Microsoft® Research TechFest 2009 the & in R&D

Tools and Services for Data Intensive Research

RS Barga, C Poulain, N Araujo, T Chou and E Viegas — MSR Redmond
Geoffrey Fox and Jaliya Ekanayake — Indiana University
Magda Balazinska and YongChul Kwon — University of Washington

Data Deluge in Scientific Research

Most scientific data analysis comprise analyzing voluminous data collected from various models and instruments. Efficient parallel/concurrent algorithms and frameworks are key to meeting scalability and performance requirements entailed in such scientific data analyses.

When a teenage boy wants to find information about his idol by [searching] with the search query "Britney Spears," he unleashes the power of several hundred processors operating on a data set of over 200 terabytes. Why then can't a scientist seeking a cure for cancer invoke large amounts of computation over a terabyte-sized database of DNA microarray data at the click of a button?

Randy Bryant (CMU), with permission.

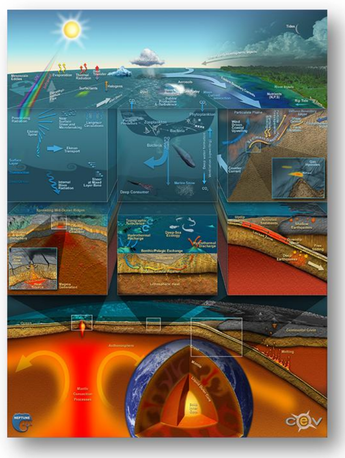
NEPTUNE/OOI

"The oceans cover 70% of the earth surface, yet we know very little about them..."

Underwater gigabit network connects thousands of sensors and instruments along 1200km of cable off Washington coast

Capture & archive 25 years of data
10^+ TBs/year

Metadata catalog, database & file storage



Pan-STARRS

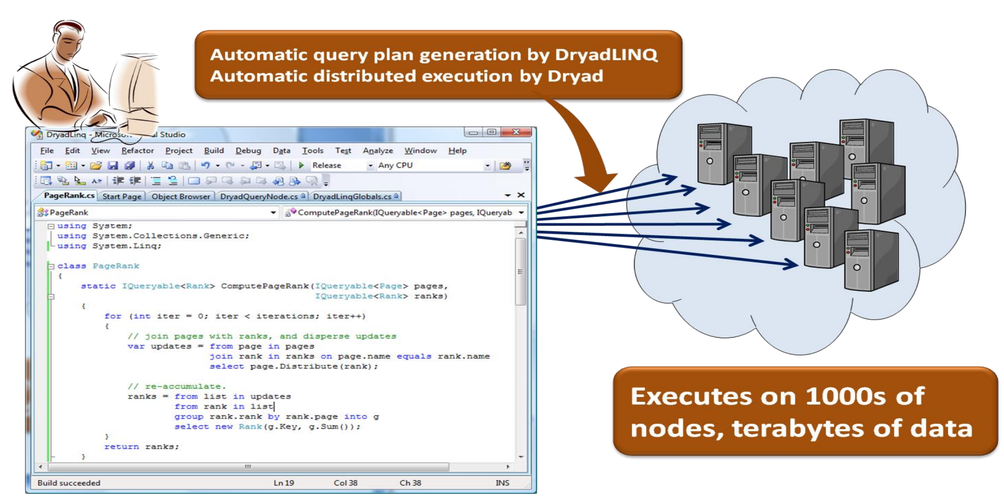
Panoramic Survey Telescope And Rapid Response System

Discover & characterize Earth-approaching objects (asteroids and comets) that might pose a danger to our planet.

Use collected scientific data for Solar System and Cosmology studies.

One of the largest visible light telescopes
- 1.4 Gigapixel camera world's largest!
- 4 unit telescopes act as one
- 1 PB raw image data per year
- 30 TB processed data per year

The External Research group of MSR, in collaboration with the Dryad team, is working with researchers to apply and evaluate the technology we use in search to tackle data intensive research challenges. We will also provide this software, with programming guides and tutorials, at no charge for academic research and education.



How Much Data?

- NOAA has ~1 PB climate data ('07)
- IRIS has ~2 PB seismic data ('09)
- Wayback machine has ~2 PB ('06)
- CERN LHC will generate 15 PB/year ('09)
- WalMart 4PB data warehouse ('07)
- Int'l Data predicts 1.8 ZB of digital data by 2011

Project Details

- *Academic release of Dryad and DryadLINQ (May '09)*
  Including programming guides, tutorials, and exercises
  Send email to dryadrel@microsoft.com for information and updates
- *Dryad as an Azure service – data centers for research*
- *Evaluate on range of scientific data analysis problems*
  Introduce enhancements, extensions as required

Make it Accessible, Useable, a Platform for Research…
http://research.microsoft.com/en-us/collaboration/tools/dryad.aspx

Microsoft® Research