Provenance-Aware Scalable Seismic Data Processing with Portability

Yinzhi Wang[1](iwang@tacc.utexas.edu), Gary L. Pavlis[2](pavlis@indiana.edu)
[1]The University of Texas at Austin, [2]Indiana University

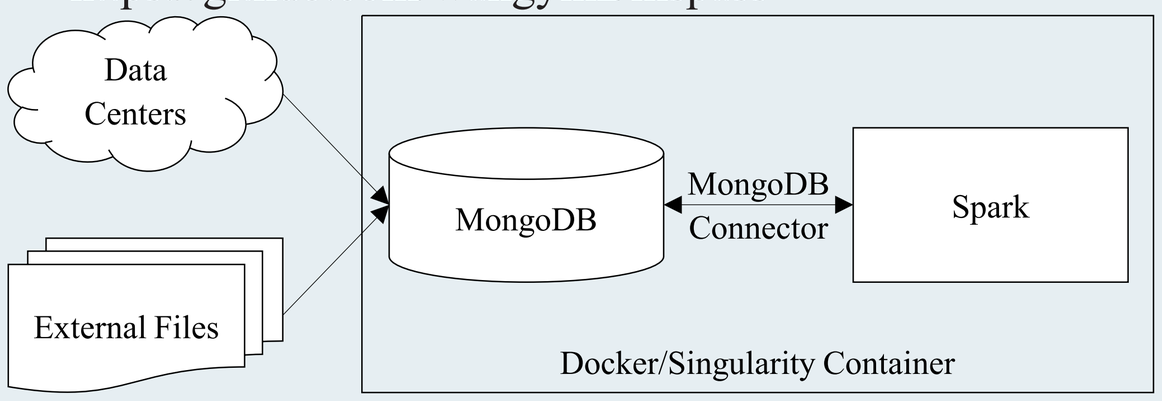
Award #: 1931352

Motivation

- Most of what we know about the Earth's deep interior comes from seismic observations
- Massive increase in data volume due to the success of large-scale projects like the USArray component of Earthscope.
- Global growth in open data through the Federation of Digital Seismic Networks (FDSN) and in the US by the Incorporated Research Institutions for Seismology
- Existing data processing infrastructure available to the seismologists is not designed to handle data volumes of this scale.

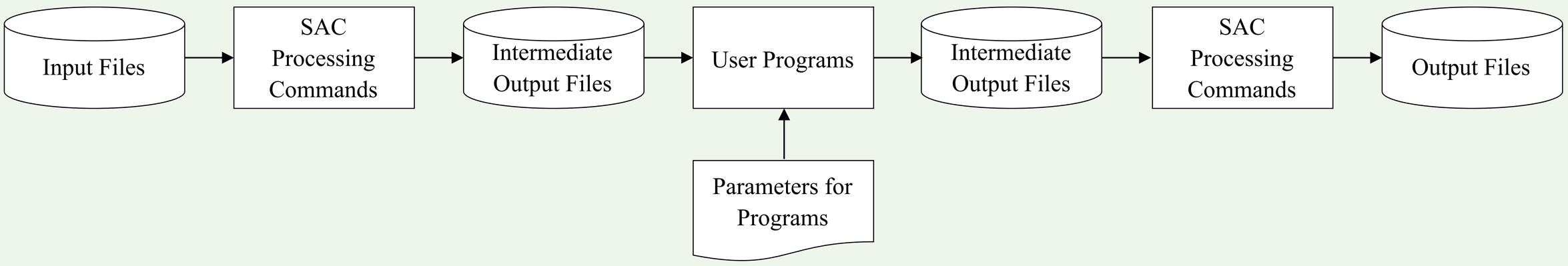
Objective

- Massive Parallel Analysis System for Seismologists: A high performance framework for data processing that can be used by all seismologists.
  https://github.com/wangyinz/mspass



A Generic Data Processing Workflow with Current Tools (Using SAC as an Example)

- Data are passed between processing steps in memory or by intermediate files.
- Many workflows use custom codes that read and write files in formats with a fixed header.
- The overall workflow is nearly always driven by a series of UNIX shell scripts



The MsPASS Architecture: High Performance and Easy to use

High Performance:

- Spark job scheduler implicitly executes the workflow in parallel.
- MongoDB's horizontal scaling make it possible to handle very large datasets and high throughput operations.
- The core Python API is wrapped around SEISPP C++ library
- Inclusion of error messages as part of the metadata makes any workflow bombproof in parallel.

Easy to use:

- Generalized headers enabled by MongoDB and YAML format.
- User interface is built on Python with provenance information automatically handled.
- Simple deployment and maintenance through the use of container technology.