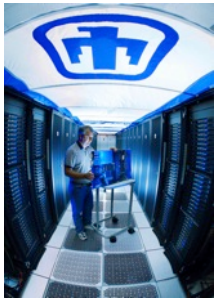
SAND2012-2580C



Big Data Analytics in a National Security Setting

CERIAS Security Symposium, Purdue University

3-4 April 2012

Jamie Van Randwyk [jvanran@sandia.gov]

R&D Technical Manager

Informatics and Systems Assessments Department

Sandia National Laboratories

Exceptional service in the national interest

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Big Data is Relative

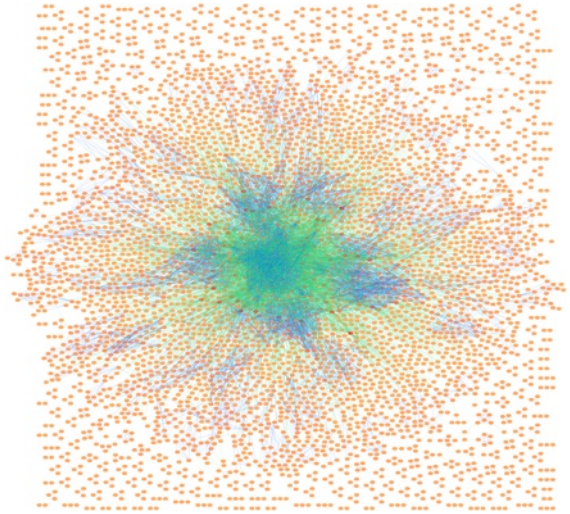
Sandia National Laboratories

- Big data means different things to different people
  - Home user - 2 terabyte hard drives!
  - Industry – low 10's of petabytes (http://gigaom.com/cloud/under-the-covers-of-ebays-big-data-operation/)
  - Gov't – Exabytes? (http://www.computerworld.com/s/article/9223677/CIA_backed_Cleversafe_announces_10_exabyte_storage_system)

Big Data Analytics are Relative

Sandia National Laboratories

- Cheap/easy (well-established algorithms and commodity hardware)
  - Internet search
  - Retail intelligence
- Complex (unknown/new algorithms and specialty hardware)
  - Netezza, EMC / Greenplum, CleverSafe
  - “Joins” across numerous datasets
  - Graph analysis at billion node scale



Is “Cloud” Big Data?

Sandia National Laboratories

- Large amounts of data are stored in “the cloud”
- Analyzing big data is best done from local storage (local cloud?)

- With terminology, we have to follow industry’s lead – so cloud == big data

Most important to understand and be hands-on with the technologies

Big Data at Sandia

Sandia National Laboratories

- Sandia produces, stores, and analyzes data for internal business functions and for our gov't customers.
- Big data involves analysis, storage, networking, and often virtualization.
- For various scientific and national security projects, we have evaluated
  - Hadoop
  - Ceph
  - Cassandra
  - Amazon Elastic Cloud running Hadoop (Amazon EMR)

Hadoop at Sandia

Sandia National Laboratories

- We like Hadoop
  - Open source
  - Large user community, constant development, production and research branches
  - Works well on cheap hardware
  - HDFS (distributed data side of Hadoop) is easy way to store terabytes, replicates automatically for safety and performance
  - MapReduce (analysis side of Hadoop) lets you write algorithms that work on HDFS data, but you don't have to worry about parallel programming issues like synchronization, deadlock, or threading

Hadoop at Sandia

Sandia National Laboratories

- We use Hadoop daily
  - Collect network traffic data (tens of terabytes)
  - Store other terabyte data sets (social network research, bioinformatics)
  - Research algorithms for graph analysis, machine learning, statistical profiling, anomaly detection
  - Tall and skinny QR matrix factorization

Nebula

$150k
500 Tbytes
64 quad-core

Future of Big Data Analytics

Sandia National Laboratories

- Eric Schmidt sound bite (paraphrased): Humans create 5 exabytes of data every 2 days.
- Needle is more difficult to find
  - Do we care about the needle?
  - Maybe we're more interested in large, unnoticed trends
- Industry innovation
  - Larger storage systems
  - Backup and availability
  - Faster retrieval
  - Smarter storage
- Government need
  - Security (IRS, SSA, TSA, etc. storing our personal information)
  - Analytics – war on terror