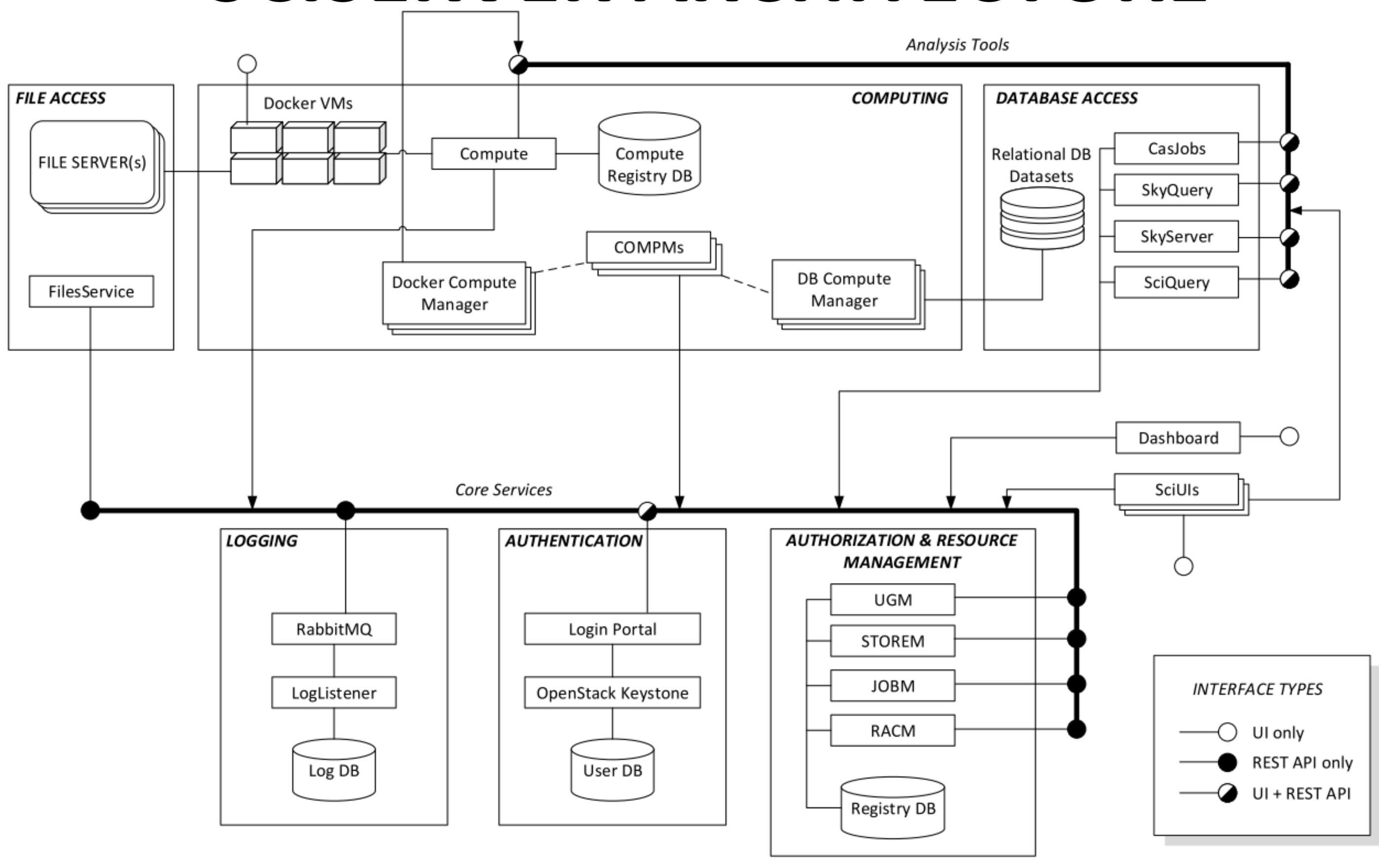
Award #: 1261715

DIBBS: Long Term Access to Large Scientific Data Sets: The SkyServer and Beyond

PI: Alex Szalay, Co-PIs: Randal Burns, Charles V. Meneveau, Michael Rippin, Ani Thakar
Institutions: Johns Hopkins University Presenter: Gerard Lemson

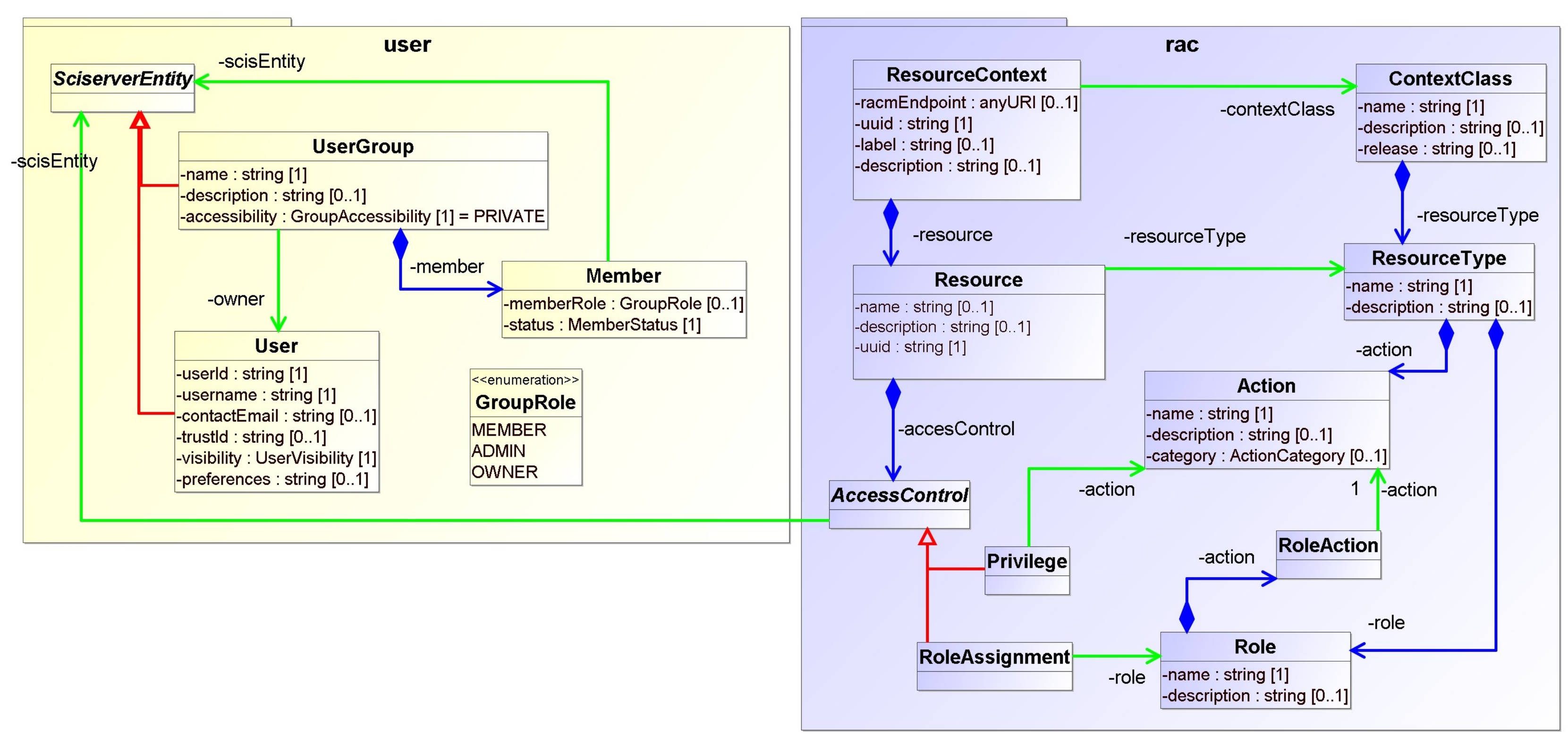
Abstract: SciServer is a *science platform* built and supported by the Institute for Data Intensive Engineering and Science at the Johns Hopkins University. SciServer extends the SkyServer system of server-side tools that introduced the astronomical community to SQL and has been serving the Sloan Digital Sky Survey catalog data to the public. SciServer uses a Docker based architecture to provide interactive and batch mode server-side analysis with scripting languages like Python and R in various environments including Jupyter (notebooks), RStudio and command-line. Users have access to private file storage as well as personal SQL database space. A flexible resource access control system allows users to share their resources with collaborators, a feature that has also been very useful in classroom environments. All these services, wrapped in a layer of REST APIs, constitute a scalable collaborative data-driven science platform that is attractive to science disciplines beyond astronomy.

SCISERVER ARCHITECTURE



The different modules provide for authentication, database and file-based storage, on-prem computation, resource access controls and logging. All are accessible through REST APIs using a simple, token based single-sign-on mechanism.

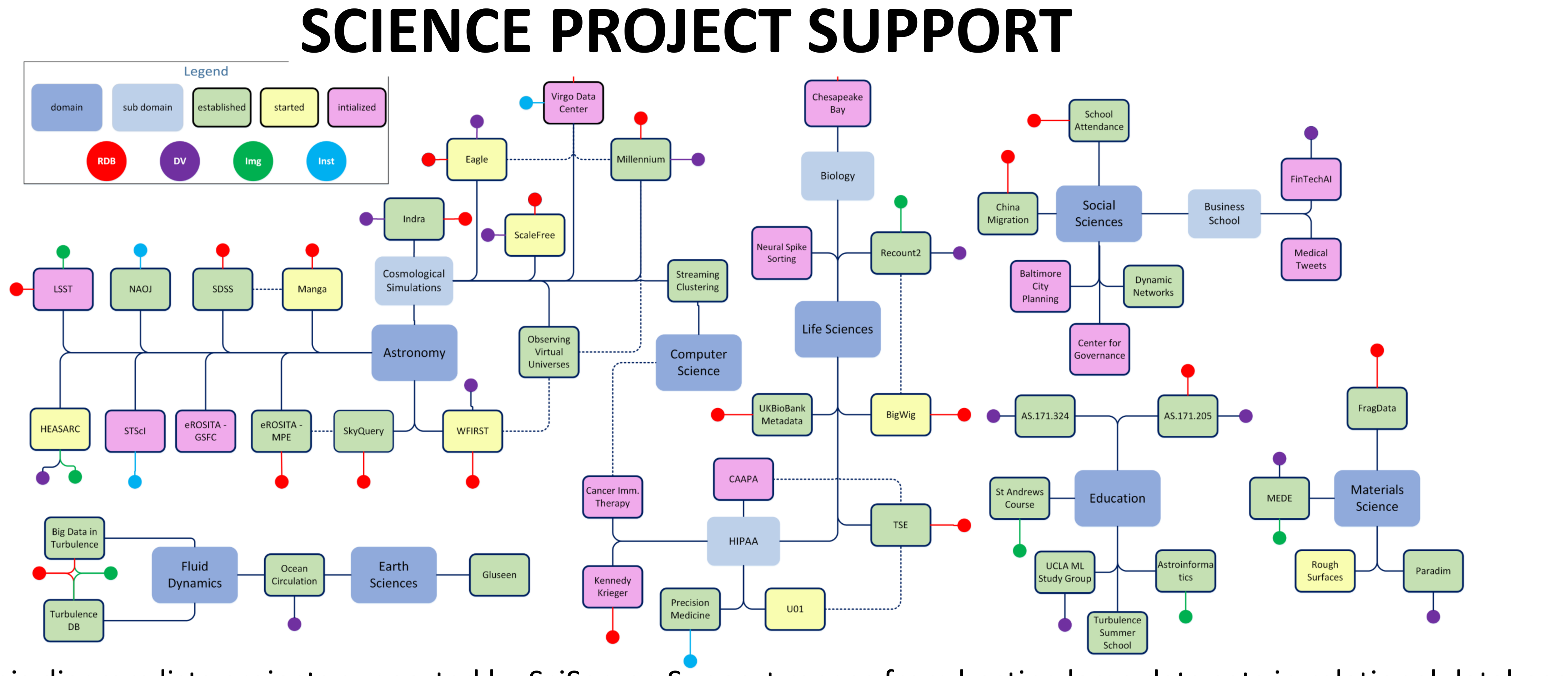
SHARING, COLLABORATION



RACM is the Resource Access Control Management component of SciServer. RACM uses a flexible data model for representing *who* is allowed to do *which* actions on *which* resource, either in groups or individually. The data model is expressed as UML and automatically mapped to a relational database model and java classes with JPA Object Relational Mapping annotation

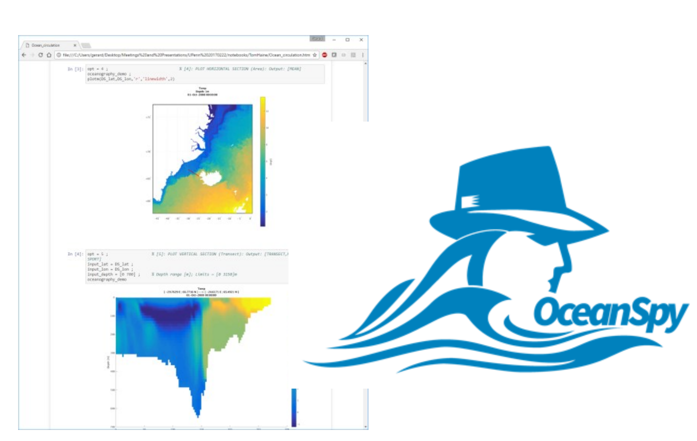
SCIENCE PROJECT SUPPORT

SciServer supports analysis of Vacant Lots in Baltimore. Michael Braverman, Tams Budavari

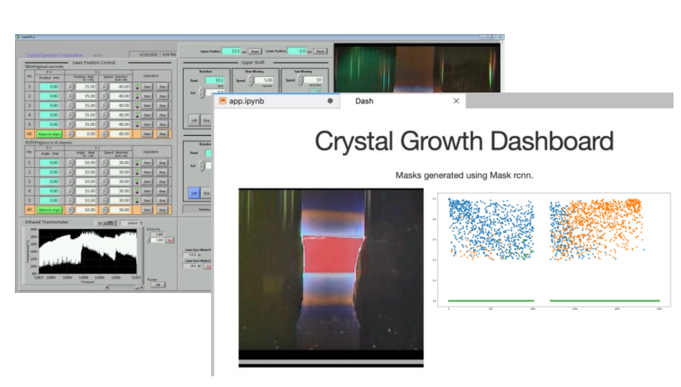


This diagram lists projects supported by SciServer. Support ranges from hosting large data sets in relational database or distributed file systems, provision of specialized compute environments, to custom deployments on site.

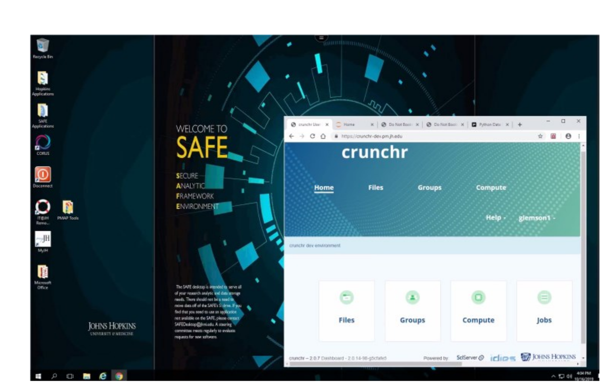
SciServer deployed at MPE, Munich, Germany Hosting data from eROSITA X-Ray satellite



SciServer supports analysis of ocean circulation simulations, NSF OAC-1835640



SciServer supports the Platform for the Accelerated Realization, Analysis, and Discovery of Interface Materials (PARADIM) NSF DMR-1539918

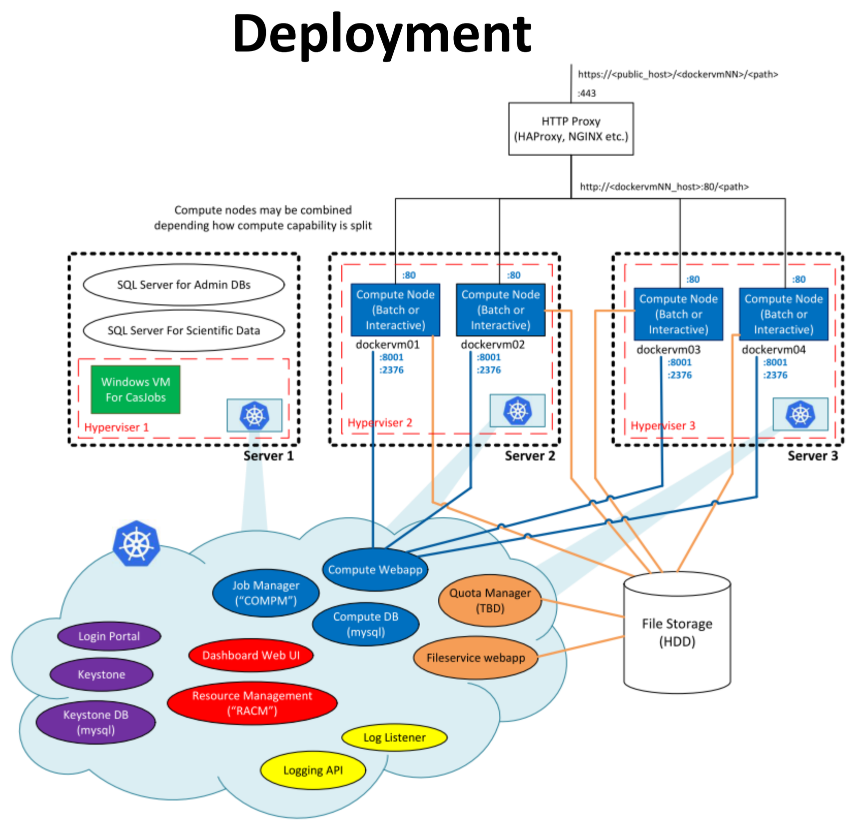


Crunchr, a SciServer fork, supports the Johns Hopkins inHealth Precision Medicine Analytics Platform

Technologies

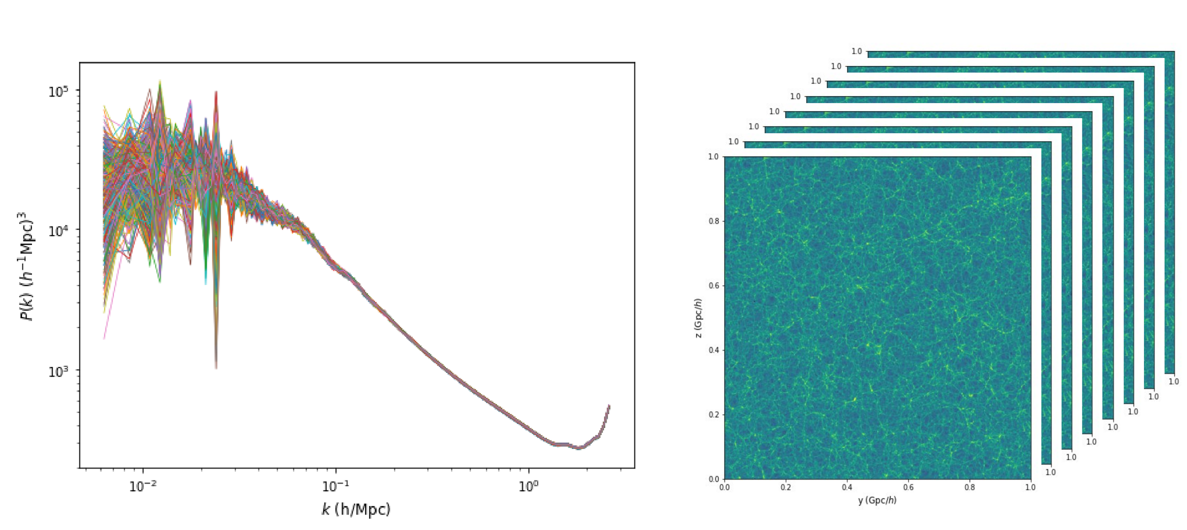
KUBERNETES BASED DEVELOPMENT: CLOUD, COMPUTE AND STORAGE

Deployment



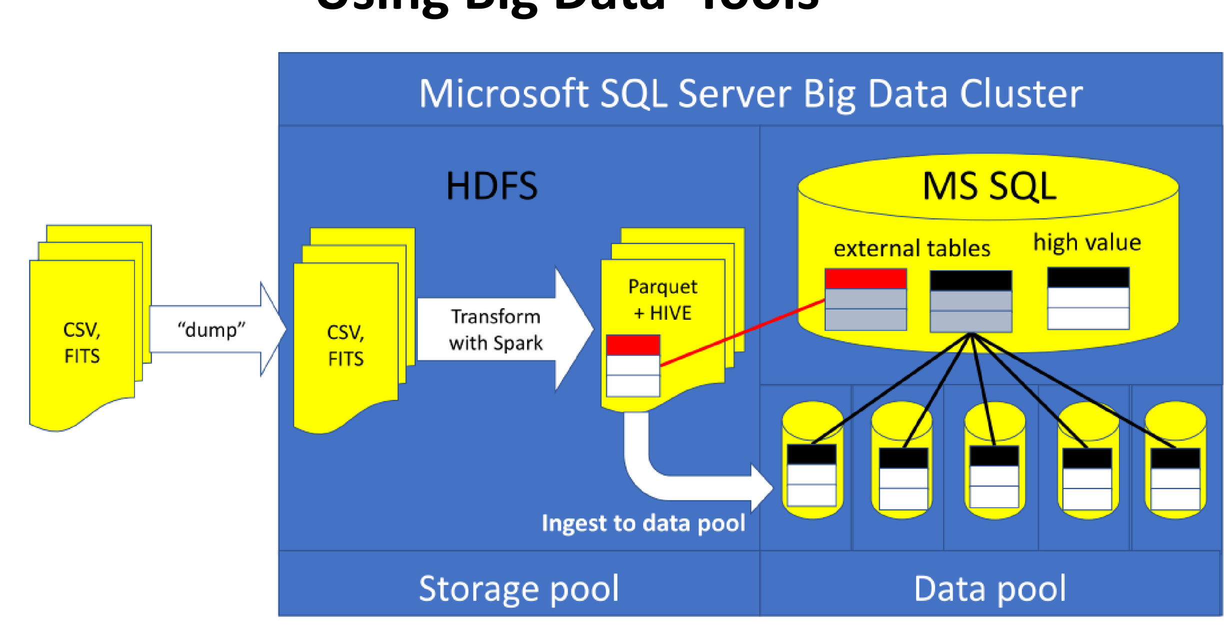
Schematic diagram of SciServer deployed on Kubernetes cluster. Thiis will greatly facilitate for example spinning up new instances in commercial clouds.

DASK, parallel, distributed python



Massive parallel analysis of cosmological simulations 448 Cloud-In-Cell density grids in 2 hours, using 8 DASK workers on distributed file system. Each simulation has 1 billion particles

Astronomy Survey Pipeline Support Using Big Data Tools



Upcoming astronomical surveys such as LSST and WFIRST exceed those hosted by SciServer by orders of magnitude. New database technologies such as MS SQL Server 2019 BDC are investigated in close cooperation with the Microsoft team and supported by a Microsoft Investigator Fellowship

Team : Alex Szalay (PI), Gerard Lemson (System Architect), Jordan Raddick (Outreach), Ani Thakar (SDSS), Sahil Hamal, Jai Won Kim, Dmitry Medvedev, Arik Mitschang, Manu Taghizadeh-Popp.

Cite: https://arxiv.org/abs/2001.08619 **URL**: http://www.sciserver.org