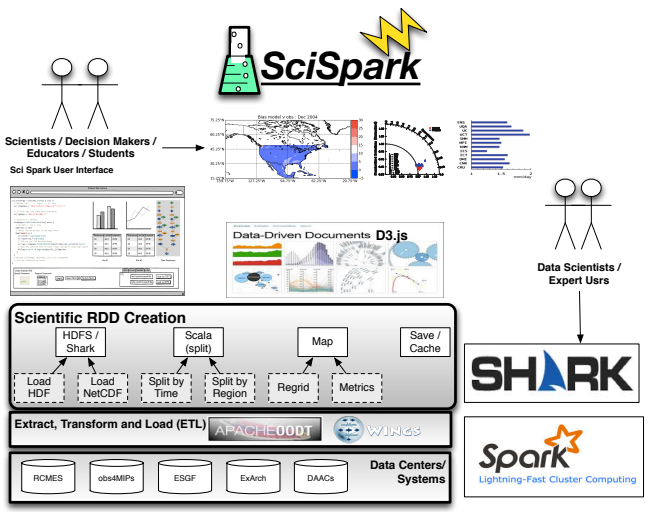
SciSpark: Highly Interactive and Scalable Model Evaluation and Climate Metrics for Scientific Data and Analysis

PI: Chris A. Mattmann, JPL

Objective

- Leverage Apache Spark big data infrastructure to perform complex model evaluation and climate metric analyses 100 times faster than state-of-the-art file systems
- Provide interactive access and analysis of data for increased understanding of regional climate systems
- Enable fast processing and analysis of highly spatially and temporally resolved observational and model datasets by reducing the number of data management operations
- Mature the SciSpark technology by exploiting the Resilient Distributed Datasets (RDD) functionality enabling operational reuse, distributed processing, and data recovery after operational failures.



SciSpark Architecture

Accomplishments

- Demonstrated SciSpark capability for significant performance improvement (greater than 100 times) for MCC extreme weather event Tera scale search and analysis from months to hours.
- Designed SciSpark 2.0 Application Programming Interface to support in-memory analytics and provenance support.
- Prototyped and demonstrated interactive (Apache Zeppelin based) notebooks to couple visualizations and analysis.
- Delivered two Spark-based implementations of iterative scientific algorithms – Mesoscale Convective Complexes (MCC) Search and Parallel Probably Density Function (PDF) clustering.
- Developed three-course curriculum/training in SciSpark given at ESIP Summer 2016 meeting.
- Software available on Github, under Apache License version 2, enabling community contribution, and sustainability.

Co-Is/Partners: Y. Gil, USC/ISI; J. Kim, UCLA JIFRESSE; H. Lee, P. Loikith, L. J. McGibbney, B. Wilson, E. Fetzer, D. Waliser, K. Whitehall, JPL

TRL_{in} = 3 TRL_{out} = 5