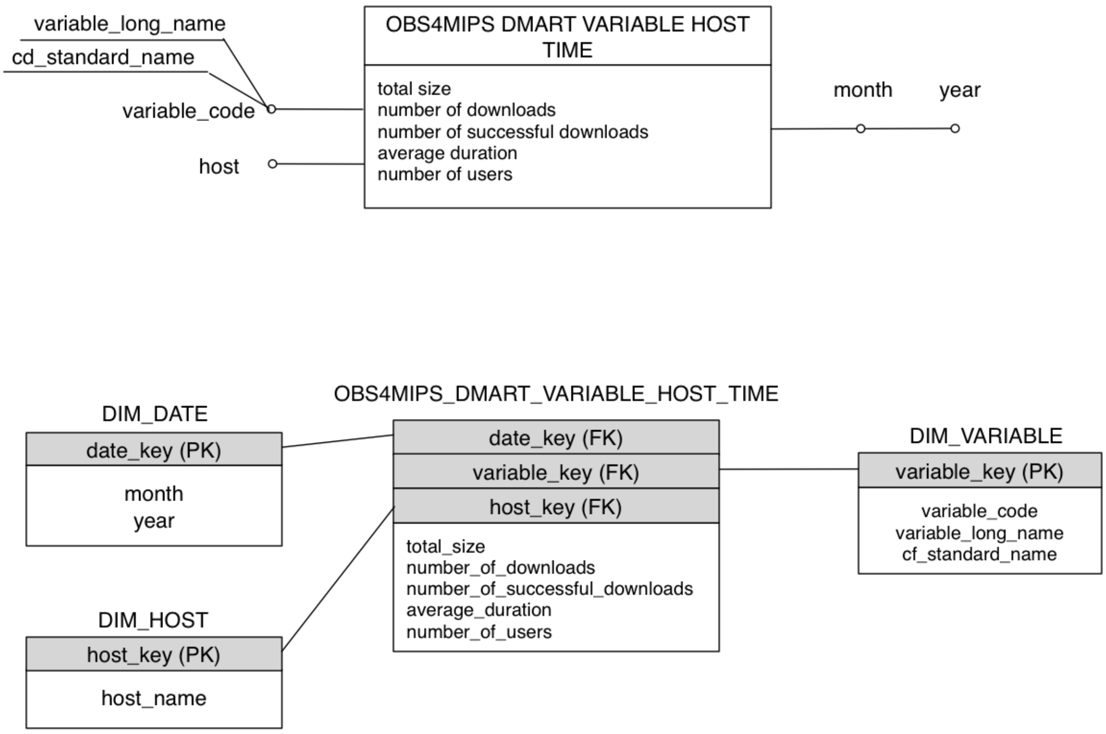



*Fig. 17: Obs4MIPs-specific data mart (correlation between variable, host and time dimensions)*

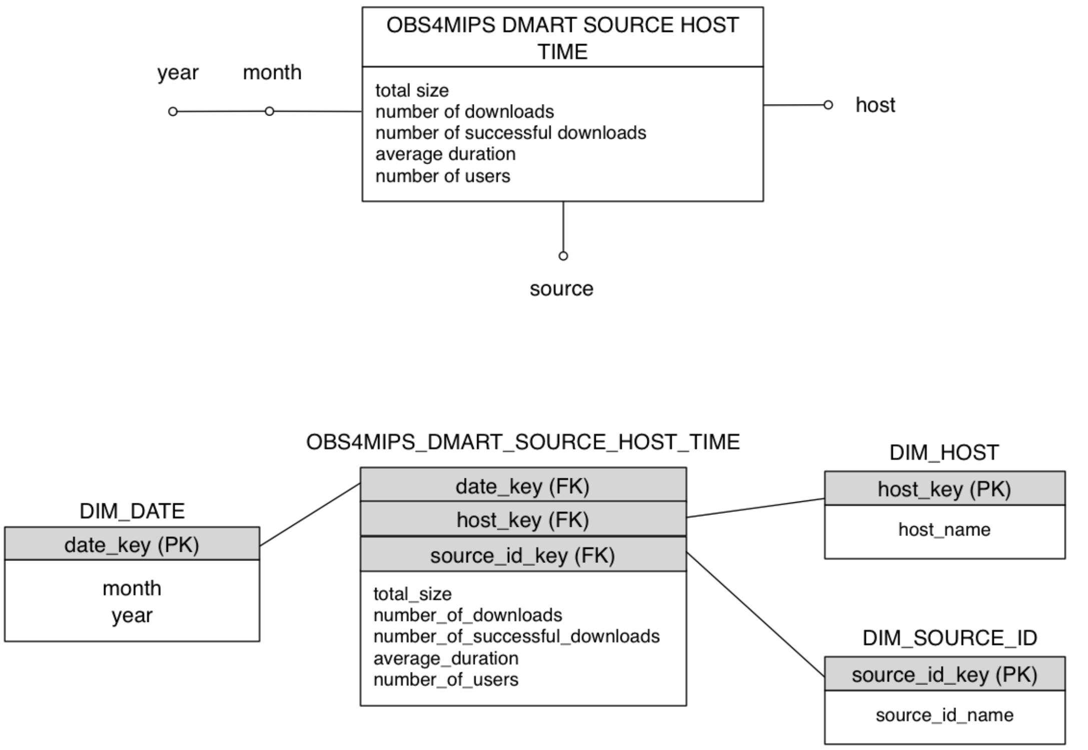



*Fig. 18: Obs4MIPs-specific data mart (correlation between source, host and time dimensions)*

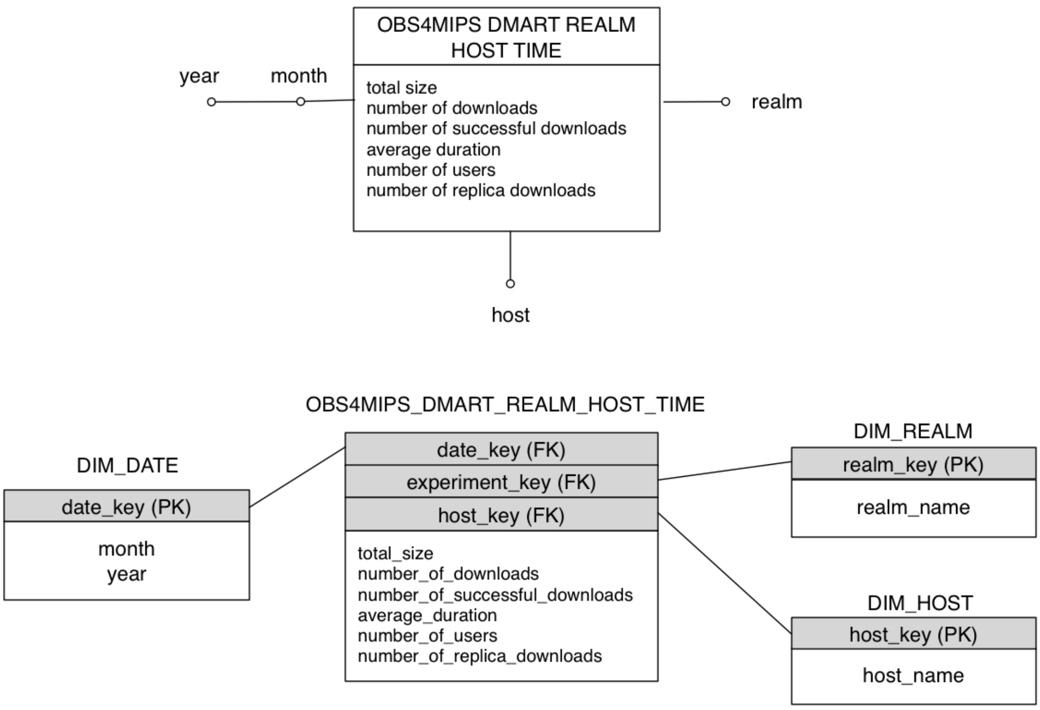



*Fig. 19: Obs4MIPs-specific data mart (correlation between realm, host and time dimensions)*

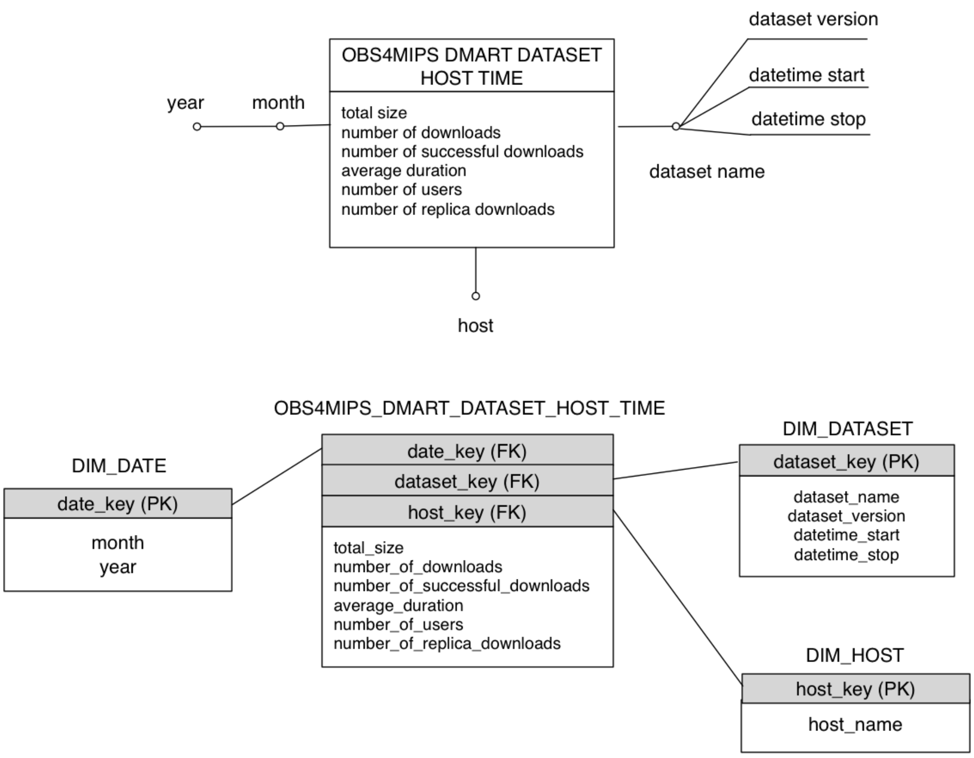



*Fig. 20: Obs4MIPs-specific data mart (correlation between dataset, host and time dimensions)*

### 3.3 FASM-N

#### a) Architecture in the small

The architecture of the FASM-N module is based on a main program and two sub-routines that are responsible to retrieve metadata from SOLR and aggregate the data into the storage database. The first sub-routine queries the *dashboard_queue* table, retrieves the URLs, queries the SOLR and stores the data into specific tables of the database. The second sub-routine aggregates the data and stores them into specific tables (data mart tables). The aggregation is made by following an upgrade at each given period. A registry table is used in order to store the last timestamp in which an update is made by the FASM-N.
The collector nodes are equipped with a third sub-routine responsible to gather the data from several data nodes. A configuration file contains the information about the data nodes to be queried in order to collect the federated statistics. During the installation of the FASM-N module, a flag is set up to distinguish between a collector or a leaf node.

#### b) SOLR interrogation

Starting from the URLs stored into the *dashboard_queue* table, the FASM-N performs the following steps:

1. in the first one, a query to the index node is executed to retrieve the list of shards:

   https://<SOLR index>/esg-search/search/?type=File&latest=true&distrib=true&
   format=application%%2Fsolr%2Bxml

   Then, for each URL,

2. to retrieve the metadata related to the downloaded file, a query to SOLR is carried out, taking into account the results of the previous interrogation, e.g:

   https://<SOLR index>/solr/files/select/?q=url:*<URL>*&
   shards=localhost:8983/solr/files,localhost:8982/solr/files

   The SOLR response is processed on the base of a configuration file containing a list of metadata for each project, as shown in Figure 21. One of the metadata retrieved is the *dataset_id*, which will be used in the third step.

```
– <projects>
  – <project name="obs4MIPs">
      <metadata occ="str">institute</metadata>
      <metadata occ="str">cf_standard_name</metadata>
      <metadata occ="str">variable</metadata>
      <metadata occ="str">variable_long_name</metadata>
      <metadata occ="str">time_frequency</metadata>
      <metadata occ="str">index_node</metadata>
      <metadata occ="str">processing_level</metadata>
      <metadata occ="str">source_id</metadata>
      <metadata occ="str">realm</metadata>
      <metadata occ="long">size</metadata>
      <metadata occ="bool">replica</metadata>
      <metadata occ="date">datetime_start</metadata>
      <metadata occ="date">datetime_stop</metadata>
    </project>
  – <project name="CMIP5">
      <metadata occ="str">model</metadata>
      <metadata occ="str">experiment</metadata>
      <metadata occ="str">cf_standard_name</metadata>
      <metadata occ="str">variable</metadata>
      <metadata occ="str">variable_long_name</metadata>
      <metadata occ="str">realm</metadata>
      <metadata occ="str">time_frequency</metadata>
      <metadata occ="str">institute</metadata>
      <metadata occ="bool">replica</metadata>
      <metadata occ="date">datetime_start</metadata>
      <metadata occ="date">datetime_stop</metadata>
      <metadata occ="long">size</metadata>
    </project>
  </projects>
```

*Fig. 21: Configuration file with metadata list for each project.*

3. the third step is in charge to gather metadata related to the dataset to which the file belongs:

   http://<SOLR index>/solr/datasets/select/?q=id:<dataset_id>&
   shards=localhost:8983/solr/datasets,localhost:8982/solr/files

### c) Single node module

This module is responsible to collect the information about the data node on which the download operation is carried out. When a download occurs, a filter of the ESGF Node Manager component records the operation in the database.
The interceptor event saves the statistics (log data) in the *dashboard_queue* table installed on the data node starting from January 1$^{st}$, 2017.

The statistics are divided in two macro categories: related to the (i) users (e.g. distribution of the clients per country, Identity Provider on which the users are registered, etc), (ii) file downloads (e.g. download number for year, data downloaded from several data nodes, the number of download distinct files, etc.).

### d) Collector module

The collector module is based on a hierarchical protocol, considering leaf and collector nodes (Figure 22). The first are the lower-level nodes, the latter are involved to gather all the information of the related leaf nodes. The leaf nodes must be installed on data nodes, where the NetCDF files are downloaded by users from the scientific community. All statistical information on the resources provided to users are contained in a data warehouse. Then, from this amount of data, extractions and combinations of information are performed, according to certain criteria and the results are stored in the data marts.

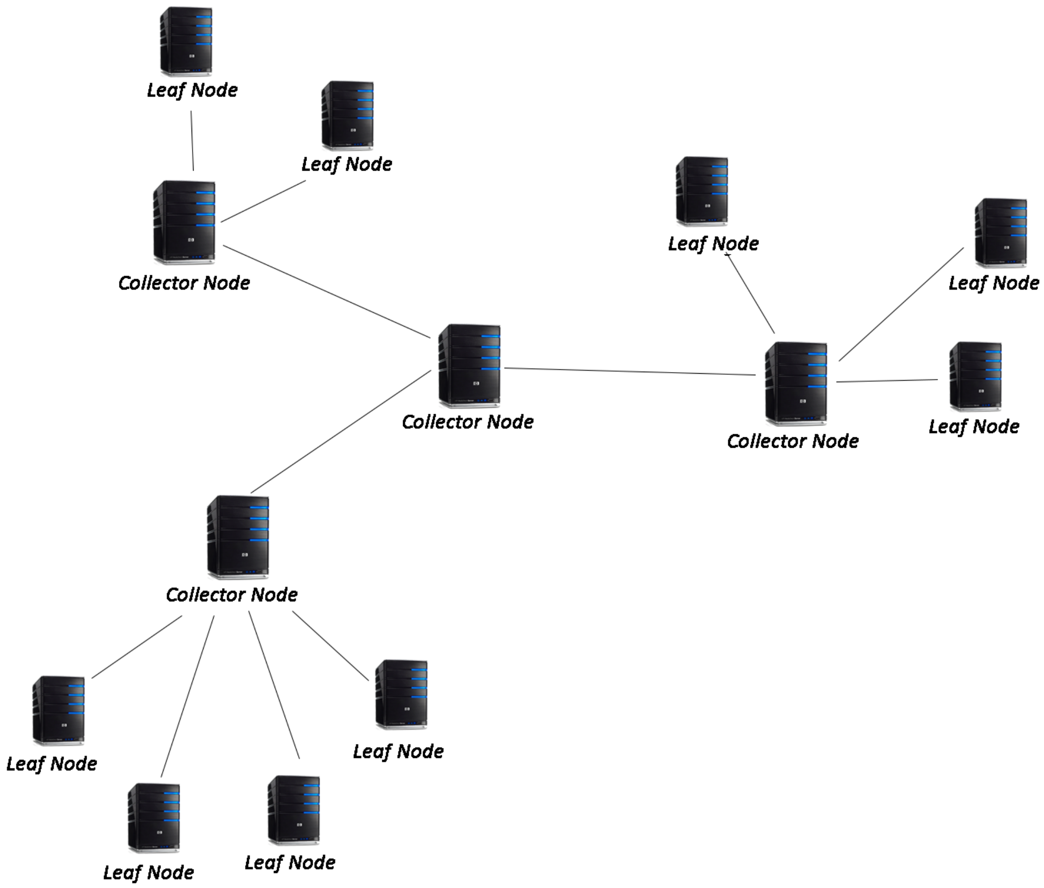



*Fig. 22: Hierarchical protocol.*

Collector and leaf nodes are mapped on two distinct data nodes (e.g. collector node on a Tier1 and leaf node on a Tier 2).
A leaf node is made up of various elements (Figure 23):

- A data warehouse where all the data related to users' downloads are stored;
- One or more data marts containing aggregated information;
- A set of REST APIs that provides access to the data of the data marts from the collector node that requires them.

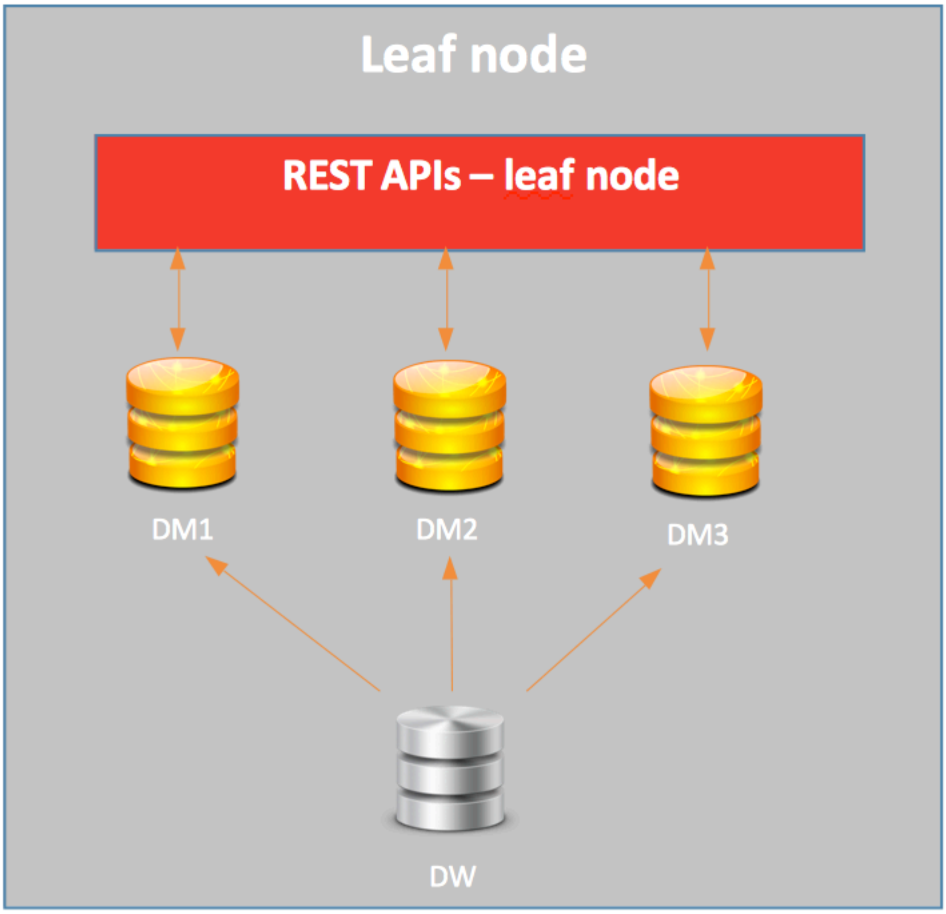



*Fig. 23: Leaf node configuration.*

The leaf nodes are able to share their information with other nodes, through the implementation of the REST APIs. When a node receives a request via its REST interface, it provides an answer through the following simple steps:

- analysing of the request (checking what data to collect and pick them up from the proper data mart);
- recovering of the data;
- serializing of the data in an XML file and sending to the node that made the request.

The collector nodes have a more complex structure because, in addition to making their information available to the collector through the REST APIs, it has the task of querying the leaf nodes associated with them.

The collector node is composed of:

- REST APIs;
- data warehouses and data marts;
- XML configuration file (Figure 24);
- federation component.

Each node consists of a unique identifier, an IP address and a port from which it can be reached. For each node, a list of datamarts is available. Each datamart has a timestamp and a path for querying the REST APIs.


```
- <configuration>
  - <node ipNodeAddress="esgf-fedtest.dkrz.de" port="8080">
      <datamart name="cross_dmart_project_host_time" path="crossproject/projecttime" timestamp="0"/>
      <datamart name="cross_dmart_project_host_geolocation" path="crossproject/projectgeolocation" timestamp="0"/>
      <datamart name="obs4mips_dmart_clients_host_time_geolocation" path="obs4mips/clients" timestamp="0"/>
      <datamart name="obs4mips_dmart_variable_host_time" path="obs4mips/variable" timestamp="0"/>
      <datamart name="obs4mips_dmart_source_host_time" path="obs4mips/source" timestamp="0"/>
      <datamart name="obs4mips_dmart_realm_host_time" path="obs4mips/realm" timestamp="0"/>
      <datamart name="obs4mips_dmart_dataset_host_time" path="obs4mips/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_experiment_host_time" path="cmip5/experiment" timestamp="0"/>
      <datamart name="cmip5_dmart_model_host_time" path="cmip5/model" timestamp="0"/>
      <datamart name="cmip5_dmart_variable_host_time" path="cmip5/variable" timestamp="0"/>
      <datamart name="cmip5_dmart_dataset_host_time" path="cmip5/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_clients_host_time_geolocation" path="cmip5/clients" timestamp="0"/>
    </node>
  - <node ipNodeAddress="esgf-data.jpl.nasa.gov" port="0">
      <datamart name="cross_dmart_project_host_time" path="crossproject/projecttime" timestamp="0"/>
      <datamart name="cross_dmart_project_host_geolocation" path="crossproject/projectgeolocation" timestamp="0"/>
      <datamart name="obs4mips_dmart_clients_host_time_geolocation" path="obs4mips/clients" timestamp="0"/>
      <datamart name="obs4mips_dmart_variable_host_time" path="obs4mips/variable" timestamp="0"/>
      <datamart name="obs4mips_dmart_source_host_time" path="obs4mips/source" timestamp="0"/>
      <datamart name="obs4mips_dmart_realm_host_time" path="obs4mips/realm" timestamp="0"/>
      <datamart name="obs4mips_dmart_dataset_host_time" path="obs4mips/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_experiment_host_time" path="cmip5/experiment" timestamp="0"/>
      <datamart name="cmip5_dmart_model_host_time" path="cmip5/model" timestamp="0"/>
      <datamart name="cmip5_dmart_variable_host_time" path="cmip5/variable" timestamp="0"/>
      <datamart name="cmip5_dmart_dataset_host_time" path="cmip5/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_clients_host_time_geolocation" path="cmip5/clients" timestamp="0"/>
    </node>
  - <node ipNodeAddress="pcmdi11.llnl.gov" port="8080">
      <datamart name="cross_dmart_project_host_time" path="crossproject/projecttime" timestamp="0"/>
      <datamart name="cross_dmart_project_host_geolocation" path="crossproject/projectgeolocation" timestamp="0"/>
      <datamart name="obs4mips_dmart_clients_host_time_geolocation" path="obs4mips/clients" timestamp="0"/>
      <datamart name="obs4mips_dmart_variable_host_time" path="obs4mips/variable" timestamp="0"/>
      <datamart name="obs4mips_dmart_source_host_time" path="obs4mips/source" timestamp="0"/>
      <datamart name="obs4mips_dmart_realm_host_time" path="obs4mips/realm" timestamp="0"/>
      <datamart name="obs4mips_dmart_dataset_host_time" path="obs4mips/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_experiment_host_time" path="cmip5/experiment" timestamp="0"/>
      <datamart name="cmip5_dmart_model_host_time" path="cmip5/model" timestamp="0"/>
      <datamart name="cmip5_dmart_variable_host_time" path="cmip5/variable" timestamp="0"/>
      <datamart name="cmip5_dmart_dataset_host_time" path="cmip5/dataset" timestamp="0"/>
      <datamart name="cmip5_dmart_clients_host_time_geolocation" path="cmip5/clients" timestamp="0"/>
    </node>
  </configuration>
```


*Fig. 24: Configuration file for the federation.*

Considering the architecture of the collector module (Fig. 25), the daemon is a process that is run on each collector node. Once in execution, it remains in background and at regular intervals launches the federation process. Other components represent the heart of architecture and contains instructions for managing the entire process of federation. The first operation is the parsing of the configuration file in order to know all the leaf nodes. Then the query generator prepares a set of queries for each leaf node. The controller runs the queries on each leaf node, through REST APIs, and the receiver receives, from leaf nodes, the query results in XML format, parses them and carries out the ingestion of the data in the data marts presents in the collector node. It updates a timestamp in the configuration file.

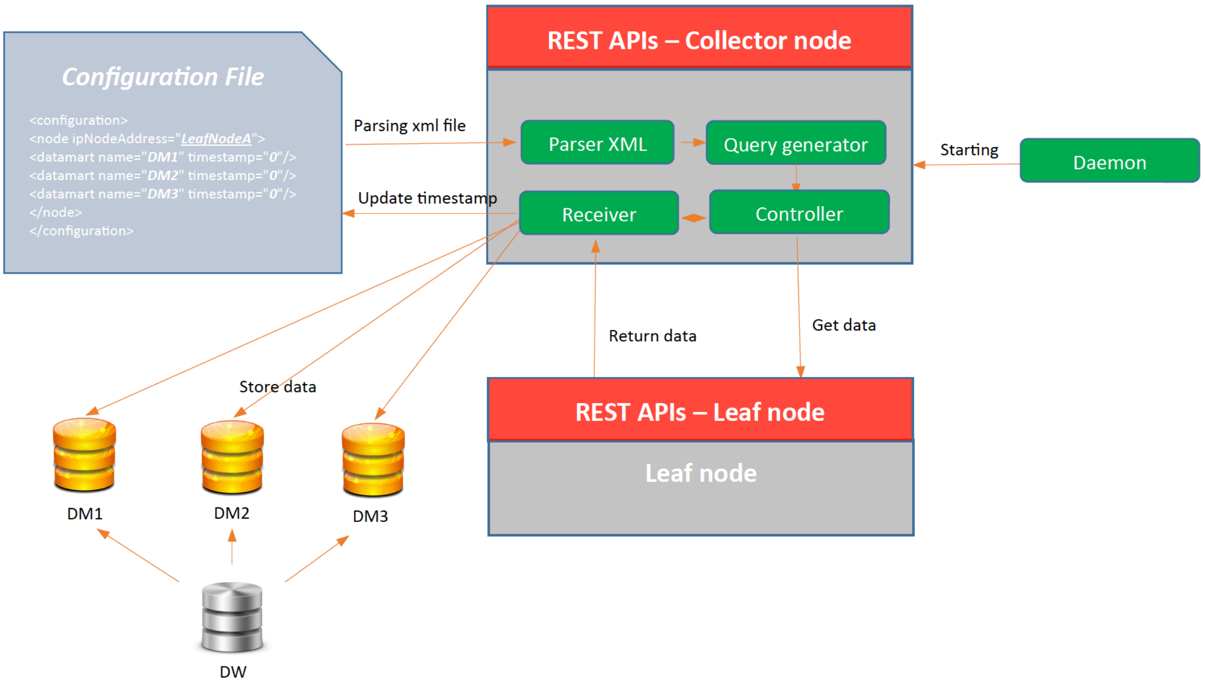



*Fig. 25: Architecture of the collector module.*

### 3.4 FASM-D

The Dashboard module of FASM has as main objective the provisioning of an integrated and usable way to visualize statistics at single site, ENES archive and ESGF federation level. Different categories of statistics, concerning the data usage and the clients activities and distribution, are provided and shown through a set of simple ad attractive graphical widgets (like charts, maps, tables and so on) giving the user a comprehensive view to access the data aggregated by the FASM-N module.

#### a) Architecture in the small

The Figure 26 shows the FASM-D GUI architecture in the small, through a detailed view of its main components.

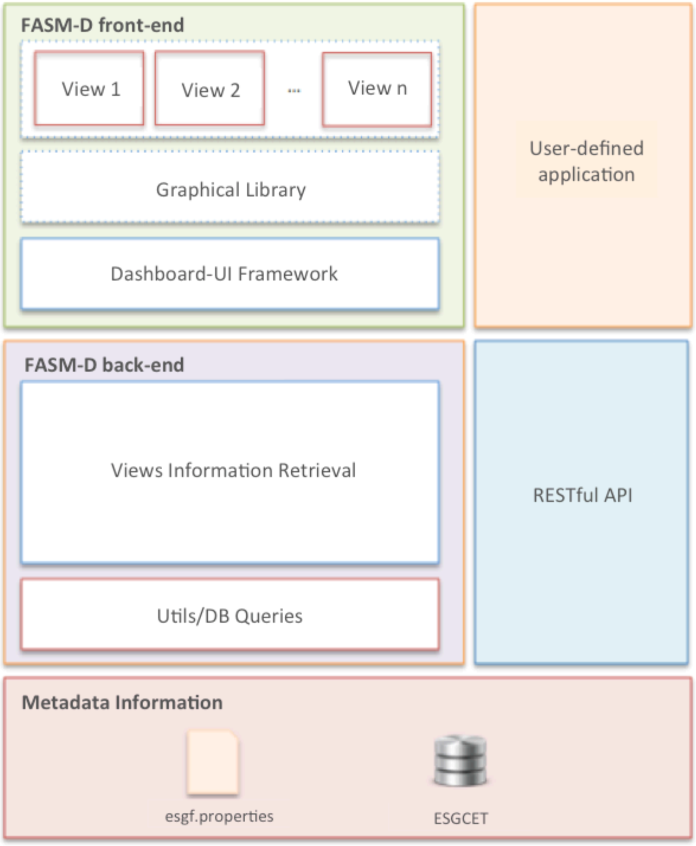



*Fig. 26 FASM-D architecture in the small*

To fullfill the requirements of modularity, configurability, robustness and scalability, the module has been designed taking in consideration a logic division of the different components which are mainly divided into:

- Metadata Information,
- FASM-D back-end,
- FASM-D front-end.

The Metadata Information section is in charge to provide data to the FASM-D back-end through the *esgcet* database, with particular reference to the *esgf_dashboard* namespace. Moreover, at this level, configuration files are placed to provide the FASM-D with additional information, such as parameters for the connection between the database and the application back-end, contained in the *esgf.properties* file.
The FASM-D back-end section represents a layer between the Metadata Information and the FASM-D front-end layer and is composed by a support layer, Utils/DB Queries, which supports the communication with the sources layer and is composed by a set of utility classes, and the Views Information Retrieval section which collects information from the Metadata Information and provide them to the FASM-D front-end.
At the same level, the REST APIs section provides the user with the system resources through a REST-like service accessible from custom web interface and optionally from individual users interested in statistics.

The FASM-D front-end represents the web interface itself the user can interact with and is composed by:

- Dashboard-UI framework representing the overall environment of the application and managing all the different front-end views;
- Graphical library: it consists of a set of base modules (trees, grids, charts, panels, ...), easily configurable and composable in order to create more complete views;
- View level: the closest level to the user, it consists of more complex and completed modules, made up of a set of the previous gadgets, in order to allow the user to make choices and visualize the desired information.

The last section (User-defined application) indicates that the user can use the Java REST APIs service to query the FASM-D back-end and define new types of applications to properly process the information.
Finally, additional information on the federated data archive are retrieved through a direct interrogation to the SOLR component by the FASM-D back-end module.

#### b) Main views

The FASM-D graphical user interface (GUI) in its final implementation moves away from the previous desktop metaphor to approach a new one closer to the dashboard concept.
Data visualization is often a difficult task for administrators and developers, as users need quick and simple access to lots of data and information.
A dashboard itself is a set of graphical component such as menu, charts, tables, and so on showing information like statistics, analytics, schedules and much more. The key challenge of such concept is to communicate the most important information in a straightforward way and allow users to go through specific details at the same time.
The FASM-D GUI can be thought as a control panel through which the user can access the different metrics/statistics at node, ENES community and ESGF federation level.
All the perspective through which the information can be analysed are grouped into several menu and sub-menu easily accessible from the left side of the GUI.
Being a centralized web-application, further requirements or customization requests should be previously discussed with the community partners and the GUI administrators.
In the following, a complete description of the different sections provided by the web-application:

- Statistics overview
- Data usage statistics
- Project-specific statistics
- Clients statistics
- Federated data archive
- Service status

##### STATISTICS OVERVIEW

The opening section of the GUI represents a sort of summary of the most relevant views and contains quick links to the most significant statistics.
In particular, Figure 27 provides a summary view of the total amount of data published on the ESGF federated archive in terms of total number of datasets and total data volume (in Terabytes) and number of datasets and data volume for CMIP5, CORDEX and Obs4MIPs projects.

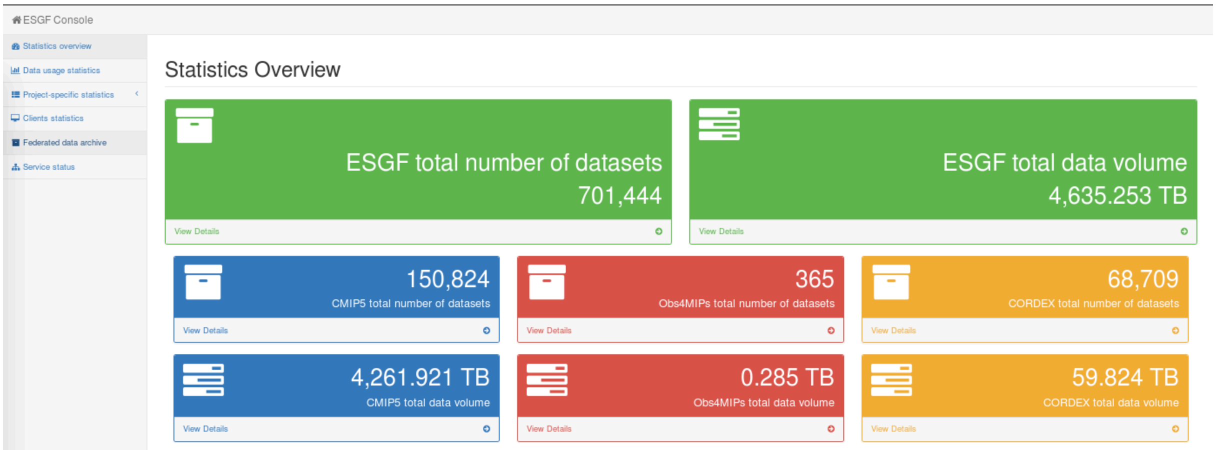



*Fig. 27: Federated data archive*

Figure 28 shows the number of registered users registered for each project, both in a donut chart and a bar chart visualizing the logarithmic scale of the values.
Same way, the number of downloads for each project is visualized, both in a donut chart and a bar chart for the logarithmic scale.


*Fig. 28: Registered users and number of downloads per project*

Going on with the views, Figure 29 offers a practical way to obtain information on the distribution of the downloads' numbers per continents on a map, while Figure 30 shows the same information but grouped by countries on different tables.

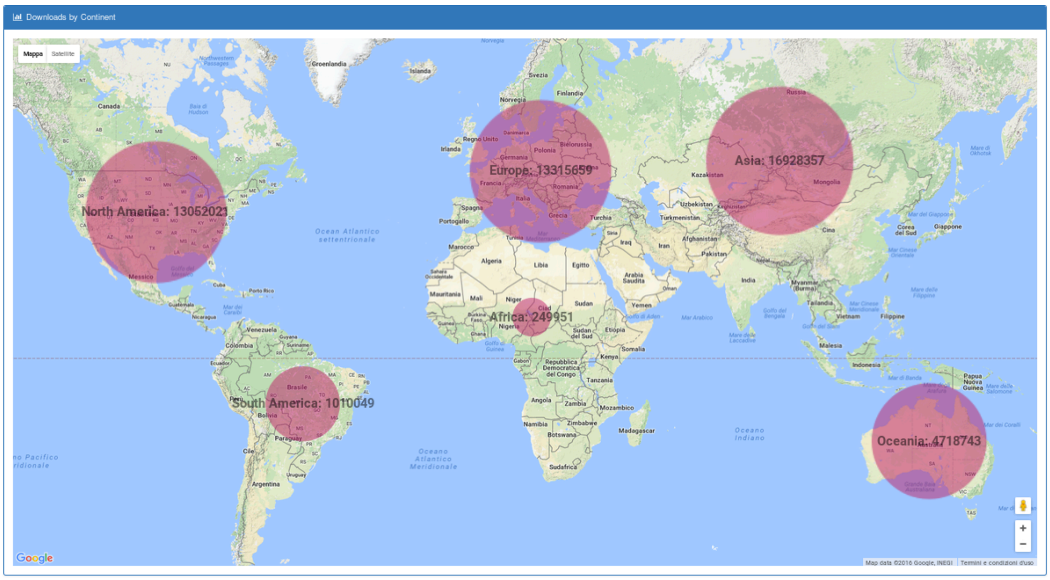



*Fig. 29: Downloads by continents*

| North America | 13,052,021 |
|---|---|
| United States of America | 12,666,874 |
| Canada | 208,352 |
| Trinidad and Tobago | 171,291 |
| Mexico | 4,432 |
| Jamaica | 1,008 |
| Costa Rica | 42 |
| Belize | 18 |
| Puerto Rico | 2 |
| Nicaragua | 2 |

| Africa | 249,951 |
|---|---|
| Kenya | 172,085 |
| South Africa | 59,502 |
| Algeria | 4,978 |
| Zambia | 2,974 |
| Senegal | 2,734 |
| Niger | 2,396 |
| Egypt | 1,426 |
| Ethiopia | 1,150 |
| Ghana | 752 |
| Morocco | 478 |
| Benin | 434 |

| Europe | 13,315,659 |
|---|---|
| Germany | 5,981,937 |
| Spain | 3,029,726 |
| Switzerland | 1,318,830 |
| United Kingdom | 1,111,048 |
| France | 735,278 |
| Netherlands | 215,692 |
| Norway | 161,670 |
| Italy | 134,776 |
| Sweden | 112,892 |
| Portugal | 102,442 |
| Greece | 51,046 |

| South America | 1,010,049 |
|---|---|
| Chile | 785,661 |
| Brazil | 121,664 |
| Colombia | 75,554 |
| Argentina | 16,734 |
| Peru | 9,886 |
| Ecuador | 420 |
| Bolivia | 66 |
| Suriname | 38 |
| Venezuela | 22 |
| Uruguay | 4 |

| Asia | 16,928,357 |
|---|---|
| China | 12,148,283 |
| Japan | 2,666,194 |
| Korea, South | 696,774 |
| Iran | 626,600 |
| Thailand | 276,832 |
| India | 211,548 |
| Taiwan | 118,082 |
| Hong Kong | 63,200 |
| United Arab Emirates | 29,088 |
| Singapore | 28,416 |
| Turkey | 19,262 |

| Oceania | 4,718,743 |
|---|---|
| Australia | 4,286,555 |
| New Zealand | 432,174 |
| New Caledonia | 14 |

*Fig. 30: Downloads by countries*

Also, information of the distribution of the registered users is provided. Figure 31 shows the number of registeres users grouped by continents on a map, Figure 32 shows the number of registered users grouped by countries on tables and Figure 33 represents the same information but grouped by Identity Provider on a map.

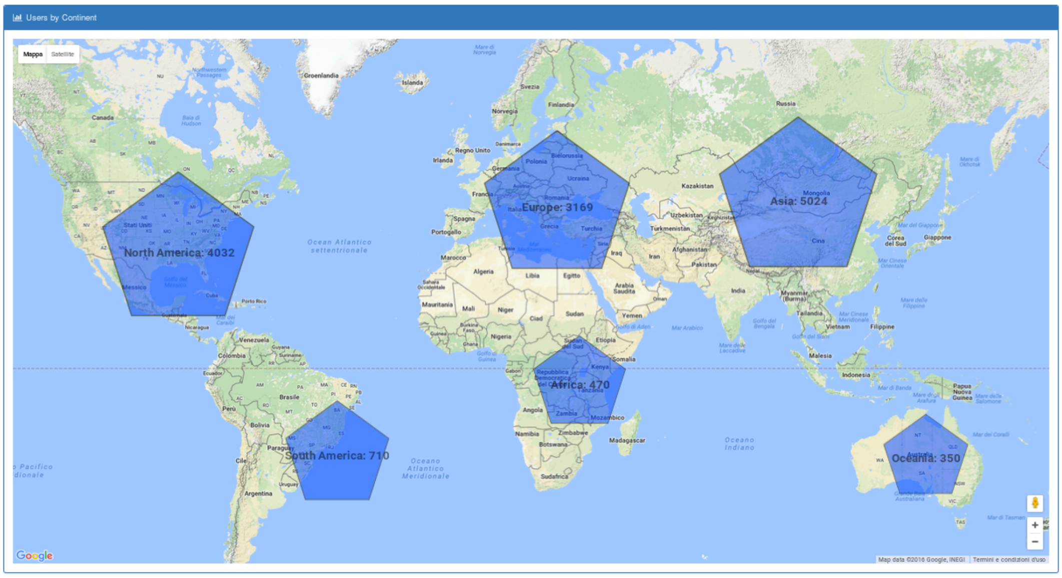



*Fig. 31: Registered users by continent*

| North America | 4,032 |
|---|---|
| United States of America | 3,486 |
| Canada | 434 |
| Mexico | 72 |
| Cuba | 10 |
| Costa Rica | 10 |
| Jamaica | 8 |
| Trinidad and Tobago | 4 |
| Guatemala | 2 |
| Barbados | 2 |
| Panama | 2 |
| Nicaragua | 2 |

| Africa | 470 |
|---|---|
| Ethiopia | 70 |
| South Africa | 66 |
| Senegal | 42 |
| Kenya | 32 |
| Zambia | 30 |
| Egypt | 26 |
| Ghana | 20 |
| Benin | 20 |
| Nigeria | 20 |
| Cameroon | 16 |
| Burkina Faso | 14 |

| Europe | 3,169 |
|---|---|
| Germany | 1,015 |
| France | 376 |
| Italy | 242 |
| Spain | 218 |
| Netherlands | 166 |
| Russian Federation | 158 |
| Norway | 150 |
| Sweden | 128 |
| Switzerland | 94 |
| Belgium | 74 |
| Denmark | 72 |

| South America | 710 |
|---|---|
| Brazil | 304 |
| Colombia | 142 |
| Chile | 112 |
| Argentina | 80 |
| Peru | 40 |
| Bolivia | 12 |
| Suriname | 6 |
| Venezuela | 6 |
| Ecuador | 4 |
| Paraguay | 2 |
| Uruguay | 2 |

| Asia | 5,024 |
|---|---|
| China | 2,162 |
| India | 920 |
| Japan | 434 |
| Iran | 342 |
| Korea, South | 270 |
| Thailand | 194 |
| Taiwan | 92 |
| Indonesia | 80 |
| Pakistan | 74 |
| Malaysia | 58 |
| Israel | 42 |

| Oceania | 350 |
|---|---|
| Australia | 308 |
| New Zealand | 38 |
| Norfolk Island | 2 |
| New Caledonia | 2 |

*Fig. 32: Registered users by countries*

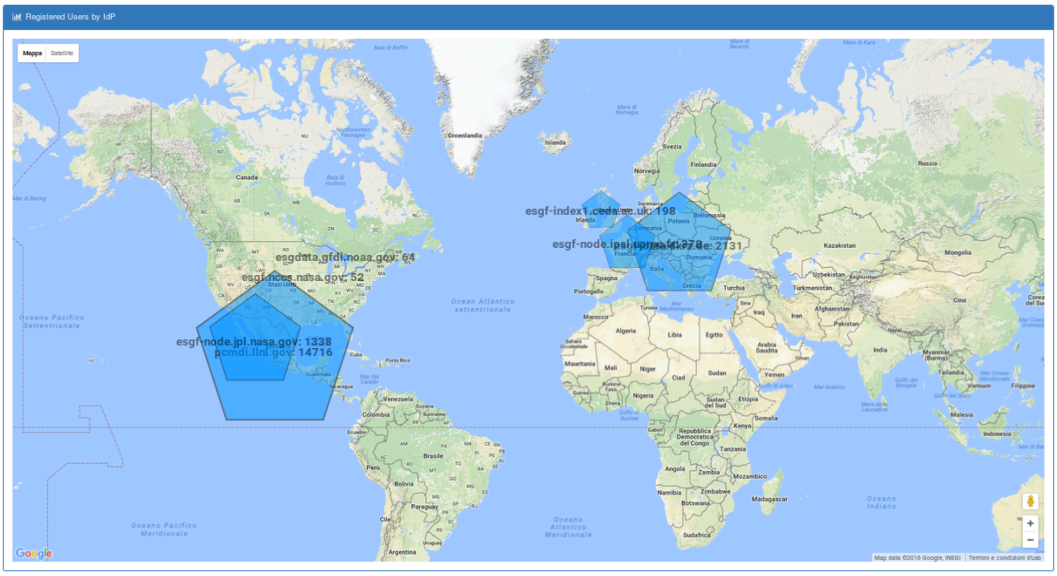



*Fig. 33: Registered users by Identity Providers*

## DATA USAGE STATISTICS

This section incorporates a set of views related to the cross-project statistics showing the same data download activities but accessed through different perspectives. Different metrics such as number of downloads, number of successful downloads, downloaded data in terms of Gigabytes are visualized in several widgets grouping the information by time, by host (as in Figure 34) and by project (Figure 35).

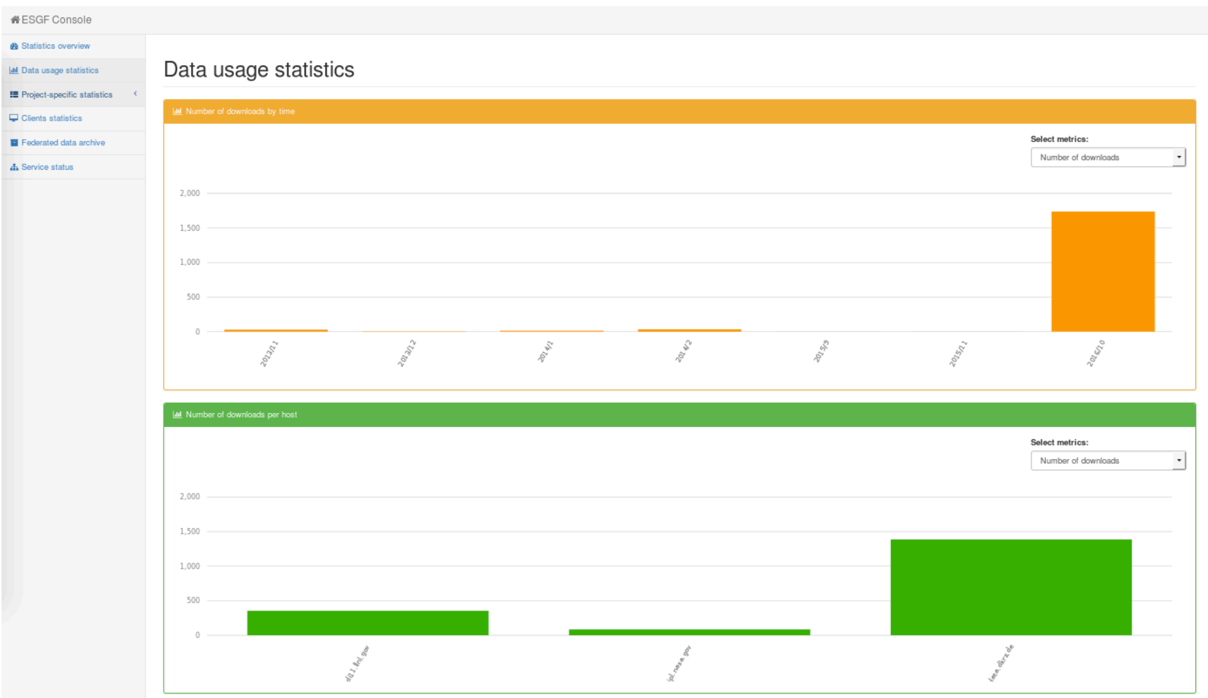


*Fig. 34: Data usage statistics by time and by host*

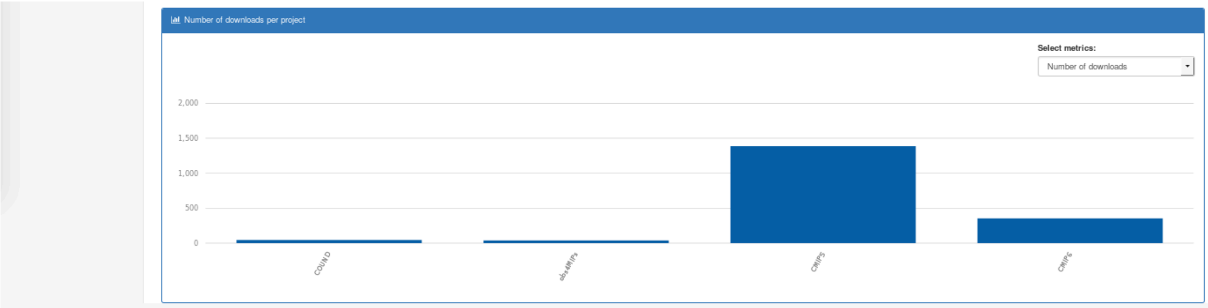



*Fig. 35: Data usage statistics by project*

## PROJECT-SPECIFIC STATISTICS

The Project-specific section gives a complete overview of different projects, like CMIP5, Obs4MIPs, CORDEX and, in particular, provides a very useful set of widgets each one representing a specific information, such as:

- Top ten datasets;
- Top ten sources;
- Top ten variables;
- Number of downloads, number of successful downloads and downloaded data (in Gigabytes) by time;
- Number of downloads, number of successful downloads and downloaded data (in Gigabytes) per host;
- Number of downloads, number of successful downloads and downloaded data (in Gigabytes) per project.

Figures 36 (a) and (b) show the previous views in the respective widgets.
The same widgets are available for CMIP5 project.

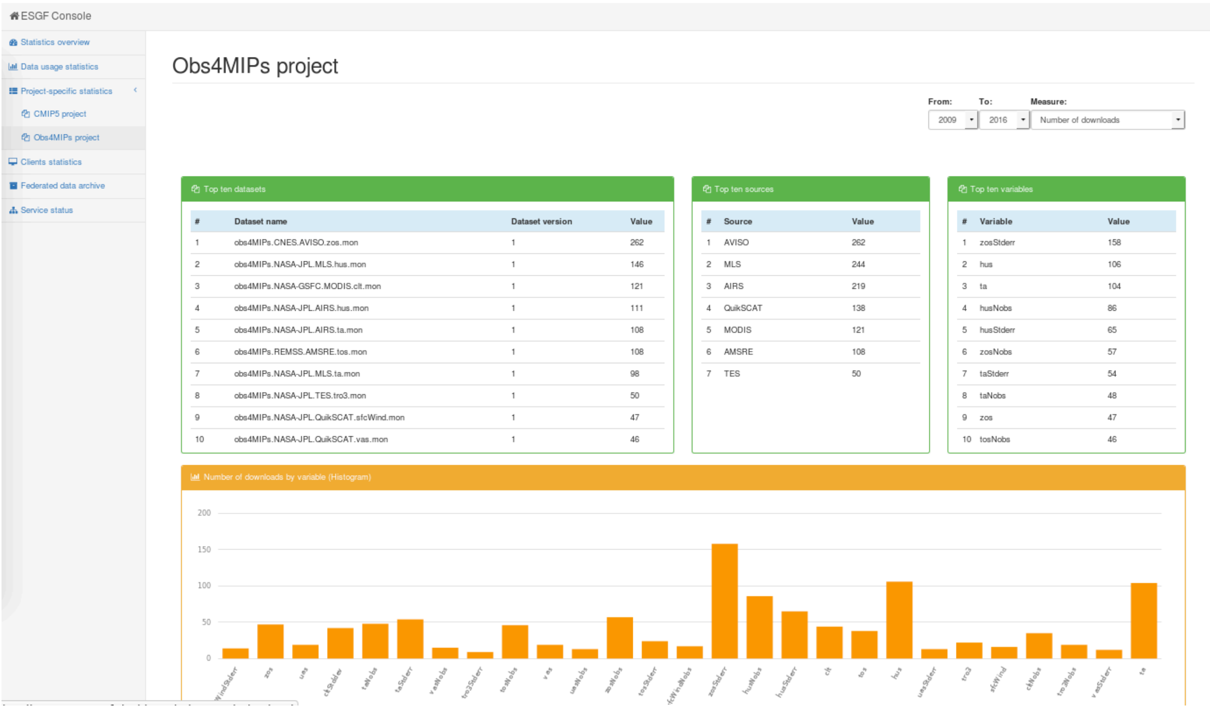



*Fig. 36 (a): Project-specific statistics*

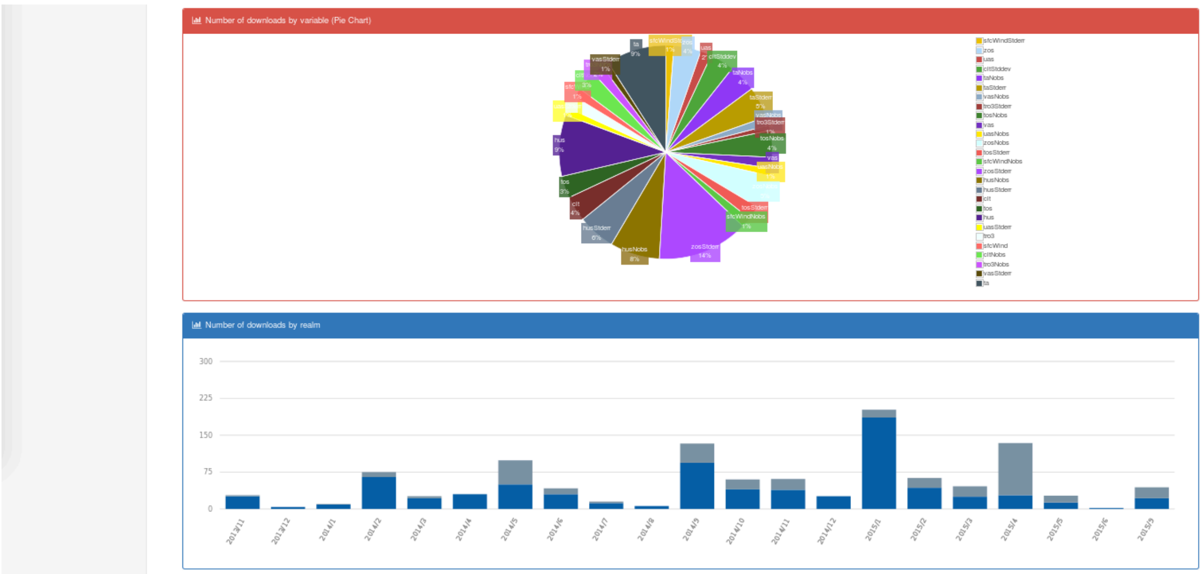



*Fig. 36 (b): Project-specific statistics*

## CLIENTS STATISTICS

The present section contains all the information related to the clients and their distribution on a map (Figure 37) and in tabular and graphical view (Figure 38).

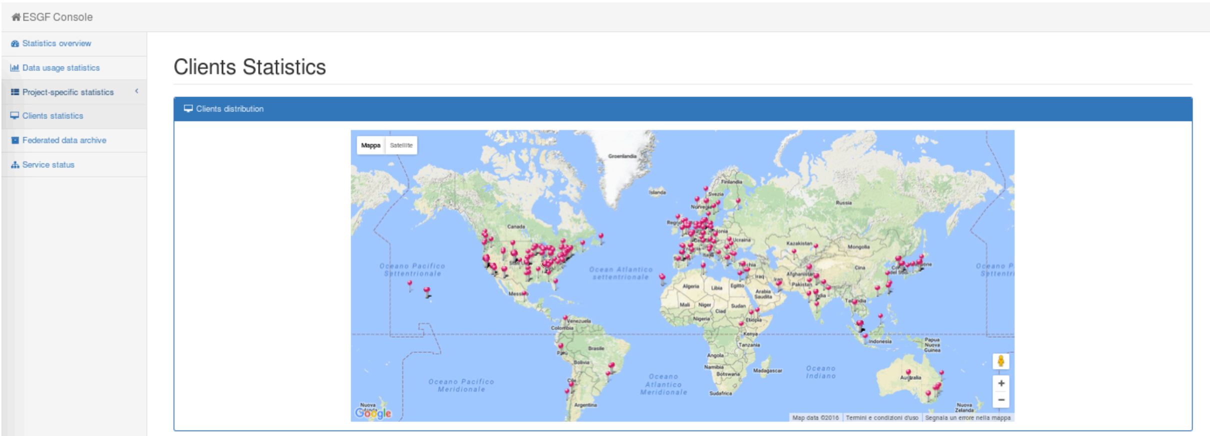



*Fig. 37: Clients statistics: map view*

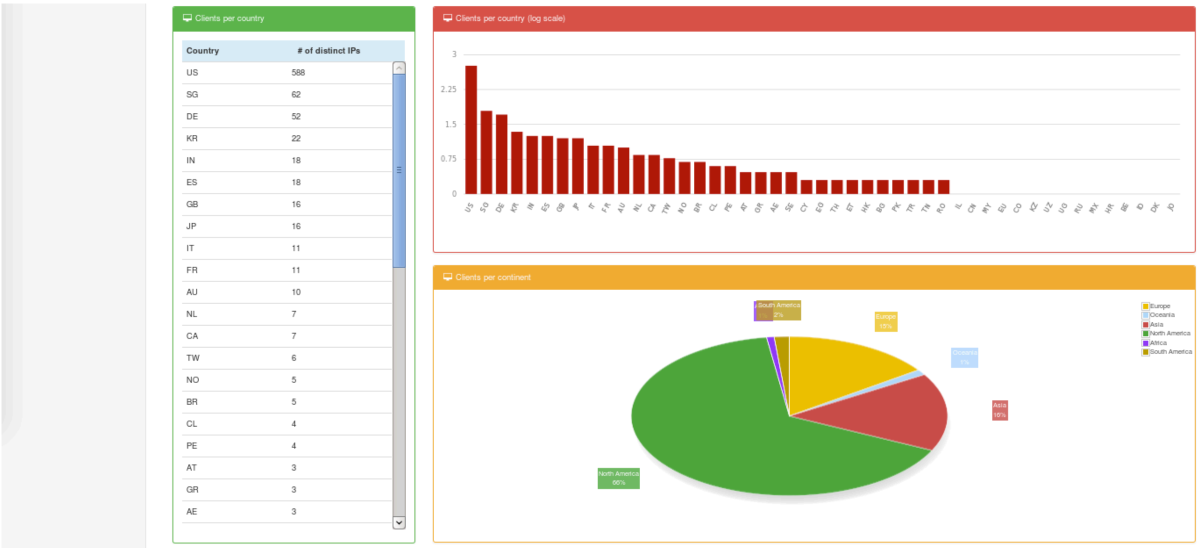



*Fig. 38: Clients statistics: tabular and graphical view*

## FEDERATED STATISTICS

Information related to the amount of data available for download on the entire federation is placed under this section.

In particular, it is possible to select a specific data node and obtain information on the total number of published dataset and total amount of data (in Terabytes) for that node (Figure 39).

Moreover, Figure 40 shows two tables listing the Models and the Modelling Institutes for CMIP5 project.

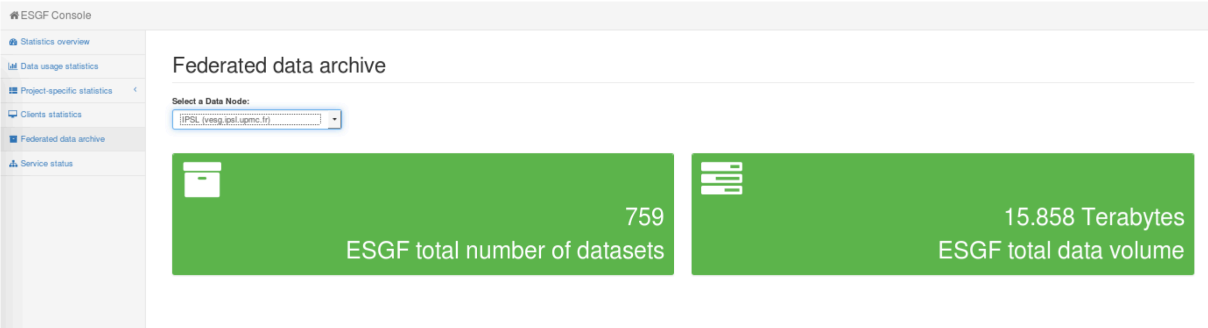



*Fig. 39: Single site data archive statistics*

Published CMIP5 data per Model

| # | Model | # of datasets | Size (TB) |
|---|---|---|---|
| 1 | ACCESS1.0 | 337 | 35.476 |
| 2 | ACCESS1.3 | 538 | 33.825 |
| 3 | BCC-CSM1.1 | 5,101 | 41.844 |
| 4 | BCC-CSM1.1(m) | 403 | 3,610.014 |
| 5 | BNU-ESM | 513 | 19.93 |
| 6 | CCSM4 | 4,956 | 166.347 |
| 7 | CESM1(BGC) | 309 | 22.716 |
| 8 | CESM1(CAM5) | 378 | 31.279 |
| 9 | CESM1(CAM5.1-FV2) | 56 | 2.928 |
| 10 | CESM1(FASTCHEM) | 51 | 3.269 |
| 11 | CESM1(WACCM) | 155 | 3.873 |
| 12 | CFSv2-2011 | 2,644 | 34.793 |
| 13 | CMCC-CESM | 87 | 1.254 |
| 14 | CMCC-CM | 992 | 144.024 |
| 15 | CMCC-CMS | 110 | 4.227 |
| 16 | CNRM-CM5 | 3,390 | 135.5 |
| 17 | CNRM-CM5-2 | 263 | 11.909 |
| 18 | CSIRO-Mk3.6.0 | 3,120 | 57.593 |
| 19 | CSIRO-Mk3L-1-2 | 26 | 0.03 |
| 20 | CanAM4 | 184 | 8.027 |
| 21 | CanCM4 | 19,118 | 23.317 |
| 22 | CanESM2 | 2,577 | 39.844 |

Published CMIP5 data per Institute

| # | Modeling institute | # of datasets | Size (TB) |
|---|---|---|---|
| 1 | BCC | 5,504 | 97.992 |
| 2 | BNU | 513 | 19.93 |
| 3 | CCCMA | 21,879 | 71.187 |
| 4 | CMCC | 1,189 | 149.504 |
| 5 | CNRM-CERFACS | 3,653 | 147.409 |
| 6 | COLA-CFS | 1,189 | 7.953 |
| 7 | CSIRO-BOM | 875 | 69.301 |
| 8 | CSIRO-QCCCE | 3,120 | 57.593 |
| 9 | FIO | 230 | 6.694 |
| 10 | ICHEC | 3,620 | 134.946 |
| 11 | INM | 486 | 21.402 |
| 12 | INPE | 24 | 7.957 |
| 13 | IPSL | 10,757 | 699.208 |
| 14 | LASG-CESS | 1,553 | 40.014 |
| 15 | LASG-IAP | 418 | 7.597 |
| 16 | MIROC | 16,791 | 823.919 |
| 17 | MOHC | 24,720 | 148.004 |
| 18 | MPI-M | 11,655 | 195.248 |
| 19 | MRI | 7,271 | 404.758 |
| 20 | NASA-GISS | 7,681 | 227.295 |
| 21 | NASA-GMAO | 2,520 | 8.631 |
| 22 | NCAR | 4,956 | 166.347 |

*Fig. 40: CMIP5 models and modeling institutes*

## SERVICE STATUS

This section lists all the widgets related to the status of the service. Figure 41, in fact, shows a map with information on the different software stack installed on each node of the federation.

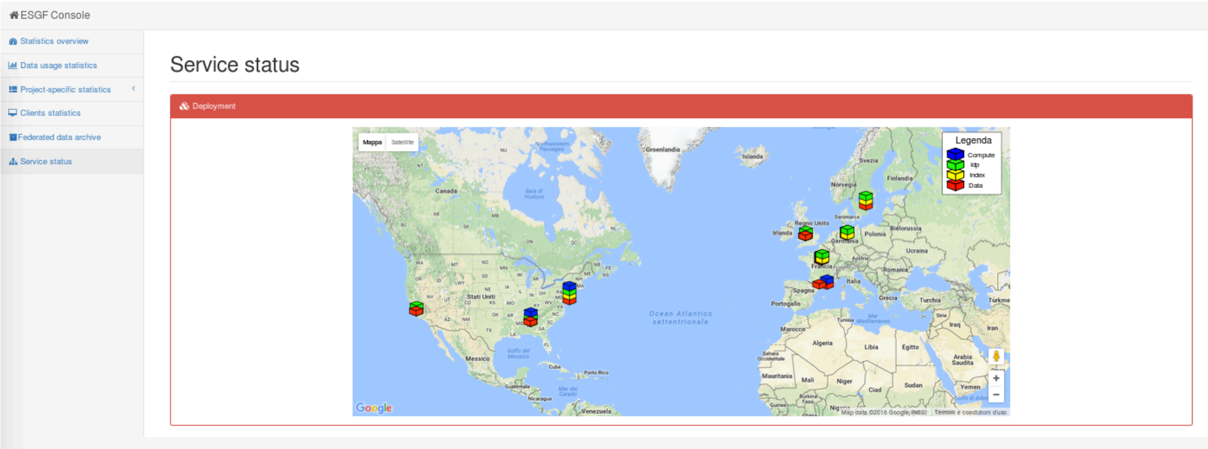



*Fig. 41: Deployment view*

## 4. Conclusions

This document provided a complete description of the implementation phase of the Federated Archive System Monitoring (FASM) package.
The architecture in the large and in the small has been presented for the FASM-N and FASM-D modules, as well as a detailed description of the entire data warehouse system, the data marts and the views. In particular, the FASM-N acts as local information provider and global collector of federated statistics at IS-ENES and ESGF levels.
The current architecture allows the collection of significant amount of data about the dissemination of Earth system models; the data warehouse behind the monitoring framework and the data marts presented in this report are only few examples of what could be extracted and visualized through the web interface. The latest version of the software is available in the related git repositories under the git project ESGF and will be officially available in the ESGF release by February 2017.
In the future, according to further user requirements or specific needs, other statistics could be inferred defining new perspectives on the same data and visualizing them through new intuitive and interactive widgets.

## 5. Glossary

WP: Work Package

JRA: Joint Research Activity

MS: Milestone

FASM-N: Federated Archive System Monitoring - Notification

FASM-D: Federated Archive System Monitoring - Dashboard

IS-ENES2: Infrastructure for the European Network for Earth System modelling - Phase 2

ESGF: Earth System Grid Federation

NFR: Non-Functional Requirements

REST API: REpresentational State Transfer Application Programming Interface (or REST interface)

CORDEX: COordinated Regional Climate Downscaling Experiment

FR: Functional Requirements

ECA: Event-Condition-Action

ETL: Extraction – Transformation - Loading

CMIP5: Coupled Model Intercomparison Project Phase 5

Obs4MIPs: Observations for Model Intercomparisons Project

GUI: Graphical User Interface

FK: Foreign Key

PK: Primary Key

IdP node: Identity Provider node

DFM: Dimensional Fact Model

CF: Climate and Forecast

DW: Data Warehouse

DM: Data Mart

URL: Uniform Resource Locator

NetCDF: Network Common Data Form

XML: eXtensible Markup Language

IP address: Internet Protocol address

Dashboard-UI: Dashboard-User Interface

DB: database

# APPENDIX

This section presents a detailed description of the tables componing the cross-project and the project-specific data warehouses. For each table all the fields are defined in terms of meaning, data type, range, storage size and other important features.

| TABLE NAME | ATTRIBUTE NAME | CONTENTS | TYPE | FORMAT | RANGE | STORAGE SIZE | REQUIRED | PK/FK | REFERENCE |
|---|---|---|---|---|---|---|---|---|---|
| cross_fact_download | download_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | size | Size of the download | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | success | Successful download | boolean | t/f | true/false | 1 byte | | | |
| | duration | Duration of the download | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | replica | Download of a replica of the dataset | boolean | t/f | true/false | 1 byte | | | |
| | user_id_hash | Hash of the user ID | character varying(64) | xxxxxxx | | variable | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | user_idp | Identity provider of the user | character varying(64) | xxxxxxx | | variable | | | |
| | hour | Hour of the day | smallint | HH | 00 - 23 | 2 bytes | | | |
| | minute | Minute of the hour | smallint | mm | 00 - 59 | 2 bytes | | | |
| | project_group_key | Reference to project group | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | cross_bridge_project |
| | geolocation_key | Reference to geolocation | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | FK | cross_dim_geolocation |
| | date_key | Reference to download date | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | cross_dim_date |
| | id_query | Reference to dashboard_queue | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | dashboard_queue |
| cross_dim_project | project_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | project_name | Project name | character varying(64) | xxxxxxx | | variable | true | | |
| cross_bridge_project | project_key | Reference to project | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | cross_dim_project |
| | project_group_key | ID for the project group | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | | |
| cross_dim_geolocation | geolocation_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | latitude | Geografic coordinate of the user location | numeric(14,11) | 999,99999 | 11 decimal digits precision | 8 bytes | | | |
| | longitude | Geografic coordinate of the user location | numeric(14,11) | 999,99999 | 11 decimal digits precision | 8 bytes | | | |
| | country_id | Reference to country | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | country |
| country | country_id | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | country_code | Code of the country | character(2) | xxxxxxx | | variable | true | | |
| | country_name | Country name | character varying(64) | xxxxxxx | | variable | true | | |
| | continent_code | Continent code | character(2) | xxxxxxx | | variable | true | FK | continent |
| continent | continent_code | Primary key of the table | character(2) | xxxxxxx | | variable | true | PK | |
| | continent_name | Continent name | character varying(255) | xxxxxxx | | variable | true | | |
| cross_dim_date | date_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | download_date | Date of the download | date | 1999-01-08 | 4713 BC - 5874897 AD | 4 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| cross_dmart_project_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_replica_downloads | Total number of replica downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | project_name | Project name | character varying(64) | xxxxxxx | | variable | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| cross_dmart_project_host_geolocation | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_replica_downloads | Total number of replica downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | project_name | Project name | character varying(64) | xxxxxxx | | variable | | | |
| | latitude | Geografic coordinate of the user location | numeric(14,11) | 999,99999 | 11 decimal digits precision | 8 bytes | | | |
| | longitude | Geografic coordinate of the user location | numeric(14,11) | 999,99999 | 11 decimal digits precision | 8 bytes | | | |

*Fig. 42: Cross-project DW data dictionary*



CAPACITIES

| TABLE NAME | ATTRIBUTE NAME | CONTENTS | TYPE | FORMAT | RANGE | STORAGE SIZE | REQUIRED | PK/FK | REFERENCE |
|---|---|---|---|---|---|---|---|---|---|
| cmip5_fact_download | download_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | size | Size of the download | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | success | Successful download | boolean | t/f | true/false | 1 byte | | | |
| | duration | Duration of the download | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | replica | Download of a replica of the dataset | boolean | t/f | true/false | 1 byte | | | |
| | user_id_hash | Hash of the user ID | character varying(n) | xxxxxxx | | variable | | | |
| | host_name | Host name | character varying(n) | xxxxxxx | | variable | | | |
| | user_idp | Identity provider of the user | character varying(n) | xxxxxxx | | variable | | | |
| | hour | Hour of the day | smallint | HH | 00 - 23 | 2 bytes | | | |
| | minute | Minute of the hour | smallint | mm | 00 - 59 | 2 bytes | | | |
| | dataset_name | Name of the dataset | character varying(n) | xxxxxxx | | variable | | | |
| | dataset_version | Publishing version of the dataset | character varying(n) | xxxxxxx | | variable | | | |
| | geolocation_key | Reference to geolocation | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | FK | cmip5_dim_geolocation |
| | date_key | Reference to download date | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | cmip5_dim_date |
| cmip5_dim_geolocation | geolocation_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | latitude | Geografic coordinate of the user location | double precision | 999,99999 | 15 decimal digits precision | 8 bytes | | | |
| | longitude | Geografic coordinate of the user location | double precision | 999,99999 | 15 decimal digits precision | 8 bytes | | | |
| | country | Country of the user location | character varying(n) | xxxxxxx | | variable | | | |
| | continent | Continent of the user location | character varying(n) | xxxxxxx | | variable | | | |
| cmip5_dim_date | date_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | download_date | Date of the download | date | 1999-01-08 | 4713 BC - 5874897 AD | 4 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| cmip5_dim_institute | institute_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | institute_name | Institute name | character varying(n) | xxxxxxx | | variable | true | | |
| cmip5_bridge_institute | institute_key | Reference to institute | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | cmip5_dim_institute |
| | institute_group_key | ID for the institute group | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | | |
| cmip5_dim_realm | realm_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | realm_name | Realm name | character varying(n) | xxxxxxx | | variable | true | | |
| cmip5_bridge_realm | realm_key | Reference to realm | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | cmip5_dim_realm |
| | realm_group_key | ID for the realm group | | | | | | | |
| cmip5_dim_model | model_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | model_name | Model name | character varying(n) | xxxxxxx | | variable | true | | |
| cmip5_bridge_model | model_key | Reference to model | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | cmip5_dim_model |
| | model_group_key | ID for the model group | | | | | | | |
| cmip5_dim_experiment | experiment_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | experiment_name | Experiment name | character varying(n) | xxxxxxx | | variable | true | | |
| cmip5_bridge_experiment | experiment_key | Reference to experiment | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | cmip5_dim_experiment |
| | experiment_group_key | ID for the experiment group | | | | | | | |
| cmip5_dim_variable | variable_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | variable_code | Variable name | character varying(n) | xxxxxxx | | variable | true | | |
| | variable_long_name | Variable name | character varying(n) | xxxxxxx | | variable | true | | |
| | cf_standard_name | Variable name | character varying(n) | xxxxxxx | | variable | true | | |
| cmip5_bridge_variable | variable_key | Reference to variable | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | cmip5_dim_variable |
| | variable_group_key | ID for the variable group | | | | | | | |
| cmip5_dim_time_frequency | time_frequency_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | time_frequency_value | Time frequency value | | | | | | | |
| cmip5_bridge_time_frequency | time_frequency_key | Reference to time frequency | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | cmip5_dim_time_frequency |
| | time_frequency_group_key | ID for the time frequency group | | | | | | | |

*Fig. 43 (a): CMIP5 DW data dictionary*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| cmip5_dmart_experiment_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_replica_downloads | Total number of replica downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | experiment_name | Experiment name | character varying(n) | xxxxxxx | | variable | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| cmip5_dmart_model_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_replica_downloads | Total number of replica downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | model_name | Model name | character varying(n) | xxxxxxx | | variable | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| cmip5_dmart_variable_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_replica_downloads | Total number of replica downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | variable_code | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| | variable_long_name | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| | cf_standard_name | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |

*Fig. 43 (b): CMIP5 DW data dictionary*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| cmip5_dmart_clients_host_time_geolocation | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_replica_downloads | Total number of replica downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | latitude | Geografic coordinate of the user location | numeric(14,11) | 999,99999 | 11 decimal digits precision | 8 bytes | | | |
| | longitude | Geografic coordinate of the user location | numeric(14,11) | 999,99999 | 11 decimal digits precision | 8 bytes | | | |
| cmip5_dmart_dataset_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_replica_downloads | Total number of replica downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | dataset_name | Name of the dataset | character varying(64) | xxxxxxx | | variable | | | |
| | dataset_version | Version of the dataset | smallint | xxx | -32768 to +32767 | 2 bytes | | | |
| | datetime_start | Start time of temporal coverage of the dataset | character varying(64) | xxxxxxx | | variable | | | |
| | datetime_stop | Stop time of temporal coverage of the dataset | character varying(64) | xxxxxxx | | variable | | | |

*Fig. 43 (c): CMIP5 DW data dictionary*

| TABLE NAME | ATTRIBUTE NAME | CONTENTS | TYPE | FORMAT | RANGE | STORAGE SIZE | REQUIRED | PK/FK | REFERENCE |
|---|---|---|---|---|---|---|---|---|---|
| obs4mips_fact_download | download_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | size | Size of the download | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | Chiudi schermo intero | | |
| | success | Successful download | boolean | t/f | true/false | 1 byte | | | |
| | duration | Duration of the download | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | user_id_hash | Hash of the user ID | character varying(n) | xxxxxxx | | variable | | | |
| | user_idp | Identity provider of the user | character varying(n) | xxxxxxx | | variable | | | |
| | host_name | Host name | character varying(n) | xxxxxxx | | variable | | | |
| | hour | Hour of the day | smallint | HH | 00 - 23 | 2 bytes | | | |
| | minute | Minute of the hour | smallint | mm | 00 - 59 | 2 bytes | | | |
| | dataset_key | Reference to dataset | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | FK | obs4mips_dim_dataset |
| | file_key | Reference to file | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | FK | obs4mips_dim_file |
| | geolocation_key | Reference to geolocation | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | FK | obs4mips_dim_geolocation |
| | date_key | Reference to download date | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | obs4mips_dim_date |
| | institute_group_key | Reference to group of institutes | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | obs4mips_bridge_institute |
| | variable_group_key | Reference to group of variables | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | obs4mips_bridge_variable |
| | time_frequency_group_key | Reference to group of time frequencies | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | obs4mips_bridge_time_frequency |
| | index_node_key | Reference to group of index nodes | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | obs4mips_dim_index_node |
| | processing_level_group_key | Reference to group of processing levels | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | obs4mips_bridge_processing_level |
| | source_id_group_key | Reference to group of sources | integer | 100000 | -2147483648 to +2147483648 | 3 bytes | | FK | obs4mips_bridge_source_id |
| | realm_group_key | Reference to group of realms | integer | 100001 | -2147483648 to +2147483649 | 4 bytes | | FK | obs4mips_bridge_realm |
| | id_query | Reference to dashboard_queue | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | | FK | dashboard_queue |
| obs4mips_dim_geolocation | geolocation_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | latitude | Geografic coordinate of the user location | double precision | 999,99999 | 15 decimal digits precision | 8 bytes | | | |
| | longitude | Geografic coordinate of the user location | double precision | 999,99999 | 15 decimal digits precision | 8 bytes | | | |
| | country | Country of the user location | character varying(n) | xxxxxxx | | variable | | | |
| | continent | Continent of the user location | character varying(n) | xxxxxxx | | variable | | | |
| obs4mips_dim_date | date_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | date_value | Date | date | 1999-01-08 | 4713 BC - 5874897 AD | 4 bytes | | | |
| | month | Month | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| obs4mips_dim_dataset | dataset_key | Primary key of the table | | | | | | | |
| | dataset_name | Name of the dataset | character varying(n) | xxxxxxx | | variable | | | |
| | dataset_version | Version of the dataset | | | | | | | |
| | datetime_start | Start time of temporal coverage of the dataset | | | | | | | |
| | datetime_stop | Stop time of temporal coverage of the dataset | | | | | | | |
| obs4mips_dim_file | file_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | file_name | Path and name of the file | character varying(n) | xxxxxxx | | variable | | | |
| | file_size | Size of the file | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |

*Fig. 44 (a): Obs4MIPs DW data dictionary*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| obs4mips_dim_institute | institute_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | institute_name | Institute name | character varying(n) | xxxxxxx | | variable | | | |
| obs4mips_bridge_institute | institute_key | Reference to institute | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | obs4mips_dim_institute |
| | institute_group_key | ID for the institute group | smallint | 99999 | -32768 to +32767 | 2 bytes | true | | |
| obs4mips_dim_variable | variable_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | variable_code | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| | variable_long_name | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| | cf_standard_name | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| obs4mips_bridge_variable | variable_key | Reference to variable | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | obs4mips_dim_variable |
| | variable_group_key | ID for the variable group | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | | |
| obs4mips_dim_time_frequency | time_frequency_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | time_frequency_value | Time frequency value | character varying(n) | xxxxxxx | | variable | | | |
| obs4mips_bridge_time_frequency | time_frequency_key | Reference to time frequency | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | obs4mips_dim_time_frequency |
| | time_frequency_group_key | ID for the time frequency group | smallint | 99999 | -32768 to +32767 | 2 bytes | true | | |
| obs4mips_dim_index_node | index_node_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | index_node_name | Index node name | character varying(n) | xxxxxxx | | variable | true | | |
| obs4mips_dim_processing_level | processing_level_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | processing_level_name | Processing level code | character varying(n) | xxxxxxx | | variable | true | | |
| obs4mips_bridge_processing_level | processing_level_key | Reference to CF standard name | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | obs4mips_dim_processing_level |
| | processing_level_group_key | ID for the CF standard name group | smallint | 99999 | -32768 to +32767 | 2 bytes | true | | |
| obs4mips_dim_source_id | source_id_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | source_id_name | Source identifier | character varying(n) | xxxxxxx | | variable | true | | |
| obs4mips_bridge_source_id | source_id_key | Reference to source | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | obs4mips_dim_source_id |
| | source_id_group_key | ID for the source group | smallint | 99999 | -32768 to +32767 | 2 bytes | true | | |
| obs4mips_dim_realm | realm_key | Primary key of the table | serial | 99999 | 1 to 2147483647 | 2 bytes | true | PK | |
| | realm_name | Realm name | character varying(n) | xxxxxxx | | variable | true | | |
| obs4mips_bridge_realm | realm_key | Reference to realm | integer | 99999 | -2147483648 to +2147483647 | 2 bytes | true | FK | obs4mips_dim_realm |
| | realm_group_key | ID for the realm group | smallint | 99999 | -32768 to +32767 | 2 bytes | true | | |

*Fig. 44 (b): Obs4MIPs DW data dictionary*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| obs4mips_dmart_clients_host_time_geolocation | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | latitude | Geografic coordinate of the user location | numeric(14,11) | 999,99999 | 11 decimal digits precision | 8 bytes | | | |
| | longitude | Geografic coordinate of the user location | numeric(14,11) | 999,99999 | 11 decimal digits precision | 8 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| obs4mips_dmart_variable_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | variable_code | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| | variable_long_name | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| | cf_standard_name | Variable name | character varying(n) | xxxxxxx | | variable | | | |
| obs4mips_dmart_source_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | source_id_name | Source identifier | character varying(n) | xxxxxxx | | variable | true | | |

*Fig. 44 (c): Obs4MIPs DW data dictionary*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| obs4mips_dmart_realm_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | realm_name | Realm name | character varying(n) | xxxxxxx | | variable | true | | |
| obs4mips_dmart_dataset_host_time | dmart_key | Primary key of the table | bigserial | 99999 | 1 to 9223372036854775807 | 8 bytes | true | PK | |
| | total_size | Total size of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_downloads | Total number of downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | number_of_successful_downloads | Total number of successful downloads | bigint | 99999 | -9223372036854775808 to +9223372036854775807 | 8 bytes | | | |
| | average_duration | Average duration of the download operations | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | number_of_users | Total number of distinct users | integer | 99999 | -2147483648 to +2147483647 | 4 bytes | | | |
| | month | Month of the download | smallint | MM | -32768 to +32767 | 2 bytes | | | |
| | year | Year of the download | smallint | YYYY | -32768 to +32767 | 2 bytes | | | |
| | host_name | Host name | character varying(64) | xxxxxxx | | variable | | | |
| | dataset_name | Name of the dataset | character varying(64) | xxxxxxx | | variable | | | |
| | dataset_version | Version of the dataset | smallint | xxx | -32768 to +32767 | 2 bytes | | | |
| | datetime_start | Start time of temporal coverage of the dataset | character varying(64) | xxxxxxx | | variable | | | |
| | datetime_stop | Stop time of temporal coverage of the dataset | character varying(64) | xxxxxxx | | variable | | | |

*Fig. 44 (d): Obs4MIPs DW data dictionary*