Surviving the Deluge (Part II): Planning the ARM "Big Data" System for Infrastructure and Research Processing of Terabyte-Scale Data Volumes

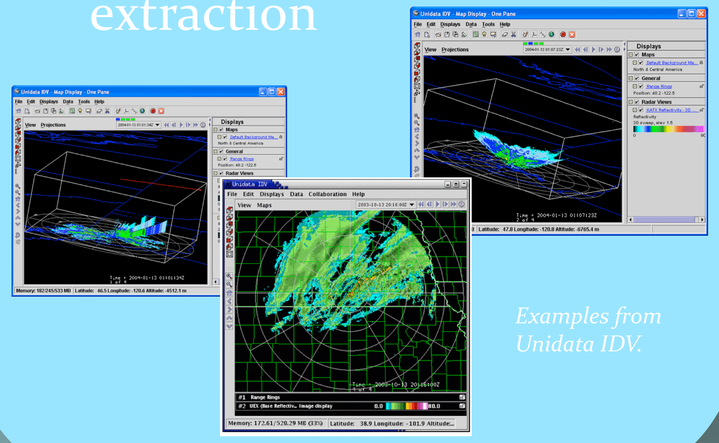
ORNL*, ARM Archive: Raymond McCord, Giri Palanisamy, W. Christopher Lenhardt, Karen Gibson
PNNL: Matt MacDuff, James Mather

Problem: 35 new instruments will each generate large amounts of data (3 - 15 GB /day; 10,000's files /instrument /month)
Question: How will ARM or the data users interact with the new, very large data?
Proposed Answer: Build and operate a "system" optimized for visualization / processing of very large data processing at the ARM Archive.

Tool 1

Interactive Data Visualization

- For large data streams (scanning radar)
- Large online data collection (~40 TB)
- Web-based access
- Library of initial views
- Nominal, embedded data extraction



Examples from Unidata IDV.

Tool 2

Data Extraction *(Deluge Part 1)*

Filter by:
- Time
- Height
- Location
- Measurement
- Frequency

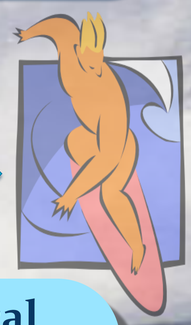
Select by:
- From multiple data streams
- *Statistical summarization?*

Tool 3

Optimized Data Processing

- Operational Policies
  - –Avoid shuffling tasks and contention; "Read only" data; parallel and independent tasks
- Alternative Hardware Configurations
  - –Few systems; small cluster; many, many systems;
- Management Assumptions
  - –Usage via Proposal Process; Data transfer by Archive; Advanced computing users.

Tool 4



Statistical Reviews

- Could be part of Tool 2
- Approach is TBD
- Most difficult computing
- Enable in 2-3 years??

4 Survival Tools

- Visualization
- Extraction
- Optimized processing
- Statistical review

Input needed:

- What saved views?
- How much data/ view?
- Initial data extraction?
- Interactive expectations?

What next?

- N-way
- Multi IN
- Functions
- Striding
- Summary

Examples needed

- Data span?
- Parallelism?
- (or not!?)
- Math complexity?
- Output size?

Examples needed

- Time span?
- Spatial span?
- What Stats?
- Batch or
- Interactive?

User Survey Results

- 300 participants, October 2009
- ~ 20% work with ARM data > 2X/year
- ~70% use their own system
- ~70% need Big Data system < 1 month/year
- 54% would use Big Data System if provided
  - Not a strong preference
- 70% reluctant to download > 100 GB per task
  - 10% download any size!
- Need data manipulation
  - Extraction, merging to common grid, and visualization
- Computational functions less attractive
  - Geophysical, radiative, SCM functions
- Most common programming languages
  - Fortran, Matlab, IDL

ARM CLIMATE RESEARCH FACILITY

U.S. DEPARTMENT OF ENERGY | Office of Science

For more information or questions please contact: ARM Archive User Services (armarchive@ornl.gov)

*ORNL (Oak Ridge National Laboratory) is managed by UT-Battelle, LLC for the U.S. Department of Energy under contract DE-AC05-00OR22725.