Understanding the Usage Model of Scientific Data Analysis Tools

Wenqin Chen, Sarah Poon, Devarshi Ghoshal , Lavanya Ramakrishnan

Usable Software Systems Group, Data Science and Technology, LBNL

Motivation & Goal

Goal:

Understand users' mental models and workflows of data analysis

Motivation:

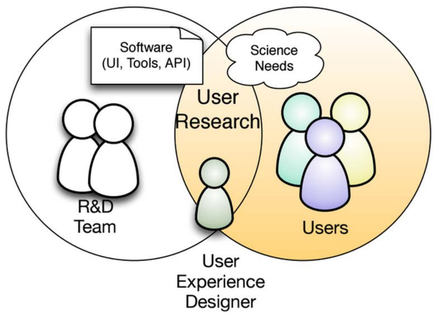
- Exascale poses new challenges such as increasing data movement costs and growing data sizes
- Important to build tools with data models that support next-generation data analysis
- First step is to understand users' mental models and workflows of data analysis

Usable Data Abstraction

Enable large-scale data analysis workflows on exascale systems using user research and ethnographic methods to

- Design and address data abstractions for next-generation exascale workflows
- Combine usability with performance, resilience and energy considerations of next-generation hardware

User Research



- User research fills in the missing link between the R&D team and their users
- Some common user research methods are user interview, participant observation, and contextual inquiry
- User research differs from requirement gathering in its holistic approach to understanding users' needs

Research Approach

- Interview and observe scientists who currently use data analysis tools on HPC to perform large-scale data analysis
- Focus on users of three data analysis tools (UV-CDAT, SPARK and R), and consider custom scripts

Tool Name	Domain	Functionality	HPC Support
UV-CDAT	Climate Science	Primarily Data Visualization	Edison, Hopper
SPARK	General-Purpose	Big Data Processing	Edison
R	General-Purpose	Statistical Computer & Graphics	Edison, Hopper

Workflow:

Extract users' data analysis workflows with focus on tasks, tools used, computing platforms, science goals, data lineage, transformation, and size.

- Interview users and talk through their workflows
- Observe users performing data analysis

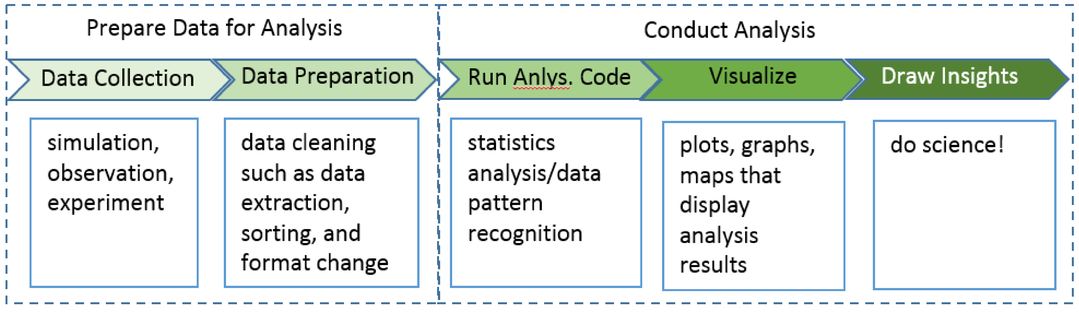
Mental Model:

Construct users' data analysis mental model by studying their experiences using existing data analysis tools on HPC.

- Reasons for switching from the previous tools
- Improvements in the workflows compared to previous tools
- Types of analysis that the current tools perform well
- Initial learning hurdles
- Adaptations made for the analysis to fit the tool
- Inefficiency in analysis workflows
- Types of analysis that are especially challenging to perform on current tools
- Recommendations for improvement

Scientific Data Analysis Workflow

An exemplar analysis workflow based on early findings:



Key Observations:

- Keeping analysis scripts organized is challenging
- Memory is analysis bottleneck
- Data management between steps is difficult
- HPC system updates cause significant productivity losses

Future Work

- The highly customized nature of scientists' data analysis requires customized interview approaches
- Continue filling out the data analysis workflow and mental model landscape
- Improve data abstraction based on user research findings
- Conduct usability evaluation on prototypes to validate proposed improvements

Acknowledgement

This work was funded by the Office of Science, Office of Advanced Scientific Computing Research (ASCR) of the U.S. Department of Energy under Contract Number DE-AC02-05CH11231.