The InterNational Committee for Information Technology Standards INCITS

Big Data

Keith W. Hare
JCC Consulting, Inc.
April 2, 2015

Who am I?

- Senior Consultant with JCC Consulting, Inc. since 1985
  - High performance database systems
  - Replicating data between database systems
- SQL Standards committees since 1988
  - Interim chair of the INCITS Big Data Technical Committee
  - Chair, INCITS Big Data Ad Hoc – the USA TAG to the JTC1 Study Group on Big Data – 2014
  - Convenor, ISO/IEC JTC1 SC32 WG3 since 2005
  - Vice Chair, ANSI INCITS DM32.2 since 2003
- Education
  - Muskingum College, 1980, BS in Biology and Computer Science
  - Ohio State, 1985, Masters in Computer & Information Science

Topics

- What is Big Data and why is it important?
- Where can standards help?
- What are the challenges?
- Big Data Efforts

What is Big Data?

- Often described using 3, 4, or 5 V's
  - Volume, Variety, Velocity, Variability, Veracity
  - Imprecise definition because the problem space is imprecise
- Paradigm Shift
- Driving Forces
- How Big is Big?

Why Big Data?

- Learn something from the data that is valuable
- Analytical and data visualization tools
- To be most effective,
  - tools need a standard interface to the data sets
  - don’t spend a lot of time custom programming an interface per data set.

Big Data: Driving Forces

- Inexpensive storage of large volumes of data
- Inexpensive compute power
- Next Generation Analytics
  - Moving from
    - Off-line to in-line embedded analytics
    - Explaining what happened to Predicting what will happen
  - Operating on
    - Data at rest – stored someplace
    - Data in motion – streaming
    - Multiple disparate data sources
- Look at available data and wonder what answers are hidden there

Big Data Soutions

Big Data Solutions are those which address problems in data analytics where the Volume, Velocity, or Variability of the source data exclude delivery of analytic results within the required timeline using conventional solutions on current technology.

Gregory Weidman
L-3 Data Tactics Big Data Insights
http://datatactics.blogspot.com/2013/10/what-is-big-data.html

How Big is Big?

- Data volumes
- Data Distribution

How Big is Big – Data Volumes

- Terabytes –1000**4
- Petabytes – 1000**5
- Exabyte – 1000**6
- Zettabyte – 1000**7
- Yottabyte – 1000**8
- Brontobyte – 1000**9
- Gegobyte – 1000**10

Notes:

- Brontobyte and Gegobyte are not recognized by the International System of Units, still subject to change.
- Geobyte is also being used (see http://www.oxfordmathcenter.com/drupal7/node/410) but it is also the name of a US corporation.

How Big is Big – Data Distribution

- Server
- Cluster
- Datacenter
- Continent
- Planet
- Solar System

Many Big Data Projects

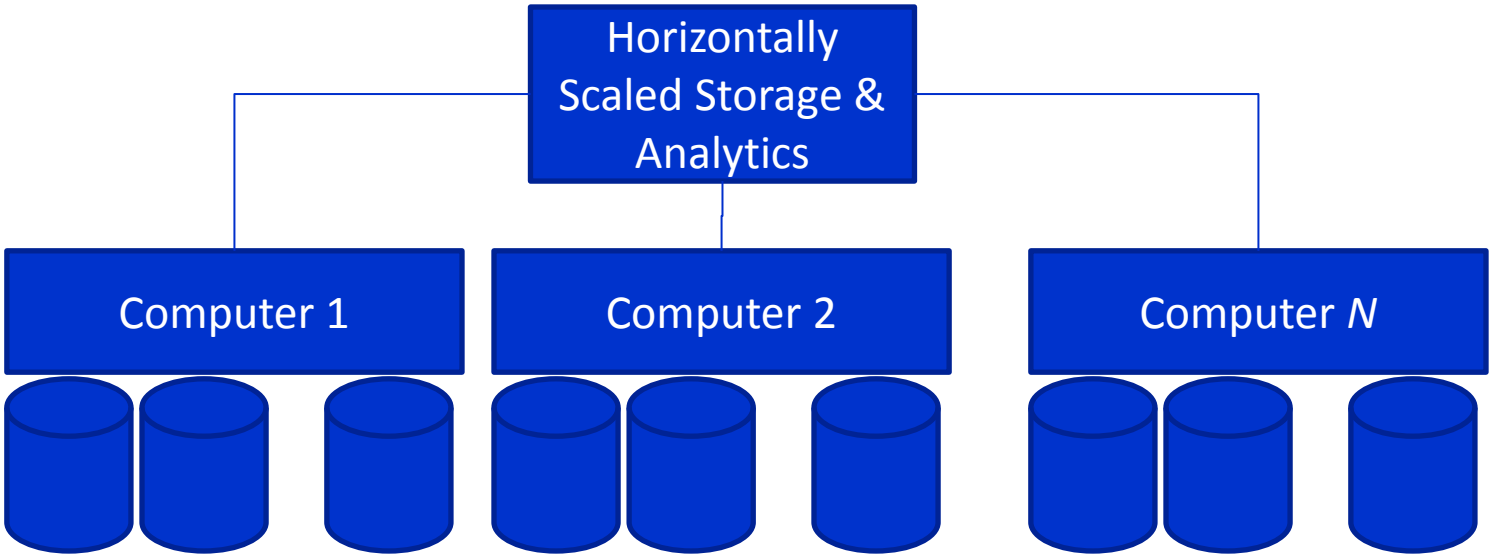
- Rush to get something done without taking the time to do upfront design
- Meaning of the data documented
  - In the custom coded interface
  - On the whiteboard in someone's cubicle
  - Intuitively obvious based on the abbreviated JSON or XML tags
- We are reinventing 1960s data technology but on a bigger scale

Where can Standards help Big Data?

Standards can assist in the Big Data arena, but we have to identify where they could be useful

- Horizontally Scaled Data Sources
- Variety of Data Sources
- Integrating Multiple Data Sources

Horizontally Scaled Data Source



Horizontal Scaling is one solution to the data volume challenge.

Horizontally Scaled Data Source

- Ease of Use
  - Language for storing data
  - Language for querying metadata
  - Language for querying data
  - Language for specifying distributed queries
  - **Potential for standardization!**
- Performance
  - Simple matter of engineering & programming
  - Language for specifying distribution
    - Likely to be product specific
    - Little potential for standardization

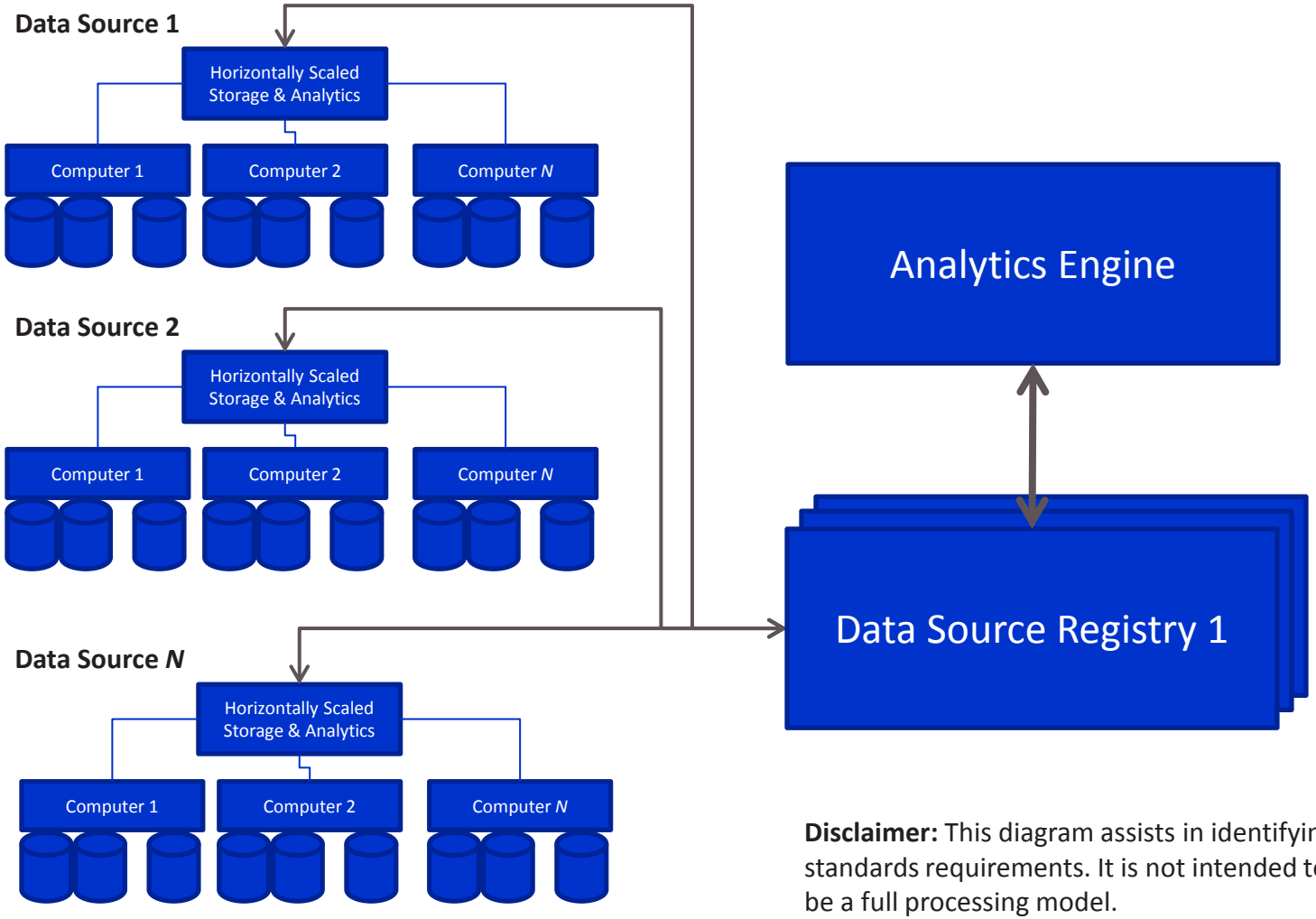
Variety of Data Sources

- Tabular data – relations
  - Designed, cleansed, curated
- Spatial data
- Images & Video
  - Well defined structures
  - Need additional domain information.
    - aerial photos, faces, stars
    - Etc.
- XML – may have well defined DTD (Document Type Definition)
- Store everything now, figure it out later
  - JSON – Java Script Object Notation
  - E.g. network packet logs
- Multiple storage models to handle the diversity

Variety of Data Source Ownership

- Self Owned
- Publically Available
- Available with Restrictions
- Data for hire
- Derived Data

Integrating Multiple Data Sources



Disclaimer: This diagram assists in identifying standards requirements. It is not intended to be a full processing model.

Data Source Registry Requirements

- Language/Interface for registering data source
- Support for discovering and identifying available data sources
  - Content of the data source
    - Semantics and Syntax of data
    - Available analytic routines
    - Security/Privacy restrictions
    - Provenance of the data
  - Information about connecting to data source
- Business agreement information
  - Costs
  - Use Restrictions
  - Service Level Agreements
- **Standards will support integration of multiple data sources**

Challenges

- Analytical and data visualization tools
- Integration of multiple data sources
- Discovery of data sources
- Description of data

Challenges – Analytical and Data Visualization Tools

- To be most effective, tools need a standard interface to the data sets
- Don’t want to spend a lot of time custom programming an interface per data set.

Challenges – Integration of Multiple Data Sources

- It's neat that my data set has eleventy-trillion records
- Can I do an analysis that integrates my data set with your data set to learn something useful?

Challenges – Data Source Discovery

- Programmatically discover that you have an interesting data set available
- Understand what it is
- Understand how to access it

To accomplish this, you need to register your data set someplace that I can programmatically query.

Challenges – Description of Data

- I need to understand what your data set contains
- Avoid accidently correlating
  - Apple crop in New Zealand with
  - Hedge Hog population in Wales
- Technical Challenge
  - Need better support for describing data – data set registries
- Social/Management Challenge
  - If you do a good job of describing your data, I get the benefit

Big Data Efforts

- ISO/IEC JTC1 WG9 Big Data
- INCITS/BigData Technical Committee
- NIST Big Data Public Working Group

ISO/IEC JTC1 WG9 Big Data

- Established by the November 2014 JTC1 Plenary
- Two Work Items approved in March, 2015
  - Definitions
  - Reference Architecture
- Meetings:
  - April 7-9, Bremen Germany
  - July 7-9, Seoul Korea (tentative)
  - November 2015 Brasilia, Brazil, (tentative)
- Convenor: Wo Chang, from the US

JTC1/WG9 April 2015 Meeting

- Representatives from 10 national bodies
  - China, Finland, France, Germany, Ireland, Japan, South Korea, Singapore, UK, USA
- ISO/IEC 20546 "*Big Data – Overview and Vocabulary*"
  - Editors: **Nancy Grady**, Lili Yang
  - FDIS March, 2017
- ISO/IEC 20547 "*Big Data – Reference Architecture*"
  - Editors: Suwook Ha, **David Boyd**, Ray Walshe
  - FDIS March, 2017
- Liaison statements to several other groups

INCITS/BigData Technical Committee

- INCITS Executive Board Established January 2015
- TAG to JTC1 Big Data Efforts
- Meetings
  - March 6, 2015
  - April 30, 2015
  - June 4, 2015 (Tentative)
  - August 6, 2015 (Tentative)
- Current Membership, 14 voting, 3 prospective
  - Users outnumber vendors
- Acting Chair: Keith Hare

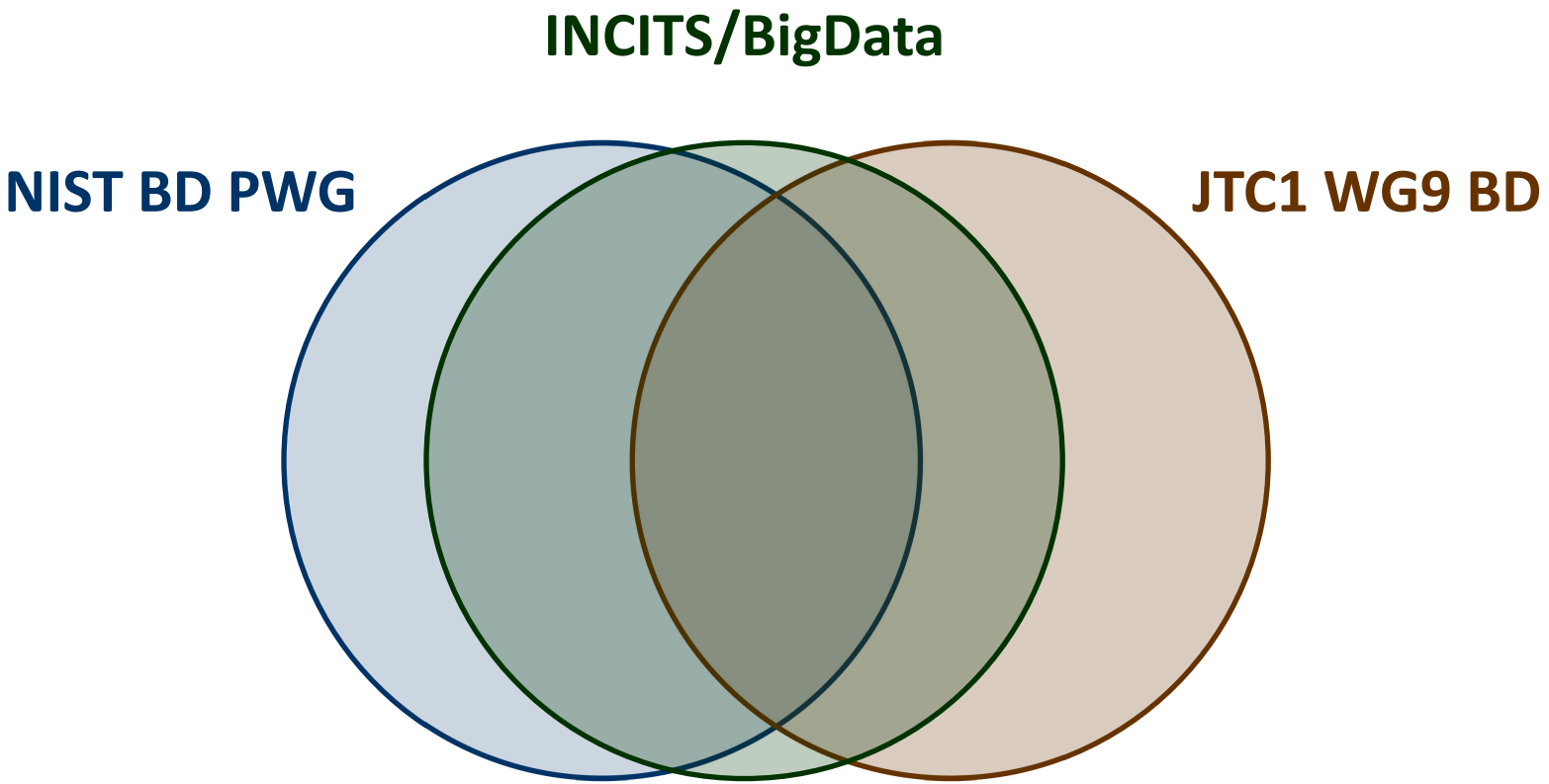
NIST Big Data Public Working Group

- Weekly web conferences since June, 2013
- Mix of industry, academic, and users
- Participation is fluid – 40 to 50
- Output Documents – NIST Special Publication 1500-1 through 1500-7
- Chaired by Wo Chang of NIST

NIST Big Data Interoperability Framework Draft Version 1 Documents

- NIST Special Publication 1500-1 Volume 1 - *NIST Big Data Definitions* http://bigdatawg.nist.gov/_uploadfiles/M0392_v1_3022325181.docx
- NIST Special Publication 1500-2 Volume 2 - *NIST Big Data Taxonomies* http://bigdatawg.nist.gov/_uploadfiles/M0393_v1_3613775223.docx
- NIST Special Publication 1500-3 Volume 3 - *NIST Big Data Use Case and Requirements* http://bigdatawg.nist.gov/_uploadfiles/M0394_v1_4746659136.docx
- NIST Special Publication 1500-4 Volume 4 - *NIST Big Data Security and Privacy* http://bigdatawg.nist.gov/_uploadfiles/M0395_v1_4717582962.docx
- NIST Special Publication 1500-5 Volume 5 - *NIST Big Data Architectures White Paper Survey* http://bigdatawg.nist.gov/_uploadfiles/M0396_v1_7656223932.docx
- NIST Special Publication 1500-6 Volume 6 - *NIST Big Data Reference Architecture* http://bigdatawg.nist.gov/_uploadfiles/M0397_v1_2395481670.docx
- NIST Special Publication 1500-7 Volume 7 - *NIST Big Data Standards Roadmap* http://bigdatawg.nist.gov/_uploadfiles/M0398_v1_1449826642.docx

Overlapping Participation



Overlaps are indicative, but not to scale.

Interactions With Other Groups

- SC32 (DM32) – Data Management & Exchange
  - WG2 – Metadata
  - WG3 – Databases and Database Languages
- SC27 (CS1) – Security
- SC38 (DAPS38) – Cloud Computing
- ISO/TC 211 (L1) – Geographic Information Systems
- …

Missing Links

- A number of Big Data related efforts are open source
  - R Project for Statistical Computing
  - Apache Software Foundation
    - Hadoop and its eco-system
    - Cassandra
    - CouchDB
  - MongoDB
- Missing Big Data Vendors include (but not limited to)
  - IBM
  - HP
  - Microsoft
  - Google
  - Amazon

Big Data in Perspective

“Space is big. Really big. You just won't believe how vastly, hugely, mind-bogglingly big it is. I mean, you may think it's a long way down the road to the chemist's, but that's just peanuts to space.”

Douglas Adams, The Hitchhiker's Guide to the Galaxy

“Don’t solve problems that you don’t have.”

Keith Gordon, in a talk to BCS Aberdeen, 2014
https://www.youtube.com/watch?v=wDU8OiRk7Ac

Summary

- Big Data
  - Lots of hype but potential and technology are real
- Standardization efforts are starting
  - NIST documents provide a good base for discussion
  - Overlap and coordination with other groups

Thank you