Big Data Challenges in the Federal Government

July 8, 2013

Big Data

A possible definition…..

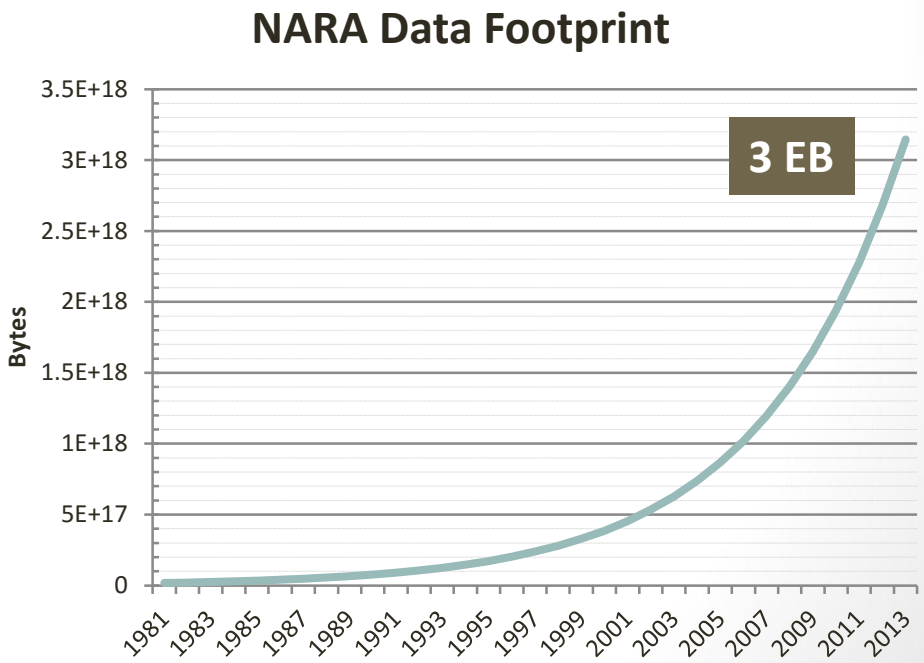
Datasets whose **size** is beyond the ability of typical software tools to **capture**, **store**, **manage**, and **analyze** within a tolerable elapsed time.

Capture

- **Presidential Directive for Records Management – August 2012**
  - By the end of this decade, the Archivist of the U.S. will no longer accept or accession records into the Archives unless they are in digital or electronic form
- **Moving large amounts of data into a central repository is an issue**
  - 2010 Census – 330 TB moved to NARA…on two trucks
  - Bush 43 electronic records – 80 TB moved physically
  - Delivery of 1940 Census – 16 TB of JPG files required transfer devices

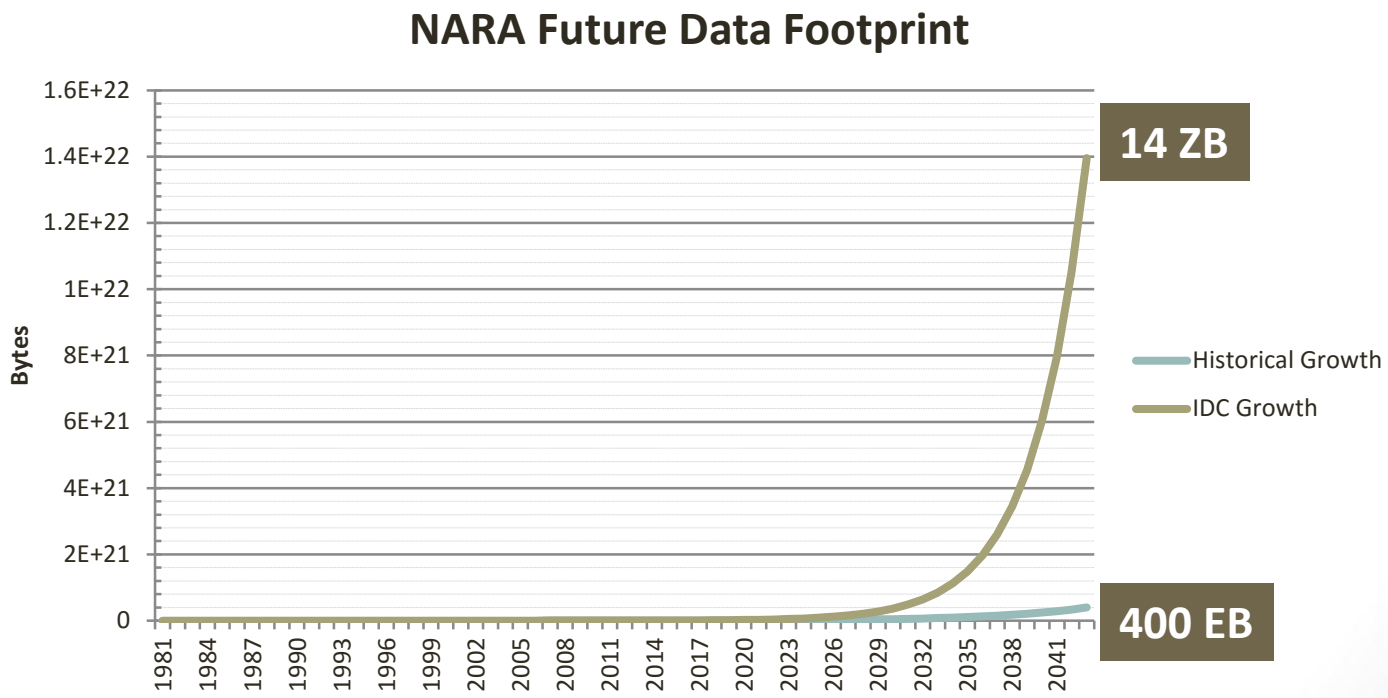
Storage -- Data!

- **NARA receives only 2-3% of the data created within the government**
- **Even at this rate, the digital equivalent of our analog holdings is big**:
  - 12 billion pages
  - 18 million maps
  - 50 million photos
  - 550 thousand artifacts
  - 360 thousand films
  - Electronic records
  - etc.



Storage – Even More Data!

- **IDC projects a compounded growth rate of >32%/year**
- **NARA's growth rate has been less than this**

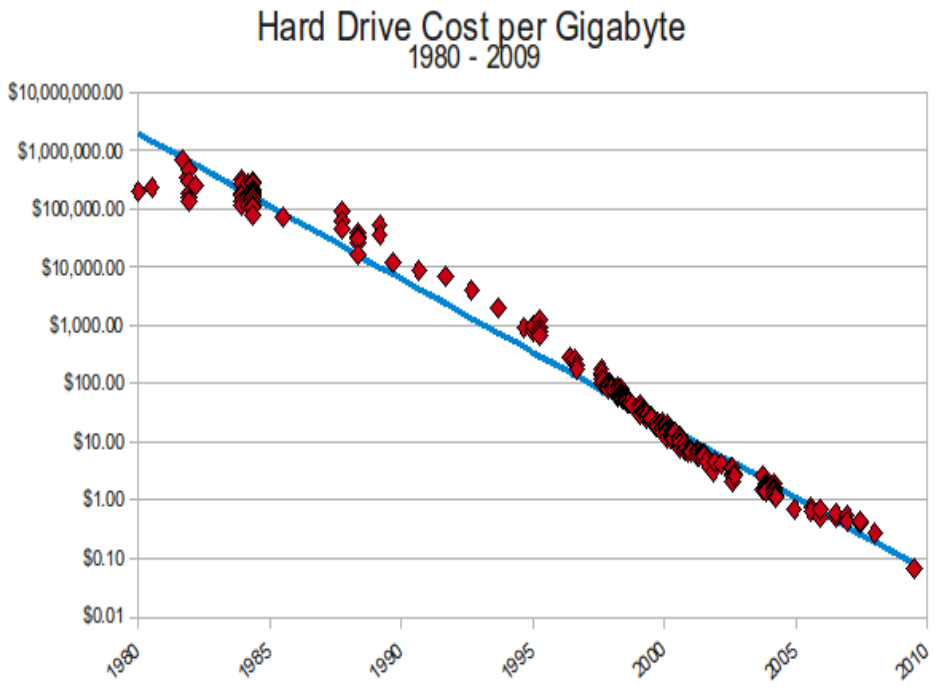


Storage -- Supporting Facts

- **The Federal government is spending ~$24B for storage in FY13**
  - ~10 EB of storage
  - At the historical accession rate, this will result in 200-400 PB of data transferred to NARA 30 years from now
- **The 2010 Census is 330 TB of data**
- **The converted 1940 Census is 120 TB of data**
- **The Bush 43 electronic records is 80 TB, half of which is images**
- **Tweets! >450M/day, 100GB/day (compressed)**

Storage -- Cost

- **Storage costs have consistently declined for decades**
- **TCO for a TB in a Federal data center is ~$2.5K/year**
- **FISMA certified clouds are more competitive**



Reference : Matthew Komorowski,
Center for Computational Research at SUNY University at Buffalo

Management

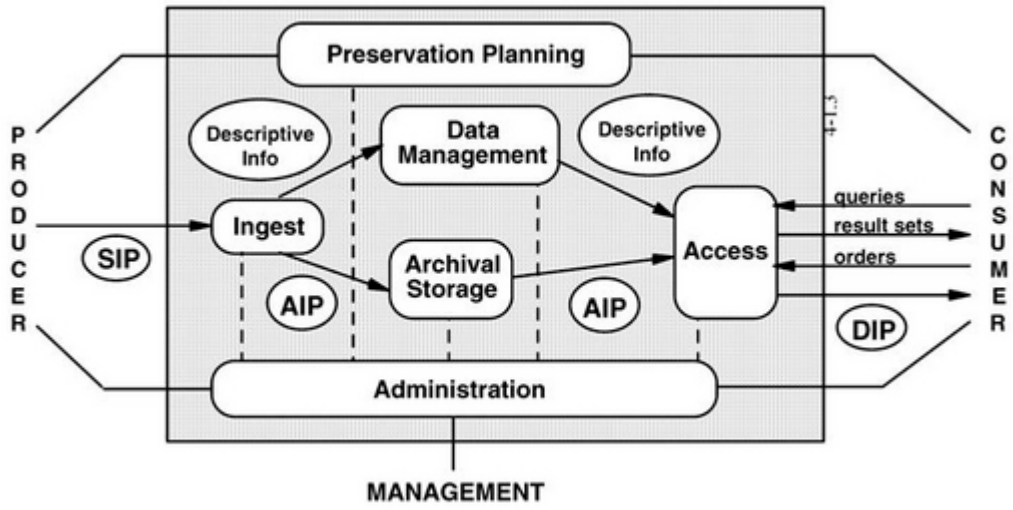
- **Storage formats becomes obsolete long before we NARA receives data**
  - **Tape:**
    - 3480, 3490, 9-track open-reel tape, 4mm, 8mm, mainframe disk-packs, 7 track open reel
  - **Magnetic disk**:
    - 8 inch floppy, 5.25 inch floppy, 3.5 inch floppy (DOS and MAC), Syquest media, Iomega ZIP
  - **Optical media:**
    - CD, DVD
  - **External Hard Drives of various types**
  - **Punch-cards**

Management

- **Applications only support 2-3 prior versions, or are discontinued**
  - Try to open a 20 year old WORD document
  - Remember WordStar?
- **The number of file formats continue to grow**
  - Droid and the Pronom registry describe and identify less than 1000
  - Estimate for the number formats is >10,000
  - NARA currently has ~100 formats

Management

- **Preservation of data needs to be anticipated from the beginning**
- **Open Archival Information System (OAIS) was developed to support long term preservation, but processing need to be nearly continuous**



Analysis/Access

- **Analysis/Access is a growing concern**
  - **Boolean search terms result in errors**
    - Big issue for FOIA requests and special research projects
  - **More and more date is unstructured**
  - **Delivery spikes on high interest data**
    - Nixon Watergate transcripts and JFK audio – 4 TB download in 3-4 days for each release
    - 1940 Census – millions of visitor and 100s of TB downloaded in the first week

Analysis/Access

- **Collaboration is key – the 1940 Census success story**
  - **The first and largest national service project of its kind**
    - Project completed in only 5 months
    - 132 million names indexed
  - **More than 160,000 volunteers**
    - Over 600 genealogical societies signed up to participate in the project
    - 5 partner organizations involved -- FamilySearch, NARA, Archives.com, Findmypast.com, ProQuest
  - **Delaware completed their index in 2-3 days**

Mike Wash
michael.wash@nara.gov
301.837.1992