GSAW 2015 Tutorial E:

Big Data Considerations for Ground System Environments

Length: Full day

Overview:

It is now widely understood that developing scalable system architectures for data intensive environments is becoming an essential skill for the next generation of space and science-driven observational systems. As more capable instruments returns significantly increasing sizes of data, organizations need work forces that are trained in software and data architectures, technologies for managing massive data, and methods for the data reduction, and visualization. Planned topics to cover, in no particular order, include: databases and data management systems, computational architectures, data mining and machine learning, data visualization, data transport, open source, web services, grids and clouds, statistics tools, semantic web, etc. A particular focus will be on addressing approaches for scalability across the entire data lifecycle for missions. The outline for the course will be as follows:

I. Introduction to Big Data
II. Architectural Considerations Across the Data Lifecycle
III. Information Models and Model-driven Approaches to Big Data
IV. Cyberinfrastructures (including data processing) and the Role of Open Source
V. Cloud Computing
VI. Data Analytics and Machine Learning
VII. Visualization
VIII. Applications of Big Data across different observing environments (spaceborne, airborne, etc)

Reference material will be provided.

Instructors: Daniel Crichton, Emily Law, J. Steven Hughes, Shan Malhotra, Thomas Huang, Thomas Fuchs, Chris Mattmann, and Scott Davidoff, NASA/Jet Propulsion Laboratory, California Institute of Technology

Biographies:

Daniel J. Crichton is a program manager and principal computer scientist at NASA JPL. He is the Director of the Center for Data Science and Technology and is actively involved in leading the development of several data intensive systems for planetary, Earth Science, and biology. He has appointments to multiple program offices in earth science, planetary science and technology, coordinating JPL's efforts in data science. He served on the National Research Council on the Committee for the Analysis of Massive Data that produced the report on Big Data and is currently serving on NASA's roadmap team for the Office of Chief Technologist covering the area of Big Data. His interest areas include novel software architectures and methods for distributed data management and analysis.

Emily Law has over 20 years of experience in the architecture and development of data systems. She joined JPL in 1996 first involved in the Deep Space Network serving as lead developer, and later on was appointed as the System Service Manager for the Network Monitor and Control Program. In 2005, she joined the Planetary Data System as the Operations Manager, and served a dual role in managing Physical Oceanographic DAAC's Operations. In 2008, she was appointed as the Deputy Manager for the

Data Systems and Technology Program. In addition, she is serving as the Chair of the NASA Earth Science Data Systems Cloud Computing Working Group, as well as the Vice President of the Earth Science Information Partners Federation.

J. Steven Hughes is a Principal Computer Scientist at the Jet Propulsion Laboratory focusing on architecting and implementing data intensive systems in complex distributed heterogeneous environments. He is currently involved in the development of digital data archives for the National Aeronautics and Space Administration (NASA) and is the lead for information model development. He has been involved in international data standards development and his current interest is information model driven system development and interoperability. He is a member of the Primary Trusted Digital Repository Accreditation Board (PTAB) and has participated in ISO 16363 audits of digital repositories. Mr. Hughes holds an MS and BS in Computer Science. He is a Senior Member of the ACM.

Shan Malhotra is a Principal Engineer at JPL. He has been a principal architect on many data system projects including the Lunar Modeling and Mapping Portal and the Service Preparation Subsystem for the Deep Space Network.

Thomas Huang is the Technologist for the Physical Oceanographic Distributed Active Archive Center. He leads the Data Intensive Systems Working Group for NASA and is an expert in cyberinfrastructures and visualization.

Thomas Fuchs is a Research Technologist at JPL. His research interests include development of new statistical learning algorithms and their application in computer vision, astronomy, computational pathology, and robotics. He runs the Caltech Machine Learning lectures and has taught introductory courses in Machine Learning.

Chris Mattmann is the Chief Architect in the Instrument and Data Systems section at JPL, an Adjunct Associate Professor in the Computer Science Department at USC, and a Director at the Apache Software Foundation. The overarching theme of his research is the design of large-scale, distributed, data intensive systems.

Scott Davidoff leads design and development of human interfaces for mission operations at NASA's Jet Propulsion Laboratory. He investigates how Data Visualization and Virtual Reality impact space exploration and tele-robotics, and is a NASA and Caltech Principal Investigator.

Description of Intended Students and Prerequisites:

Students should have a familiarity with computer programming and databases. These are fundamental skills that will be applied in the context of big data. Students who have a background in large-scale software architectures, particularly as applied to mission systems, will also have a good background.

What can Attendees Expect to Learn:

Attendees will be given an introduction to Big Data and associated capabilities including software and data architectures, cyberinfrastructures, cloud computing, machine learning and visualization. The course will also discuss the application of these capabilities to mission systems, including the challenges, tradeoffs, technologies, and considerations in developing scalable, Big Data systems.