ECERFACS

CENTRE EUROPÉEN DE RECHERCHE ET DE FORMATION AVANCÉE EN CALCUL SCIENTIFIQUE.

Les nouveaux défis de l'analyse de données climatiques

État actuel et perspectives

Christian Pagé

Ingénieur de recherche, gestionnaire de projets

Xavier Pivan Ingénieur de développement Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS)

> Forum TERATEC 2017 27-28 juin 2017

Ecole Polytechnique, Palaiseau, France

www.cerfacs.fr

CERFACS/Toulouse: *High Performance Computing Research Centre*

- Develop scientific and technical researches in order to improve advanced computing methods
- Access to computers with new architecture
- Transfer this scientific knowledge and technical methods for application to big industrial sectors
- Train high qualified people

 2015: The tenth computer set at CERFACS since 1996 occupies the 388° place in the top 500 delivering a peak power of 242 Tflop/s



Forum TERATEC 2017, 27-28 juin 2017, Ecole Polytechnique, Palaiseau, France

Outline

- Motivations: Scientific, Technical, Societal
- Current Situation and Issues: we have to improve
- "Standard" Solutions
- Background
- Some Solutions
 - Building Blocks
 - Putting it all together
- Big Data Technologies and Analytics
- Summary and Perspectives



Scientific

Research data lifecycle

- Perform efficient Data Analysis
 - Large number of realizations (ensemble of scenarios)
 - Uncertainties range estimation
 - Process Higher spatial and temporal resolution
 - Easily share intermediate results with collaborators
- Achieve a more robust and flexible Data Life Cycle
 - More robust experiments setup
 - Explore several experiment configurations to answer scientific questions
 - Reproducible experiments



Scientific

- Every development work in climate modelling comprises comparison of realizations
 - I introduced this small change....
 - What happened to my model?
 - Does it work?
 - Does it work in the expected way?
 - Are there consequences I did not expect?





Technical

- Process large data volumes, ideally near(er) the data storage
 - Data Analytics
 - Data Life Cycle
- Streamline the data processing workflow
- Proper metadata description of the data objects
- Properly track provenance information
- Interconnect e-infrastructures and research infrastructures services, interfaces & platforms



Societal

- Provide climate projections data to climate change impact researchers, facilitators, practitioners
 - Ease access with better intuitive interfaces
 - Provide more common data formats
 - Generate tailored products from data processing workflows









Lars Bärring, SMHI Rossby Centre, Circle-2 Conference on European Climate Change Adaptation Research and Practice, Lisbon, 10-12 March 2014



Climate Research Community Data available for scientific analysis: a very large trend

 Limitations in data access means limitations in data analytics and scientific results

Download locally then Analyze: a workflow that

cannot be sustained _N

- Climate researchers
- Impact researchers





Practical Example: Climate Community



- Temperature at 850 hPa field (Aggregated files 30 levels)
- 10 climate models
- 1960-1990 & 2040-2070 = 60 years = 21 915 days
- Daily fields = 1 field per day
- Global spatial scale 100 km resolution

Data reduction...

TOTAL: 6 754 500 fields to download ~100 Kb per 2D field = 626 Gb

After the analysis post-processing

- Anomaly of the average of the two periods over a specific country for each climate model
- **Result**: 10 times 2D fields over a small domain
 - Estimated datasize after post-processing: 1 Mb



ESGF Data Nodes 2015:

- 40 worldwide
- 18 in Europe (coordinated in IS-ENES)

IS-ENES ESGF Portals IS-ENES climate4impact Portal

- BADC (UK)
- DKRZ (Germany)
- IPSL (France)
- SMHI (Sweden)
- CMCC (Italy)
- DMI (Denmark)

- KNMI (Netherlands)
- Interlinked with Uni. Cantabria downscaling portal (Spain)

CLIPC Portal

 Climate Information Portal for Copernicus

Ack: Michael Lautenschlager, DKRZ

ECERFACS

Downloaded data volumes – European ESGF data nodes





Status CMIP5 data archive:

1.8 PB for 59000 data sets stored in 4.3 Mio Files in 23 ESGF data nodes CMIP5 data is about 50 times CMIP3

Extrapolation to CMIP6:

CMIP6 has a more complex experiment structure than CMIP5. Expectations: more models, finer spatial resolution and larger ensembles Factor of 20: 36 PB in 86 Mio Files Factor of 50: 90 PB in 215 Mio Files



"Standard" Solutions

IPSL e-infrastructure to sustain climate analysis





"Standard" Solutions



Fig. 13. CEDA's JASMIN analysis platform. JASMIN integrates cloud architecture, container technologies, and virtual machines to improve flexibility and performance and track maintenance.



"Standard" Solutions

Snow on/off – Length of snow season

Snow off



Background

ESGF Future Computing Nodes: API

• Goal: perform data analysis near the data storage

- Better data access
- Move away from the download/analyze workflow





Background: EUDAT





Background: EGI





Some Solutions: Building Blocks



- Federation of Peer-to-Peer Data Nodes
- Computing Nodes API (Using ISO-OGC WPS)
- OpenID Authentication/Authorization

EUDAT

- API for deploying calculations (workflows)
- B2 Services for orchestration and storage

IS-ENES

Data Analytics Services => climate4impact.eu platform

Solutions: Putting it all together

Bridge EUDAT / EGI / ESGF / IS-ENES

- EUDAT Workflow API (GEF) ⇔
 - ESGF Computing API WPS
 - EGI Federated Cloud
 - IS-ENES Data Analytics Services => climate4impact.eu platform
- Challenge: Common Authentication and Authorization



Solutions: Putting it all together





Big Data?

What about Big Data Technologies and Analytics??

Big Data Landscape 2016 (Version 3.0)									
Infrastructure Analytics				Applications					
Hadoop On-Premise Cloudera Pivotal IBM InfoSphere B-blavdata jethro	Spark Liatabricks dGain/4 CENTON EXTURN CONTRACTOR Statuto Statuto Statuto	Analyst Platforms Palatorns Palatorns Quid engma Retworks Context retwart Quid engma Retworks Context retwart Quid engma Retworks Context retwart Quid engma Botteriose interions Analytics Data Science Platforms Context retwart Alpine Botteriose Science retwart Alpine Data Science Platforms Context retwart Alpine Science Science Science retwart Alpine Science Science Science retwart Alpine Science Science Science retwart Alpine Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science Science	Visualization bet Oracide Construction Olik © Loöker Broambi Baseves Screate Chartio	Sales & Marketing RADIUS' Gainsight' () bloomrach Zetäi blueyonder 2 Lattice blueyonder 2 Lattice Status Antone Cuck Fox persäde Autor Actionice Command Autor	Legal Capital Gild nonnectifier Cextice entelo				
NoSQL Databases Cocode Coud Patform CRACLE Microsoft Azure MarkLogic mongoDB Detrastrix Cocode Coud Patform MarkLogic mongoDB Detrastrix Cocode Coud Patform CRACLE mongoDB Detrastrix Cocode Coud Patform CRACLE Microsoft Azure MarkLogic Coude Coud Patform Cocode Coud Patform CRACLE Microsoft Azure MarkLogic Coude Coud Patform Cocode Coud Cocode Coud Patform Cocode Coud Patform C	NewSQL Databases Clustrix Pivotol. Paradigm4 MoricoD8 Post Trafetien © Cockroach uss	BI Platforms Dever BI Computing SSSS Computing SSSS District Platform Catcale Computing SSSS Computing SSSS Market Market Market Computing SSSS Market Market Computing SSSS Market Market Computing SSSS Market Computing SSSS Market Computing SSSS Market Computing SSSS Market Computing SSSS Market Computing SSSS Market Computing SSSS Market Computing SSSS Market Computing SSSS Market Computing SSSS Computing Computing SSSS Computing Computing Computing Computing Computing Computing SSSS Computing	Social Analytics Hootsuite NETBASE ATASIFT track bitly synthesio simplereach	Ad Optimization AdpoNexus CriteOL Phone Station CriteOL Methodebek Metho	rdian nce				
Graph Databases Databases Decomposition Construction Cons	Data Transformation alteryx talend tamerate talend tamerate talend tamerate talend tamerate talend tamerate talend tamerate	Real-Time With the second sec	Horizontal Al Contana e ventent Contana e ventent Merida al anti- Merida al anti- misponel Sumal al anti- sumal al anti- sum	Data 20 Oppier MOAT (Produces feedoal esco Publisher Tools Outbrain Taboela guancast Chartbeat Steamin Steamin Steamin Clearning KINSHCO Clever Geclara Production Clever Sector	IIII Cumiata IIII Cumiata IIII Cumiata IIII Cumiata IIII Cumiata IIII Cumiata IIII Cumiata IIII Cumiata IIIII Cumiata IIIII Cumiata IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII				
			ACLE NetApp	Open Source					
Framework Windows YARN Windows Figure CODPD Framework Windows Figure CodPD Figure CodPD Figure CodPD Figure CodPD Figure CodPC Figure CodPC Figu									
Data Sources & APIs Incubators & Schools									
Health JAWBONE GARMIN Withings Value relation Withings Value relation Withings All All All All All All All All All Al	Financial & Bloomber ThingWerx helium is ansora	LCONOMIC Data TOTAS DTERS DEPENSE CARLON SPACE CARLON C	Crimson Hexagon	Copie / Entudes Construction Stratecture STREETLURE © Construction Facture Facture RASIS © Series Construction Constructi	PLURALSIGHT INSIGHT DataElite he Data Incubator METIS				
Last Updated 3/23/2016	© Matt Turck (@	mattturck), Jim Hao (@jimrhao), & First	Mark Capital (@fi	rstmarkcap) FIRST	MARK				



Big Data: Hadoop and Climate Data @NASA



Northward wind component



Air temperature



Specific humidity



Wei, et al.

- ~8.4 TB transferred from archive to local workstation (weeks)
- Clipping, averaging performed by Fortran program on local workstation (days)

MERRA/AS

- Clipping, averaging performed by MERRA/AS (~28 hrs)
- Only ~35 GB final product transferred to local workstation (minutes)
- Significant time savings in data wrangling,
- rapid screening over monthly means files takes minutes, and
- there's a possibility of folding Dr. Wei's modeling algorithm back into the CDS API ...

Applying Apache Hadoop to NASA's Big Climate Data: Glenn Tamkin, John Schnase, Dan Duffy, Hoot Thompson, Denis Nadeau, Scott Sinno, Savannah Strong,



Big Data: Hadoop and Climate Data @NASA

The original MapReduce application utilized standard Hadoop Sequence Files. Later they were modified to support three different formats called Sequence, Map, and Bloom.

Dramatic performance increases were observed with the addition of the Bloom filter (~30-80%).

Read a single parameter ("T") from a single sequenced monthly means file	Standalone VM	6.1	1.2	1.1	+81.9%
Single MR job across 4 months of data seeking "T" (period = 2)	Standalone VM	204	67	36	+82.3%
Generate sequence file from a single MM file	Standalone VM	39	41	51	-30.7%
Single MR job across 4 months of data seeking "T" (period = 2)	Cluster	31	46	22	+29.0%
Single MR job across 12 months of data seeking "T" (period = 3)	Cluster	49	59	36	+26.5%

Applying Apache Hadoop to NASA's Big Climate Data: Glenn Tamkin, John Schnase, Dan Duffy, Hoot Thompson, Denis Nadeau, Scott Sinno, Savannah Strong,



Big Data: Spark & Hadoop / CDAS @NASA



Climate Data Services Framework (CDAS). Thomas Maxwell and Dan Duffy. NASA.



Big Data: Spark & Hadoop / CDAS @NASA



Climate Data Services Framework (CDAS). Thomas Maxwell and Dan Duffy. NASA.



Big Data: Spark & Hadoop / CDAS @NASA



Climate Data Services Framework (CDAS). Thomas Maxwell and Dan Duffy. NASA.



Big Data Analytics on Climate Data





Weather fronts (left) and Tropical Cyclones (right) as detected by a convolutional neural network.

Liu et al. KDD 2016 August 13-17, San Francisco, CA, USA



Summary and Perspectives

• Infrastructure to access relevant climate data

- Global & Regional Climate Model projections
- Observations: satellite, reanalyses, surface

Community Services with standard interfaces

- On-demand downscaling services
- On-demand product calculations
 - » Climate indices and indicators
 - » Bias-correction
- On-demand calculations
 - » Reduce datasize to be transferred over the network
 - » Ease access to calculations to end users
- Support to heterogeneous users

Bridge e-infrastructure to research infrastructures

- Provide Data Analytics & Tailoring to users
- Enhance and Ease Data Sharing & Discovery
- Provide support for "Long Tail of Science" (LToS)

• Big Data Techniques

- Data Mining for Geophysical Data
- Neural Networks, Hadoop/Spark





Questions! 😳



