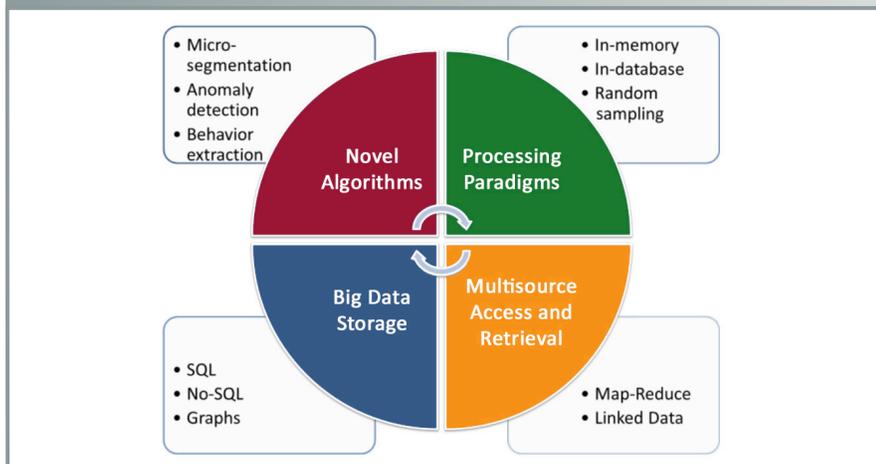


Big Data App Store: Democratizing Data-Science with On-Demand Analytics

UT-B ID 201202953, 201403307



Technology Summary

Web 2.0, the Internet as we know it today, was not constructed to handle the growing amounts of data and increasingly large data sets known as “Big Data.” Further complicating things, many of the data sets are either unstructured (80% by one estimate) or stored in distributed resources with different schemas, formats, and structures. Thus we aren’t realizing all of the advantages possible with Big Data and high-performance computing. Data-driven enterprises, from academia to government to private industry, need tools to seamlessly integrate intra- and inter-enterprise data sets to extract “nuggets” of data (patterns and behaviors) previously, not discovered because of format or location, for analysis.

Researchers at ORNL have created a suite of what they call “knowledge catalytic” tools to seamlessly integrate intra- and inter-enterprise data sources to extract these nuggets and create actionable information for decision making. Using the techniques of the emerging Web 3.0, such as semantic interfaces; powerful models and algorithms for manipulating data; and the power of high-performance parallel computing, these tools speed up the pattern discovery (feature extraction) process and the pattern recognition (learning) process across disparate data sources. The tools can be viewed as an app store of algorithms that can be combined to create virtual databases for data-driven discoveries and decisions.

Based on algorithms that scale to multiple platforms, the various tools include a fuzzy logic toolkit for matching schema attributes; graph-processing tools (“crosswalk keys”) for field-level linkages across databases and data sources; tools for transforming unstructured data into structured forms that can then be analyzed; tools for automated clustering of data elements across different data sources; and models and algorithms for sensing connections, anomalies, and emergent behaviors. Two key elements are (1) visual display capabilities for analyzing various potential links between data/data sets and (2) a knowledge recorder tool that captures search information and analyst comments, which ultimately will save enterprise resources and time by avoiding analytic silos and re-searching/rediscovering what other analysts have already learned from data.

In the era of Big Data, not everyone can afford to have an operational information technology department, and not everyone can afford expensive data scientists. The ORNL approach targets those customers that have data and cannot afford Big Data systems but can benefit from the insights. The app-store model allows people to rent computational tools and analytics on demand, paying for what they need.

Advantages

- Data and domain agnostic
- Takes advantage of the latest advances in high-performance computing
- Builds on the evolving technology of semantic web interfaces and Linked Data

Potential Applications

- Enterprise data sharing and discovery
- Virtual schema for distributed data sets
- Intelligence and law enforcement
- Counterterrorism
- Social network analysis
- Fraud detection and forensics

Patents

Sreenivas R. Sukumar, Regina K. Ferrell, and Mallikarjun Shankar. *Knowledge Catalysts*, US Patent Application 14/089,395, filed November 25, 2013.

Sreenivas R. Sukumar. *Scalable Pattern Search in Multi-Structure Data*, US Patent Application 62/106,342, filed January 22, 2015.

Inventor Point of Contact

Sreenivas R. Sukumar
Computational Sciences and Engineering Division
Oak Ridge National Laboratory

Licensing Contact

David L. Sims
Technology Commercialization Manager
Technology Commercialization
UT-Battelle, LLC
Oak Ridge National Laboratory
Office Phone: 865.241.3808
E-mail: simsd@ornl.gov

To view this and other ORNL inventions, visit ORNL Partnerships at <http://www.ornl.gov/connect-with-ornl/for-industry/partnerships/technology-licensing/available-technologies>