

[SciSpark 201: Searching for MCCs](#) [1]



Submitted by ChrisMattmann on Fri, 2016-02-12 13:01 Friday, February 12, 2016 - 13:00

Event: [2016 ESIP Summer Meeting](#) [2]

Session Type: [Breakout](#) [3]

Expertise Level: [Beginner](#) [4]

Collaboration Area: [Cloud Computing](#) [5]

[Data Management Training](#) [6]

[Geospatial](#) [7]

[Science Software](#) [8]

Abstract/Agenda:

We introduce a 3 part course module on [SciSpark](#) [9], our AIST14 funded project for Highly Interactive and Scalable Climate Model Metrics and Analytics. The three part course session includes 101, 201, and 301 classes for learning how to use Spark for science.

SciSpark 201 is a 1.5 hour session in which we will use the search for Mesoscale Convective Complexes (MCCs) in Satellite Infrared data to show a real world example of how SciSpark enables real time response to both search queries and modifications to the underlying code. This task is representative of the motivation behind SciSpark - iterative data-reuse algorithms that share information between multiple stages.

- Whitehall, Kim, et al. "Exploring a graph theory based algorithm for automated identification and characterization of large mesoscale convective systems in satellite datasets." Earth Science Informatics 8.3 (2015): 663-675.
- [Implementation of Grab Em', Tag Em', Graph Em' \(GTG\) algorithm](#) [10] in Python.

Notes:

Note for SciSpark 201

A two-pronged approach to Spark

1. The goal of scientific RDD(sRDD)? The scientific Resilient Distributed Dataset (sRDD), exploits Apache Spark's concept of RDDs for multi-dimensional data representing a scientific measurement that can be subset by time, or by space. The sRDD supports multidimensional data and processing of scientific algorithms in the MapReduce paradigm within a distributed environment.

2. sciTensor

The sciTensor datatype is a self-documented array that keeps a list of arrays for a variable arrays and maintains associated metadata in a hashmap. The sciTensor is read into the sRDD and the data within is operated on via arithmetic and relational operations. sciTensor can load data from: HDFS, OpenDap, and local FS.

3. Demo

Data visualization:

Scala RDD -> Python RDD -> python visualization

Use case: Mesoscale convective complexes

- Data: brightness temperature data
- Nodes: areas with a given brightness temperature value and a given size

SciSpark 201: Searching for MCCs

Published on Commons (<http://commons.esipfed.org>)

- Edges: determined by area overlaps between nodes within consecutive time periods
- Identify nodes and edges
- Find cloud elements and connect the cloud elements between frames.
- Find the subgraphs of cloudy areas that have evolve in time

Attachments/Presentations:  [SciSpark 201.pdf](#) [11]

Presenters:

Name: [Chris Mattmann](#) [12]
Organization(s): [Jet Propulsion Lab](#) [13]
Email: chris.a.mattmann@jpl.nasa.gov
[14]

Name: [Brian Wilson](#) [15]
Organization(s): [Jet Propulsion Lab](#) [13]

Notes takers:

Name: [Fei Hu](#) [16]
Organization(s): [George Mason University](#) [17]
Email: fhu@gmu.edu [18]

Creative Common License: Creative Commons Attribution 3.0 License

Teaser: SciSpark 202: MCC graph search and PDF k-means clustering of Big Data!

Accepted:

Keywords: [SciSpark Apache Spark Mattmann MCC search GTG Whitehall Loikith](#) [19]

Source URL: <http://commons.esipfed.org/node/9029>

Links:

- [1] <http://commons.esipfed.org/node/9029>
- [2] <http://commons.esipfed.org/2016SummerMeeting>
- [3] <http://commons.esipfed.org/session-type/breakout>
- [4] <http://commons.esipfed.org/taxonomy/term/260>
- [5] <http://commons.esipfed.org/collaboration-area/cloud-computing>
- [6] <http://commons.esipfed.org/collaboration-area/data-management-training>
- [7] <http://commons.esipfed.org/collaboration-area/geospatial>
- [8] <http://commons.esipfed.org/taxonomy/term/1310>
- [9] <http://sciscspark.jpl.nasa.gov>
- [10] <https://github.com/kwhitehall/grab-tag-graph>
- [11] <http://commons.esipfed.org/sites/default/files/SciSpark%20201.pdf>
- [12] <http://commons.esipfed.org/node/8937>
- [13] <http://commons.esipfed.org/taxonomy/term/197>
- [14] <mailto:chris.a.mattmann@jpl.nasa.gov>
- [15] <http://commons.esipfed.org/node/252>
- [16] <http://commons.esipfed.org/node/8920>
- [17] <http://commons.esipfed.org/taxonomy/term/836>
- [18] <mailto:fhu@gmu.edu>
- [19] <http://commons.esipfed.org/taxonomy/term/2153>