Big Data: Perspectives & Initiatives at NASA

Sasi K. Pillay

Chief Technology Officer for IT

March 5, 2013
Big Data – What Is It

“We don’t have a data problem… we have an analysis problem…”

Or

“We have a filtering problem…”

Or

“We have an opportunity…”
**Big Data** is a term applied to data sets whose size or complexity is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time.
Federal Networking and Information Technology R&D Program

<table>
<thead>
<tr>
<th>DOC</th>
<th>DHHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>- NIST</td>
<td>- AHRQ</td>
</tr>
<tr>
<td>- NOAA</td>
<td>- NIH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DoD</th>
<th>DoI - USGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>- DARPA</td>
<td>- Independent Agencies</td>
</tr>
<tr>
<td>- NSA</td>
<td>- DHS</td>
</tr>
<tr>
<td>- OSD</td>
<td>- EPA</td>
</tr>
<tr>
<td>- Service Research Orgs</td>
<td>- NASA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DOE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- NNSA</td>
<td>- NARA</td>
</tr>
<tr>
<td>- SC</td>
<td>- NSF</td>
</tr>
<tr>
<td></td>
<td>- Treasury</td>
</tr>
</tbody>
</table>
• Describe what a Big Data National Initiative would look like, its vision, goal, scope.

• Develop four main areas:
  – Core Technologies
  – Domain Research Projects
  – Challenges/Competitions
  – Workforce Development
A future in which the ability to analyze and extract information from large, diverse, and disparate data sets accelerates the process of scientific discovery and innovation; promotes new economic growth; and leads to new fields of research and new areas of inquiry that would otherwise be impossible.
Goals

• Promote New Science…by harnessing big data.

• Exploit big data to address national needs and agency missions.

• Support stewardship of federal data.

• Develop the necessary workforce and infrastructure to advance data science.
Core Technologies

- Solicitation: *Core Technologies and Technologies for Advancing Big Data Science & Engineering (BIGDATA) NSF 12-499*
  - 9 NSF Directorates and 7 NIH Institutes

- Announced on March 29, 2012
Challenges

• Series of Challenges: Ideation -> New Tool

• Workshops to decide the parameters

• NASA Center of Excellence for Collaborative Innovation -> Plan (April 12)

• BD SSG Approval and Funding
Description of Challenge

• Develop tools or solutions that allow heterogeneous collections of data to become more homogeneous and searchable “on the fly” or “at first presentation.”

• Involves more than one agency and the resulting tool could be easily generalized for use by multiple agencies.

• Incorporates agency mission statements.

• Involves more than one data set, of which:
  – At least one must be a “big data” data set
  – At least one must be an active or streaming data set (this could be a requirement or an option)

• Promotes Data to Knowledge to Action
Domain Research Projects

• Propose new cross-agency projects
  – Identify obstacles to collaboration
  – Establish foundation for future info sharing
• Near-term effort to identify current projects that could benefit from cross-agency collaboration.
  – Health Effects of Climate Change with USGCRP
  – Earth Science Data & Models
  – Electronic Heath Records
• Long-term includes multi-party solicitations
  – Apply evolving core technologies
  – Test beds under consideration
Workforce Development

- 20 projects across 7 agencies that may be suitable for adapting to Big Data, (Grants, Fellowships, Summer Internships, Scholarships, etc.)

- Undergraduate, Doc, Post-Doc, Mid-Career

- Building a Data Science Community: Meetings at annual conferences, professional associations etc.
Contacts


- Wendy Wigen, Technical Coordinator
  - wigen@nitrd.gov

- Co-Chairs:
  - George Strawn, NITRD
  - Suzanne Iacono, NSF
Current Activities

• Access to large datasets, external & internal
  – Analysis
  – Experimentation
  – Learning

• Large Scale Data Management & Analysis
  – Modeling & Simulation
  – Information Technology & Processing
  – Intelligent Data Understanding
Current Activities (Contd)

• Applications for Science & Mission needs
  – Research Opportunities in Space & Earth Sciences

• Human Exploration
  – Lunar Mapping & Modeling
  – NASA Center of Excellence for Collaborative Innovation

• Aeronautics
  – DASHLink Virtual Laboratory

• Big Data Working Group

• Inter Agency Collaboration
Current Activities (Contd)

• NASA Earth Exchange (NEX)
  A virtual laboratory that integrates high performance computing, data system, data visualization, large amount of online data, models and algorithms, with social network and collaborative technology

• Atmospheric Science Data Center (ASDC)

• Planetary Data System (PDS)

• Mission Data Processing & Control System (MPCS)
  MPCS interfaces with NASA’s deep-space network, and in turn the Mars Reconnaissance Orbiter, to relay data to and from Curiosity and process the raw data in real time
Current Activities (Contd)

• NASA Center for Climate Simulation (NCCS)
  – Used by NASA’s Global Modeling and Assimilation Office and the Goddard Institute for Space Studies
  – focuses on climate and weather data
  – Currently houses 32 petabytes of data, with a total capacity of 37 petabytes
  – 17-by-6-foot visualization wall which allows for one high-resolution surface on which scientists can display still images, video and animated content from data housed in the system.
Current Activities (Contd)

- **Aviation Safety**
  - Extracting relevant information from structured and unstructured data to help predict and prevent safety problems
  - Open-source Multiple Kernel Anomaly Detection (MKAD) algorithm to analyze two continuous data streams or networks are similar, and then analyzed them using a single framework
  - Detection of patterns to automatically discover precursors related to adverse events while an airplane is in flight.
Commercial Cloud Computing Services

- Engineered and migrated legacy content management system and websites to Amazon Web Services
- Mars Science Laboratory relied heavily on mission-critical applications that could sustain failure of over a dozen data centers
- Delivered over 150 Gigabits per second of traffic to a global community of operators, scientists, and general public
- Downloaded raw images and telemetry directly from Curiosity and place them into Amazon S3 storage buckets
- Images from Mars were uploaded, processed, stored, and delivered from the cloud
- Data was catalogued in highly available and scalable databases and exposed to applications and users
- Allowed the content managers for the Mars Web sites to easily create informative Web pages with powerful real–time images.
- Delivered 120 TB of dynamic content and 30 TB of static content the first night
- Met the demands when over 8 million hits were requested of their
Approaches to Big Data

• **Pattern recognition**
  – Statistical Methodologies
  – Clustering
  – Decision Logic

• **Visualization**
  – Picture
  – Video
  – Two Dimensional
  – Three Dimensional
  – Static
  – Dynamic/ Adaptive
Information Content

• Shannon’s Information Theory
  – Define Entropy of an environment
  – Estimate the a priori probability of outcomes
  – Calculate the probability of outcomes based on real data
  – Reduce the Entropy based on the calculations
Information Content (Contd)

\[ H(C) = \text{Sum of all } p(c) \times \log p(c) \]

\[ I(C/X) = H(C) - H(C/X) \]

- **H(C):** Total Entropy in a Domain
- **p(c):** Probability of an event, c within the Domain C
- **H(C/X):** Entropy that remains after measuring X
- **I(C/X):** Information conveyed by measuring X
Approaches

- Parallel Processing
  - Clusters
  - Semantic Processing
- Structured
- Unstructured
Trends for 2013

• Adoption of Commercial Cloud Computing Services
• Public Release of Big Data Sets
• Utilization of Open Innovation Platforms to Solve Challenges Related to Big Data
• Adoption of Open Source Solutions
• Role of the Citizen Scientist