Big Data

George O. Strawn
NITRD.gov
Caveat auditor

The opinions expressed in this talk are those of the speaker, not the U.S. government
Outline

• What do we already know about Big Data?

• What do we still need to learn about Big Data?

• What are Big Data needs in Business and Science?
What is Big Data?

• A term applied to data whose size, velocity or complexity is beyond the ability of commonly used software tools to capture, manage, and/or process within a tolerable elapsed time.

• volume, velocity, variety, etc
Why now for Big Data?

• "Moore's laws" for cpu's, networks, sensors and disks

• eg: disk storage cost has gone from about a dollar per byte (IBM 305 Ramac in 1956) to less than a dollar per ten gigabytes today. A dollar per terabyte soon?

• Sensors: cheap remote sensing, video surveillance, environmental sensing, scientific instruments (not always cheap), etc

• The Internet: Five billion gigabytes and counting (estimated by Eric Schmidt)
What is different about Big Data?

• "Existence precedes essence" (J-P Sartre)

• ie, big data may be created before the specification of (many of) its uses

• Standard IT paradigm: conceive an app; create/collect the data; process the data

• Bid Data paradigm: create/collect data; conceive apps; process the data
Volume: big data requires big computing

• These days, supercomputers aren't actually bigger: they're broader (thousands of cpu's)

• Server farms are "loosely coupled" supercomputers (thousands of servers)

• Big volume data resides on supercomputers or server farms (or at least on clusters)
Volume: big data requires new database architectures

- Relational database architecture don't scale
- NoSQL databases limit functionality and do scale
- eg, BigTable, Document- and Column-oriented databases
The CAP Theorem

- Consistency, Accessibility, Partitionability
- Relational Databases can have all three
- Big Data architectures can only have two out of three!
Velocity: fast big data

- Success with OLTP (online transaction processing) such as google and amazon, but sensors pose a bigger challenge

- Need more "smart sensors" like the LHC, which generates a petabyte of data per second but "only" saves a petabyte per month (take the processing to the data)
Variety: diverse big data

• In traditional programming, the meaning of the data being processed is encoded in the program

• In the Web, the meaning of the data can be deduced by humans reading web pages

• In the Semantic Web (and other metadata schemes) the meaning of the data can be deduced by software
Big Data processing

- Phase 1: Ingest
- Phase 2: Store
- Phase 3: Analyze (three options)
- Phase 4: Visualize
- Phase 5: Insight/Decide
Analyze phase options

- Distributed Memory Architecture (cluster/server farm); e.g., Hadoop
- Shared-Memory Non-Coherent Architecture (supercomputer used in a non-standard way)
- Shared-Memory Coherent Architecture (supercomputer used in a standard way)
And now a word from your sponsor
NITRD
Networking and IT R&D

• Reports to the White House Office of Science and Technology Policy

• A 22-year-old interagency program to enhance coordination of and collaboration among the IT R&D programs of a number of Federal agencies
NITRD Member Agencies

- DoC
  - NOAA
  - NIST
- DoD
  - OSD
  - DARPA
  - AFOSR, ARL, ONR
- DoE (SCI, NNSA, OE)
- DHS
- EPA

- HHS
  - AHRQ
  - NIH
  - ONC
- NARA
- NASA
- NRO
- NSA
- NSF (CISE, OCI)
NITRD PCAs
(program component areas)

- Cyber Security and Information Assurance
- High Confidence Software and Systems
- High-End Computing
- Human Computer Interaction and Info Mgmt
- Large Scale Networking
- Social, Economic, and Workforce Implications
- Software Design and Productivity
NITRD SSGs
(sector steering groups)

- Cybersecurity
- Health IT R&D
- Wireless Spectrum Efficiency
- CyberPhysical Systems
- Big Data
NITRD’s Big Data Initiative

- Core Technologies
- Domain Research Data
- Challenges/Competitions
- Workforce Development
Core Tech I: Collection, Storage

- Data representation, storage and retrieval
- New parallel data architectures, including clouds
- Data management policies, including privacy and access
- Communication and storage devices with extreme capabilities
- Sustainable economic models for access and usage
Core Tech II: Data Analytics

• Computational, mathematical, statistical and algorithmic techniques for modeling high dimensional data

• Learning, inference, prediction and knowledge discovery for large volumes of dynamic data sets

• Data mining to enable automated hypothesis generation, event correlation and anomaly detection
Core Tech III: Data Sharing and

- Tools for distant data sharing, real time visualization and software reuse of complex data sets

- Cross disciplinary model, information and knowledge sharing

- Remote operation and real time access to distant data sources and instruments
Business and Big Data

- Bigger Data
- Unstructured Data
- Distributed Data
- Distributed Computing
Business Analytics

• The use of statistical analysis, data mining, forecasting, and optimization to make critical decisions and add value based on customer and operational data.

• Critical problems are often characterized by massive amounts of data and the need for rapid decisions and high performance computing

  • Eg, modeling customer lifetime value in banks

  • Eg, reducing adverse events in health care

  • Eg, managing customer relationships in hospitality industry
Science and Big Data

• Analyzing output from supercomputer simulations (eg, climate simulations)

• Analyzing instrument (sensor) output

• Creating databases to support wide collaboration (eg, human genome project)

• Creating knowledge bases from textual information (eg, Semantic Medline)
Clouds

- Economy of scale is clear

- Commercial clouds are too expensive for Big Data--smaller private clouds with special features are emerging

- May become regional gateways to larger-scale centers

- The “Long Tail” of a huge number of small data sets (the integral of the “long tail” is big)

- Facebook brings many small, seemingly unrelated data to a single cloud and new value emerges. What is the science equivalent?
Science and Big Data

- Science is increasingly driven by data (large and small)
- Large data sets are here, COTS solutions are not
- From hypothesis-driven to data-driven science
- We need new instruments: “microscopes” and “telescopes” for data
- There is also a problem on the “long tail”
- Similar problems present in business and society
- Data changes not only science, but society
- A new, Fourth Paradigm of Science is emerging…
From Bits to Its?

• After newton, the world consisted of matter in motion

• After the steam engine came thermodynamics and the world consisted of matter and energy

• After the computer, perhaps comes a science of information and the world may then consist of matter, energy and information
What the future may hold

• Data intensive science appears to be revolutionary science

• Data analytics and other big data services are major opportunities for business and government

• Big Data may also be the basis of new services for people, perhaps as significant as the Web, Google and Facebook