Discovery Informatics at PNNL

Kerstin Kleese van Dam

Over the past few years Pacific Northwest National Laboratory has focused its research and development activities in particular on data intensive science, we have therefore a strong interest particularly in knowledge discovery in very large or very complex data sets, repositories or community data collections (incl. distributed ones). It is our aim to create research environments in which users can easily discover and assess existing data and analyze it in the context of their own research results, both individually or as part of a wider group or community. We see that the demand for tools that support multidisciplinary research will strongly increase over the coming years and have started to investigate in particular themes such as multi-context data representation and multidisciplinary analysis in response. We currently support data intensive projects in a wide variety of domains such as chemistry, biology, climate, future power grid and high energy physics, to explore observational measurement, experimental and computational data. Next to software research and development activities, PNNL is also heavily vested in the exploitation of new architectures such as Cray XMT, Netezza Twin Fin or Niagra2 that are more conducive to data intensive science knowledge discovery applications through its CASS-MT center http://cass-mt.pnnl.gov/ .

Key Science Drivers for PNNL e.g.:

Chemical Imaging – Science is moving towards ever more complex and detailed measurement techniques to determine the structure and property of samples from single molecules to microbial communities. Major challenges are the exponentially increasing data volumes, a need for real time analysis of the data to enable experimental steering and the ability to analyze the results from different imaging techniques together.

Climate and Energy – An increasing set of projects is investigating strategies for a secure energy provision against the background of climate change scenarios. Major challenges are decreasing uncertainty in climate models, characterization of uncertainty in results, coupling of multidisciplinary codes and the collaborative analysis of their results.

Future Power Grid – Scientists are investigating how a future power grid with many more sensors can be operated more safely and efficiently with the help of IT. Major challenges are the real time data transfer and analysis requirements of the system, security, fast access to past scenarios that match a specific situation.

High Energy Physics – PNNL is the US lead of the international BELLE II experiment, which is expected to produce 250PB by 2019 alone. Major challenges are data transfer, management and distributed analysis.

Biology - Systems Biology has a long tradition of utilizing diverse research results from various experimental and computational methods, stemming from a variety of site. Unfortunately these sources of scientific knowledge are mostly characterized by their diversity of data formats, representation and access methods, making it difficult indeed to identify all relevant data sources for a given topic, assess their quality, match them with suitable analysis and integrate them into the scientific research process.

To be able to address these challenges above we are currently engaged in the following ***broad research activities:***

- Orchestration of in situ tasks through scalable workflow engines
- Scalable provenance for complex and data rich environments – incl. Exascale, Provenance driven Resilience, Provenance Propagation
- Uncertainty quantification research
- Collaborative Framework for Accelerating Science and Innovative Solution – collaborative analysis, knowledge management, workflows, provenance
- Data management and analysis for experimental facilities – real time analysis, co-analysis of heterogeneous and distributed results, simulation aided experimental planning and analysis
- Semantics driven analysis and inferencing in scientific data
- Analytic Repositories – Annotated Data repositories with User driven analysis capabilities
- Scalable Analytics – Visual analytics, Analysis at Scale (100's of PB – e.g. BELLE II - High Energy Physics), Distributed Analysis, Push Analytics, Analysis of complex networks, including scalable graph algorithms and analysis of both streaming and historical power grid data
- Scalable coupling of multi-scale, cross domain scientific codes as part of iRESM and other activities

These research activities are supported by the development of ***core tools and capabilities***:

- OASCIS - Open source framework for Accelerating Scientific Collaboration and Innovative Solutions combining VELO, CAT, PROVEN and MeDICi
- Velo: a collaborative platform for modeling and simulation, incorporating the ability to plug in domain specific data types and analysis tools (very flexible), share data across Web and desktop environments, and capture full provenance of data and analyses
- CAT: Collaborative Analytical Tool Box, knowledge management system driven data sharing and analysis
- PROVEN: Provenance Environment, scalable collection and analysis of provenance information from experimental and computational sources.
- MeDICi/Kepler: workflow systems – MeDICi is particularly developed for data intensive applications
- ISDE – Interactive Software Development Environment - Library based development tool for the flexible generation of analysis codes and pipelines
- VESPA – Visual Analytics platform for proteomics
- ROM Builder: a desktop tool for creating reduced order models from high fidelity simulations. Can be extended to work with any simulator (e.g. CFD, climate, subsurface), and includes multiple sampling and regression methods for building and validating ROMs