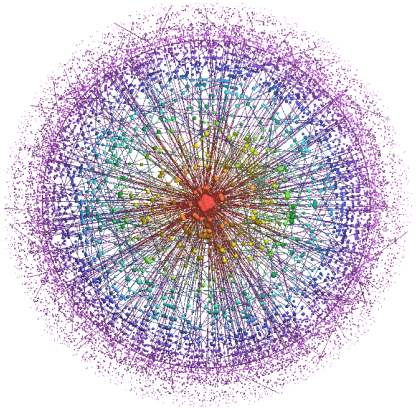
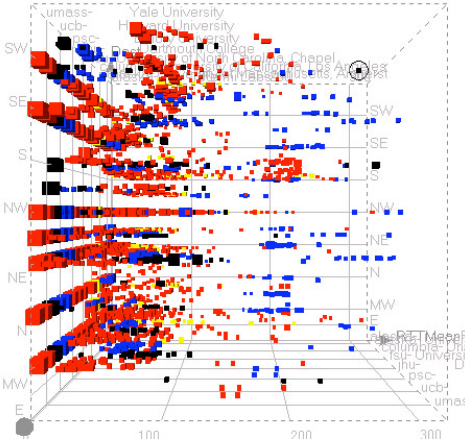
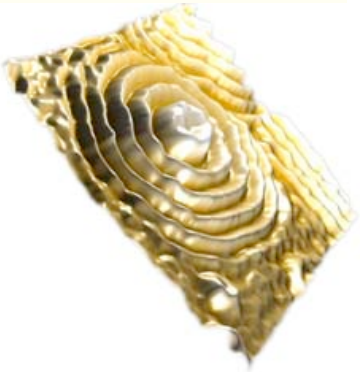
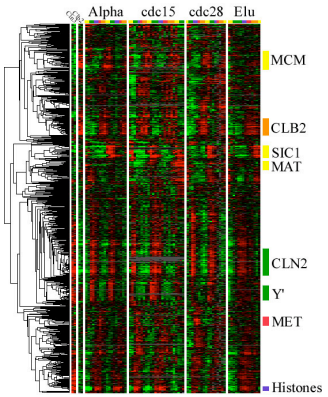




Scientific Data Formats

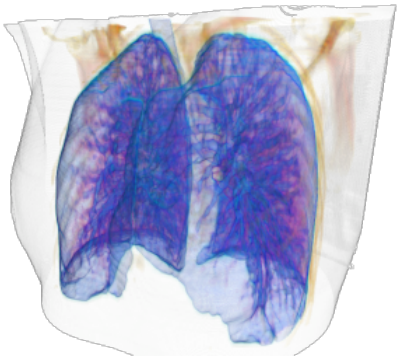




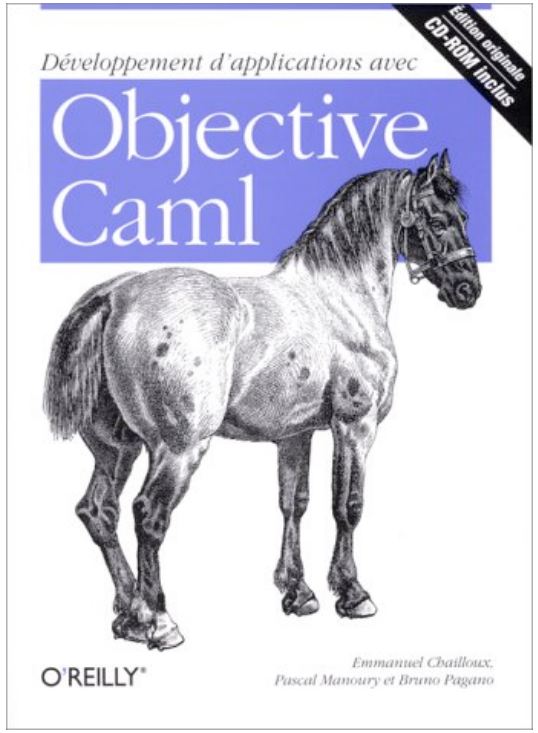
Meelis Kull

University of Tartu

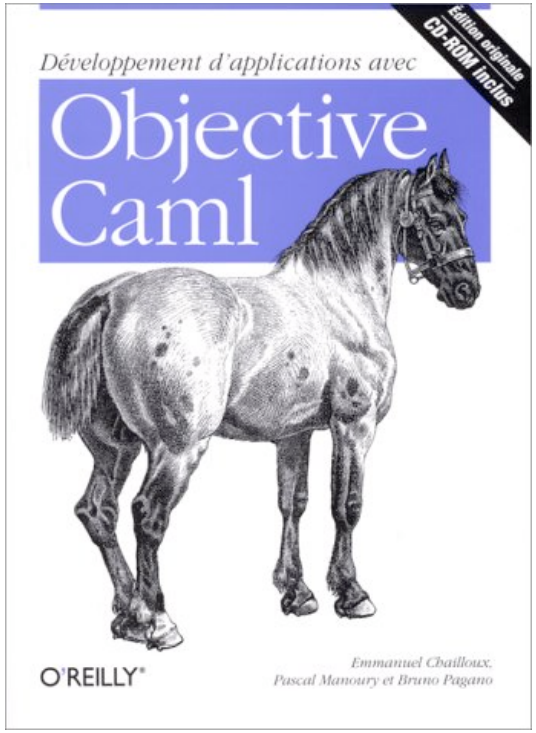
BIO IT group



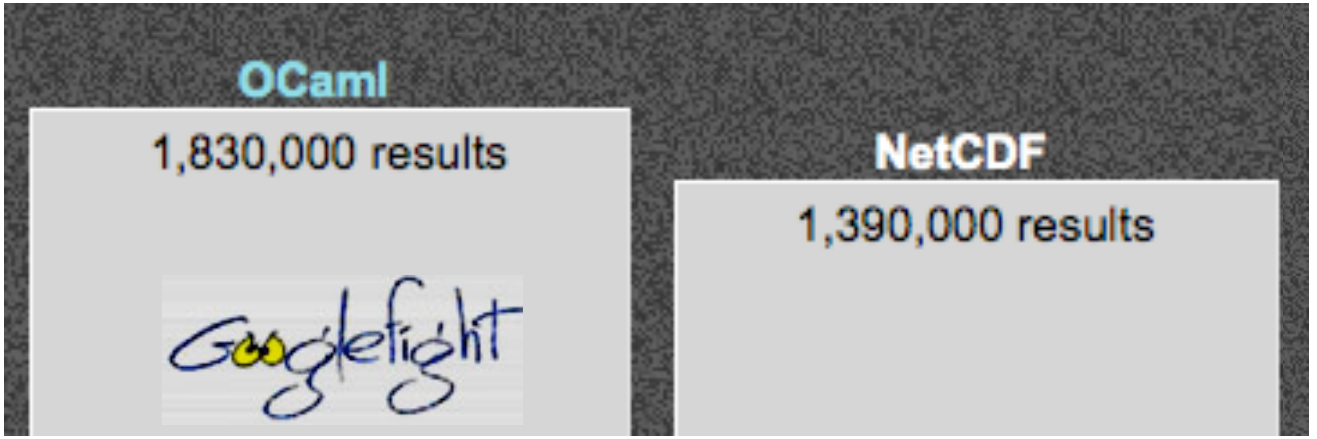
netCDF4 HDF5


Développement d'applications avec
Objective Caml
Édition originale
CD-ROM inclus
O'REILLY
Emmanuel Chailloux,
Pascal Manoury et Bruno Pagano

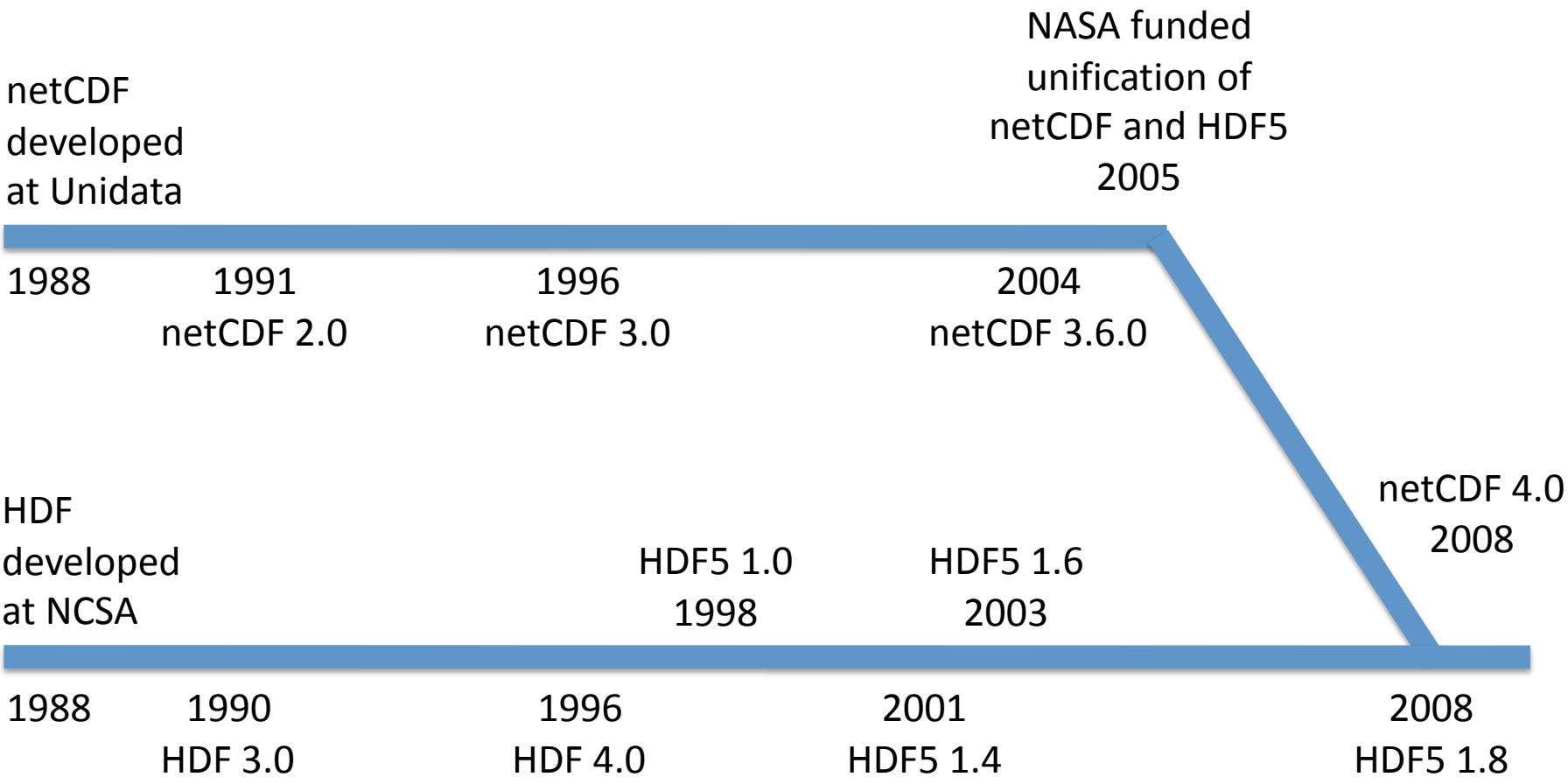
OCaml


Développement d'applications avec
Objective Caml
Edition originale
CD-ROM inclus
O'REILLY
Emmanuel Chailloux,
Pascal Manoury et Bruno Pagano


netCDF


OCaml
1,830,000 results
NetCDF
1,390,000 results
Googlefight

History of netCDF and HDF



Outline

- Scientific data
- Requirements
- NetCDF
- TabCDF
- HDF5
- Future

What is scientific data?

N-dimensional arrays + metadata:

- Measurements at specific time, location, condition
  - Physics: temperature, pressure
  - Chemistry: reaction speed
  - Biology: type (species, cell types, nucleotides)
  - Economics: price
  - Algorithmics: program time and space
  - Networking: network activity
  - Robotics: movements

Example dataset

Monthly average air temperatures (°C) in some capitals of the world

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Tallinn	−3	−5	−1	3	10	13	16	15	10	6	1	−2
Beijing	−3	0	6	13	20	24	26	25	20	13	5	−1
Berlin	0	−1	4	7	12	16	18	17	14	9	4	1
Buenos Aires	23	22	20	16	13	10	10	11	13	16	18	22
Cairo	13	15	17	21	25	27	28	27	26	23	19	15
Canberra	20	20	17	13	9	6	5	7	9	12	15	18
Cape Town	21	21	20	17	15	13	12	13	14	16	18	20
Helsinki	−5	−6	−2	3	10	13	16	15	10	5	0	−3
London	3	3	6	7	11	14	16	16	13	10	6	5
Moscow	−8	−7	−2	5	12	15	17	15	10	3	−2	−6
Ottawa	−10	−8	−2	6	13	18	21	20	14	7	1	−7
Paris	3	4	7	10	13	16	19	19	16	11	6	5
Riga	−3	−3	1	5	11	15	17	16	12	7	2	−1
Rome	8	8	11	12	17	20	23	23	21	17	12	9
Singapore	27	27	28	28	28	28	28	28	27	27	27	26
Stockholm	−2	−3	0	3	10	14	17	16	11	6	1	−2
Washington D.C.	2	3	7	13	18	23	26	25	21	15	9	3

Requirements

- Compact storage - compression
- Fast I/O – parallel, partial, random access
- Mobility – transporting data between computers
- Tools for manipulating data – reorganizing, aggregating, subsetting, converting, visualizing
- Easy API in many languages – C, C++, Fortran, Java, Matlab, Perl, Python, R, …

Example task

Monthly average air temperatures (°C) in some capitals of the world

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Tallinn	−3	−5	−1	3	10	13	16	15	10	6	1	−2
Beijing	−3	0	6	13	20	24	26	25	20	13	5	−1
Berlin	0	−1	4	7	12	16	18	17	14	9	4	1
Buenos Aires	23	22	20	16	13	10	10	11	13	16	18	22
Cairo	13	15	17	21	25	27	28	27	26	23	19	15
Canberra	20	20	17	13	9	6	5	7	9	12	15	18
Cape Town	21	21	20	17	15	13	12	13	14	16	18	20
Helsinki	−5	−6	−2	3	10	13	16	15	10	5	0	−3
London	3	3	6	7	11	14	16	16	13	10	6	5
Moscow	−8	−7	−2	5	12	15	17	15	10	3	−2	−6
Ottawa	−10	−8	−2	6	13	18	21	20	14	7	1	−7
Paris	3	4	7	10	13	16	19	19	16	11	6	5
Riga	−3	−3	1	5	11	15	17	16	12	7	2	−1
Rome	8	8	11	12	17	20	23	23	21	17	12	9
Singapore	27	27	28	28	28	28	28	28	27	27	27	26
Stockholm	−2	−3	0	3	10	14	17	16	11	6	1	−2
Washington D.C.	2	3	7	13	18	23	26	25	21	15	9	3

Extract the temperatures for summer months for 10 first capitals

Solution 1: Text file database

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Tallinn	-3	-5	-1	3	10	13	16	15	10	6	1	-2
Beijing	-3	0	6	13	20	24	26	25	20	13	5	-1
Berlin	0	-1	4	7	12	16	18	17	14	9	4	1
Buenos Aires	23	22	20	16	13	10	10	11	13	16	18	22
Cairo	13	15	17	21	25	27	28	27	26	23	19	15
Canberra	20	20	17	13	9	6	5	7	9	12	15	18
Cape Town	21	21	20	17	15	13	12	13	14	16	18	20
Helsinki	-5	-6	-2	3	10	13	16	15	10	5	0	-3
London	3	3	6	7	11	14	16	16	13	10	6	5
Moscow	-8	-7	-2	5	12	15	17	15	10	3	-2	-6
Ottawa	-10	-8	-2	6	13	18	21	20	14	7	1	-7
Paris	3	4	7	10	13	16	19	19	16	11	6	5
Riga	-3	-3	1	5	11	15	17	16	12	7	2	-1
Rome	8	8	11	12	17	20	23	23	21	17	12	9
Singapore	27	27	28	28	28	28	28	28	27	27	27	26
Stockholm	-2	-3	0	3	10	14	17	16	11	6	1	-2
Washington D.C.	2	3	7	13	18	23	26	25	21	15	9	3

```
> head -11 data.txt | tail -10 | cut -f7-9
```

Solution 2: Relational database system

Temperature

Capital ID	Month ID	Temp °C
0	0	-3
0	1	-5
…	…	…
0	11	-2
…	…	…
16	11	3

Capital

ID	Name
0	Tallinn
1	Beijing
…	…
16	Washington D.C.

Month

ID	Name
0	Jan
1	Feb
…	…
11	Dec

```
SELECT temp FROM Temperature
    WHERE (capital_id<10) and (month_id>=5) and (month_id<8)
    ORDER BY capital_id, month_id;
```

Solution 3: NetCDF

- NetCDF (Network Common Data Form) – binary array-oriented data file format
- “Variable” – a multi-dimensional array of data, of any of 6 types (char, byte, short, int, float, or double).
- “Dimension” – information about an axis: it’s name and length.

Solution 3: NetCDF

```
netcdf temperatures {
dimensions:
    months = 12;
    capitals = 17;

variables:
    float temperature(capitals, months);

data:
    temperature =
    -3,-5,-1, 3,10,13,16,15,10, 6, 1,-2,
    -3, 0, 6,13,20,24,26,25,20,13, 5,-1,
    . . .
     2, 3, 7,13,18,23,26,25,21,15, 9, 3;
}
```

```
> ncks input.nc –d months,5,7 –d capitals,0,9 output.nc
```

Solution 3: NetCDF

Language	Implementation

C++	`NcFile nc("temperatures.nc");`
	`NcVar *v = nc.get_var("temperature");`
	`v->set_cur(0,5);`
	`float *f = new float[10*3];`
	`v->get(f,10,3);`

Java	`Netcdf nc = new NetcdfFile("temperatures.nc");`
	`Variable v = nc.get("temperature");`
	`MultiArray ma = v.copyout(new int[] {0,5},new int[] {10,3});`
	`float [] f = (float [])ma.toArray();`

Perl	`$nc = PDL::NetCDF->new("temperatures.nc");` `$f = $nc->get("temperature",[0,5],[10,3]);`
R	`nc = open.ncdf("temperatures.nc");` `f = get.var.ncdf(nc,"temperature",start=c(0,5),count=c(10,3));`

Solution 3: NetCDF

```
netcdf temperatures {
dimensions:
    months = 12;
    capitals = 17;
    month_strlen = 4;
    capital_strlen = 16;
variables:
    float temperature(capitals, months);
    char month_name(months, month_strlen);
    char capital_name(capitals, capitals_strlen);
data:
    month_name =
      J, a, n,\0,
      . . .
      D, e, c,\0;
    capital_name =
      T, a, l, l, i, n, n,\0,\0,\0,\0,\0,\0,\0,\0,\0,
      B, e, i, j, i, n, g,\0,\0,\0,\0,\0,\0,\0,\0,\0,
      . . .
      W, a, s, h, i, n, g, t, o, n,  , D, ., C, .,\0;
    temperature =
      . . .
```

How to generate a NetCDF file?

- Write the textual version and apply ncgen
- Generate from a program using NetCDF API
- Use TabCDF (http://biit.cs.ut.ee/tabcdf)

TabCDF

- Converts from tabular text files to NetCDF and back using a simple data layout description language
- Authors: Tambet Arak, Meelis Kull
- Example script:

```
{
  string corner;
  string month_name[months];
}
capitals * {
  string capital_name[];
  float temperature[][months];
}
```

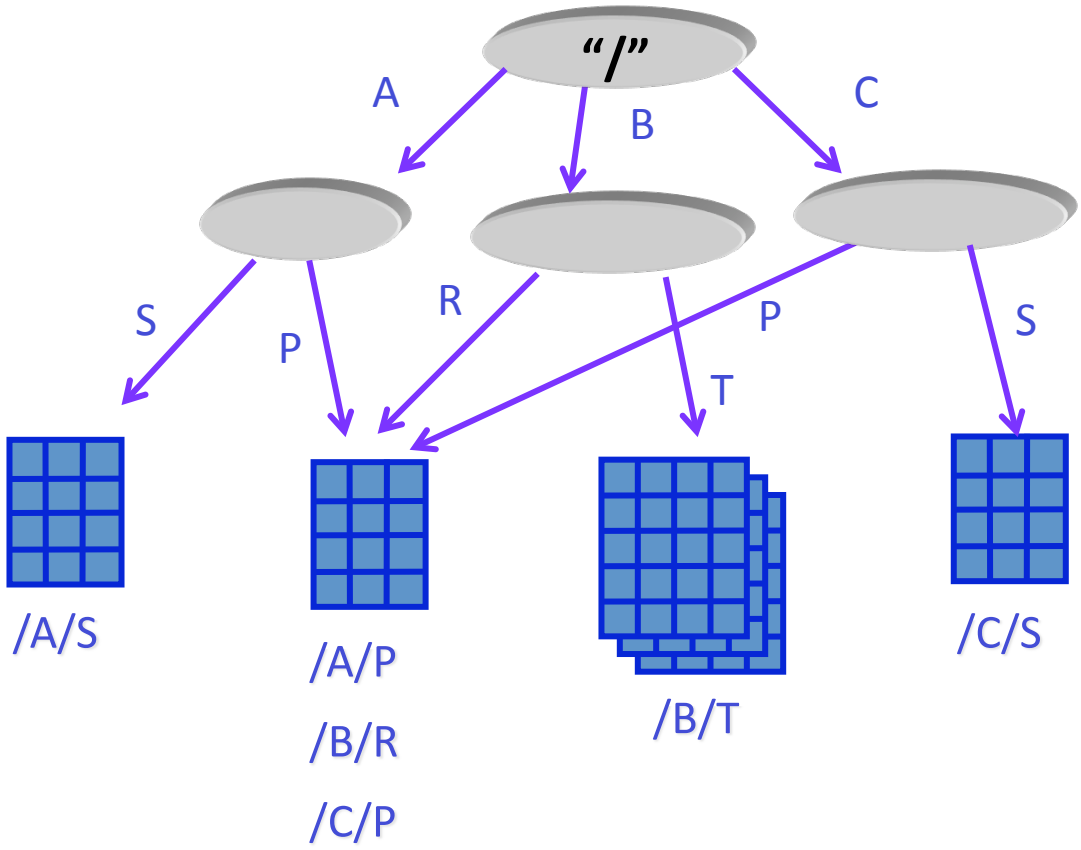
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Tallinn	-3	-5	-1	3	10	13	16	15	10	6	1	-2
Beijing	-3	0	6	13	20	24	26	25	20	13	5	-1
Berlin	0	-1	4	7	12	16	18	17	14	9	4	1
Buenos Aires	23	22	20	16	13	10	10	11	13	16	18	22
Cairo	13	15	17	21	25	27	28	27	26	23	19	15
Canberra	20	20	17	13	9	6	5	7	9	12	15	18
Cape Town	21	21	20	17	15	13	12	13	14	16	18	20
Helsinki	-5	-6	-2	3	10	13	16	15	10	5	0	-3
London	3	3	6	7	11	14	16	16	13	10	6	5
Moscow	-8	-7	-2	5	12	15	17	15	10	3	-2	-6
Ottawa	-10	-8	-2	6	13	18	21	20	14	7	1	-7
Paris	3	4	7	10	13	16	19	19	16	11	6	5
Riga	-3	-3	1	5	11	15	17	16	12	7	2	-1
Rome	8	8	11	12	17	20	23	23	21	17	12	9
Singapore	27	27	28	28	28	28	28	28	27	27	27	26
Stockholm	-2	-3	0	3	10	14	17	16	11	6	1	-2
Washington D.C.	2	3	7	13	18	23	26	25	21	15	9	3

HDF5 – Hierarchical Data Format

HDF5 adds to NetCDF Classic the support of:

- Groups for organizing variables (hierarchy)
- More data types, including compound types
- Variable length arrays
- Unlimited data sizes
- Chunking
- MPI
- and more…

HDF5 hierarchy



http://www.hdfeos.org/workshops/ws12/presentations/day1/bxj.ppt

The HDF Group

R&D 100

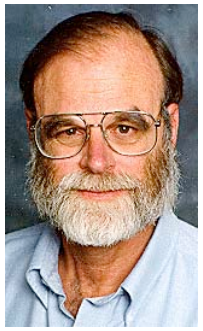
2002

"If you find yourself designing a Star schema to fit your data into SQL, then you might want to investigate HDF5 as a simpler, faster alternative storage mechanism"

http://en.wikipedia.org/wiki/Hierarchical_Data_Format

SIGMOD Record, Vol. 34, No. 4, Dec. 2005

Scientific Data Management in the Coming Decade



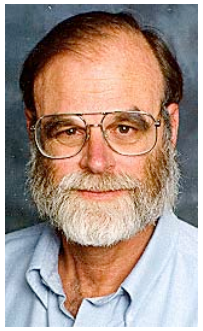
Jim Gray, Microsoft
David T. Liu, Berkeley
Maria Nieto-Santisteban & Alex Szalay, Johns Hopkins University
David J. DeWitt, Wisconsin
Gerd Heber, Cornell

Jim Gray (1944-2007) - Turing Award 1998

- "While the commercial world has standardized on the relational data model and SQL, no single standard or tool has critical mass in the scientific community."
- "In the next decade, as data interchange among scientific disciplines becomes increasingly important, a common HDF-like format and package for all the sciences will likely emerge."

SIGMOD Record, Vol. 34, No. 4, Dec. 2005

Scientific Data Management in the Coming Decade



Jim Gray, Microsoft
David T. Liu, Berkeley
Maria Nieto-Santisteban & Alex Szalay, Johns Hopkins University
David J. DeWitt, Wisconsin
Gerd Heber, Cornell

Jim Gray (1944-2007) - Turing Award 1998

- "Database systems have not traditionally supported science's core data type: the N-dimensional array."
- "… we expect HDF and other file formats to be added as types to most database systems."
- "We believe this database, file system, and programming language integration will be the key to managing and accessing peta-scale data management systems in the future."

Questions?