BIG DATA STANDARDS

Keith W. Hare

JCC Consulting, Inc.

September 29, 2014

Introduction

- ISO/IEC JTC 1 Study Group on Big Data
- What is Big Data?
  - Driving Forces
  - Working Definition
  - How big is big?
- Where can Standards help with Big Data?
- Standards Gaps Identified by SGBD
- SGBD Recommendations
- Path Forward
  - SC32
  - New Working Group

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
  - Convenor, ISO/IEC JTC1 SC32 WG3 since 2005
  - Chair, INCITS Big Data Ad Hoc – the USA TAG to the JTC1 Study Group on Big Data
  - 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

Identifying Standardization Areas

ISO/IEC JTC1 Study Group on Big Data (SGBD)

- Established at November 2013 JTC1 plenary
- Four face to face & two electronic meetings
- Report completed September 3, 2014
  - ISO/IEC JTC 1 SGBD N0095 "*Final SGBD Report to JTC 1*"

Integrates ideas from multiple standards development groups, organizations and national bodies

Standards Hierarchy Translation

ISO/IEC JTC1 SC32 WG3

- ISO – International Organization for Standardization (l'Organisation internationale de normalisation)
- IEC – International Electrotechnical Commission
  - JTC1 – Joint Technical Committee 1 – computer related standards
    - SC32 – Sub Committee 32 – Data Management and Interchange standards
      - WG1 – eBusiness
      - WG2 – Metadata
      - **WG3 – Database Languages – SQL Standards**
      - WG4 – SQL/Multimedia and application packages

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
- Dimensions of Big Data
- Paradigm Shift
- Driving Forces
- Working Definition
- How Big is Big?

Big Data: Dimensions

- Characteristics:
  - Quantity, size
  - Complexity
  - Rate of change
  - Varieties*
  - Availability
  - Persistence
  - Integrity*
  - Location*
  - Relevance*
  - *Etc.*
- Aspects
  - Data, *per se*
  - Metadata, models*
  - Privacy & Security*
  - Storage, reliability
  - Query & Analysis*
  - Transport, interchange*
  - Life cycle*
  - Accessibility
  - Integration*
  - *Etc.*

*** Areas where SC 32 has expertise**

Big Data: Paradigm Shift

Database users are attempting to escape the restrictions of the current SQL databases and database vendors

- Distributed
- Replicated
- Highly Available
- Large data volumes
- Reduced up-front development costs
- Minimal upfront licensing costs

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
  - 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: Working Definition

- Variety, Volume, Velocity, Variability, Availability
  - Imprecise terms, but useful for understanding problem space
  - All relative – Yesterday's impossible is today's Big Data and will be tomorrow's trivial
- Requirements cannot be met on a single computer
  - Distribute data storage to support volume & velocity
    - Replicate data storage to provide availability
  - Distribute processing
    - Apply compute power in parallel
    - Avoid moving data across the network – move the answers

Big Data: Personnel

- Problem space is too big and/or complicated for a single person
- Need some combination of:
  - Domain Expert – understands the problem space
  - Systems Expert – understands the technology for storing and retrieving data
  - Data Scientist – understands domain, algorithms, statistics, programming, etc.
- **Standards make it easier!**

How Big is Big?

Data Volume

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

Data Distribution

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

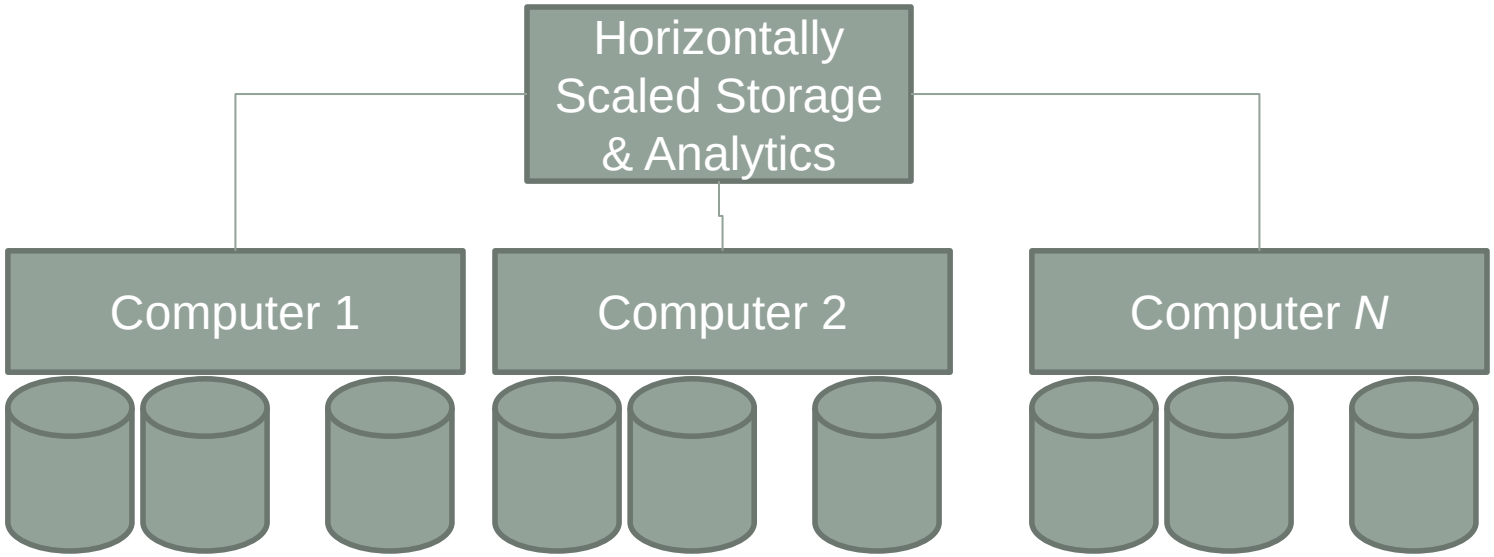
*This terminology is still subject to change.

Where can Standards help with 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

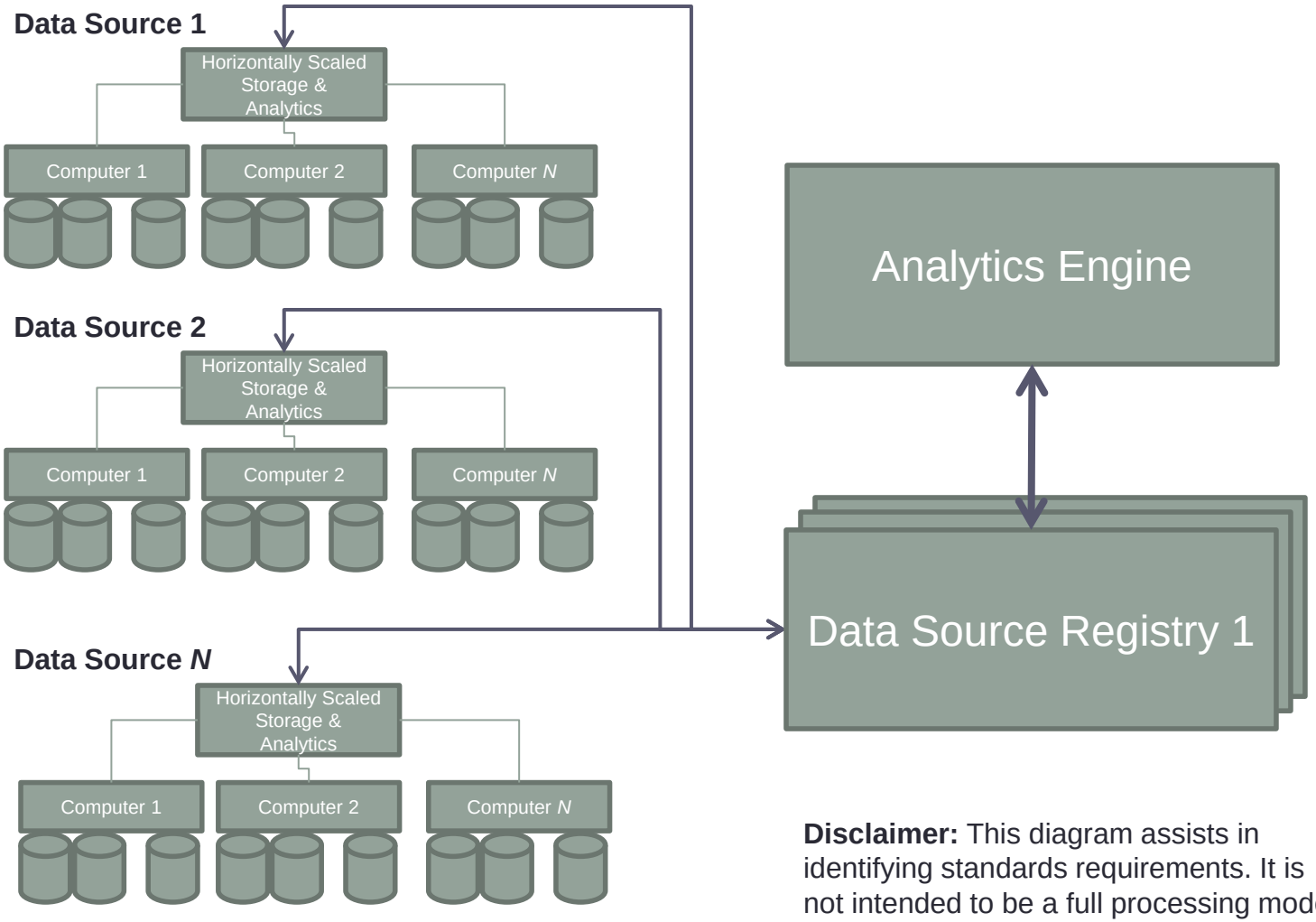
Variety of Data Sources

- 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
- Store everything now, figure it out later
  - JSON
  - 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 support integration of multiple data sources**

Standards Gaps Identified by SGBD

1. *Big Data use cases, definitions, vocabulary and reference architectures (e.g. system, data, platforms, online/offline, etc.)*
2. *Specifications and standardization of metadata including data provenance*
3. *Application models (e.g. batch, streaming, etc.)*
4. *Query languages including non-relational queries to support diverse data types (XML, RDF, JSON, multimedia, etc.) and Big Data operations (e.g. matrix operations)*
5. *Domain-specific languages*
6. *Semantics of eventual consistency*
7. *Advanced network protocols for efficient data transfer*
8. *General and domain specific ontologies and taxonomies for describing data semantics including interoperation between ontologies*

Standards Gaps Identified by SGBD

9. *Big Data security and privacy access controls.*
10. *Remote, distributed, and federated analytics (taking the analytics to the data) including data and processing resource discovery and data mining*
11. *Data sharing and exchange*
12. *Data storage, e.g. memory storage system, distributed file system, data warehouse, etc.*
13. *Human consumption of the results of big data analysis (e.g., visualization)*
14. *Energy measurement for Big Data*
15. *Interface between relational (SQL) and non-relational (NoSQL) datastores*
16. *Big Data Quality and Veracity description and management*

SGBD Recommendations

- Assign standards gaps to existing JTC1 SCs:
  - SC 6 – Telecommunication and information exchange between systems
  - SC 22 – Programming Languages
  - SC 24 – Computer Graphics, Imaging Processing, and Environmental Data representation
  - SC 27 – IT Security techniques
  - SC 32 – Data management and interchange
  - SC 34 – Document description and processing languages
  - SC 38 – Distributed application platforms and services (DAPS)
  - SC 39 – Sustainability for and by Information Technology
- Create a new JTC1 level Working Group for:
  - Definition and Vocabulary of Big Data
  - Big Data Reference Architecture
  - Coordination of efforts across other SCs

Tasks Assigned to SC32

- Definition of standard interfaces (e.g., language, API) to support non-relational datastores (4)
- Definition of SQL extension to support exchange and integration between SQL and non-SQL datastores (11, 15)
- Metadata and provenance standards (2,9)
- SQL and NoSQL standards for data mining (10)
- Support for large complex data structures in SQL and/or SQL/MM (4,11)
- Support for operations on complex data structures and defined operations on such structures (e.g. add, multiply union) (4,5)
- Standards for eventual consistency and acceptable consistency (6)

Standard interfaces for non-relational datastores (4)

4. Query languages including non-relational queries to support diverse data types (XML, RDF, JSON, multimedia, etc.) and Big Data operations (e.g. matrix operations)

- Diverse data types
  - XML – Supported since 9075:2003
  - JSON – In next edition of 9075 (2015 or 2016)
  - RDF – Resource Description Framework – graph data structures
  - Multimedia – SQL/MM standard
  - Etc.
- Big Data operations
  - Matrix operations – new 9075 part, SQL Multi Dimensional Arrays
  - Etc.

SQL extension for exchange and integration (11, 15)

11. Data sharing and exchange

15. Interface between relational (SQL) and non-relational (NoSQL) datastores

- APIs for data access
  - SQL/CLI
  - JDBC
- Support for registering data sources
- Support for identifying potentially useful data sources
- Identify requirements for data exchange

Metadata and provenance standards (2,9)

2. Specifications and standardization of metadata including data provenance

9. Big Data security and privacy access controls.

- ISO/IEC 11179
- Review provenance requirements
- Data Source registry using 11179
- Security & Privacy
  - Similar topic space with differences
  - Cost benefit tradeoff
  - Difficult problems

SQL and NoSQL standards for data mining (10)

10. Remote, distributed, and federated analytics (taking the analytics to the data) including data and processing resource discovery and data mining

- SQLMM Data Mining
- Support for registering data sources
- Support for identifying potentially useful data sources

Support for large complex data structures in SQL and/or SQL/MM (4,11)

4. Query languages including non-relational queries to support diverse data types (XML, RDF, JSON, multimedia, etc.) and Big Data operations (e.g. matrix operations)

11. Data sharing and exchange

- Operations on Images, video, sound, etc.

Operations on complex data structures (e.g. add, multiply, union) (4,5)

4. Query languages including non-relational queries to support diverse data types (XML, RDF, JSON, multimedia, etc.) and Big Data operations (e.g. matrix operations)

5. Domain-specific languages

- External technologies
  - R Statistical Computing Platform
  - W3C specifies SPARQL Query Language for RDF
- SC32 WG3 Technologies
  - Support for matrices under development – SQL/Multi-Dimensional Arrays
  - SQL support for graph data structures has been discussed, but no concrete change proposals

Standards for eventual consistency and acceptable consistency (6)

6. Semantics of eventual consistency

- ISO/IEC 9075 currently specifies details for ACID transactions
  - Atomic, Concurrent, Isolated, Durable
  - Supports specification of transaction consistency
    - Serializable, Repeatable Read, Read Committed, Read Uncommitted
- Need to investigate
  - BASE Transactions
    - Basically Available, Soft state, Eventual consistency
  - Brewer's CAP theorem
    - Consistency, Availability and Partition Tolerance – pick two
- ACID transaction consistency not the same as BASE transaction consistency

Data Source Registry Requirements

- Ability to register a database in a metadata registry
  - Create a new metadata model
  - Tie source data to the registry entries
  - Match/map to an existing metadata model – might be manual
- Ability to create tables in a database using components from a metadata registry
  - Tie the database metadata back to the source registry
    - Registry URL & Unique identifier/certificate
    - Unique identify for each table and column
    - Timestamp
  - Could be syntax on Create Table statement
- “Semantics is not computable – approximate semantics with syntax”

SC32 Summary

- Lots of existing work to support bits
- Existing standards need review and enhancement

New Working Group

- Tasks
  - Definition and Vocabulary of Big Data
  - Big Data Reference Architecture
  - Coordination of efforts across other SCs
- Terms of Reference
  1. Identify, develop, and maintain JTC 1 deliverables in the field of Big Data Definitions and Vocabulary and a Big Data Reference Architecture
  2. Investigate the requirements of Big Data and address the technology gaps for the evaluation and development of new work within the scope of JTC 1
  3. In conjunction with JTC1 coordinate Big Data efforts among JTC 1 SCs
  4. Support JTC 1 goals and respond to requests pertaining to Big Data, JTC 1 and external Liaison organizations
  5. Liaise with SDOs and consortia related to Big Data as appropriate
  6. Maintain future JTC 1 PAS and Fast Track submissions in the area of Big Data which are not within the scope of existing JTC1 SCs
- Must be approved by JTC1 Plenary

Summary

- Suitably crafted standards can guide construction of scalable implementations
- Community experimentation and understanding are evolving rapidly
- Need complete eco-system to make this all work
- Standards are essential – niche solutions lead to vendor lock-in
- Wherever possible, use existing standards as foundation for new efforts
- Coordination required between SCs

Acknowledgements

- All errors, misunderstandings, misleading statements, and idiotic comments are mine and mine alone.

Additional Resources and Discussions

The following slides have useful information but are beyond the time available for the current presentation.

Sources – JTC1 Study Group on Big Data

- N0028 *Volume 1: NIST Big Data Definitions*
- N0029 *Volume 2: NIST Big Data Taxonomies*
- N0030 *Volume 3: NIST Big Data Use Case & Requirements*
- N0031 *Volume 4: NIST Big Data Security and Privacy Requirements*
- N0032 *Volume 5: NIST Big Data Architectures White Paper Survey*
- N0033 *Volume 6: NIST Big Data Reference Architecture*
- N0034 *Volume 7: NIST Big Data Technology Roadmap*
- N0095 *Final SGBD Report to JTC 1*

Big Data Analysis Challenges

A number of challenges in both data management and data analysis require new approaches to support the big data era. These challenges span generation of the data, preparation for analysis, and policy-related challenges in its sharing and use, including the following:

- Dealing with highly distributed data sources,
- Tracking data provenance, from data generation through data preparation,
- Validating data,
- Coping with sampling biases and heterogeneity,
- Working with different data formats and structures,
- Developing algorithms that exploit parallel and distributed architectures,
- Ensuring data integrity,
- Ensuring data security,
- Enabling data discovery and integration,
- Enabling data sharing,
- Developing methods for visualizing massive data,
- Developing scalable and incremental algorithms, and
- Coping with the need for real-time analysis and decision-making.

National Research Council. 2013. *Frontiers in Massive Data Analysis*. Washington, D.C.: The National Academies Press.

References

- National Research Council. 2013. *Frontiers in Massive Data Analysis*. Washington, D.C.: The National Academies Press. http://www.nap.edu/catalog.php?record_id=18374

- MAY 2014, *BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES* Executive Office of the President, http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf