Astro-DISC: Data-Intensive Analytics for Astrophysics

Julio López, Bin Fu, Eugene Fink, Swapnil Patil, Wittawat Tantisiriroj, Milo Polte, Lin Xiao, Vijay Vasudavan, Garth Gibson

Overview

- New discoveries and science driven by large data analytics
- Increasingly larger datasets: Terabytes to Petabytes
- Help scientists analyze their data to shorten time to science
- Enable quick analysis of massive datasets
- Need large distributed resources
- Analysis across multiple data sources

Astro Analytics

- Properties of large-scale structures in the universe
- Scalable group finding
- Multi-tree N-point correlation functions
- Merger trees
- Fourier domain decomposition
- Matched filter searches
- Tracking particle history in cosmology datasets
- Finding patterns of anomalies

Astrophysics Datasets

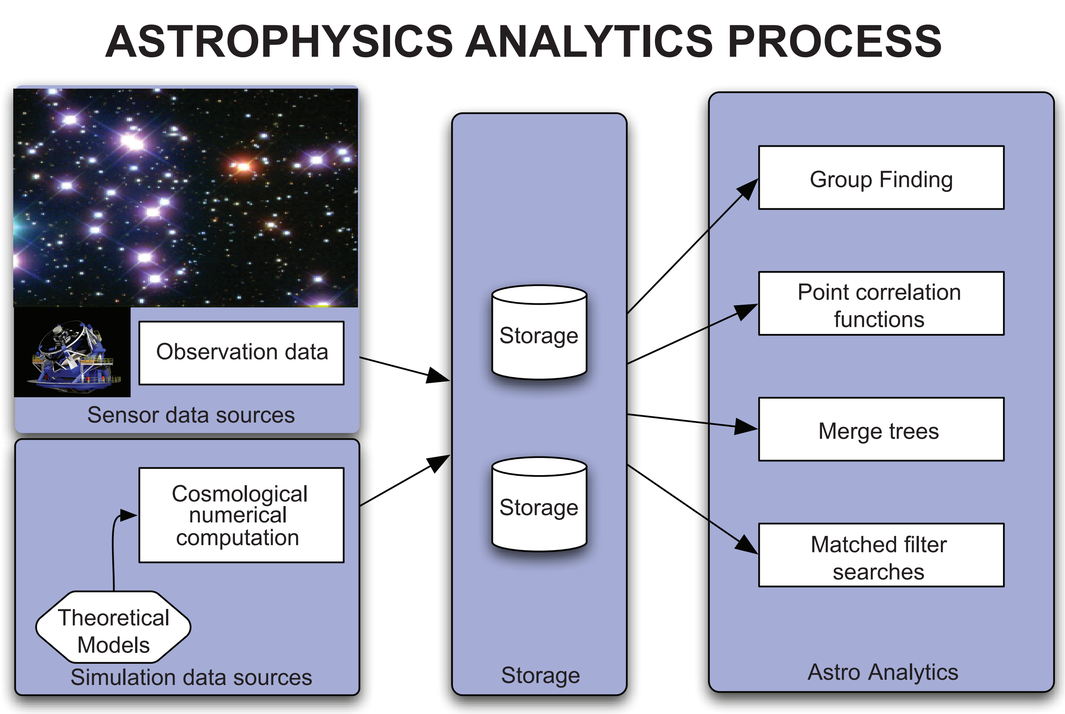
Name	Size	Source
Sloan Digital Sky Survey (SDSS)	50 TB	Obs
Bigben BHCosmo	30 TB	Sim
Millennium Simulation	20 TB	Sim
Coyote Universe	50 TB	Sim
Pan-STARRS	30 TB/yr	Obs
Murchison Widefield Array	2 GB/s	Obs
Roadrunner Universe	10 PB	Sim
Bluewaters BHCosmo	50 PB	Sim
Large Synoptic Survey Tel (LSST)	40 PB	Obs
Ultimate Dark Matter Sim.	100 PB	Sim
Square Kilometer Array	10 PB/yr	Obs

Requirements

- Challenges: Scalability, programmability
- Scale to large datasets
- Easy-to-use programming interface
- Deal with processing failures

Approach

- Commonly used operations and structures
- New scalable algorithm for astro analytics
  - E.g., group finding, point correlation functions
- Simple-to-use scalable abstractions
- Scalable spatial indexing support atop table software
  - E.g., HBase, Cassandra, HyperTable
- Data-Intensive Scalable Computing architecture
  - Clusters with many computers
  - Designed for data-intensive operations
- Leverage open-source distributed processing frameworks:
  - E.g., parallel FSs, Hadoop, HDFS, performance libs



Supported by a Google Grant

Carnegie Mellon

Parallel Data Laboratory

R09