Dynamic Data-Driven Applications using Workflows as a Programming Model for Scalable and Reproducible Streaming and Steering

Ilkay Altintas
San Diego Supercomputer Center, UC San Diego
altintas@sdsc.edu

Over the last decade, scientific workflows and dataflow systems have emerged as a successful model for stream processing, especially in scenarios where a scalable and reusable integration of streaming data, analytical tools and computational infrastructure is needed. There are many applications currently using workflows successfully to process streaming data, often defined as Big Data due the *volume*, *velocity* and/or *variety* of the data to be processed. However, some requirements still remain unsatisfied for wider utilization of automation methods for dynamic steering of data-driven applications:

- The volume, rate and variability of Big Data makes traditional Extract, Transform and Load (ETL) methods not dynamic enough. Effective interfaces to streaming data systems and middleware are needed for dynamic and scalable management and integration of streaming data sources. A more in-depth understanding of data systems, formats, standards and metadata needs to be built into workflow systems, at the level of planning data operations along with the planning for computing of downstream analytical processes. In addition, the ability to dynamically convert streams into scientific and analytical data types within workflows makes automated processing of such data possible at the rates the data is being produced.
- Our ability to process such data depends our ability to understand the Big Data ecosystem including the platforms and tools used for Big Data management and analysis. Big Data processing platforms, like Hadoop and Spark, and their associated stacks of tools, become indispensible components in the overall system to provide efficient and robust processing. A provenance-aware view over these Big Data platforms is needed for in-depth tracking of data being processed in these platforms. Programmability and dynamic execution scalability over Big Data and cloud systems along with traditional high-performance computing (HPC) systems is needed. Timespan, cost or energy usage optimization of workflow-driven applications based on data or compute needs is a growing demand in many application domains. Workflow systems must construct the best execution plan, based on rules and cost estimates, and then execute it by scheduling the appropriate available compute resources for the task.
- Data and process provenance as well as data quality, often referred to as the *veracity* of Big Data, is critical to many capabilities for streaming data processing including experiment validation, reusability and reproducibility, fault tolerance, process optimization and performance prediction. Many challenges remain for modeling, recording, storing and querying provenance information for streams.

An example project that uses such workflows is the NSF-funded (1331615) WIFIRE project (wifire.ucsd.edu) that is building an end-to-end cyberinfrastructure for real-time and data-driven simulation, prediction and visualization of wildfire behavior. As shown in Figure 1, using Kepler workflows (kepler-project.org) as the underlying programming and system integration interface, the project integrates networked observations, e.g., heterogeneous satellite data and real-time remote sensor data with computational techniques in signal processing, visualization, modeling and data assimilation to provide a scalable, technological, and educational solution to monitor weather patterns and to predict a wildfire's Rate of Spread.

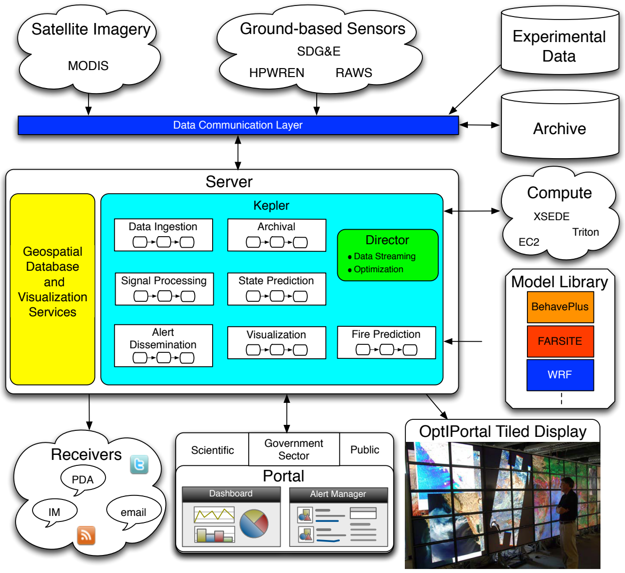

Figure 1. Integrated real-time data processing and programming environment in WIFIRE for monitoring, modeling and prediction of wildfire spread.