Benchmark Comparison of Cloud Analytics Methods Applied to Earth Observations

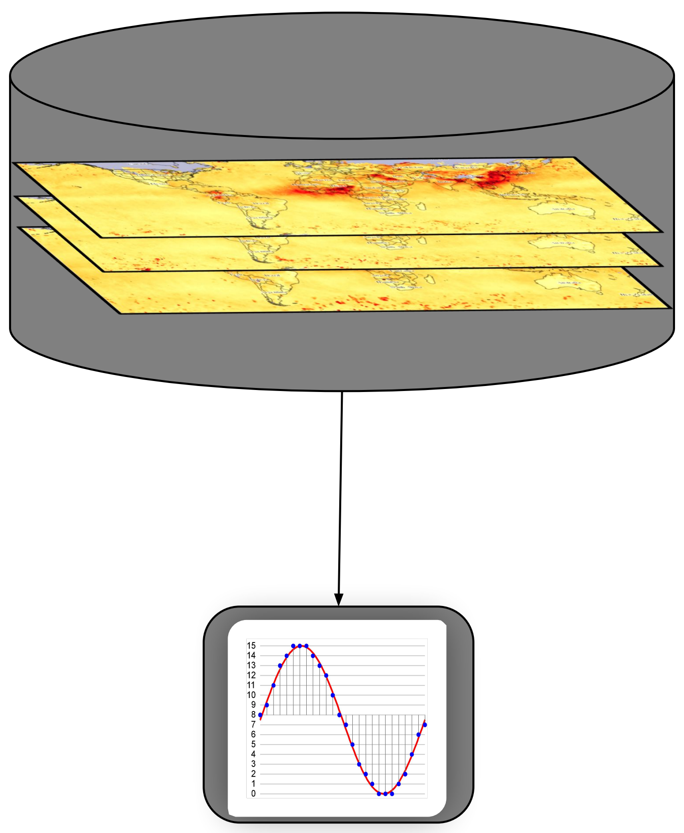
IN13B-1658
AGU Fall Meeting, 2016, San Francisco

Chris Lynnes, NASA/GSFC
Mike Little, NASA/HQ
Thomas Huang, NASA/JPL
Joseph Jacob, NASA/JPL
Phil Yang, George Mason University
Kwo-Sen Kuo, NASA/GSFC

Key Points:

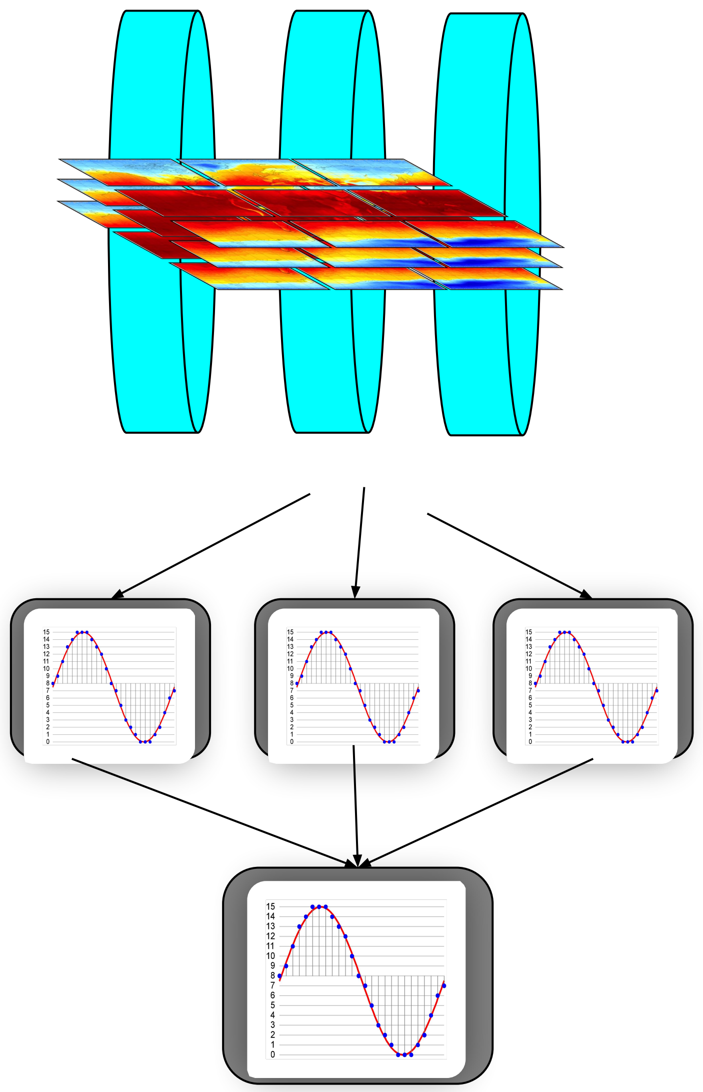
- Cloud computing has the potential to bring high performance computing capabilities to the average science researcher.
- In order to take full advantage of cloud capabilities, the science data used in the analysis must often be reorganized across multiple nodes via cloud-based filesystems or cloud-enabled databases to enable relatively fine-grained parallelism.
- Since storing an extra copy of data leads to increased cost and data management complexity, NASA is interested in determining the benefits and costs of various cloud analytics methods for real Earth Observation cases.
- NASA's Earth Science Technology Office and Earth Science Data and Information Systems project have teamed with cloud analytics practitioners to run a benchmark comparison on cloud analytics methods using the same input data and analysis algorithms.
- We have particularly looked at analyses that use long time series, because these are particularly intractable for many Earth Observation datasets, which typically store data with one or just a few time steps per file.

Giovanni
(Conventional Architecture)



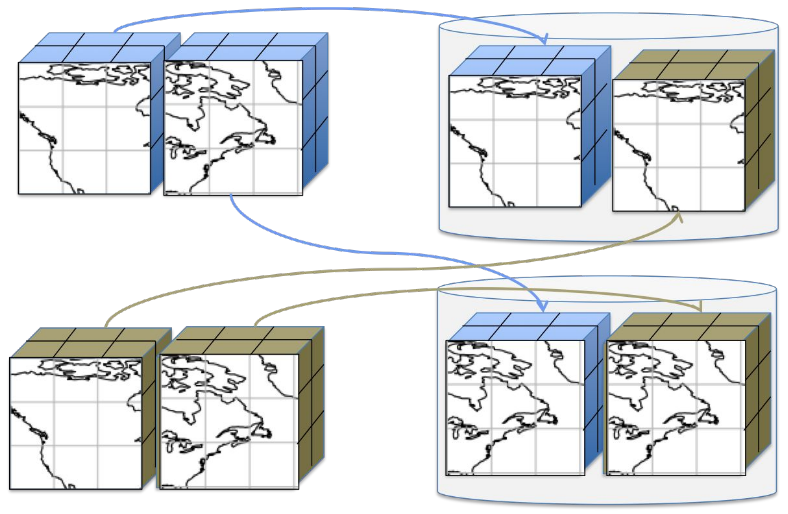
The current Giovanni system provides web-based analysisi in a conventional server architecture. Data are stored as time slices of the whole globe, usually one per file.

NEXUS
Spark + Cassandra



The NEXUS system breaks global datasets up into spatio-temporal tiles, storing them in a Cassandra Databse. Spark is used a highly parallel MapReduce-style algorithm.

SciDb



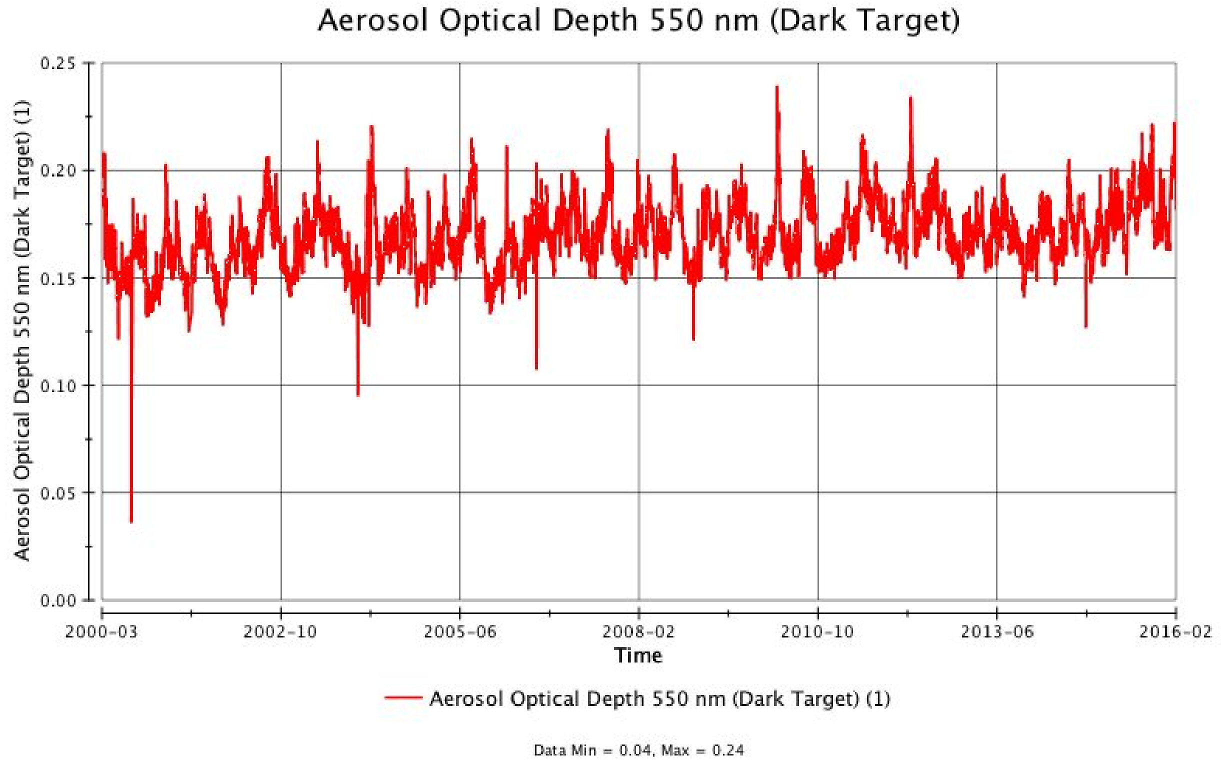
SciDb is an open-source column databaes supported commercially. Similarly to NEXUS, It stores adjacent data (coordinate-wise) in chunks (tiles). For multi-variable files,

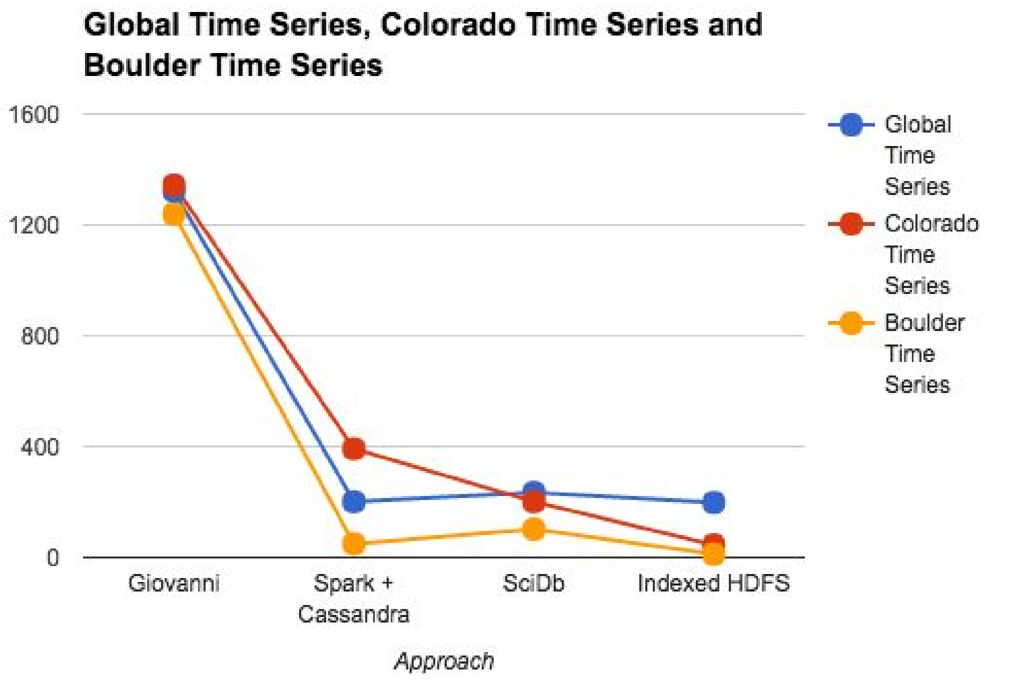
Dataset: MODIS Terra Daily L3 Atmospheres, Collection 6
Variable: Aerosol_Optical_Depth_Land_Ocean_Mean
Date Range: 1 March 2000 to 29 Feb 2016
Task #1: Time Series

Long time series are especially problematic because both models and observational typically output data by timestep. There will be 3 subtasks:
1a) Point Time series for Boulder, CO
1b) Area-averaged time series for the state of Colorado
1c) Area Averaged time series for the globe

For this case, Area Averaging means computing weighted averages, where weights = cos(lat).

Task #2: Climatological Map: For each month compute a global map of the average values of the chosen variables over the time period in question.


Aerosol Optical Depth 550 nm (Dark Target)
Aerosol Optical Depth 550 nm (Dark Target) (1)
Time
Aerosol Optical Depth 550 nm (Dark Target) (1)
Data Min = 0.04, Max = 0.24


Global Time Series, Colorado Time Series and Boulder Time Series
Giovanni
Spark + Cassandra
SciDb
Indexed HDFS
Approach
Global Time Series
Colorado Time Series
Boulder Time Series