Conceptual Underpinnings of Ultra-Large Scale, Unified Data-Space (unified data persistence and information search & retrieval) Management

Andrew Loebl; loeblas@ornl.gov

Needs of stakeholders

The business processes and policies that shape Exascale community interactions are part of an operational environment. Engagement directly among many disciplines of science, computer science, data management, policy and operational community (stakeholders) offers the best hope for real and positive improvement in effective technology utilization and methods outcomes. An engagement effort would be greatly aided by a unified data space that takes advantage of data attributes to enable data fusion while freeing the data from idiosyncratic model and storage constraints. Such a unified data space would become a tool for stakeholders, providing access to data for further testing, development of analysis, operations tools, and visualization capabilities. Such a solution is designed to be evolutionary, flexible and takes advantage of growing commercial interest and capabilities in data integration tools.

Data integration aims at maintaining valuable data complexity while overcoming accidental complexity caused by data silos. This accidental complexity takes the form of "physical, representational, structural, and semantic barriers among data sources, types and domains."[i] At its core, successful data integration enables improved service and operations.

Data access, integration, and security are processes characterized by a sound, high quality and sustainable resource. Coherent data integration offers stakeholders the opportunity to ensure that data (in whatever form) is strong and authoritative for its intended uses and allows stakeholders to make best use of capabilities and resources. An added consideration is the growing capacity of stakeholders to access and integrate diverse data, which puts increasing power in the hands of superempowered individuals.[ii] The discipline of data integration can be understood as "practices, architectural techniques and tools for achieving consistent access to, and delivery of, data across the spectrum of data subject areas and data structure types in the enterprise to meet the data consumption requirements of all applications and business processes."[iii]

Exa-scale futures and power

It is understood that humans lack the attention span, response time and memory (data recall) required to monitor data, to recognize important variations, and respond. While there is broad awareness of data and system integration challenges, state-of-practice solutions and approaches invariably balance the necessity of pulling all of the data into centralized repositories or dictating a specialized structure that does not meet the needs of all of users. In reality, stakeholders need to use data at multiple levels in multiple ways. A single solution is rarely sufficient to meet stakeholder needs. Current practice dictates a large expense (financial, complexity increases, ADP maintenance, etc.) for maintaining multiple schemas. Dr. Jim Gray described this challenge as the "Fourth Paradigm." Gray's first three paradigms were; experimental, theoretical and computational science. The Fourth Paradigm involves an "exaflood of observational data" which is threatening to overwhelm stakeholders and forces counterproductive actions on the part of curators. A new generation of computing tools to "manage, visualize and analyze the data" is required.[iv] The goal is not succumbing to Moore's Law with regard to computer power, but getting all of the data of whatever form and mode online and interoperable.

Large amounts of data offer unique challenges and opportunities. Industry is responding to the same, inadequate technological and data volumes pressures as the government to effectively employ large amounts of heterogeneous data for the greater good. Handling this "big data" requires "a row-based data store powered by massively parallel processing (MPP) engines, or -- even better, according to some -- MPP-based columnar data stores."[v] State-of-Practice, electro-mechanical processing may become more powerful as the power of MPP grows and develops commensurate with improvement in the performance of is data stores.[vi] In short, more diverse data, coherently integrated and holistically managed trumps any other curatorship now possible. Add to such a responsive and holistic curatorship the advanced analytics which build the system into an information generator and knowledge creation becomes possible.

Research contributions to the State-of-the-Art

Data models currently form the backbone of data architecture and are essential in the 5th Generation of information management and the 6th Generation of information technology. An additional key challenge is an agnostic form of data integration for effective and user-oriented 7th Generation use of information technology and its concomitant scale of data. Much current and past effort at integration has focused on

ontology mapping or designing universal ontologies.[vii] These efforts had some success but have not been able to overcome the real need for data to be bound in specific ways to enable certain processes, varied needs of different users, and the tendency of people to employ unique semantics. More recently automated metadata tagging, modularized and reusable processes, and data analytics have been developing to address this problem. Master Data Management (MDM) products can be employed to match "entities" across data sources to the same identity. [viii] It is also important to understand that advanced data capabilities can offer increased security while exposing appropriate data by making data about the user part of every transaction.

Stakeholders (particularly in the private sector) are realizing the potential of entity-to-identity matching and are pursuing an integration approach called Ultra-Large-Scale (ULS) Systems. The ULS System concept is built on the concepts of Dr. Jim Gray's Fourth Paradigm. [ix]

ULS Systems can be defined as "…interdependent webs of software-intensive systems, people, policies and economics." They are designed to operate at large scale, be decentralized, be developed and operated by various entities with different or even conflicting needs, and are built to evolve. "Stakeholders will not just be users of a ULS system; they will be elements of the system. The acquisition of a ULS System will be simultaneous with its operation and require new methods for its control."[x] ULS Systems, whether known by this name or another, are the operating environment of the future.

A Precursor to ULS

The U.S. Army has already made ULS a key focus area for the Distributed Common Ground System-Army (DCGS-A) of the future.[xi] DCGS-A uses a database management approach which is becoming the backbone of intelligence databases for command and control among DOD branches. The value of this solution is not limited to DCGS-A; it is useful for stakeholders, analysts and operators who can come to terms with the technical and operational advantages DCGS-A illustrates for data fusion.

Data fusion must:

- Present minimal barriers to incorporating new data and semantics
- Embrace all data "sources, types, models, and modalities"
- Support diverse processing by which "structural and semantic barriers are overcome to yield information and knowledge"
- Allow reuse of data, information, and knowledge from diverse perspectives[xii] of users and experts.

To Advance the DCGS-A Precursor

To achieve this operational data integration flexibility, data models must be considered from a higher level of abstraction.[xiii] The growth in data virtualization offers a window into the need to abstract data from its original data model and data storage containers.

Successful data-integration solutions fit the business processes of users. Operations processes "include data collection, semantic enhancement, fusion from data to information to knowledge, and communication/collaboration to create understanding."[xiv] A "Unified Data Space" facilitates climbing the knowledge pyramid to enable data to exist unmodified by the shape of the data storage container while retaining its key identifying information (the data about the data or the Metadata). In this construct, data is not just integrated, it is unified. This solution preserves the sources' original data and semantics, uses diverse data of any type, can modify sources readily for evolutionary flexibility, and supports powerful processing "without limitations." Current solutions require intense "pre-integration processing (schema harmonization and data normalization) and usually entail loss/distortion of original data and semantics."[xv] This heavy processing limits data fusion due to forcing the data back into a new data schema.

Current state-of-practice in data integration is inefficient and limited to the data structures employed. This state slows analysis. At worse, these solutions cost too much to maintain, yield to difficulties in detection and prevention of corruption of data, and result in decisions with no measurable outcomes. Fusion centers or clouds with access to many discreet data stockpiles are unusable or unused. One need only review the example cited by *Information Week*[xvi], reflecting on the Department of Homeland Security's Inspector General Report. In a fiscally constrained environment, it is irresponsible to ignore planning for integration of the most valuable data in a manner more elegant and powerful than e-mailing or posting briefings for happenstance retrieval among constituents. Analysts need much more powerful data discovery and integration capabilities. This is the essence of data mining as an expression of science and knowledge accumulation.

[i] M. Andrew Eick and Suzanne Yoakum-Stover, "Fixing Intel and Operationalizing Data – The Data & Processing Syndicate," www.imintel.org, accessed 6 Oct 2010.
[ii] Daniel Goure, "Wikileaks Dilemma: How Does a Nation Fight a Superempowered Person?" Lexington Institute Early Warning Blog, 6 Dec 2010, http://www.lexingtoninstitute.org/, accessed 8 Dec 2010. This article references Thomas Friedman's *The Lexus and the Olive Tree* definition of a superempowered individual and asserts that Julian Assange may be the "first truly superempowered individual."
[iii] Ted Friedman, Mark Beyer, and Eric Thoo, "Magic Quadrant for Data Integration Tools," Gartner, 19 Nov 2010, http://www.gartner.com/technology/mediaproducts/reprints/sas/vol7/ article4/article4.html, accessed 28 Nov 2010.

[iv] Dr. Jim Gray as quoted by John Markoff, "A Deluge of Data Shapes a New Era in Computing," *New York Times* (December 15, 2009): D2.

[v] Stephen Swoyer, "Crunching the Numbers on Big Data," The Data Warehousing Institute (TDWI), 1 Dec 2010, 1, http://tdwi.org/articles/2010/12/01/crunching-big-data-numbers.aspx?admgarea=news, accessed 4 Dec 2010.
[vi] P.W. Singer, *Wired for War: The Robotics Revolution and Conflict in the 21st Century (*NewYork, NY: The Penguin Group, 2009), 102.

[vii] Leo Orbst, Mitre Corporation, e-mail interview 6 Dec 2010. An ontology is an organization of some knowledge domain that contains all relevant entities. Ontology mapping links the individual entities to each other. A universal ontology seeks to identify all possible entities of interest across knowledge domains. The challenge is pre-identifying all possible ways in which data entities interrelate or even creating fully exhaustive ontology.
[viii] Arnon Rosenthal, MITRE Corporation, 6 Dec 10 e-mail interview. His concerns were echoed by many others, but were the clearest depiction of the potential that we will throw data integration efforts out with the Wikileaks response. Data integration does involve exposing data, but this author posits that such exposure if combined with security solutions, to include data about the user and what the user is doing with the data, can provide a better solution overall for knowledge creation. A detailed discussion of security implications and data integration would be a worthwhile paper in its own right. Dr Rosenthal identified development of "rational ways to justify and manage risk/reward as a basis for access decisions" as an area worthy of further focus.

[ix] *Ultra-Large-Scale Systems: The Software Challenge of the Future*. Study lead Linda Northrup. Pittsburgh, PA: Carnegie Mellon Software Engineering Institute, June 2006, ix-3, http://www.sei.cmu.edu//library/abstracts/books/0978695607.cfm, accessed 23 Sep 2010.

[x] "Ultra-Large-Scale Systems Overview," Software Engineering Institute, Carnegie Mellon. http://www.sei.cmu.edu/uls, accessed 9 Sep 2010.

[xi] Suzanne Yoakum-Stover, "Trends in Infrastructure: Commercial vs Military." Lecture. National Association of Broadcasters Military & Government Summit, Las Vegas, NV, 13 April 2010.

[xii] All characteristic of the ideal data integration future are from Norbert Antunes, Tatiana Malyuta, and Suzanne Yoakum Stover, "A Data Integration Framework with Full Spectrum Fusion Capabilities," August 2009, 2-3

[xiii] Ibid, 3.
[xiv] Yoakum-Stover, "DDF 2009," 2.

[xv] Norbert Antunes, Tatiana Malyuta, and Suzanne Yoakum Stover, "A Data Integration Framework with Full Spectrum Fusion Capabilities," August 2009.
[xvi] Alice Lipowicz, "Fusion Centers Hampered by Limitations of DHS nets, IG says," *Federal Computer Week*, 16 Nov 2010, http://fcw.com/article/2010/11/16/dhs-fusion-centers.aspx, accessed 5 Jan 2011.