CMIP6 Datenmanagement

(AP5 „Nationales CMIP6 Datenarchiv“)

- Status CMIP Datenpool
- Status ESGF Datenknoten

Das DKRZ ESGF Datenmanagement Team:
Stephan Kindermann, Katharina Berger, Hans Dieter Hollweg,
Carsten Ehbrecht, Tobias Weigel

Der DKRZ MIP Datenpool

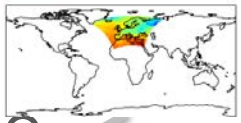
show details about tasmax

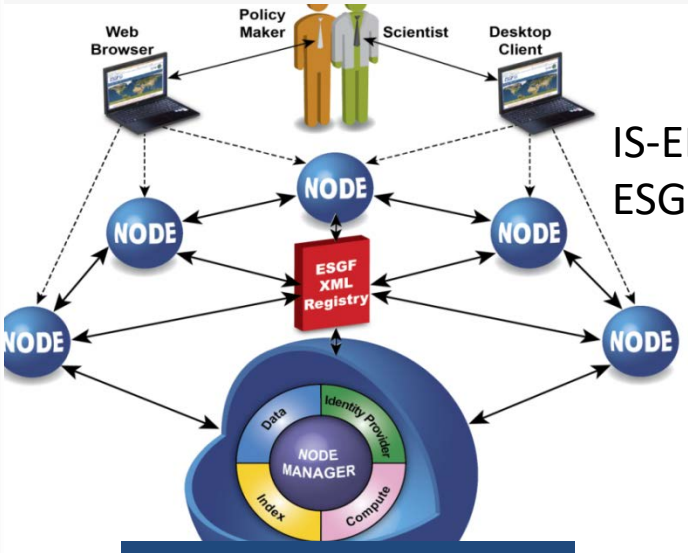
```
In [13]: ds.variables['tasmax']

Out[13]: <type 'netCDF4._netCDF4.Variable'>
         float32 tasmax(time, rlat, rlon)
             standard_name: air_temperature
             long_name: Daily Maximum Near-Surface Air Temperature
             comment: daily-maximum near-surface (usually, 2 meter) air temperature.
             units: K
             cell_methods: time: maximum
             history: 2014-02-01T11:17:01Z altered by CMOR: Treated scalar dimension: 'height'.
             coordinates: height lat lon
             missing_value: 1e+20
             _FillValue: 1e+20
             associated_files: gridspecFile: gridspec_atmos_fx_MPI-CSC-REM02009_historical_r0i0p0.nc
             grid_mapping: rotated_latitude_longitude
             _ChunkSizes: [  1 412 424]
         unlimited dimensions:
         current shape = (20453, 412, 424)
         filling off
```

Create a plot of the first timestep using cartopy

see http://scitools.org.uk/cartopy/docs/latest/matplotlib/advanced_plotting.html?highlight=netcdf

```
In [14]: import matplotlib.pyplot as plt
         from cartopy import config
         import cartopy.crs as ccrs

         timestep = 0

         tasmax = ds.variables['tasmax'][timestep, :, :]
         lats = ds.variables['lat'][:]
         lons = ds.variables['lon'][:]

         ax = plt.axes(projection=ccrs.PlateCarree())
         ax.coastlines()
         ax.set_global()

         fig = plt.contourf(lons, lats, tasmax, 60, transform=ccrs.PlateCarree())
         #plt.show()
```





IS-ENES Datenföderation
ESGF Datenföderation

DKRZ ESGF Knoten

Daten / Compute Schnittstellen

Opendap
ESGF search API
OGC WPS
....

DKRZ MIP Datenpool

CMIP5
CORDEX
CMIP6
Abgeleitete Produkte
Analyse Input

DKRZ Mistral Rechner

Gemountetes Filesystem

DKRZ Langzeitarchiv

(WDC Climate IPCC DDC)

Der DKRZ MIP Datenpool

Dateneintrag in den Datenpool:

- Allgemeine Anfragen, Wünsche etc. rund um Datenpool:

  data-pool@dkrz.de

- Ablieferung von CMIP6 Daten zur Speicherung, ESGF Puplikation und Archivierung:
  - Speziell ESGF Publikation über esgf-publication@dkrz.de
  - Unterstützenden Formulare und https://data-forms.dkrz.de:8080
    - Abfrage der für die Datenverwaltung wichtigen Charakteristika der Datenablieferung (Umfang, Zuordnung, Qualitätsstatus, …)

Der DKRZ MIP Datenpool

CMIP6 Dateneintrag in den Datenpool: Replikation

- Replikation von CMIP6 Daten
  - Voll Automatisierter Teil:
    - Replikation im Rahmen der IS-ENES und ESGF weiten Umsetzung der CMIP6 Replikationsstrategie
    - Replikation der für das ESMValTool benötigten Daten
  - Nutzer getriebener Teil: Anfrage an esgf-replication@dkrz.de

    Spezifikation der Anfragedetails wird wieder unterstützt durch Formulare

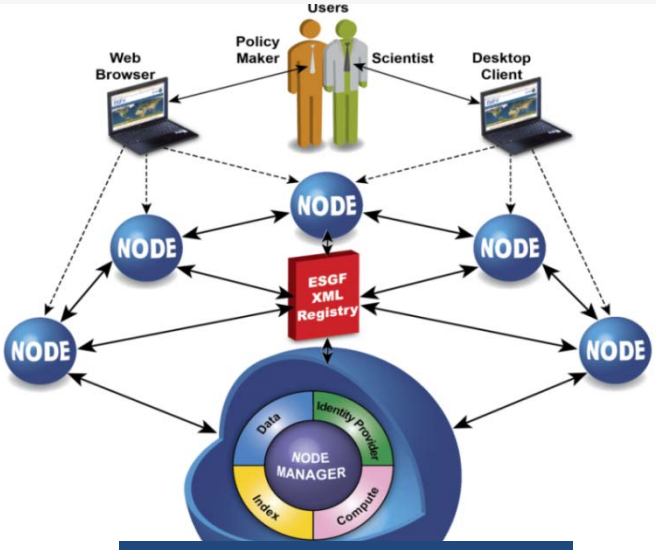
Der DKRZ MIP Datenpool

Status der Planung bzgl. Volumen und Inhalt

- Erste Planungsgrundlage bietet Sammlung und Auswertung der am häufigsten angefragten Variablen im Kontext von CMIP5 (Dokumente siehe redmine)
  - ✓ Liste der häufigsten benutzten CMIP5 Variablen (WDCC, ESGF)
  - ✓ ESMValTool Variablenliste
  - ✓ Diskussion in der ENES Data Task Force (April 2017) – WCRP WIP in Diskussion einbezogen
  - ✓ Diskussion mit IPCC WGs, Workshop am DKRZ mit WG1 (Sept. 2017)
  - → Aktuelle Volumenschätzung bleibt unverändert: „Core der am häufigsten angefragten Variablen dürfte ca 2 Pbyte umfassen“

- Sammlung der deutschen CMIP6 Beiträge: bisher noch sehr unvollständige Auskünfte der Modellierungsgruppen – das Bild wird erst klarer wenn die ersten Läufe gestartet sind

ESGF und der DKRZ MIP Datenpool



IS-ENES Datenföderation
ESGF Datenföderation
Copernicus
....

DKRZ ESGF Knoten

DKRZ MIP Datenpool

- CMIP5
- CORDEX
- CMIP6
- **Abgeleitete Produkte**
- **Analyse Input**

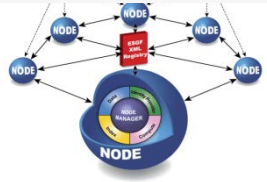
DKRZ Langzeitarchiv

(WDC Climate IPCC DDC)

ESGF und der DKRZ MIP Datenpool

esgf-data.dkrz.de

DKRZ ESGF portal



Copernicus

cp4cds-index1.dkrz.de

DKRZ ESGF PID services

Handle.Net®

DKRZ ESGF PID services

Handle.Net®

esgf1.dkrz.de

esgf3.dkrz.de

esgf2.dkrz.de

external ESGF data nodes publishing to DKRZ ESGF portal:

- DLR (eridanus.eoc.dlr.de)
- DWD (planed)
- Others .. (CMCC, DMI, ICHEC, KNMI)

DKRZ MIP Datenpool

CMIP5

CORDEX

CMIP6

Abgeleitete Produkte

Analyse Input

DKRZ ESGF data nodes:

- esgf1: CMIP5, CORDEX, ..
- esgf2: WDCC (CMIP5)
- esgf3: CMIP6 (new)

DKRZ LZA (WDC Climate IPCC DDC)

CMIP Datenpool: ESGF Replikation

Ziel: Weitgehend automatisierte und zeitnahe lokale Bereitstellung von internationalen CMIP6 Daten als Teil des CMIP Datenpools.

Workflow:

(1) Angabe der zu replizierenden Datenkollektionen (synda selection files)
(2) Replikation (http, gridftp) mithilfe von synda
(3) Generierung von „Publikations-Listen“ (map-files)
(4) ESGF Publikation basierend auf den map-files

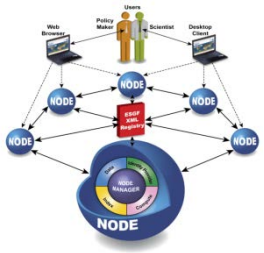
Status: Dieser workflow ist vollständig automatisiert und wurde in ESGF Test-Föderation getestet

Automatisierte Replikations-Pipeline

Web Formulare, oder
esgf-replication@dkrz.de

synda selection files

```
project=CMIP5
model=CNRM-CM5 CSIRO-Mk3-6-0
experiment=historical amip
ensemble=r1i1p1
variable[atmos][mon]=tasmin tas psl
variable[ocean][fx]=areacello sftof
variable[land][mon]=mrsos nppRoot nep
```



Daten Transfer Knoten
- Synda tool

http, gridftp

esgf-data.dkrz.de

DKRZ ESGF portal

esgf1.dkrz.de

esgf3.dkrz.de

DKRZ MIP Datenpool

CMIP5

CORDEX

CMIP6

Publikations Knoten
- Synda pp Publikations-script

Status ESGF Datenknoten

- Neuer CMIP6 Datenknoten installiert und operationell (esgf3.dkrz.de)
- CMIP6 PID Dienste operationell (message queue, PID server, API)
  - In letztes ESGF release (2.5) integriert
- Testinfrastruktur verfügbar (Test-Datenknoten + Test-Portal, integriert in ESGF Test-Föderation, Replikations-Tests mit internationalen Partnern)
- Erster Datentransferknoten (zum Daten-ingest) operationell, wird Ende 2017/Anfang 2018 um weitere Knoten erweitert
  - Jan – August 2017 Replikations-Tests in ESGF ICNWG working group
  - Seit Sept 2017 Start der Optimierung der Produktions-Netzinfrastruktur (lokale Netzkonfiguration, Netzanbindung, WAN Anbindung ..)

„Datenpool nahes“ Prozessieren

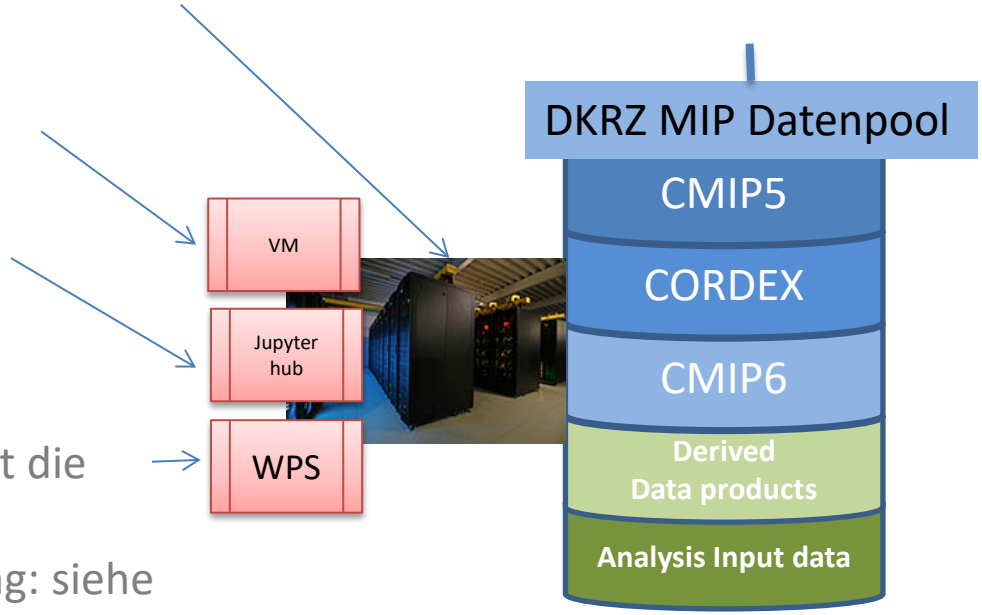
Unterstützung verschiedener Optionen

A) Nutzer login in „mistral“

B) Nutzer „mietet“ eine VM

C) Nutzung von „jupyterhub.dkrz.de“

D) Bereitstellung von Prozessierungsdiensten
- see z.B. mouflon.dkrz.de
- Das birdhouse framework[1)] unterstützt die Bereitstellung von WPS konformen Prozessierungs Web Services .. (hosting: siehe Option B oder im Institut oder in der cloud …)



Optionen C) and D) sind aktuell in Entwicklungs / Ramp up Phase

[1)] http://bird-house.github.io/

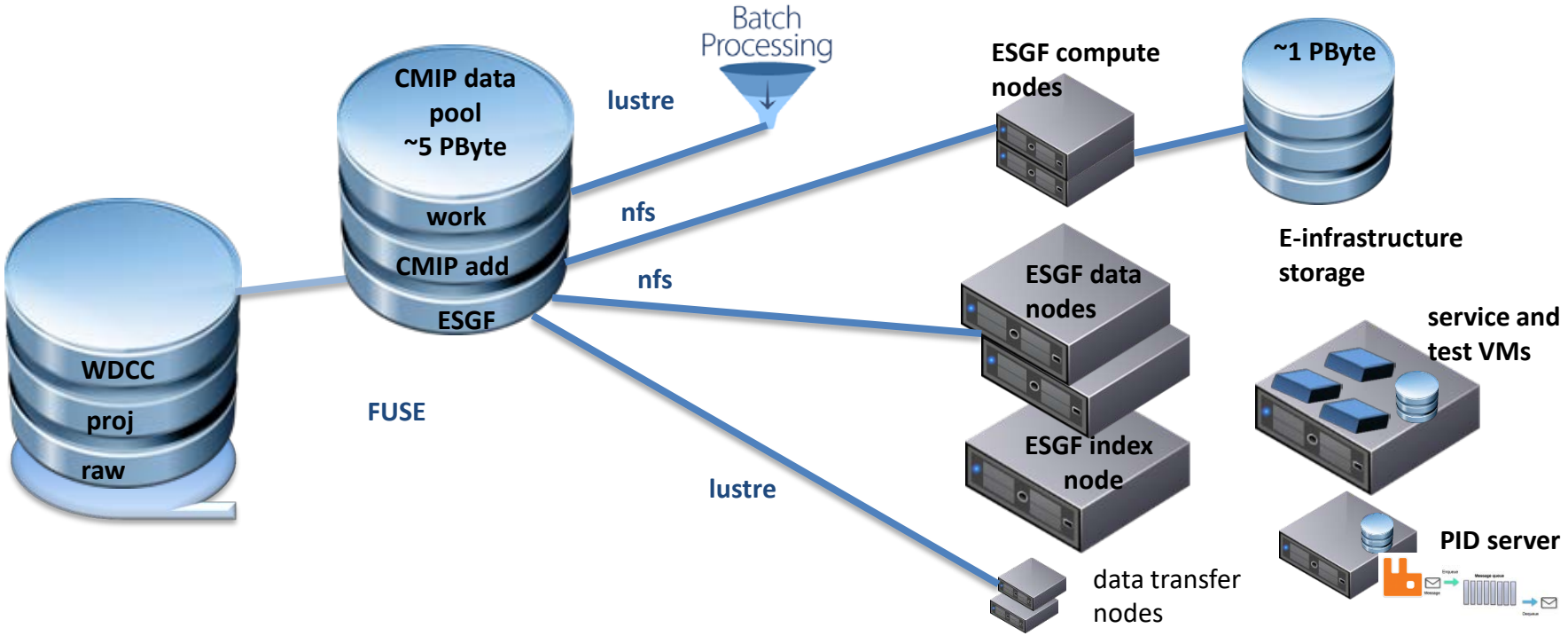
The technical ESGF infrastructure at DKRZ

HPSS tape
- 190 Pbyte capacity

Lustre file system
- 54 PByte

„Mistral“ HPC
- 3.6 Pflops
- ~100.000 cores

Compute/ storage cluster
- VM servers, database servers
- Openstack cloud storage



Diskussion

- ....